
ASHPC26
Austrian-Slovenian
HPC Meeting 2026
Vienna,
7–10 April 2026

Austrian-Slovenian HPC Meeting 2026 – ASHPC26

Vienna, 7–10 April 2026

https://ashpc.eu

Welcome to ASHPC26

Austrian-Slovenian HPC meetings (ASHPC) are key events for the Austrian and Slovenian High-Performance Computing (HPC) communities, bringing together researchers, practitioners, and infrastructure providers from both countries to discuss the latest developments in HPC, AI, Quantum Computing, and data-intensive computing. It plays an increasingly important role for the wider Central European HPC ecosystem as well, by fostering cross-border collaboration, enabling knowledge exchange, and encouraging the sharing and joint development of advanced computing infrastructures, services, and scientific projects. In this way, the conference also strengthens the network of national HPC competence centres and national HPC communities, and helps align scientific use cases, training activities, and technology roadmaps across the region.

The ASHPC26 program features 5 keynote talks that provide strategic and scientific perspectives on HPC, AI, simulation, and quantum computing, complemented by 10 technical-track contributions focusing on architecture, operations, and software stacks for modern supercomputing systems. Across the main sessions, more than 40 contributed talks highlight a broad spectrum of applications and methods, from large-scale simulations and numerical methods to machine learning workflows and novel HPC-enabled services. The conference also includes 18 posters, each preceded by a short lightning talk, showcasing success stories, emerging projects, and innovative ideas from both established and early-career contributors.

I would like to express my sincere thanks to the Organizing Committee for their dedicated work in handling the many logistical details, finalizing and communicating the scientific and social program, and creating a welcoming environment that will make ASHPC26 a vibrant meeting point for all participants.

Next, I would like to express my sincere thanks to the Program Committee, which carefully selected the keynote speakers, reviewed all submitted abstracts, and made the final decisions on the accepted talks and posters. The dedicated efforts of the members of the Program Committee are essential for maintaining the high scientific quality of ASHPC meetings.

Finally, I would also like to thank the Steering Committee, which has closely monitored all activities and provided guidance and support whenever needed. All these efforts are crucial to keeping ASHPC an attractive meeting point for the Central European high-performance and quantum computing community.

Welcome to ASHPC26 – let's make it a memorable and fruitful experience!

Janez Povh (Program Committee Chair)

Schedule

Tuesday, 7 April 2026

Start	Title
	Central European NCCs Workgroup Meeting – CE-NCCs-WG
14:00	Registration & Get Together & Coffee Break
14:30	CE-NCCs-WG Plenary Session 1
16:00	Coffee Break
16:30	CE-NCCs-WG Plenary Session 2
18:00	End of Session
18:30	Dinner in a nearby restaurant

Wednesday, 8 April 2026

Start		Title
09:00	Registration & Get Together	
09:30		Welcome by PC & OC chairs
09:45	Dieter Kranzlmüller	Optimizing Energy Efficiency for HPC-AI Deployment – The Path to AI Gigafactories (Keynote)
10:30	Michael Iro	Bridging the AI Knowledge Divide: The AI Factory Austria AI:AT Learning Center
10:45	Iulia-Georgiana Rinea	Managed ML Inference on Shared HPC Infrastructure
11:00	Coffee Break	
11:30	Séverine Habert	Optimizing and Scaling LLM Inference: A Review of State-of-the-Art Techniques
11:45	Andreas Lindner	Training deep learning models for identification of Austrian butterflies and moths on more than half a million images
12:00	Georg Heiler	From Laptop to Supercomputer: Reproducible ML Pipelines with dagster-slurm and metaxy
12:15	Andreas Ravazzolo-Mehrle	Scaling AI Systems development from desktop to HPCs using MATLAB & Simulink
12:30	Lunch Break	
14:00	Michael Bussmann	Beyond exascale – When data becomes more important than FLOPs/s (Keynote)
14:45	Sylvia Plöckinger	The multi-phase interstellar medium of COLIBRE galaxies
15:00	James McKevitt	Simulating Magnetic Reconnection To Support Next-Generation Space Missions
15:15	Lukas Winkler	Scaling Differentiable Simulations in Cosmology to Multiple GPUs
15:30	Coffee Break	
16:00	Jesper Larsson Träff	Circulant Graphs for Collective Communication
16:15	Ruben Laso	To `ncclsee`, or Not to `ncclsee`: That is the Profiling Question
16:30	Jonas Sieberer	Topology-Aware Communication Optimization for CFD Simulations in OpenFOAM
16:45	Marko Ferme	Bridging HPC and Cloud: OpenStack-Based Infrastructure for Efficient AI Computing
17:00	Mladen Borovič	MENTHOS-spam: High-Throughput Detection for Phishing and SMS Spam
17:15	Zoé Lloret	Kilometer-scale Climate Modeling of TRAPPIST-1e Using ICON-Sapphire: Peering through the high clouds
19:00	**ASHPC26 Dinner at Rathauskeller**	

Technical Track – System Architecture, Configuration, and Operations of HPC and AI-Optimized Supercomputers

Wednesday, 8. April 2026

Start	@ BA 10A	Title
16:00	Katrin Muck	ASC Cluster Admin & Infrastructure Service Modernization
16:15	Adam McCartney	The MUSICA software stack
16:30	Florian Goldenberg	Moving stuff around: Rucio & FTS
16:45	Hernan Picatto	dagster-slurm: Connecting Data Orchestration to HPC Resources
17:00	Gent Rexha	Managed Multitenant ML Workflows on HPC

Thursday, 9. April 2026

Start	@ BA 10A	Title
11:00	Orlenys Troconis	Organizing the software stack in CINECA's clusters. Towards LISA.
11:15	Florian Atzenhofer-Baumgartner	Design and Operation of a Federated GPU Cluster for Digital Humanities within DHinfra.at
11:30	Thomas Rattei	LiSC software catalog: a software installation framework for Life Sciences
11:45	Ümit Seren & Leon Schwarzäugl	Zero-Touch HPC Nodes: NetBox, Tofu and Packer for a Self-Configuring SLURM Cluster
12:00	Sebastian Sitkiewicz	HPC Info: Enhanced SLURM Job Resource Monitoring

Start		Title
13:00	@BA 10B	**SLING Meeting**
13:15	@BA 10A	**ASC User Meeting**

Thursday, 9 April 2026

Start		Title
09:00	Latha Venkataraman	From Atoms to Current: Predicting Function in Single-Molecule Circuits (Keynote)
09:45	Darin Lah	EPICURE project - practical example
10:00	Florian Goldenberg	RI-Scale – bringing data holdings and HPC together
10:15	Victoria Döller	EVITA – EuroHPC Virtual Training Academy
10:30	Coffee Break	
11:00	Silvio Heinze	From City-Scale Street-View Imagery to Building-Level Urban Indicators: A Precompute Layer for GeoAI ML Models on the MUSICA HPC System
11:15	Davor Davidović	NCC Croatia Success Story: Setting up Photogrammetry Workflow on an HPC cluster - A Cultural Heritage Use Case
11:30	Gregor Molan	HPC-Driven Dimension-Aware Neural Architecture Search for Cryocooler Lifetime Prediction
11:45	Jelena Joksimović	Simple is better? HPC-enabled neural architecture search for energy demand forecasting
12:00	Max Hodapp	Automated learning of multiscale models
12:15	Manuel Hofbauer	Extracting Metallurgical Graphs using Reasoning LLMs on HPC
12:30	Lunch Break	
14:00	Gerhard Hummer	Learning from molecular simulations (Keynote)
14:45	Maša Lah	Open-Boundary Molecular Dynamics of Red Blood Cell Suspensions
15:00	Thomas Haschka	SIMD, GPU, and MPI Acceleration of Density-Based Tree Construction for Large-Scale Sequence Data
15:15	Anže Hubman	Efficient inference of overdamped Langevin models from projected molecular dynamics trajectories
15:30	Coffee Break	
16:00	Relindis Rott	General Matrix-Matrix Multiplication and NVIDIA Tensor Cores Applied to the Lattice Boltzmann Method
16:15	Lea Enzenberger	GPU algorithm for efficient runtime detection of coalescence and breakup events in phase-field multiphase flows
16:30	Amina Tahreen	Numerical analysis of electrocoagulation using computational fluid dynamics for sustainable water treatment
16:45	Antonija Rajic	HPC-Enabled AI-Agent-Driven Digital Twin Framework for Real-Time Exergoeconomic Optimization of PV-Supported Thermal Energy Systems
17:00	Muhammad Mizan	High-Fidelity Flow Simulations empowered by HPC
17:15	Poster Lightning Talks	
17:35	Posters & Pizzas	
20:00	End of Session	

Poster Session

Poster		Title
Poster 1	Andreas Lindner	EuroCC Austria: The Austrian Competence Centre for Supercomputing
Poster 2	Jurica Špoljar	Integration of Advanced Computing service and e-learning platform Merlin
Poster 3	Teo Prica & Samo Lorenčič	HPC Vega: Supporting Services
Poster 4	Žiga Zebec	EPICURE: Unlocking European-level HPC Support
Poster 5	Tina Marc	FFplus: Bridging SMEs with State-of-the-Art HPC and Generative AI
Poster 6	Tristan Pahor	NOUS: Advancing Europe's Sovereign Cloud through HPC, Edge Computing and Data Spaces
Poster 7	Bernd Saurugger	Privacy Preserving Generative AI and Model Sanitization
Poster 8	Marie Czuray	HPC meets EOSC: Emerging Tasks, Opportunities and Concerns
Poster 9	Martin Thaler	MUSICA - From Scratch To A Fully Functional Server Room @UIBK
Poster 10	Ezhilmathi Krishnasamy	Quantum Computing for Scientific Computing
Poster 11	Michael Otto & Florian Atzenhofer-Baumgartner	A Domain-Aware Controller for Managed LLM Inference on Shared HPC Infrastructure in Digital Humanities
Poster 12	Maizura Ibrahim	High-Performance Computing for AI-Based Insider Threat Analytics: An Experimental Study
Poster 13	Till Kahlke	Training Alternative Large Scale Representations on Current High-Performance Computers
Poster 14	Aleksander Grm	Open-source framework for parametric study of hydrofoil profiles and motivation for using Physics-Informed Neural Networks (PINN)
Poster 15	Leon Kos	VIVID-DTE — Verification-oriented Interactive Visualisation and Decision Support for the EUROfusion Digital Twin Environment
Poster 16	David Lajevec	Heterogeneous Exascale Particle-in-Cell
Poster 17	Luis Casillas-Trujillo	LEONARDO data centric general purpose partition at AURELEO
Poster 18	Atul Singh	Spreading the Word — ASC Outreach

Friday, 10 April 2026

Start		Title
09:00	Bojan Žunkovič	Beyond Isolated Quantum Computing Paradigms: Hybridization and Supercomputing (Keynote)
09:45	Janez Povh	Quantum and Simulated Annealing-Based Iterative Algorithms for QUBO Relaxations of the Sparsest k-Subgraph Problem
10:00	Peter Kandolf	Quantum computer integration in multi-site HPC infrastructure
10:15	Jens Stücker	JZ-TREE: Lightning fast neighbor search and friends-of-friends with dual tree traversal in JAX and CUDA.
10:30	Coffee Break	
11:00	Ivona Vasileska	Performance-Portable Particle-in-Cell with Multigrid Solvers on Heterogeneous CPU–GPU Node
11:15	Matthias Weber	LAMMPS Molecular Dynamics Simulations of Laser-induced Periodic Surface Structure Formation: Removal of desorbed atoms between Laser Shots
11:30	Márk Dénes	Time-Series Forecasting and Alert Classification for Proactive IT Infrastructure Monitoring
11:45	István Tamás	Supporting file intensive AI Workloads on High Performance Computing
12:00	Tristan Pahor	Innovative Expansion of HPC Infrastructure for Scalable AI Inference Using MACx GPUs
12:15	Aleksandr Trklja	HPC-Enabled LLM Fine-Tuning and Machine Translation of Legal Texts on LEO5
12:30		Closing by PC chair & ASHPC27 announcement
12:45	Lunch	
14:30	End of ASHPC26	

Contents

Keynote Talk:

Optimizing Energy Efficiency for HPC-AI Deployment – The Path to AI Gigafactories

Dieter Kranzlmüller

Leibniz Supercomputing Centre (LRZ)

The rapid expansion of Artificial Intelligence (AI) is fundamentally reshaping the requirements for High-Performance Computing (HPC) infrastructures. As AI models grow in scale and complexity, their energy demand increasingly dominates data center design, operational costs, and sustainability strategies. The concept of AI Gigafactories – large-scale facilities dedicated to training and deploying advanced AI systems – raises critical questions about how to balance computational performance with energy efficiency.

This talk examines the technological and architectural pathways toward energy-optimized HPC-AI deployments. Building on the pioneering aspects of the Leibniz Supercomputing Centre (LRZ) it explores advances in technologies and cooling solutions. Particular attention is given to our holistic approach that align hardware, software, and facility infrastructure, as well as integrating renewable energy sources and waste heat reuse. By outlining both current best practices and emerging innovations, the presentation highlights how energy-efficient HPC-AI platforms can form the foundation of scalable, sustainable AI Gigafactories.

Bridging the AI Knowledge Divide: The AI Factory Austria AI:AT Learning Center

Michael Iro[a,b], Daniel Lehner[a,c], and Claudia Blaas-Schenner[a,b]

[a] *AI Factory Austria AI:AT*
[b] *ASC Research Center, TU Wien*
[c] *Austrian Academy of Sciences (ÖAW)*

The AI Factory Austria AI:AT [1] is an initiative to strengthen the Austrian and European AI ecosystem, upskill the workforce, bridge the AI talent gap, and accelerate the adoption of trustworthy AI across industries. The project is being carried out jointly by Advanced Computing Austria ACA GmbH and AIT Austrian Institute of Technology GmbH, with participation from BOKU, EODC, INiTS, ISTA, JKU, ÖAW, TU Graz, TU Wien, University of Innsbruck, and University of Vienna, and is equally funded by the EuroHPC Joint Undertaking (EuroHPC JU) and the Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) via Österreichische Forschungsförderungsgesellschaft (FFG).

As such AI:AT is part of a network of 19 European AI Factories [2,3] and 13 AI Factory Antennas, advancing the EU's strategic goal of "Bridging AI Innovation and Trust" by linking supercomputing capacity, data, and talent across borders.

In this talk we want to highlight the activities of the AI:AT Learning Center integrating training, education, curriculum development, and practical exchange programs to empower businesses, public institutions, research, and academic partners with the skills needed to develop, deploy, and govern trustworthy AI-driven solutions.

The AI:AT Learning Center strategy is based on four pillars: (I) a comprehensive training offer, (II) a learning platform for self-learning, (III) an AI Academy aligning education with industry needs by supporting the development of secondary and tertiary curricula covering AI topics, and (IV) exchange programs and internships. An overview of the current and planned training offer will be given covering topics across the entire AI landscape, ranging from legal and ethical aspects of AI to technical content, such as training and fine-tuning Large Language Models (LLMs) on High-Performance Computing (HPC) systems, and extending to AI applications in the Internet of Things (IoT). We will highlight how these initiatives prepare the next generation of AI professionals while fostering collaboration between academia and industry and how the AI:AT Learning Center implements measures of quality assurance to guarantee high quality trainings. This includes an overview over relevant achieved key performance indicators (KPIs, e.g., number of trainings, number of participants).

References

[1] https://ai-at.eu

[2] https://digital-strategy.ec.europa.eu/en/policies/ai-factories

[3] https://www.eurohpc-ju.europa.eu/ai-factories_en

Managed ML Inference on Shared HPC Infrastructure

Iulia-Georgiana Rinea

AI Factory Austria AI:AT

HPC centers are increasingly expected to offer real-time machine learning inference alongside traditional batch computing. While batch workloads follow a well-established lifecycle of submission, scheduling, and collection, inference services are fundamentally different: they are long-lived, latency-sensitive, and consumed by multiple tenants who expect immediate access through standard APIs. Deploying such services on shared GPU infrastructure raises questions that the conventional HPC software stack does not necessarily address, from per-tenant resource accounting to model lifecycle management and fair access despite varying workload costs.

At AI Factory Austria AI:AT, we are building a platform that integrates managed inference into an HPC environment running on Kubernetes. The core of the inference pipeline relies on KServe as the control plane for model lifecycle and given the current rise of LLMs as the primary serving workload, we use vLLM as the default serving engine. KServe abstracts model deployment into Kubernetes-native custom resources, allowing new models to be onboarded declaratively across the cluster without requiring privileged access. vLLM provides the high-throughput inference runtime with optimizations for memory-constrained accelerators. The platform is not restricted to a single engine or modality: the serving runtime abstraction is designed to accommodate different backends, adapter-based model variants, and non-LLM prediction services as the platform evolves.

A model-aware API gateway sits at the entry point, routing requests based on model identity and enforcing token-level rate limits per tenant. Token-based metering is essential in this context because inference cost varies significantly with input and output length, making flat request-rate limits inadequate for fair resource sharing. The gateway also provides the authentication and tenant isolation layer, where each consumer gets scoped credentials and individual usage tracking.

Observability covers both the gateway layer and the engine layer. The gateway tracks token usage per tenant and per model and on the engine side, vLLM exposes metrics on cache utilization, scheduling queues, and generation latency. Together, these two streams give operators a full view of the inference path without having to manage the serving runtime separately.

This contribution shows the architectural decisions and early operational experience from deploying the platform. It includes challenges around model cold-start times, GPU scheduling on heterogeneous node pools, and tenant isolation strategies in a shared Kubernetes environment.

Optimizing and Scaling LLM Inference: A Review of State-of-the-Art Techniques

Séverine Habert

NVIDIA

Large Language Models (LLM) are nowadays not only used for text generation, but also for reasoning and agents, which leads to an exponential increase in inference compute needs and requirements. As models grow to hundreds of billions of parameters and context windows can reach millions of tokens, traditional monolithic serving strategies struggle to maintain high throughput and low latency. In this talk, we will provide a technical overview of the current state of the art in LLM inference optimization techniques, from node-level to cluster-scale, starting from GPU-level optimization until disaggregated serving architectures, and showcasing these concepts with concrete implementations in inference engines and distributed serving frameworks.

We will first observe the distinct execution characteristics of the prefill and decode phases, which are respectively compute-bound and memory-bound. Building on this, we will survey node-level optimizations that are very common across today's inference engines and essential for optimal latency and throughput. We will cover quantization to low-precision formats (FP8 and FP4), model compression techniques, batching techniques (chunked prefill and inflight), speculative decoding and prefix caching that reuses KV-cache beyond a single prompt. Together, these techniques reduce memory footprint, improve effective bandwidth utilization, and increase throughput and GPU utilization for single-node deployments.

In the second part, we will discuss how those principles extend to cluster level and multi-node environments through disaggregated serving [1], where prefill and decode workers can be optimized and scaled independently, and KV-cache becomes a shared, distributed resource that can be used to avoid recomputing and reduce latency. We will outline key architectural principles for such systems such KV-cache aware routing, tiered KV-cache offloading and management, and dynamic reallocation of GPUs and will showcase how these principles are realized in recent distributed inference frameworks [2]. Then, we will present benchmark results demonstrating the benefits of disaggregated serving over traditional aggregated approaches, the benefits of KV-cache aware routing vs round-robin routing, as well as the benefit of offloading the KV cache in terms of Total Cost of Ownership. Finally, we will provide practical design guidelines for HPC/AI systems who need to build large-scale LLM inference stacks and have them ready for production.

References

[1] Zhong, Y., Liu, S., Chen, J., Hu, J., Zhu, Y., Liu, X., ... and Zhang, H., DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving, in 8th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24) (pp. 193-210) (2024)

[2] NVIDIA Developer Blog, "NVIDIA Dynamo, A Low-Latency Distributed Inference Framework for Scaling Reasoning LLMs" (March 2025)

Training deep learning models for identification of Austrian butterflies and moths on more than half a million images

Andreas Lindner[a,b], Friederike Barkmann[c], and Johannes Rüdisser[c]

[a] *Advanced Computing Austria ACA GmbH, Karlsplatz 13, A-1040 Wien, Austria*
[b] *University of Innsbruck, Research Area Scientific Computing, Technikerstrasse 13, A-6020 Innsbruck, Austria*
[c] *University of Innsbruck, Department of Ecology, Sternwartestraße 15, A-6020 Innsbruck, Austria*

Deep learning models can accelerate the processing of image-based biodiversity data and provide educational value by giving direct feedback to citizen scientists. Butterflies and many moth species, which serve as important biodiversity indicators, are particularly well-suited for image-based identification. Moreover, butterflies are perceived positively by the public and are well-suited for observation in citizen science projects.

Training such models requires large amounts of labeled data. We use a high-quality dataset of more than 540,000 images covering 185 species occurring in Austria, collected via the citizen-science smartphone application "Schmetterlinge Österreichs" and verified by an expert entomologist [1]. As is typical for species-record datasets, the numbers of images per species are highly imbalanced, ranging from single observations to nearly 30,000. Rare species with less than 50 occurrences in the dataset were omitted to ensure meaningful training. Using this dataset – the largest published collection of butterfly and moth images to date – and computing power from various Austrian and European HPC systems, we evaluated more than 40 deep learning models on the fine-grained task of species identification. Ten percent of all samples were set aside as the test set, while the remaining 90 % were divided into training and validation subsets using an 80–20 split. All splits were stratified to preserve the original species distribution throughout the dataset.

Several models achieved overall validation accuracies above 97 %. Subsequent parallelized hyperparameter optimization (HPO) scans further improved performance, see the training history of a MaxViT model – a CNN–Transformer hybrid – in Fig. 1. This architecture surpassed the other models in accuracy while maintaining low training cost. While class weights were used in the initial training stage to address the class imbalance, oversampling of minority classes in the training data substantially increased recall and precision for underrepresented species in the HPO stage. A Top-1 accuracy of 98.14 % and a Top-5 accuracy of 99.53 % were ultimately reached on the test data, at a mean precision of 97.57 % and a mean recall of 95.01 %.

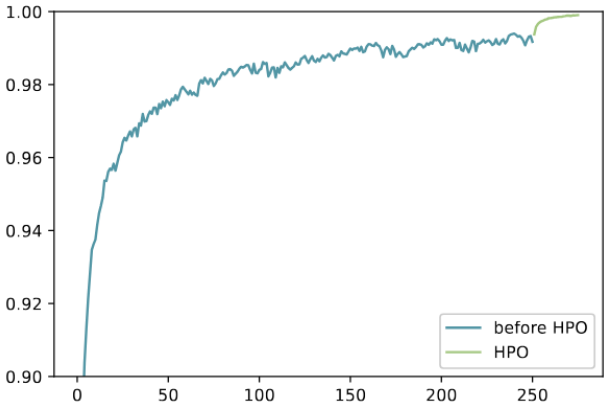


Fig. 1: Training history of a MaxViT model. The x-axis shows the epochs; the y-axis, the accuracy on the training data.

This dataset and the benchmarking effort provide a strong foundation for advancing automated species identification in ecological research and citizen-science applications.

References

[1] Barkmann, F., Lindner, A., Würflinger, R. et al., Sci Data **12**, 1369 (2025).

From Laptop to Supercomputer: Reproducible ML Pipelines with dagster-slurm and metaxy

Georg Heiler[a,b], **Daniel Gafni**[c], **and Hernan Picatto**[b]

[a] *Complexity Science Hub Vienna (CSH), Austria*
[b] *Austrian Supply Chain Intelligence Institute (ASCII), Austria*
[c] *Anam, Austria*

Motivation and challenge: European ML teams increasingly move between laptops, CI systems, and sovereign Slurm clusters [3], but still maintain separate scripts and deployment logic for each environment. This fragmentation hurts reproducibility, slows iteration, and reduces operational visibility. The problem is strongest in multimodal pipelines, where expensive GPU steps are often rerun unnecessarily after small upstream changes. We present a practical integration of dagster-slurm[1] and metaxy[2] to address both portability and selective recomputation in one workflow [1,2].

Technical approach: dagster-slurm extends Dagster's asset model to Slurm [3] through a unified compute resource with execution modes for local development and scheduler-backed production runs. Assets remain unchanged while execution targets shift from local processes to one-job-per-asset Slurm submission. Environments are packaged reproducibly with Pixi, payloads are submitted through Slurm and Dagster Pipes, and logs plus scheduler metadata are returned to the orchestration UI for monitoring and debugging [1].

Metaxy complements this by tracking field-level dependencies at record granularity. Instead of invalidating full tables, it computes exactly which records and downstream fields are affected by a code or dependency change. The resulting diffs are consumed by orchestrators (e.g., Dagster/Ray) to schedule only the necessary work [2].

Use case and expected impact: We apply this pattern to a document-processing workflow and introduce the tooling in three stages: plain processing, distributed Ray execution, and metaxy-aware incremental updates integrated into Dagster assets. The same user code can be tested locally, validated in containerized Slurm CI, and deployed to academic HPC systems with minimal configuration changes [1]. This provides an actionable blueprint for reproducible ML operations on Slurm: asset-oriented orchestration for end-to-end visibility, plus metadata-driven record-level updates to reduce redundant GPU computation. The approach targets research software engineers, data engineers, and ML teams that need higher throughput per GPU-hour without forcing every contributor to become a scheduler specialist.

References

[1] Picatto, H., Heß, M., Heiler, G., and Pfister, M., Discovering the SUPER in computing – dagster-slurm for reproducible research on HPC, software manuscript in preparation (2026).

[2] Gafni, D., and Heiler, G., Metaxy: Record-Level Feature Metadata Management for GPU-Accelerated ML Pipelines, software manuscript in preparation (2026).

[3] Yoo, A.B., Jette, M.A., and Grondona, M., Slurm: Simple Linux Utility for Resource Management, in Job Scheduling Strategies for Parallel Processing, Springer, 44–60 (2003).

[1] Public draft: https://github.com/ascii-supply-networks/dagster-slurm/blob/main/docs/paper.md.
[2] Public draft: https://github.com/anam-org/metaxy/blob/main/publications/2026-introducing-metaxy/paper.md.

Scaling AI Systems development from desktop to HPCs using MATLAB & Simulink

Akhil Gopinath and Andreas Ravazzolo-Mehrle

The MathWorks

The development of modern engineered systems—ranging from autonomous robotics to quantum-scale sensors—increasingly relies on the convergence of High-Performance Computing (HPC), Model-Based Design, and Artificial Intelligence. While the desktop remains the primary environment for algorithm development and initial prototyping, the transition to large-scale clusters is now a critical requirement for handling the "computational explosion" inherent in high-fidelity simulations and complex AI training.

This presentation explores the workflow of scaling MATLAB and Simulink applications across HPC environments (such as the Vienna Scientific Cluster or the Vega supercomputer). We focus on the technical challenges of moving beyond the local workstation, specifically addressing:

- **Massive Parallelization:** Utilizing MATLAB Parallel Server to run thousands of concurrent Simulink simulations for data generation, sensitivity analysis, and Monte Carlo studies.
- **AI at Scale:** Leveraging distributed GPU resources for deep learning and reinforcement learning, where the complexity of the "environment" requires the robust modeling capabilities of Simulink.
- **Advanced Numerical Methods:** How specialized fields—such as quantum research or physics-informed AI—utilize automatic differentiation and tensor-based operations to overcome memory bottlenecks via distributed memory architectures.
- **Transition to Hardware Prototyping and Deployment:** Make use of MATLAB and Simulink coder capabilities to generate and deploy target specific C/C++, HDL or CUDA code automatically. Hence networks can easily be deployed to a variety of hardware platforms like embedded systems, PLCs, FPGAs or PCs.

By showcasing a variety of successful implementations from our user stories, including quantum research [1], industrial digital twins and arts and social sciences [2], we demonstrate how a unified software platform allows researchers to focus on domain-specific innovation while abstracting the complexities of job scheduling (SLURM), containerization, and inter-node communication. The session concludes with a vision of the "AI Factory," where HPC resources transform simulation from a validation tool into a high-throughput engine for discovery.

References

[1] Enhancing Tensor Network Algorithms for Many-Body Quantum Systems with MATLAB: The MathWorks, https://www.mathworks.com/company/technical-articles/developing-new-approaches-for-quantum-computing-research.html

[2] AI Unveils the Secrets of Ancient Artifacts: The MathWorks, https://www.mathworks.com/company/mathworks-stories/ai-for-digital-preservation-of-ancient-artifacts.html

Keynote Talk:

Beyond exascale – When data becomes more important than FLOPs/s

Michael Bussmann

Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany

Exascale computing has been a technological breakthrough in shifting the attention of single large-scale simulations to more complex workflows. The reason for this is two-fold. First, instead of just the raw calculation power the memory available in these systems has enabled simulations of unprecedented fidelity and resolution. In plasma physics, this now means that we can run atomic resolution simulations of high energy density matter in hours or days instead of weeks. Physics processes that could only be included in post processing such as atomic physics can now be computed on the fly, making more use of the available computation power than the original algorithms can even after performance optimization.

This, secondly, opens the path towards multi-simulation, multi-scale, multi-physics approaches. However, each simulation and the orchestration of many simulations now poses new challenges in data preservation, analysis and optimum system use. We argue that optimizing simulation performance has now to be augmented by optimizing complex workflows that focus on optimum data extraction, intelligent reduction and feedback mechanisms. The paradigm of post-Exascale computing shifts from classical HPC tasks to coupling simulations with machine learning and online data processing and analysis and even connects to sources other than simulation data.

We showcase examples from laser-driven high energy density physics where we have implemented online streaming of Exascale simulation data into training a large-scale invertible AI model connecting ab initio atomic dynamics to observables available in experiment. We also showcase orchestrating many Petascale simulations using optimized patterning of phase space and then discuss future approaches to couple online analysis of experimental data with online steering of HPC simulations for experimental augmentation and optimized operation. We close with the vision of AI-augmented research facilities that integrate HPC resources with on-site data analytics in a federated way, focusing on HPC systems as sources of high-fidelity, high-resolution data augmenting experimental observations.

The multi-phase interstellar medium of COLIBRE galaxies

Ploeckinger Sylvia and the COLIBRE team

Department of Astrophysics, University of Vienna, Türkenschanzstrasse 17, A-1180 Vienna, Austria

The Colibre project [1] is a suite of highly detailed hydrodynamic cosmological simulations, modelling the formation and evolution of 100,000s of galaxies (more than 800,000 in the largest simulation) for more than 13.5 billion years. The individual Colibre flagship simulations used up to 72 million core hours on cosma8 in Durham (UK), running on 160 nodes with 128 cores each.

The galaxy formation model developed for Colibre uses novel prescriptions for star formation and stellar feedback as well as black hole growth and feedback from active galactic nuclei. In addition, it models—for the first time in these large cosmological volumes—the multi-phase structure of the interstellar medium (ISM), following a complex network of chemical reactions, as presented in [2]. This leads to an extremely realistic "virtual" galaxy population (Fig. 1), which has been demonstrated to reproduce key observed scaling relations of galaxies in the "real" Universe. The large amount of Colibre data is utilized by researchers across the globe for in-depth comparisons with a large variety of observations in an effort to test the leading theories on galaxy formation. First results show remarkable agreement to observations, from the abundance of massive, quiescent galaxies in the early Universe, to the relation between the gas content and the star formation rate of galaxies across 13 billion years.

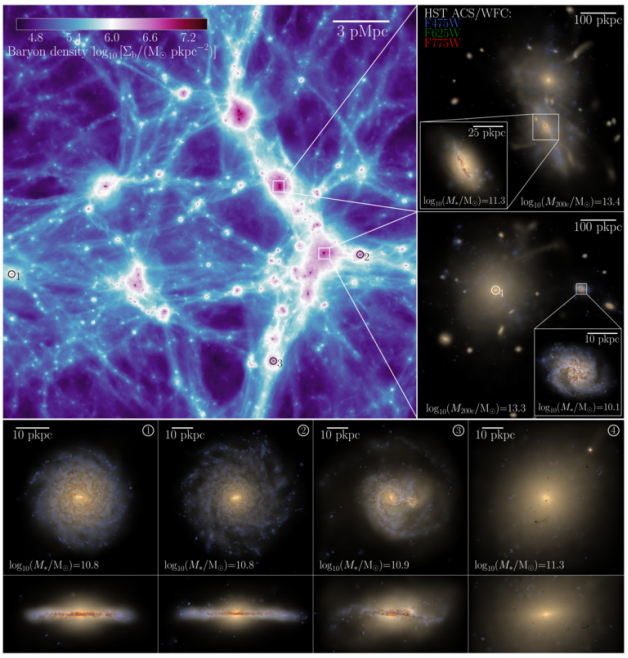


Fig. 1: A large number of galaxies self-consistently form and evolve (see small panels for examples) within the large-scale structure (large panel). Figure from [1].

Colibre spin-off projects testing various dark matter models, such as warm or self-interacting dark matter are currently developed. Together with the detailed ISM model from [2], we are able to perform a detailed analysis of the interplay between the assumed dark matter micro-physics and the observable Universe, aiming at tight constraints on viable dark matter models. In this contribution, I will introduce the Colibre project, and show the first results, focusing on the multi-phase ISM. Finally, I will present an outlook on Colibre spin-off projects and discuss options to realize them with ASC resources.

References

[1] Schaye J., Chaikin E., Schaller M., Ploeckinger S., Huško F., McGibbon R., Trayford J. W., et al., arXiv, arXiv:2508.21126 (2025).

[2] Ploeckinger S., Richings A. J., Schaye J., Trayford J. W., Schaller M., Chaikin E., MNRAS, 543, 891 (2025).

Simulating Magnetic Reconnection To Support Next-Generation Space Missions

James McKevitt[a,b]

[a] *University College London, Mullard Space Science Laboratory, Holmbury St Mary, Dorking Surrey, RH5 6NT, UK*
[b] *University of Vienna, Institute of Astrophysics, Türkenschanzstrasse 17, 1180 Vienna, Austria*

Magnetic reconnection is a fundamental plasma process that operates throughout the universe, from laboratory plasmas to planetary magnetospheres and astrophysical environments. On the Sun it powers solar flares that drive space weather, impacting satellite operations, communication and navigation, and power-grid operations. Magnetic reconnection rapidly restructures magnetic fields in a plasma and converts stored magnetic energy into heat, flows, and particle acceleration [1,2]. In the solar atmosphere, some of the most informative signatures of reconnection are observed in the extreme ultraviolet (EUV) and appear as changes in spectral line widths, shifts, and asymmetries. Next-generation EUV spectroscopy is expected to provide a powerful tool for understanding reconnection through observations of these emission lines, provided we can translate such observations into complex real three-dimensional plasma dynamics.

The SOLAR-C/EUVST mission, launching in the late 2020s, will deliver high-cadence, high-resolution EUV spectra across the solar atmosphere to track energy release and plasma transport in space and time. To maximise the science return from this mission, I present radiative magnetohydrodynamic (RMHD) simulations and instrument forward modelling that capture how plasma behaves and accounts for how such a spectrometer performs observations. This work is directly informing the design of the instrument, including the tuning of electronics that are impossible to change once the instrument is in space.

In my parallel RMHD simulations of reconnection, I form a self-consistent stratified solar atmosphere, and quantify where reconnection is localised, how fragmented the current-sheet region becomes, and how energy is partitioned between heating, bulk flows, and turbulent motions. I then use my forward-modelling code `ECLIPSE`[3] to convert these RMHD outputs into hyper-realistic slit-scan spectrometer measurements, allowing direct, like-for-like comparison between observable line emission and the underlying 3D plasma dynamics. My results show that EUVST will resolve fine structure in solar plasma that previous-generation EUV spectrometers cannot, and that EUVST will already observe strong plasma flows ∼30-minutes before a flare which current instruments are unable to see [3]. These findings, powered by and impossible without, HPC are an exciting preview of a mission poised to revolutionise our understanding and forecasting of solar flares, with wide-reaching implications across physics and space weather operations.

References

[1] McKevitt, J., Jarolim, R., Matthews, S., Baker, D., et al., Astrophys. J. Lett. **961**, L29 (2024).

[2] McKevitt, J., Harra, L., Valori, G., Baker, D., et al., Astrophys. J., in press (2026).

[3] McKevitt, J., Matthews, S., Brooks, D.H., Shimizu, T., et al., Publ. Astron. Soc. Japan, in review (2026).

[3] https://github.com/jamesmckevitt/eclipse

Scaling Differentiable Simulations in Cosmology to Multiple GPUs

Lukas Winkler[a], Florian List[b], Thomas Flöss[c], Jens Stücker[a], Alejandro Estrada[c], Adrian G. Adame[a], and Oliver Hahn[a,c]

[a] *Department of Astrophysics, University of Vienna, Austria*
[b] *Max Planck Institute for Astrophysics, Garching, Germany*
[c] *Department of Mathematics, University of Vienna, Austria*

A fundamental question in cosmology is how the large-scale structure we observe in our universe, the cosmic web, formed from primordial perturbations. In the coming years, the ongoing and next generation of instruments such as Euclid, LSST, DESI, and SPHEREx will map tens of billions of galaxies. Using these observations to solve the inverse problem of inferring cosmological parameters, models and initial conditions requires forward models that combine the accuracy of large-scale N-body simulations with the computational speed and efficient gradient evaluation needed for methods such as Hamiltonian Monte Carlo (HMC).

Recent libraries such as JAX enable GPU-efficient, automatically differentiable physical models, propagating gradients through the entire simulation and enabling more efficient sampling and optimisation. Building on this, the Disco-DJ framework [1], developed at the University of Vienna, provides a fully differentiable forward model for cosmological inference. It includes a linear Einstein–Boltzmann solver [2] and non-linear structure formation models such as Lagrangian perturbation theory (LPT) and fast particle-mesh (PM) N-body simulations using LPT-inspired time integrators [3]. In addition, it can output light-cone data (as seen by a present-day observer) and post-process Friends-of-Friends particle groups directly on the GPU.

Originally, Disco-DJ was limited to lower resolutions by the memory constraints of a single GPU. However, the vast fields of view of recent surveys demand extremely large simulation box volumes, while at the same time resolving small, non-linear scales requires high particle resolutions. To address this, the latest version of Disco-DJ can distribute simulations across multiple GPUs. Using 32 MUSICA GPU nodes, we can simulate 4096^3 particles in about 85 seconds, with further improvements currently in development.

Even on HPC clusters with fast inter-node connections, such as MUSICA or Leonardo, the vast majority of this runtime is spent on data communication. Achieving this performance therefore requires a combination of numerical and algorithmic improvements, reducing unnecessary communication and understanding and profiling all computational layers below the simulation code (JAX, XLA, CUDA kernels, NCCL, HPC cluster setup).

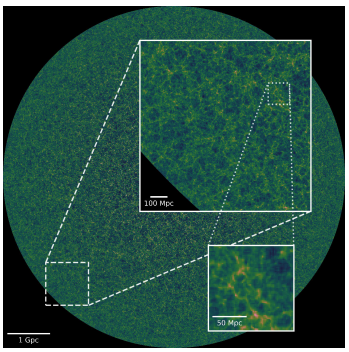


Fig. 1: A slice through a 4096^3-particle light-cone simulation with a diameter of 8 gigaparsecs

References

[1] List, F., Hahn, O., Flöss T., Winkler, L., submitted, arXiv:2510.05206

[2] Hahn, O., List, F. and Porqueres, N., JCAP 06(2024), doi:10.1088/1475-7516/2024/06/063, arXiv:2311.03291

[3] Rampf, C., List, F. and Hahn, O., JCAP 02(2025), doi:10.1088/1475-7516/2025/02/020, arXiv:2409.19049

Circulant Graphs for Collective Communication

Jesper Larsson Träff

TU Wien, Faculty of Informatics, Institute of Computer Engineering, Research Group Parallel Computing

A circulant graph $C_p^{s_0,s_1,\ldots,s_{q-1}} = (V, E)$ is a $2q$-regular graph over p vertices $V = \{0, 1, \ldots, p-1\}$ with edges $(r, (r \pm s_k + p) \bmod p)$ for a given sequence of skips (jumps) $s_k, k = 0, 1, \ldots q-1$. Circulant graphs, with certain skip sequences, have been used as communication structures for algorithms for some of the common collective communication operations, as is well-known [1], and many of these algorithms are among the standard, default implementations in MPI libraries (and elsewhere).

I will argue for the use of circulant graphs as a communication structure for *all* of the common collective communication operations (MPI, NCCL), both as a vehicle to teach, understand, and analyze collective communication algorithms and for actual implementation in MPI libraries etc. Based on two recent papers [2,3], I will outline communication round and volume optimal algorithms for

1. n-block broadcast (where for pipelining the input is divided into n blocks), generalizing to the (irregular) all-gather operation
2. n-block reduction, generalizing to the (irregular) reduce-scatter operation
3. regular reduce-scatter (and allgather, as is well-known [1]) with application to allreduction
4. regular alltoall communication (surprisingly, perhaps, better than [1], for some p)
5. specializations to regular gather and scatter operations

for connected circulant graphs that fulfill the simple property that $s_{k+1} - s_k \leq s_k$ (take $s_q = p$ and $s_0 = 1$). There are also interesting and relevant circulant graph algorithms for the scan operations.

The circulant graph algorithms, in contrast to somewhat similar hypercube algorithms, by design work for any number of processes p, and for many operations, a single algorithm can cover the full range of input sizes. This can possibly make the algorithm selection and tuning problems more manageable.

Current discussions with ANL may lead to adoption of some of the algorithms as baseline implementations in `mpich`.

References

[1] Bruck, J., Ho, C.-T., Kipnis, S., Upfal, E., and Weathersby, D. Efficient algorithms for all-to-all communications in multiport message-passing systems. *IEEE Transactions on Parallel and Distributed Systems*, 8(11):1143–1156 (1997).

[2] Träff, J. L. Optimal broadcast schedules in logarithmic time with applications to broadcast, reduction, all-broadcast and all-reduction. *ACM Transactions on Parallel Computing*, 12(3):1–21 (2025).

[3] Träff, J. L. Optimal, non-pipelined reduce-scatter and allreduce algorithms with an application to all-to-all communication. *ACM Transactions on Parallel Computing*, 12(4):1–23 (2025).

To `ncclsee`, or Not to `ncclsee`: That is the Profiling Question

Ruben Laso[a], Majid Salimi Beni[b], Ioannis Vardas[b], Siegfried Benkner[a], and Sascha Hunold[b]

[a] *Faculty of Computer Science, University of Vienna, Austria*
[b] *Faculty of Informatics, TU Wien, Austria*

Distributed deep learning has become the backbone of modern HPC systems, in which multiple accelerators (typically GPUs) continuously exchange data during model training and inference. For such tasks, NCCL (and related libraries such as RCCL, OneCCL, etc.) is the most widely used library for GPU-GPU communication. Despite NCCL being a well-established technology, profiling tools are still in an early stage of development. In this work, we present the latest version of `ncclsee` [1, 2] and compare it against two other profilers, NVIDIA Inspector [3] and Google CoMMA.

We evaluate the three profilers using micro-benchmarks and a real-world DDL application, training a DenseNet121 model using PyTorch and Distributed Data Parallel (DDP). We run our experiments on two nodes of the *Leonardo* supercomputer, each equipped with 4 NVIDIA A100 GPUs interconnected via NVLink, and the nodes are connected via 200 Gbit/s InfiniBand.

In our experiments with micro-benchmarks, we find that Inspector and CoMMA miss collective-operation events, while `ncclsee` records them all (Fig. 1). Therefore, `ncclsee` provides more accurate timing information that closely matches the communication time reported by the micro-benchmarks (Fig. 2); and thus, allows us to get accurate profiling information of AI workloads.

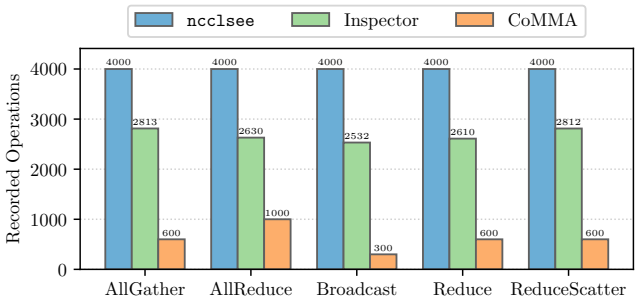


Fig. 1: Amount of collective-communication events recorded for different profilers. Expected value 4000.

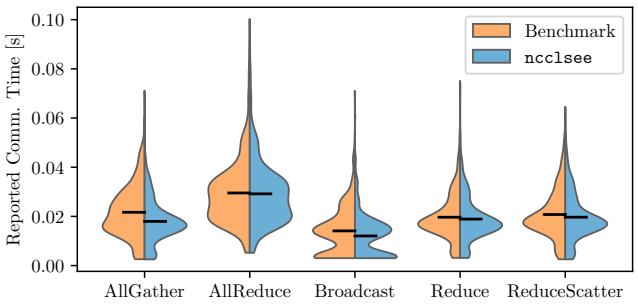


Fig. 2: Communication times reported by micro-benchmarks and `ncclsee`. Message size of 32 MiB.

References

[1] Vardas, I., Laso, R., and Salimi Beni, M. (2025). `ncclsee`: A Lightweight Profiling Tool for NCCL. In ASHPC25 (p. 39). https://doi.org/10.34726/10426.

[2] Laso R., Vardas, I., `ncclsee`, https://github.com/parlab-tuwien/ncclsee.

[3] Das, S. *et al.* Enhancing Communication Observability of NCCL Inspector. NVIDIA Technical Blog (2025). https://developer.nvidia.com/blog/enhancing-communication-observability-of-ai-workloads-with-nccl-inspector/.

Topology-Aware Communication Optimization for CFD Simulations in OpenFOAM

Jonas Sieberer[a], Clemens Gößnitzer[b], Andreas Schröder[a], and Robert Elsässer[a]

[a] *Universität Salzburg, 5020 Salzburg, Austria*
[b] *Large Engines Competence Center LEC, 8010 Graz, Austria*

OpenFOAM is an open-source C++ software package for simulating computational fluid dynamics (CFD). It uses the finite-volume method, which approximates the solution of partial differential equations by solving large (and sparse) systems of linear equations. To solve the system on a parallel machine, these equations are distributed across several MPI processes. OpenFOAM uses iterative linear solvers, which repeatedly perform sparse matrix–vector multiplications, vector updates, and global reductions. These operations need frequent nearest-neighbor data exchange across process boundaries (local communication) as well as the collection of data on a master process (global communication), resulting in significant communication overhead.

On modern HPC systems, each compute node forms a shared-memory hierarchy, which consists of multiple CPU sockets, NUMA nodes, and multi-level cache subsystems. Furthermore, the most costly communication occurs when data packages are exchanged between different compute nodes. Although all cores within a node share a global address space, memory-access latency, and communication costs depend on their placement within the hierarchy described above. OpenFOAM's default MPI rank assignment, however, follows an index-based mapping that is independent of both the communication structure induced by the system of linear equations and the hardware topology. Consequently, neighboring subdomains may be assigned to cores that have to communicate via slower inter-socket or inter-node paths.

We present a topology-aware communication optimization for OpenFOAM that focuses on local point-to-point as well as global communication. First, the communication structure induced by the decomposition of the system of linear equations is modeled as a weighted graph. A hierarchical graph-partitioning approach is then applied to assign MPI ranks to compute nodes, NUMA domains, and NUMA nodes such that heavily communicating processes are mapped to physically close cores. Based on this assignment, a rank file is generated to enforce topology-aware process placement. Second, the default index-based reduction tree used by OpenFOAM for global operations (such as scalar products and norm evaluations) is replaced by a hierarchical reduction schedule aligned with the hardware topology. According to this improved schedule, reductions are performed first within NUMA nodes, then across NUMA domains, and finally between compute nodes, thereby deploying most communication to faster local links before engaging slower interconnects.

We evaluated our methods on two distinct HPC clusters using several two- and three-dimensional OpenFOAM benchmark cases under strong-scaling conditions. Our optimizations were able to substantially redistribute the data exchange towards fast intra-NUMA paths. Across all benchmark cases, the combined optimization of local and global communication reduces the runtime compared to the unmodified OpenFOAM implementation. The improvement depends on both the hardware architecture and solver characteristics, while the largest gains were observed in communication-intensive cases.

Our results show that topology-aware communication strategies can significantly improve the efficiency of parallel CFD simulations without modifying the underlying numerical algorithms. The computational results were partially achieved using the Austrian Scientific Computing (ASC) infrastructure. The related paper will appear in the proceedings of the 26th International Conference on Computational Science (ICCS'26).

Bridging HPC and Cloud: OpenStack-Based Infrastructure for Efficient AI Computing

Marko Ferme, Vid Kranjec, Tobias Korže, and Mladen Borovič

Laboratory for Heterogeneous Computer Systems, Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

Introduction: Small and medium-sized enterprises (SMEs) increasingly require access to GPU-capable infrastructure for AI training and inference, but their operational expectations are shaped by public-cloud platforms: self-service provisioning, API-driven automation, and reproducible environments. Classical HPC access models (accounts, queues, static software stacks) can therefore become a usability barrier, even when sufficient compute capacity exists.

Approach: We present an engineering-and-research effort that converts a subset of an HPC cluster into an OpenStack-based private cloud for AI. The design targets (i) cloud-like resource access through a GUI and REST APIs, (ii) support for both VM-based and container-based AI workflows, and (iii) efficient packing of heterogeneous resources (CPU-only and GPU nodes) to reduce fragmentation and improve utilization. OpenStack services are used for identity, image management, virtual networking, and compute orchestration; GPU nodes are exposed via PCI passthrough to enable accelerator access in tenant-isolated workloads.

System and Deployment: The prototype is deployed on the HPC RIVR – MAISTER [1] infrastructure at the University of Maribor using 10 compute nodes (7 CPU nodes and 3 GPU nodes). Existing nodes are repurposed as OpenStack compute resources, enabling on-demand provisioning of VMs with pre-built AI images as well as containerized execution of AI pipelines (e.g., training jobs, inference services, notebook-style development). Container support is treated as a first-class requirement to align with modern MLOps practices and to improve portability across environments.

Resource Efficiency: We propose provisioning policies that translate user resource requests (vCPU, memory, GPU, and storage locality) into efficient hardware allocations through bin-packing-based scheduling and constraint-aware placement. Special attention is given to GPU allocation to reduce fragmentation and eliminate stranded capacity. The platform supports both low-latency provisioning for interactive workloads and efficient execution of long-running AI jobs, with configurable overcommitment and placement strategies to satisfy diverse service-level requirements

Evaluation and Lessons Learned: We evaluate usability (time-to-first-workload), provisioning latency, and the ability to sustain GPU-accelerated workloads with near-native performance. We summarize operational lessons for integrating cloud control planes into HPC environments (networking, images, multi-tenancy, and day-2 operations) and outline how such an approach can lower the adoption threshold for SMEs while keeping the performance characteristics expected from HPC hardware.

References

[1] HPC RIVR (MAISTER) system description, University of Maribor, https://www.hpc-rivr.si/sistem/, (online, accessed 2026).

MENTHOS-spam: High-Throughput Detection for Phishing and SMS Spam

Mladen Borovič, Tom Li Dobnik, Vid Kranjec, and Marko Ferme

Laboratory for Heterogeneous Computer Systems
Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

High-Performance Computing (HPC) systems are becoming increasingly critical for addressing the rising complexity and volume of modern cybersecurity threats. Traditional network monitoring and threat detection tools often struggle to process the massive throughput of real-time communication data such as emails and SMS, while maintaining the deep semantic understanding required to identify sophisticated phishing and spam attacks. This work introduces **MENTHOS-spam**, a high-efficiency model within our MENTHOS suite (ModernBERT Embedded Network Threat Operational Suite), designed for rapid, large-scale threat detection in network environments. By leveraging recent advances in the encoder-only transformer architectures, MENTHOS-spam achieves industry-leading predictive metrics while providing a significant leap in inference performance compared to existing solutions such as the NVIDIA Morpheus suite.

The MENTHOS-spam model was trained on the HPC Maister cluster using the configuration consisting of 4 NVIDIA Tesla V100 PCIe 32GB GPUs. To optimize training efficiency and scale across multiple accelerators, we employed Distributed Data Parallel (DDP) supported by PyTorch Lightning. The model was trained on a comprehensive dataset of 83980 samples, comprising a public phishing email collection [1] and a balanced, downsampled version of the UCI SMS Spam Collection [2]. Using an 80/20 train-validation split, the model reached convergence in just 26 minutes on the specified HPC hardware.

Table 1: Comparative performance and cost analysis of MENTHOS-spam and NVIDIA Morpheus.

Model	Accuracy	Recall	ROC AUC	Median latency (p50) [ms]	Throughput [samples/s]	Operational cost [€/1M samples]
NVIDIA Morpheus	0.9901	0.9799	0.9997	1396.18	22.6	$\approx$ 22.86
MENTHOS-spam	0.991	1.0	0.9997	16.28	170.7	$\approx$ 3.02

We benchmarked (Table 1) MENTHOS-spam against the NVIDIA Morpheus suite on a MacBook M3 Pro with 12 CPU, 18 GPU and 16 Neural Engine cores (running on ONNX). In terms of predictive quality, our model proved equally robust, and in some areas slightly superior to the Morpheus equivalent, indicating that despite the architectural optimizations for speed, our model maintains superior sensitivity to threats. Our model achieves a 7.5x increase in throughput and an 86 % reduction in operational costs. Further work will expand the MENTHOS suite with models for sensitive information detection, flexible log parsing and root cause analysis, as well as benchmarking using the TensorRT inference engine.

References

[1] Al-Subaiey, A., Al-Thani, M., Alam, N. A., Antora, K. F., Khandakar, A., and Zaman, S.A.U., https://arxiv.org/abs/2405.11619 (2024).

[2] Almeida, T. and Hidalgo, J., https://doi.org/10.24432/C5CC84 (2011).

Kilometer-scale Climate Modeling of TRAPPIST-1e Using ICON-Sapphire: Peering through the high clouds

Zoé Lloret and Aiko Voigt

Department of Meteorology and Geophysics, University of Vienna, Austria

Recent advances have made kilometer-scale Earth climate modeling possible, yet global exoplanet simulations still rely on coarse-resolution models (>100 km) that require explicit parametrization of convection and clouds that introduce significant uncertainties. These parameters are critical for tidally locked planets, where we can only observe the terminator: the boundary between the day and night side. The presence of high clouds at this location can determine our ability to characterize a planet's atmosphere [1].

In this work, we focus on TRAPPIST-1e, a rocky planet slightly smaller than Earth orbiting in the habitable zone of an ultra-cool red dwarf star 40 light-years away. We simulate its atmosphere at 5 km horizontal resolution using ICON-Sapphire, a kilometer-scale model previously applied only to Earth's climate [2]. To do so, we adapted the model to reflect TRAPPIST-1e's size, rotation, stellar irradiation, and an idealized atmospheric composition consistent with the THAI model intercomparison project [3]. To ensure long-term stability, we incorporated modifications, including the use of artificial ozone heating to stabilize the temperature of the stratosphere.

Development and model spin-up were done at around 100 km resolution on the VSC-5 system of Austrian Scientific Computing (ASC) running on 20 to 48 CPU nodes at a decadal timescale to reach a steady atmospheric state from which to start the high resolution simulations. The computationally demanding kilometer-scale simulations are being executed on GPUs on Leonardo, the pre-exascale EuroHPC supercomputer, hosted by CINECA. To facilitate this cross-platform development and ensure performance portability, we developed and deployed containerized versions of ICON, enabling seamless compilation and execution on diverse CPU and GPU architectures.

We examine how planetary parameters shape the simulated climate of a tidally locked exoplanet, with emphasis on high clouds at the terminator. Comparing our convection-resolving simulation with lower-resolution simulations from the existing literature, we assess how kilometer-scale modeling alters atmospheric circulation and cloud processes. This work highlights the potential of high-resolution exoplanet climate modeling to help refine the interpretation of future observational data and shows how an already existing complex earth system model can be reshaped and used for new applications with a relatively low development effort, while showcasing the effectiveness of containers for enabling portability across computing environments.

References

[1] Komacek, Thaddeus D., Thomas J. Fauchez, Eric T. Wolf, and Dorian S. Abbot. The Astrophysical Journal Letters **888**, no. 2 (2020).

[2] Hohenegger, Cathy, Peter Korn, Leonidas Linardakis, et al. Geoscientific Model Development **16** (2): 779–811 (2023).

[3] Fauchez, Thomas J., Martin Turbet, Eric T. Wolf, et al. Geoscientific Model Development **13** (2): 707–16. (2020).

Technical Track:

ASC Cluster Admin & Infrastructure Service Modernization

Katrin Muck and Adam McCartney

ASC Research Center, TU Wien, Austria

We present a modernization of our internal Django-based **Cluster Admin** tool and adjacent services through a unified workflow orchestration layer that consolidates previously existing heterogeneous service integrations. The system connects and manages user on-boarding (Authentik), resource access rights (Cluster Admin), storage provisioning (GPFS, WEKA), management of compute node assignments (Netbox), and the dynamic configuration of our Slurm systems across multiple MUSICA sites.

By decomposing existing directly integrated service connections into containerized workers, orchestrated by *Temporal* [1], we established a unified abstraction layer for complex multi-step infrastructure-related workflows. The architecture fundamentally improves logical scalability: workflows can be extended, modified, and parallelized without modifying the core application. Independent service scaling is enabled through flexible worker deployment patterns.

Fault tolerance is significantly enhanced through *Temporal*'s built-in retry mechanisms, circuit breaking, and state persistence. Failed workflow steps automatically recover without manual intervention, and service outages are gracefully handled through configurable backoff policies. The internal state of the system remains consistent across transient failures. **Observability** gains are realized through centralized workflow execution tracking, providing end-to-end visibility into request lifecycle and service interactions.

Deployment of the Cluster Admin website, Temporal & the Temporal Workers on *Kubernetes* [2] ensures high availability through multiple service instances as well as direct health monitoring and automated restarts. The container orchestration furthermore simplifies updates and enables seamless capacity expansion.

During the **MUSICA Test phases** the system has already demonstrated flexibility and reliability. Due to the lack of personnel the path to full transformation remains ongoing - however we can already present lessons learned from the existing implementation.

Major future goals are to expand and complete the existing solution, integrate the workflow engine into our monitoring stack as well as employing "workflow versioning" to improve the handling of upgrades.

In summary this transformation decouples the web interface from infrastructure dependencies while improving reliability, scalability, and observability. The architecture supports natural evolution toward multi-cluster deployments and provides a solid foundation for future services *ASC* wants to provide and integrate.

References

[1] https://temporal.io

[2] https://kubernetes.io

Technical Track:

The MUSICA software stack

Adam McCartney

ASC Research Center, TU Wien, Austria

The ASC Research Center delivers a uniform software stack to the three sites of MUSICA (Innsbruck, Linz and Vienna). Although this does not present a novel engineering problem, distributing a software stack across multiple HPC clusters was new within the context of the organization. Instead of developing a custom solution, we leveraged an existing approach, that used by the European Environment for Scientific Software Installations (EESSI). The talk will first present a general overview of EESSI and how to use it as a basis for building additional software. The rest of the talk will provide a critical reflection on the decision to use EESSI as a base for the software stack.

The filesystem layer: An architectural diagram of the setup used to distribute software across the sites shows some of the operational overhead required to set up an maintain the infrastructure used to run the components of the filesystem layer.

Software is published to the stratum0 server in Vienna. Each site retains a local copy of the repositories on its' locally deployed stratum1 server. Additionally, two caching proxies are set up at each site. The clients (compute nodes) are configured to talk to the proxies.

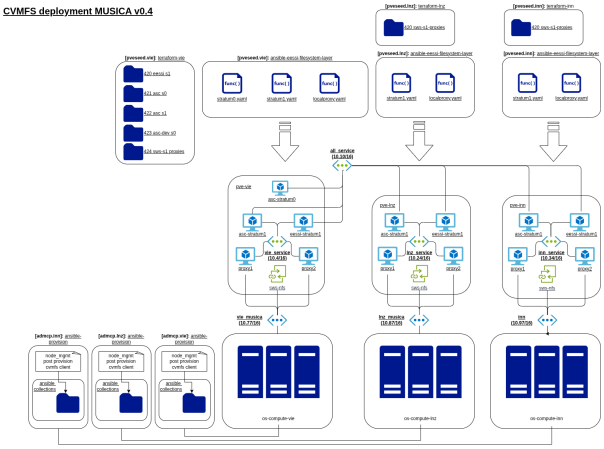


Fig. 1: View of the filesystem layer.

The software layer: The core software is provided by a set of around 1796 environment modules that enable the user to load sets of binaries built for zen4. This software is built centrally and delivered via the CVMFS repository 'software.eessi.io'. The distribution is strictly limited to software carrying an open source license. Any software requiring the acceptance of a end user license agreement, must be installed with an additional step. This affects software packages such as CUDA and NVHPC, that are essential for users who want to leverage the underlying GPU hardware on MUSICA. These additional software packages are delivered using a similar mechanism. At the time of writing, the 'software.asc.ac.at' repository contains around 120 environment modules and their respective software packages.

The compatibility layer: Its use brings benefits to organizations whose software is required to run on multiple clusters and be portable among operating systems. Whether these same requirements are applicable to smaller organizations, such as the ASC Research Center, is unclear. Installing additional software is not as simple as accepting the license agreement. Often, it is necessary to provide the build tool with additional information about the compatibility layer, a process that involves patching binaries to modify their linker and insert rpath values.

Tools to build software: Finally we will present the build pipeline used to build software at the ASC Research Center. How containers and union mount filesystems are used to create a writeable overlay of the 'software.asc.ac.at' repository. How Slurm is leveraged as a build back end, and how build jobs are triggered with dynamic configuration.

Technical Track:

Moving stuff around: Rucio & FTS

Florian Goldenberg

ASC Research Center, TU Wien, Austria

The implementation of a multi-site cluster (MUSICA), as well as the involvement in cross European collaborative projects, such as RI-Scale and EOSC (European Open Science Cloud), created need for an efficient mode of data transfer as well as a method to manage such data. Due to the heterogeneous nature of the various projects, the large amount of data and the often very long distances between sites, the tool needs to be independent from the storage sites and must be able to accommodate various common storage protocols. Furthermore, some form of federated authentication is needed to allow user access from the different participating entities.

For the actual movement of data, FTS (File Transfer Services) will be used [1]. This tool is developed and maintained by CERN as free and open source software. It consists of a web based server and various endpoints at the storage sites. The web-server provides a graphical user interface as well as a REST-API for commandline or software access, a monitoring connection to Grafana and spawns the file transfer agents, which in turn initiate the actual data movement.

The main operating mode is via third party copy, which means a direct transfer between 2 sites over the WebDAV protocol is initiated using up to 1000s of parallel streams. To enable this functionality, involved storage sites need appropriate endpoints, which can be either a fully compatible storage system, such as dCache or EOS, or a node that connects a storage system to FTS. We will be using the latter in the form of a protocol node running xrootd to expose the storage site to the public via WebDAV and to mount our Spectrum Scale storage via standard POSIX functionality.

FTS supports various federated authentication methods, thus eliminating the need for users to have accounts on both ends of a transfer. It also follows all necessary GDPR regulations and can integrate new sites 'on the go'.

In order to provide a high-level data management and orchestration method, Rucio will be implemented [2]. Like FTS, it is a free and open source tool maintained by CERN. The main objective is to catalog and manage data over different, heterogeneous, world-wide distributed sites. The web based tool provides federated authentication, data transfer tools (of which FTS is one option), extensive monitoring and CLI or API interfaces. Rucio is not a distributed file system, it connects existing storage infrastructure over the network. In contrast to the actual transfer tools, no software or endpoints are needed at the storage sites.

We will present our ideas and concepts how to implement these tools in the context of MUSICA and several European project collaborations.

References

[1] https://fts.web.cern.ch/fts/

[2] https://rucio.cern.ch/

Technical Track:

dagster-slurm: Connecting Data Orchestration to HPC Resources

Hernan Picatto[a] **and Georg Heiler**[a,b]

[a] *Austrian Supply Chain Intelligence Institute (ASCII), Austria*
[b] *Complexity Science Hub Vienna (CSH), Austria*

Data scientists and research software engineers are increasingly building pipelines that span multiple compute tiers, including cloud preprocessing, on-premise databases, and GPU-intensive training on HPC clusters. However, this multi-tier reality is often handled poorly, with the HPC step typically living in a custom Slurm submission script that is disconnected from the rest of the pipeline. This results in no shared lineage, no unified observability, and no way to automatically trigger downstream work when the job finishes.

Dagster-slurm is an open-source integration that connects Dagster [1], a widely used data orchestration framework in the data engineering and ML communities, to Slurm-managed HPC resources. The same pipeline code that runs on a developer's laptop can be easily redirected to a supercomputer by changing a single environment variable, without modifying the workflow logic. Environment packaging is handled automatically using pixi and pixi-pack, producing self-contained and reproducible bundles that deploy cleanly in air-gapped HPC environments (systems that are physically and logically isolated from the public internet to ensure high security).

In real-time, logs, Slurm job IDs, CPU efficiency, memory usage, and structured metadata are streamed back to the Dagster UI, providing teams with the observability they expect from cloud-native tooling, even on infrastructure they don't own.

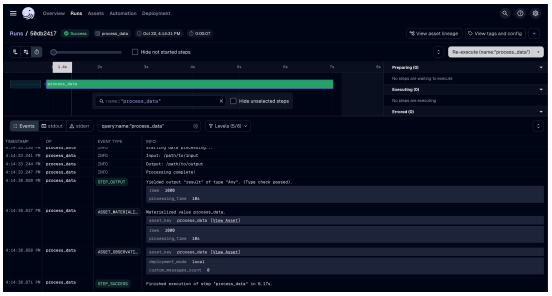


Fig. 1: Real-time observability of HPC jobs streamed directly into the Dagster UI using the dagster-slurm integration.

This talk will cover the architecture of the integration, including Compute Resource, Dagster Pipes over SSH, and pixi-pack deployment. It will also discuss key design decisions around environment portability and job lifecycle management, as well as lessons learned from production deployments on VSC-5 and CINECA Leonardo.

Additionally, we will explore how dagster-slurm positions itself relative to existing HPC workflow frameworks, such as Parsl [2], executorlib, and PSI/J, which it complements rather than replaces, targeting teams already invested in Dagster for the non-HPC parts of their data platform.

The project was developed with support from the Austrian Scientific Computing community and the EUROCC AI Hackathon 2025, and has been released with an accompanying JOSS paper [3].

References

[1] Dagster Labs., *Dagster: The data orchestration platform*, https://dagster.io (2024).

[2] Babuji, Y., et al., *Parsl: Pervasive Parallel Programming in Python*, 28th ACM International Symposium on High-Performance Parallel and Distributed Computing (2019).

[3] Picatto, H., *dagster-slurm: Bridging Cloud-Native Data Orchestration and HPC*, Journal of Open Source Software **volume**, firstPage (2025).

Technical Track:

Managed Multitenant ML Workflows on HPC

Gent Rexha and Endri Deliu

AI Factory Austria AI:AT

Motivation: High-Performance Computing (HPC) clusters provide the raw computational power required for training and running large-scale machine learning (ML) models, yet they typically lack the self-service workflow tooling that modern ML teams expect. In practice, this means researchers must understand low-level batch scheduler semantics, manually manage container images, and coordinate multi-step pipelines without standardized abstractions. When multiple research groups share the same cluster, the problem compounds: each team needs isolated environments, access controls, and visibility into its own jobs without interfering with others. This gap between the low-level interfaces that HPC systems provide and the self-service workflow experience that researchers need motivates a managed platform that lets domain scientists define and execute ML workflows without becoming HPC experts.

Architecture: We present the AI Factory Austria AI:AT Software Platform, which delivers managed, multi-tenant ML workflows on HPC infrastructure through three layers.

Compute abstraction. A unified API sits in front of the HPC batch scheduler (SLURM [2] in our deployment) and translates high-level job requests into scheduler-native submissions. It automatically selects a suitable container runtime, maps GPU resource requests to the scheduler's native allocation model, and configures per-user execution environments. By encapsulating these details behind a single endpoint, the platform frees researchers from writing scheduler scripts or managing container images by hand, while remaining portable to other batch schedulers that expose a programmatic interface.

Pipeline orchestration. A co-located Kubernetes cluster hosts a pipeline orchestration engine [1] that provides declarative, DAG-based ML workflow execution. Each tenant is isolated in a dedicated namespace with its own visualization service and artifact storage prefix; a shared workflow controller schedules and monitors pipeline steps across tenants. Users define and submit pipelines through a Python SDK without requiring direct cluster access, lowering the barrier for research groups that need reproducible, multi-step training and evaluation workflows.

Operations. Job logs from HPC compute nodes are collected with per-tenant and per-job tagging, then ingested into a centralized search engine that supports full-text querying across all workflow executions. Platform operators and tenants alike can trace individual pipeline runs from submission through completion. The entire infrastructure is managed declaratively through a GitOps workflow, enabling auditable, version-controlled changes to both platform services and tenant configurations, and reducing the operational burden of running a shared ML environment.

References

[1] Kubeflow Project, Kubeflow Pipelines – a platform for building and deploying ML workflows, https://www.kubeflow.org/docs/components/pipelines/ (accessed 2026).

[2] SchedMD, SLURM Workload Manager Documentation, https://slurm.schedmd.com/documentation.html (accessed 2026).

Keynote Talk:

From Atoms to Current: Predicting Function in Single-Molecule Circuits

Latha Venkataraman

Institute of Science and Technology Austria, Klosterneuburg, Austria
and
Lawrence Gussman Professor of Applied Physics, Professor of Chemistry Columbia University, New York, New York, United States

Over the past decade, the field of single-molecule electronics has witnessed remarkable advances in the measurement, modeling, and interpretation of structure–function relationships in molecular circuits. The development of reliable and reproducible single-molecule junction techniques has been central to this progress, enabling increasingly precise exploration of charge transport at the molecular scale [1,2]. Despite these achievements, the quantitative prediction of single-molecule conductance remains a formidable challenge [3].

Single-molecule circuits typically comprise organic backbones built from light elements, chemically bonded to heavy-metal electrodes—most commonly gold—where subtle interfacial effects critically shape transport properties. In this keynote, I will first outline the standard ab initio framework based on density functional theory (DFT) combined with non-equilibrium Green's functions (NEGF) for computing electronic transmission through molecular junctions, see Figure 1 for an illustration of a single-molecule device. I will then highlight two case studies demonstrating how this theoretical approach yields new physical insight into charge transport mechanisms and guides the rational design of functional single-molecule devices.

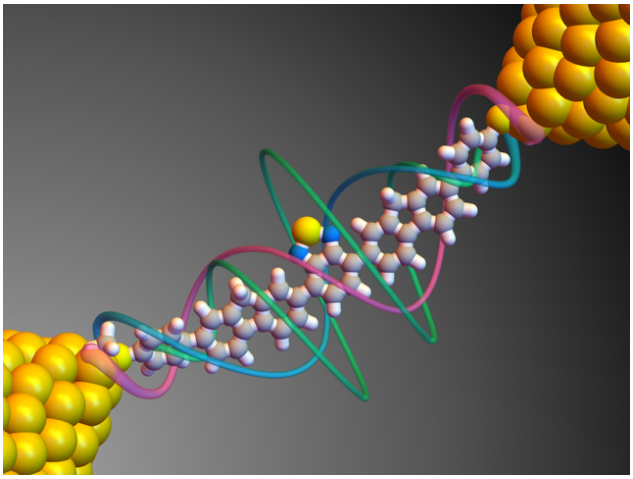


Fig. 1: Illustration of a single-molecule device.

References

[1] L. Venkataraman et al, Nature 442, 904–907 (2006).

[2] L. Li et al, Nature Chemistry 14, 1061–1067 (2022).

[3] F. Evers et al, Rev. Mod. Phys. 92, 035001 (2020).

EPICURE project – practical example

Darin Lah and Samo Miklavc

Institute of Information Science (IZUM), Slovenia

The EPICURE project is funded by the EuroHPC Joint Undertaking (JU) and delivers specialized technical support services to researchers granted access to EuroHPC supercomputers throughout the continent [1]. The project provides several services, including code enablement and scaling, performance analysis and benchmarking, code refactoring, and code optimization [2]. EPICURE also includes close collaboration with the different Competence Centers and Centers of Excellence. The project consortium includes 16 partners from 14 different European countries, including the Institute of Information Science (IZUM) and the Jozef Stefan Institute (JSI) from Slovenia. The only requirement to get EPICURE support is the EuroHPC allocation number. EPICURE is designed to support any High Performance Computing (HPC) executable software without limitation. Whether written in FORTRAN, Python, or other languages, and regardless of open-source or proprietary licensing, all software is eligible for support. EPICURE provides dedicated Level 2 (installation and porting) and Level 3 (performance tuning and code optimization) support, ensuring that user applications achieve optimal readiness and efficiency on target HPC architectures. EPICURE measures support success by tangible outcomes: improved runtime performance, lower energy-per-calculation footprints, and optimized allocation of computational HPC resources.

Presentation overview

Our presentation is divided into two parts. The first part introduces the EPICURE project, its consortium members, and the comprehensive support services it offers to EuroHPC JU users. The second part showcases concrete support cases delivered by IZUM and JSI to researchers who requested EPICURE assistance. Our presentation will showcase detailed, hands-on support delivered through concrete EPICURE projects, offering prospective users practical insight into how expert assistance translates computational challenges into successful scientific outcomes. We will walk through a representative case study where our team installed specialized software—including Boltz1 and Chai1-on the Vega supercomputer, followed by systematic benchmarking to establish performance baselines. Subsequently, we implemented several optimizations such as SLURM parameter tuning (GRES allocation, memory provisioning, and time-limit adjustments) to enable efficient large-scale experimentation. By presenting this end-to-end support journey-from software deployment through scaling optimizations—we illustrate EPICURE's value in transforming ambitious research visions into executable, high-performance workflows on EuroHPC infrastructure.

References

[1] EPICURE project, "About Us", *EPICURE* https://epicure-hpc.eu/about/about-us/ (accessed Jan. 29, 2026).

[2] David Vicente, "How an EPICURE Project Is Supported (with an Example from BSC)" *EPICURE blog*, https://epicure-hpc.eu/2025/08/06/how-an-epicure-project-is-supported-with-an-example-from-bsc/ (accessed Jan. 29, 2026).

RI-Scale – bringing data holdings and HPC together

Florian Goldenberg and Andreas Rauber

ASC Research Center, TU Wien, Austria

As the size and complexity of research infrastructure (RI) data repositories grow, there is a lack of available computational infrastructures for large scale analysis of such data. Solutions for replicating and managing RI data on third party infrastructures (like HPC clusters) are also very limited. Therefore, the EC funded project RI-Scale aims to empower RIs producing big-data with the technical means to extend their services with scalable computational platforms suited for the development and delivery of applications to improve, analyze and exploit their data at large [1].

To achieve the project goal, a Data Exploitation Platform (DEP) is being developed. This DEP will include the required data lifecycle management, including data transfer (using Rucio and FTS), data preparation procedures, data exploration tools, and the integration of data holding sites and computing infrastructure. The ASC Research Center clusters, especially MUSICA, will be one of the computing sites, with two additional clusters in Turkey (TUBITAK) and the Czech Republic (Masaryk University).

An authentication and authorization framework is being designed to enable interoperability and security, based on federated principles. This ensures common secure access to both data holdings and computing sites for all involved parties. A credit management system is added to track and account resources used.

Enabling the RIs to utilize the DEP without the need for their own software is another major goal, achieved by a technology and application stack that is being developed based on a scalable AI computing framework and an AI model hub. These will be provided ready to use for all participants.

In order to validate the whole concept, 8 scientific and 4 technical use cases will be tested within the scope of the RI-Scale project. The scientific cases are further subdivided into 4 from the area of environmental and climate sciences, and 4 from life sciences. The technical cases include, among other topics, developments in image compression and the use of TPU (Tensor Processing Units) in AI inference.

The ASC Research Center is primarily involved in providing a compute site and enabling data transfer and lifecycle management, as well as in the test of TPU systems.

References

[1] https://www.riscale.eu/

EVITA – EuroHPC Virtual Training Academy

Victoria Döller and Claudia Blaas-Schenner

ASC Research Center, TU Wien, Austria

The EVITA (EuroHPC Virtual Training Academy) project is a pan-European initiative dedicated to advancing qualification and training in High-Performance Computing (HPC) [1]. Funded by the EuroHPC Joint Undertaking and coordinated by the Barcelona Supercomputing Center, EVITA is a collaboration of leading universities, research institutions, supercomputing centers, and training providers with the distinguished goal of a comprehensive, high-quality learning ecosystem and training platform for the European HPC community.

The centerpiece of EVITA is the Competence and Qualification Framework (CQF) that defines the components, topics, and professional profiles required for the empowerment of a competitive HPC workforce. This framework defines quality guidelines and drives the development of aligned training material, that is divided into concise learning units – the EVITA Modules.

Fig. 1: The cornerstones of EVITA.

The definition of EVITA Modules is based on competencies articulated within the Skill Tree, a hierarchical structure that decomposes knowledge into granular skills from the understanding of foundational concepts to the application of specialized expertise [2]. EVITA Modules are designed for flexible composition, enabling the creation of tailored courses and personalized learning pathways. All materials are openly accessible, supporting both training providers seeking to adopt mature resources and individuals pursuing self-paced professional development. For an international recognition of passed examinations in HPC EVITA aims at a European-wide certification framework.

Call for Training Material The EVITA Modules will be solicited through a Cascade Funding mechanism comprising three open calls. Proficient HPC training providers are invited to submit proposals for EVITA Modules and courses and will receive financial support for contributing to EVITA's pool of high-quality training resources. The first call, to be launched in May, will focus on core HPC topics. Following a rigorous review process, selected providers will develop and align their modules with the CQF quality guidelines and the prescribed structural requirements using the module template. Subsequent calls will broaden the scope to include advanced topics and domain-specific applications of HPC.

Acknowledgement The project is supported by the European High Performance Computing Joint Undertaking and its members. Funded by the European Union. Grant agreement No. 101196394.

References

[1] https://www.evitahpc.eu

[2] https://www.hpc-certification.org/wiki/skill-tree/b

From City-Scale Street-View Imagery to Building-Level Urban Indicators: A Precompute Layer for GeoAI ML Models on the MUSICA HPC System

Silvio Heinze

Institute for Urban and Regional Research, Austrian Academy of Sciences, Austria

The MOSAIK project investigates micro-scale socioeconomic transformation in Vienna by automatically extracting built-environment indicators from large image archives, with a particular focus on the Vienna Kappazunder street-view dataset. The Kappazunder mobile mapping data comprise georeferenced high-resolution imagery, point clouds, navigation trajectories, and camera orientations, collected in 2020 and 2023 and released as Open Government Data (OGD). In contrast to increasingly restrictive street-view sources for machine-learning use (e.g., Google Street View), Kappazunder provides a rare combination of image quality, spatial precision, and licensing suitability for large-scale GeoAI workflows.

We develop neural-network pipelines to detect façade structure and condition, and supervised models to classify ground-floor uses. This enables, for the first time, comprehensive building-level mapping of built-functional characteristics across Vienna's densely built urban area, producing consistent city-wide layers. These layers support downstream geostatistical analysis (hotspot and cluster detection) and can be linked with socioeconomic registers (e.g., education, income, migration background) to study patterns of appreciation and depreciation and to evaluate whether social structures and dynamics can be inferred from image-derived signals. The computational demands of processing city-wide, high-resolution street-view data position MOSAIK as a distinctive HPC use case in urban studies.

This contribution focuses on a reproducible HPC "precompute layer" that standardizes the transformation of raw mobile mapping data into ML-ready artifacts for scalable recognition and classification. We implement (i) geometry-aware view extraction (façade-aligned patches), (ii) systematic image quality assessment (occlusion, blur, illumination, seasonal effects) to reduce bias in spatial inference, and (iii) deterministic sharding and versioned provenance to ensure reproducibility across model iterations. The pipeline produces building-linked image subsets, compact semantic summaries, and optional cached feature embeddings that accelerate downstream training, change analysis, and retrieval.

The workflow is executed on MUSICA (Multi-Site Computer Austria), leveraging heterogeneous CPU/GPU resources and high-throughput storage for large-scale preprocessing and inference. Building on earlier work on automated façade inventorying and interpretation of Vienna's Gründerzeit building stock and on extracting social structures from street-view imagery ([1,2]), the resulting indicator layers bridge computer vision and urban studies, enabling fine-grained spatial statistics and integrative socioeconomic analysis at a resolution unavailable from conventional data sources.

References

[1] Heinze, S., Guinand, S., and Musil, R., AI based evaluation of the Viennese Gründerzeit facades for their complete inventorying, understanding and preserving", XXXII International Seminar On Urban Form, Turin (2025).

[2] Heinze, S.,Musil, R., Unveiling urban dynamics: Extract social structures from Street View imagery with machine learning, 10th EUGEO Congress, Vienna (2025).

NCC Croatia Success Story: Setting up Photogrammetry Workflow on an HPC cluster - A Cultural Heritage Use Case

Vinko Đurić[a], Branimir Kolarek[b,c], Nenad Mijić[b], and Davor Davidović[b]

[a] *Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia*
[b] *Centre for Informatics and Computing, Ruđer Bošković Institute, Croatia*
[c] *Faculty of Graphic Arts, University of Zagreb, Croatia*

Photogrammetry, particularly the Structure-from-Motion (SfM) pipeline, has become a key technology for 3D documentation and digital preservation of cultural heritage sites and artefacts. While modern workstations provide a reliable baseline for standard datasets, they often become limiting when processing ultra-high-resolution data. Such datasets are crucial for high-fidelity digitisation of intricate artefacts and large architectural sites requiring millimetre-level precision. High-Performance Computing (HPC) environments address these demands, yet widely used software such as Agisoft Metashape often lacks native integration with HPC job schedulers.

In this work, we present a successful deployment of Agisoft Metashape on the national supercomputer "Supek". The primary technical challenges involved developing a command-line wrapper, automating the setup of Metashape's network processing environment, and managing software licences. The transformation of standard GUI-based workflows into a headless, scriptable executable allows seamless integration of photogrammetric tasks into the PBSPro scheduler, enabling automated resource allocation, multi-GPU affinity management, and efficient job queuing. This configuration enables the use of HPC resources even for less experienced users.

A critical component of the study involved fine-tuning configuration parameters, specifically mapping Agisoft's workers to MPI ranks and assigning GPUs to specific tasks to prevent hardware oversubscription. To validate the efficiency of this pipeline, we benchmarked the workflow using a high-resolution dataset comprising 330 photos of a sculpture by the Croatian sculptor Ivan Meštrović. Our preliminary performance analysis compares execution times on a high-end workstation equipped with a single GPU with those obtained on the HPC cluster. The analysis indicates that performance is strongly influenced by the mapping of workers and GPUs, as well as data partitioning strategies. Significant speedups were achieved, particularly during the mesh and model building phases, confirming that a well-orchestrated HPC environment can substantially reduce processing time.

References

[1] Kolarek, B., Davidović, D., Maričević, M. *Energy efficiency in photogrammetry: a comparative analysis of dataset, hardware, and resolution effects in Agisoft Metashape.* Proceedings 28th International conference on printing, design and graphic communication.

[2] Kolarek, B., Gamulin, Lj., Davidović, D. *Cultural Heritage on HPC - Creating High Resolution 3D Models Using Photogrammetry.* 47th ICT and Electronics Convention (MIPRO), IEEE, 2024. pp. 1121-1127.

[3] Kolarek, B., Davidović, D., Gamulin, Lj. *Applying Advanced Computational Infrastructure in Creating a High-Resolution 3D Model of the Small Fountain of Onofrio.* International Conference Digital Art History, Institute of Art History, Zagreb, 2025. pp 39-41.

HPC-Driven Dimension-Aware Neural Architecture Search for Cryocooler Lifetime Prediction

Gregor Molan[a] **and Martin Molan**[a,b]

[a] *Comtrade 360 d.o.o., Ljubljana, Slovenia*
[b] *Comtrade AI GmbH, Zug, Switzerland*

Introduction: We address non-destructive lifetime prediction for satellite cryocoolers under small-data constraints, with a focus on *high-performance computing (HPC)* methodology that enables capacity scaling and systematic architecture exploration at an industrial scale. To enrich neural architecture search (NAS), we introduce **dimension-aware NAS (da-NAS)**, which evaluates embedding sizes from 2–512.

Design: Experiments ran on the **EuroHPC Leonardo Booster** (CINECA) under grant **EHPC-DEV-2025D06-042**. The workload orchestrates a **family of foundation models (FFM)**, CNN1D, LSTM, GRU, Transformer, trained via self-supervised sequence-to-sequence objectives, integrated with da-NAS for embedding optimization. Two scheduling principles maximize throughput:

1. Low-dimension-first phase (dimensions 2, 4, 8, 16) to identify trends and prune branches quickly.
2. Concurrent CNN1D/Transformer sweeps for optimal node utilization.

Data filtering, normalization, and staging use the Leonardo filesystem to support high-throughput NAS job submission. Data filtering/normalization and dataset staging are performed on the Leonardo filesystem prior to training to enable high-throughput submission of independent NAS jobs.

What we ran at scale: The allocation supported massive parallel exploration of encoder families and embedding capacities; each NAS branch (model × dimension) trains a self-supervised encoder and evaluates downstream performance, with early-stop signals propagated across dimensions to avoid wasteful runs. This parallel design turns a prohibitively long serial search into a bounded wall-clock process within the allocation window.

Results and benefits of HPC: The Booster's parallelism substantially *reduced NAS time* versus local compute and enabled *broader coverage* of configurations critical for small-data reliability tasks. Early low-dimension results provided quick "capacity diagnostics", guiding which high-dimension branches to pursue. This led to *faster convergence* to robust architectures: Transformer variants reached peak performance in richer data regimes, while CNN1D/LSTM remained more stable under severe imbalance and minimal labels. The HPC-backed sweeps also enabled *systematic sensitivity analysis* to embedding size and *stress-testing* under label scarcity, analyses that need large-scale parallelism.

Conclusion: EuroHPC resources were instrumental in executing distributed self-supervised training and dimension-aware NAS at scale, yielding a practical, non-destructive pathway for cryocooler lifetime prediction. The *HPC-first* design, dimension scheduling, concurrent architecture sweeps, and filesystem-local preprocessing generalize to industrial time-series applications that require compact models, thorough capacity exploration, and rapid iteration under tight compute windows.

References

[1] Modi, A., et al., "Towards modular machine learning pipelines," ICML LLW, 2023.
[2] Smithson, S.C., et al., "Neural networks designing neural networks:…" ICCAD, 2016.

Simple is better? HPC-enabled neural architecture search for energy demand forecasting

Jelena Joksimović

Rudolfovo - Science and Technology Centre Novo Mesto, Novo Mesto, Slovenia

Accurate one-day energy demand forecasting at high temporal resolution is a key enabler for advanced energy management systems in buildings. This paper presents the results of an HPC-enabled neural architecture search (NAS) results for multi-horizon (96-step, 24-hour ahead) electricity consumption forecasting at 15-minute resolution using Bidirectional Long Short-Term Memory (BiLSTM) neural networks. The input data consists of historical energy consumption (kWh) recorded at 1-minute granularity and weather data corresponding to the geographical location of the building. From these sources, a feature set of 31 variables was constructed to capture temporal patterns and external influences on energy demand. The workflow is designed for scalable execution on high-performance computing infrastructure to overcome the computational limitations of on-premise systems when training and optimizing deep recurrent models. Neural architecture search was performed using Keras Tuner with the Hyperband algorithm over a broad hyperparameter space, including sequence length ∈ [192, 768] (step 48), number of units ∈ [60, 600] (step 60), number of layers ∈ 1, 2, 3, dropout ∈ [0.0, 0.5] (step 0.1), and learning rate ∈ [1e-5, 1e-2] sampled logarithmically. Candidate models were trained for up to 50 epochs and evaluated using validation Huber loss. HPC resources enabled the practical execution of NAS by allowing parallel exploration of candidate architectures. On the original CPU-only infrastructure, such hyperparameter search would require more than one month of computation, whereas HPC resources enabled efficient experimentation and model selection within a feasible timeframe. Despite the broad search space, the NAS identified a relatively simple architecture, a single-layer BiLSTM with moderate capacity as the best-performing solution, indicating that increased model complexity does not necessarily improve forecasting accuracy for building energy demand. The selected model reduced the 15-minute forecasting error (WMAPE) from 47.88% to 10.49% on a single building case study, showcasing the strong potential of HPC-driven model selection for scalable deployment in real-world energy management systems.

Acknowledgements: The authors' research was co-funded by the Republic of Slovenia, the Ministry of HE, Science and Innovation and the Slovenian Research and Innovation Agency of the European Union - NextGenerationEU through the DIGITOP project and by the Slovenian Research and Innovation Agency (ARIS) through the annual work program of Rudolfovo. The work is also funded through the business experiment Advanced Multi-objective Optimization of Energy Management - AIMED-HPC, which received funding through the FFplus project, which is financed by the European High-Performance Computing Joint Undertaking (JU) under the funding agreement No. 101163317. The JU receives support from the Horizon Europe program of the European Union.

Automated learning of multiscale models

Max Hodapp[a] **and Guillaume Anciaux**[b]

[a] *Christian Doppler Laboratory for Digital material design guidelines for mitigation of alloy embrittlement, Materials Center Leoben Forschung GmbH (MCL), Leoben (AT)*
[b] *Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne (CH)*

Machine-learning-based multiscale modeling connects different length and timescales by training a coarse-grained, but cheap, model on data coming from a fine-scale, but expensive, model. Corresponding training protocols can be very complex as they usually involve different simulation codes and contain many different stages (e.g., pre-training, active learning, fine-tuning, distillation, …) that are difficult to implement efficiently in a massively parallelized manner, and cumbersome to interpret, analyze, and reproduce.

For the problem of training machine-learning interatomic potentials (MLIPs) on quantum-mechanical simulations, we propose AutoPot [1], a software for automating the construction and archiving of MLIPs. AutoPot is based on BlackDynamite, a software operating parametric tasks, i.e., 1) producing configurations, 2) selecting configurations based on the MLIP's uncertainty, 3) single point evaluation of energy, forces, and stresses, and 4) training the potential. Interactions between the tasks are realized using a novel event-based workflow orchestrator, Motoko, that keeps every information produced by such tasks, order metadata in object-database, while providing an asynchronous python description of the global orchestration. As a demonstration, we present a fully functional workflow that supports selection of training configurations from large training candidate sets, on-the-fly selection from molecular dynamics simulations by active learning based on the D-optimality criterion, using Moment Tensor Potentials, a class of MLIPs, as implemented in MLIP-2, and single-point calculations of the selected training configurations using VASP. All details of the orchestration are accessible within an open-source repository [2].

Another strength of AutoPot is its flexibility: BlackDynamite tasks and orchestrators are Python functions to which own existing code can be easily added and manipulated without writing complex parsers. Therefore, it will be straightforward to add other MLIP and ab initio codes, and manipulate the Motoko orchestrators to implement other training protocols. Moreover, functionalities of AutoPot are not limited to training MLIPs but could be applied to other problems, such as learning continuum models from atomistic data.

In general, automation tools struggle to adapt to user-level permissions and constraints appearing when using HPC facilities. To that end, we also share our experiences and some of the pitfalls that we encountered while setting up AutoPot on the Vienna Scientific Cluster (VSC-5) provided by the Austrian Scientific Computing (ASC) facilities.

References

[1] Hodapp, M., & Anciaux, G. (2026). AutoPot: Automated and massively parallelized construction of Machine-Learning Potentials. arXiv preprint arXiv:2601.01185.

[2] Autopot: https://gitlab.com/mhodapp/autopot

Extracting Metallurgical Graphs using Reasoning LLMs on HPC

Manuel Hofbauer, Lukas Pichlmann, Johannes Kronsteiner, and Johannes A. Österreicher

LKR Light Metals Technologies, AIT Austrian Institute of Technology, Ranshofen, Austria

Introduction: Accelerating materials discovery requires unlocking experimental knowledge from unstructured literature. Standard tabular extraction often fails in metallurgy, where properties like yield strength depend on sequential processing steps (tempers) rather than composition alone [1]. We propose using reasoning-enhanced Large Language Models (LLMs) to extract data into Directed Acyclic Graphs (DAGs). Graphs offer an optimal compromise: enforcing strict typing for machine readability while preserving the topological timeline of processes—from homogenization to aging—that define the material's final state.

Methodology: We benchmarked open-source LLMs using a prompting strategy for lineage-aware JSON extraction. Node types (Source, Base Material, Process, Attribute) follow strict schemas, though models may flag new keys for missing parameters. To bypass high variance in expert ground truths, we assessed accuracy via head-to-head tournament rankings of "border-case" papers, prioritizing topological integrity over string matching.

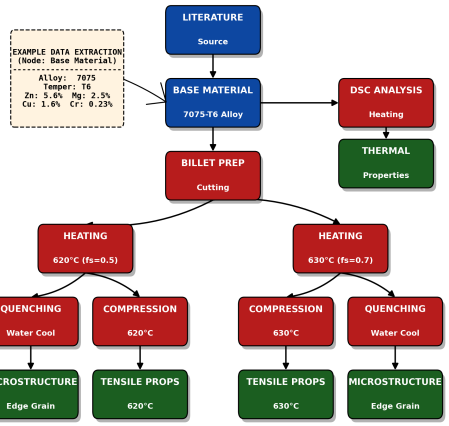


Fig. 1: Extracted Graph Example showing the lineage from source to attributes.

HPC Deployment Challenges: A major aspect of this work was deploying quantized reasoning models on the Leonardo supercomputer. After facing persistent compatibility issues with vLLM, we transitioned to the SGLang inference engine [2], though still encountering severe friction with INT4 quantization on Ampere architecture and NCCL networking timeouts.

Results: Standard models frequently failed to maintain valid graph topology or hallucinated relationships. Only reasoning-oriented models provided consistent, topologically valid outputs, with Kimi K2 Thinking (temperature 0.6) achieving the highest accuracy in reconstructing complex processing chains.

HPC Contribution: We stabilized the pipeline on 4 nodes of the Leonardo Booster partition, achieving ∼20 tokens/s across 8 parallel extractions. This demonstrates that A100 hardware can efficiently run next-gen reasoning models given the correct configuration. Our optimized launch scripts and comprehensive SGLang/NCCL documentation have been released on Codeberg [3].

References

[1] Tshitoyan, V., et al., Nature **571**, 95 (2019).

[2] Zheng, L., et al., arXiv preprint arXiv:2312.07104 (2023).

[3] Project Repository: *LLMonHPC*, Codeberg, https://codeberg.org/LKR/LLMonHPC, 2026.

Technical Track:

Organizing the software stack in CINECA's clusters. Towards LISA.

Orlenys Troconis

Cineca - HPC Department, Casalecchio di Reno (BO), Italy

The Italian Supercomputing Centre CINECA is continuously growing as part of an European network, hosting and managing cutting-edge technologies for HPC and Artificial Intelligence (AI). The Tier-0 HPC system **Leonardo** is hosted at the Bologna Technopole in Italy and it is supplied by Bull. It started its production on 2022 and currently counts a Data Centric General Purpose (DCGP) and a Booster partition: the first one is equipped with 1536 nodes, with Intel Sapphire Rapids CPUs; the second one with 3456 nodes accelerated by 4 customized A100 Nvidia GPUs per node.

In few months **LISA** will enhance Leonardo's capabilities, by adding 2.5 exaflops of FP8 performance optimized for AI and machine learning, thanks to 166 compute nodes with 8 H100 Nvidia GPUs per node. LISA will also introduce 8-way GPU nodes for greater acceleration in AI workloads. Its fat-tree non-blocking NDR InfiniBand network fabric will enhance the performance of Large Language Models (LLMs) training and AI applications by ensuring fast and efficient inter-node communication and data transfer. The infrastructure and services of the Italian AI Factory, IT4LIA, will then grow and provide increasing support to the Italian and Europen academic and industrial research.

During the presentation I am going to give you a brief introduction about the main steps that allow us to organize the software stack of our HPC systems relaying on Spack package manager in such a way to satisfy the needs of the different communities and users. I would like to describe the main organization steps from the CINECA's point of view and also provide hints from the user point of view who would like to install its own stack making use of Spack as a tool.

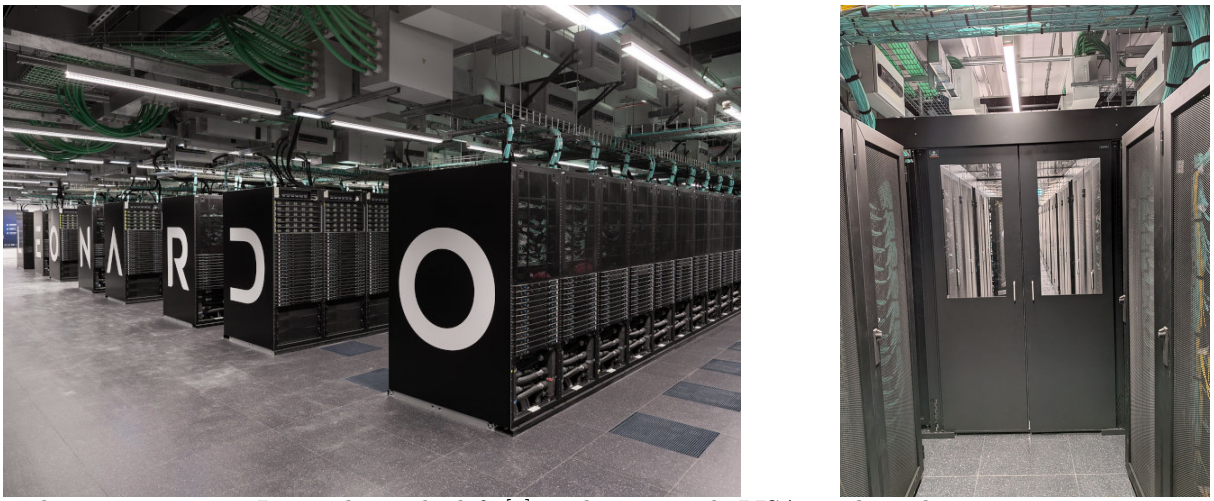


Fig. 1: The supercomputer Leonardo on the left [1], and its upgrade LISA on the right.

References

[1] https://leonardo-supercomputer.cineca.eu/

Technical Track:

Design and Operation of a Federated GPU Cluster for Digital Humanities within DHinfra.at

Florian Atzenhofer-Baumgartner[a], David Fleischhacker[a], Max Resch[b], Lukas Waldhofer[a], and Michael Otto[a]

[a] *Department of Digital Humanities, University of Graz, Austria*
[b] *Center for Cultures and Technologies of Collecting, University for Continuing Education Krems, Austria*

Context: Earlier plans for a federated GPU cluster for Digital Humanities (DH) were presented at ASHPC24 [1]. This talk reports on progress since. The DHinfra.at project now operates infrastructure accessible to nine ACOnet institutions, with planned CLARIAH-AT expansion. Computational demands—LLMs, handwritten text recognition, corpus analysis—continue to grow; governance and usability requirements differ from traditional HPC settings.

Hardware and Network: The primary site (Graz) comprises a head node and compute nodes with NVIDIA H200 (MIG-capable), RTX Pro 6000, and L40S GPUs, interconnected via 400 Gbit InfiniBand (NFS/RDMA, NCCL). ZFS provides home, project, model, and scratch volumes. A second site at Krems offers additional capacity with Ceph storage and dedicated VMs. Real-world constraints—power budget, network provisioning, emergency power signalling—shaped deployment choices.

Software Stack and Access: The stack uses open-source components: Slurm for scheduling, Proxmox and Apptainer for (rootless) GPU containers, Prometheus/Grafana with DCGM exporters for monitoring. We broker ACOnet logins via Authentik, supporting SAML and OpenID Connect. Users access the cluster through JupyterHub, SSH over VPN, or REST APIs. We experiment with NixOS for reproducible configuration. Traefik serves as reverse proxy. Inspired by multi-tenant research platforms [2], we aim for sustainable operations at smaller scale.

Operations: The clusters operate in dedicated networks, independent of but complementary to central university HPC, leveraging institutional resources such as network and identity federation. As a small team, we prioritize simplicity, security, and documentation, pursuing infrastructure as code. Early experience shows that prioritized access exposes optimization needs in user applications—we address this through monitoring and user engagement rather than strict enforcement, accepting lower peak utilization as a trade-off for accessibility.

Outlook: Apart from DH projects, pilots span web security analysis, adaptive AI tutoring, privacy-preserving cybersecurity, video retrieval, and dependency resolution. Open questions remain: coupling federated identity with institutional quotas, caching strategies, and batch scheduling differentiation. We seek community feedback.

References

[1] Atzenhofer-Baumgartner, F. HPC and DHInfra. In *Austrian-Slovenian HPC Meeting 2024—ASHPC24*, Grundlsee, Austria (2024).

[2] Weitzel, D. et al., The National Research Platform: Stretched, Multi-Tenant, Scientific Kubernetes Cluster. In *PEARC '25*, ACM (2025).

Technical Track:

LiSC software catalog: a software installation framework for Life Sciences

Thomas Rattei[a], Robert Happel[b], Jan-Lukas Hodics[a], Michael Neumayer[a], and Marcel Rennig[a]

[a] *Centre for Microbiology and Environmental Systems Science, Centre for Microbiology and Environmental Systems Science*
[b] *Max Perutz Labs, University of Vienna, Austria*

The Life Science Compute Cluster (LiSC) is a shared data management and high-performance computing infrastructure for bioinformatics and computational life science at the University of Vienna. It is mainly, but not exclusively, used by students and members of four organizational units of the University of Vienna: Centre for Microbiology and Environmental Systems Science, Faculty of Chemistry, Faculty of Life Science, Max Perutz Labs.

LiSC users share data management and scientific compute needs that are typical for Life Sciences. These include the heterogeneous and dynamic method and software space as well as the need for AI and statistical methods applied to large data volumes. Nevertheless, most LiSC users have little or no computing background.

In 2025, the LiSC team has replaced the previous, proprietary framework for the central installation of software at LiSC. We have replaced it by a new, efficient and user-friendly framework that is based on established standards. To handle the dependencies of installed software efficiently, we use two main installation methods:

1. EasyBuild, a software build and installation framework especially for HPC systems
2. Conda, an open-source, cross-platform and language-agnostic package manager and environment management system, which is widely used in Life Sciences.

EasyBuild and Conda allow to install software reproducibly and efficiently. We prefer EasyBuild installations, whenever possible, to allow the user to load multiple modules at once. Conda is only used for software with very specific dependencies. Our Conda platform can be loaded as module and does not change the configurations of users' shells.

In addition to the repository of EasyConfigs and Conda environment files, the LiSC software catalog contains an automatic workflow that generates a user-friendly and searchable webpage for each installed software, including links to the original documentation and usage recommendations. A weekly maintenance script checks all installed software for new versions and allows the LiSC team to install them pro-actively.

By end of January 2026, 756 modules with 2560 versions and 87 Conda environments with 121 versions have been installed. By that date, the LiSC software catalog comprised 302 pages. The feedback by LiSC users about the new software installation framework was very positive so far. The LiSC team is ready to share this framework with other HPC sites.

Technical Track:

Zero-Touch HPC Nodes: NetBox, Tofu and Packer for a Self-Configuring SLURM Cluster

Ümit Seren and Leon Schwarzäugl

Vienna BioCenter

Over the last five years, we ran an HPC system for life sciences on top of OpenStack, with a deployment pipeline built from Ansible and involving manual steps. It worked—but it wasn't something we could easily rebuild from scratch or apply consistently to other parts of our infrastructure.

As we designed our new HPC system (coming online in 2026), we set ourselves a goal: treat the cluster as something we can declare and then recreate, not pet and nurture. The result is a "zero-touch" style pipeline where a new node can go from "just racked" to "in SLURM and running jobs" with no manual intervention.

In this talk, we walk through the end-to-end workflow:

1. NetBox as DCIM and source of truth: racking a server and adding it to NetBox is the trigger; MACs, serials and IPs are automatically imported from vendor tools and IPAM/DNS into our automation.
2. Using Tofu/Terragrunt (instead of Openstack's Heat orchestration service) to provision OpenStack/Ironic, SLURM infrastructure and network fabric across three environments (dev plus two interchangeable prod clusters for blue/green rollouts).
3. Image-based deployment with Packer and Ansible: we split roles into "install" and "configure". Packages and heavy setup are baked into images, while an ansible-init service runs locally on first boot to apply configuration and join the cluster.
4. Making nodes self-sufficient, including fetching the secrets they need via short-lived credentials and a minimal external dependency chain.

Come and see how we built a reproducible HPC/Big-Data cluster on open-source tooling, reusing as much of the stack as possible for the rest of our infrastructure.

Technical Track:

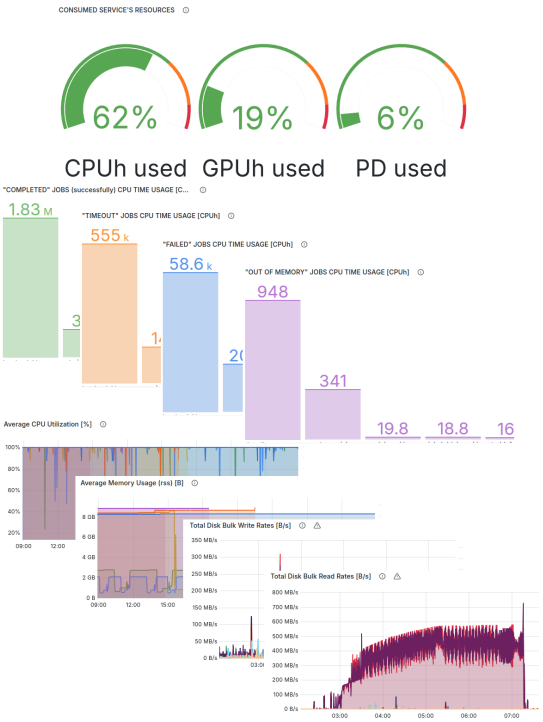
HPC Info: Enhanced SLURM Job Resource Monitoring

Sebastian Sitkiewicz

Wroclaw Centre for Networking and Supercomputing (WCNS),
Wroclaw University of Science and Technology, Wroclaw, Poland

High-performance computing (HPC) centers provide critical resources for scientific research and industry, yet users and administrators often face challenges in monitoring computational resource utilization. The widely used SLURM queueing system [1], while very effective for job scheduling and resource management, offers limited insight into resource consumption by users' computational jobs. The SLURM built-in solutions do not provide the required fine-grained reporting on the SLURM account usage, nor the time series of resource usage within the users' jobs.

To address this issue, we present the *HPC Info* suite, a job statistics monitoring system based on VictoriaMetrics [2] that is tailored for the SLURM environments. HPC Info empowers users to track their computational resource usage with ease through Grafana's graphical interface [3], where they can safely browse their job history and check the resources on their SLURM accounts. Moreover, HPC Info provides users direct access to the granular job performance metrics such as the time series of utilization of CPU cores, GPU cards, IO devices and the Lustre filesystems. HPC Info aids the generation of detailed summaries and reports, streamlining account management and oversight—the computational grant owners have straightforward information on how much computational hours were spent on completed and failed jobs by each user under their grant, whereas for the HPC administrators it facilitates the detection of users that incorrectly utilize the supercomputer resources.



This presentation will outline the features of the HPC Info suite and its implementation at the production HPC system at WCNS. We will present showcases from different perspectives—regular HPC users, computational grant and service managers, and HPC administrators, as well as discuss extensions of the system to further enhance its functionality.

References

[1] Jette, M.A. and Wickberg, T. *"Architecture of the Slurm Workload Manager"* in "Job Scheduling Strategies for Parallel Processing", 3 (2023) (eds Klusáček, D. *et al.*), Springer Nature Switzerland

[2] VictoriaMetrics software, https://victoriametrics.com (accessed on 22.01.2026)

[3] Grafana software, https://grafana.com (accessed on 22.01.2026)

Keynote Talk:

Learning from molecular simulations

Gerhard Hummer

Max Planck Institute of Biophysics, Frankfurt, Germany

Molecular dynamics (MD) simulations allow us to probe complex molecular processes, from the nucleation of crystals to the function of the molecular machineries of life. In my presentation, I will showcase the power of traditional and emerging ways of using MD simulations. By taking advantage of powerful HPC setups, we can now tackle systems that approach organellar scales. By integrating MD with artificial intelligence (AI), we autonomously construct quantitative mechanistic models of complex molecular events, validate the models in real time, and use the validated models to guide and accelerate the sampling in a closed loop of simulation and learning.

Open-Boundary Molecular Dynamics of Red Blood Cell Suspensions

Maša Lah[a,b], Tilen Potisk[a,b], and Matej Praprotnik[a,b,c]

[a] *Laboratory for Molecular Modeling, National Institute of Chemistry, SI-1001 Ljubljana, Slovenia*
[b] *Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, SI-1000 Ljubljana, Slovenia*
[c] *Universitat de Barcelona Institute of Complex Systems, 08028 Barcelona, Spain*

Blood is a complex suspension of cells whose macroscopic flow properties are governed primarily by red blood cells (RBCs), which account for more than 95% of the cellular content. The deformability of RBC membranes, hematocrit, defined as the volume fraction of RBCs in blood, and the wide range of shear rates encountered in the circulation strongly influence the apparent viscosity of blood. Capturing such behavior requires cell-resolved simulations capable of reproducing both the nonlinear rheology of bulk blood and the heterogeneous microstructures that arise in flow.

Among mesoscale methods, dissipative particle dynamics (DPD), lattice Boltzmann methods, and smoothed particle hydrodynamics have been applied to study RBC deformation, aggregation, and suspension rheology. While these approaches successfully reproduce many aspects of blood flow, they typically rely on periodic boundary conditions (PBCs) to approximate bulk behavior. PBCs are efficient, since direct simulation of macroscopic systems is prohibitive, but they are not well suited for non-equilibrium conditions as they impose artificial correlations at the boundaries and require continuous external forcing to sustain flows. In contrast, real flows are maintained through ongoing mass and momentum exchange with the surroundings.

We present the first application of open-boundary molecular dynamics (OBMD) to RBC suspensions, with explicit control of flux exchange across the open boundary [1]. We introduce a novel membrane insertion algorithm that enables simulations at high hematocrit. RBC membranes are inserted at a reduced size and smoothly expanded to their full dimensions while traversing the boundary region toward the bulk, allowing cells to deform and rearrange as they enter the dense suspension. The solvent and cytosol are modeled using DPD fluids, while RBC membranes are represented as triangulated surfaces with node forces derived from continuum elasticity theory. To address the high computational demands of these simulations, we implemented the OBMD functionality within Mirheo [2], a GPU-accelerated code efficiently designed for large-scale HPC simulations. Benchmarking on simple DPD fluids demonstrates that the OBMD functionality introduces an overhead of approximately 15% compared to standard PBC simulations. However, the computational cost of the particle insertion logic depends on specific simulation conditions including the system density, the interaction potentials, and the particle flux. Here, we performed simulations of up to 55% hematocrit, involving more than 500 individual RBCs.

The framework reproduces experimentally established bulk hemorheological properties, including shear-thinning and hematocrit-dependent viscosity. These results demonstrate the applicability of OBMD to blood rheology and establish a computational foundation for future studies of ultrasound–blood interactions and other phenomena where periodic boundaries constrain natural dynamics, such as pressure-driven flows, transient inflows, and cell-free layer formation.

References

[1] Lah M., Potisk T., and Praprotnik M., *J. Chem. Phys.* 164, 104107 (2026).

[2] Alexeev D., Amoudruz L., Litvinov S., Koumoutsakos P., *Comput. Phys. Commun.* 254 (2020).

SIMD, GPU, and MPI Acceleration of Density-Based Tree Construction for Large-Scale Sequence Data

Thomas Haschka

E020-04 Service Unit of High Performance Computing, DataLab, Campus IT, Technische Universität Wien, Operngasse 11, 1040 Vienna, Austria

MNHN-Tree-Tools[1] is a high-performance software library for the construction and analysis of large hierarchical tree structures in computational biology, taxonomy, and phylogenetics. Tree construction is based on density-driven algorithms derived from large-scale sequence distance relationships rather than classical recursive traversal or topology-driven methods. This approach enables robust inference of hierarchical relationships from large collections of biological sequences, but also leads to a substantial computational burden dominated by pairwise distance evaluations. The rapid growth of biological datasets, combined with the increasing availability of heterogeneous high-performance computing (HPC) platforms, motivated a systematic parallelization effort during the implementation and further development of MNHN-Tree-Tools.

This work presents the porting of MNHN-Tree-Tools to modern HPC systems, introducing parallelism at three complementary levels: processor-level SIMD vectorization, accelerator-based parallelism, and distributed-memory scalability. At the CPU level, data representations and computational kernels were redesigned to enable explicit SIMD vectorization using AVX instruction sets. These optimizations specifically target sequence comparison, distance evaluation, and density estimation, which form the computational core of the tree-building process. As a result, MNHN-Tree-Tools now contains what are arguably some of the fastest bioinformatics implementations for standard tasks such as k-merization and, more generally, DNA pattern detection.

To further accelerate the dominant distance computations, GPU acceleration was introduced. In particular, the Smith–Waterman sequence alignment and distance metric was implemented using OpenCL, allowing efficient execution on a wide range of accelerator hardware. The GPU implementation is designed to expose fine-grained parallelism inherent in sequence alignment while minimizing data transfer overheads between host and device. Similar algorithmic considerations apply to the optimized CPU implementation, ensuring consistent behavior across heterogeneous architectures.

At the multi-node level, MNHN-Tree-Tools was extended with MPI-based distributed parallelism to enable scalable execution on HPC clusters. Sequence datasets and associated distance computations are partitioned across processes, allowing density estimation and tree construction to scale beyond a single node with limited synchronization and communication overhead. This enables the analysis of datasets that would otherwise exceed the computational capacity of a single system.

Performance results demonstrate substantial speedups and good strong and weak scaling on multi-core, GPU-accelerated, and multi-node systems. This work shows how density-based phylogenetic tree inference algorithms can be systematically adapted to modern heterogeneous HPC architectures, enabling large-scale sequence analysis beyond single-node limits.

References

[1] Haschka, T., Ponger, L., Escudé, C. and Mozziconacci, Bioinformatics **37 (21)**, 3947-3949 (2021).

Efficient inference of overdamped Langevin models from projected molecular dynamics trajectories

Anže Hubman[a,b] **and Franci Merzel**[a]

[a] *Laboratory for Molecular Structural Dynamics, Theory department, National Institute of Chemistry, Slovenia*
[b] *Faculty of Mathematics and Physics, University of Ljubljana, Slovenia*

Molecular dynamics (MD) simulations have become an indispensable tool in biophysics, materials science, and drug design. The development of efficient open-sourced simulation software packages, together with the widespread availability of high-performance computing (HPC) resources, have made it possible to model complex phenomena across multiple spatial and temporal scales. Exciting applications include the characterization of viral capsid dynamics [1] as well as understanding the thermodynamic forces driving self-assembly of biomolecules [2].

Modern MD simulations generate enormous amounts of data in the form of particle trajectories, making the analysis and interpretation of such high-dimensional datasets a significant challenge. A particularly effective strategy, rooted in the Langevin formalism, is to project the trajectories onto a low-dimensional set of collective variables q, defined as functions of the generalized coordinates that resolve key states of the system in configuration space. Assuming that the time evolution of q follows the overdamped Langevin equation, constructing an appropriate model then reduces to determining the potential of the mean force $F(q)$ (i.e. the free-energy) acting on q and the corresponding position-dependent diffusion coefficient $D(q)$. A key advantage of this approach is that it enables a robust estimation of transition rates between states in configuration space, which is difficult to achieve with standard MD simulations. Furthermore, a suitable physics-based choice of relevant collective variables results in a low-dimensional interpretable model of the underlying dynamics.

In this work [3], we present an efficient numerical scheme for parameterizing overdamped Langevin models along selected collective variables from projected MD trajectories. Our main contribution is the design of a loss function that maximizes the agreement between the analytical short-time propagators of an overdamped Langevin model and those estimated directly from projected MD trajectories. In contrast to maximum-likelihood-based methods, evaluation of this loss function is independent of the trajectory length, making it particularly well suited for large datasets. We further show how it can be naturally combined with adaptive Monte Carlo moves for efficient optimization. To illustrate the robustness of the method, we apply it to two model systems undergoing diffusive dynamics under both equilibrium and nonequilibrium conditions, as well as to water transport across the interface of a biomolecular condensate. Finally, we discuss extensions of the approach to multidimensional examples, which is a numerically nontrivial task.

References

[1] Perilla, J.R., and Schulten, K., Nat. Commun. **8**, 15959 (2017).

[2] Benayad, Z., von Bulow, S., Stelzl, L.S., and Hummer, G., J. Chem. Theor. Comput. **17**, 525 (2020).

[3] Hubman, A., and Merzel, F., bioRxiv, 2026.01.01.697292 (2026).

General Matrix-Matrix Multiplication and NVIDIA Tensor Cores Applied to the Lattice Boltzmann Method

Relindis Rott[a], René Prieler[b], Michael Landl[b], Siegfried Höfinger[c], and Christoph Hochenauer[b]

[a] *Virtual Vehicle Research GmbH, Graz, Austria*
[b] *Institute of Thermal Engineering, Graz University of Technology, Austria*
[c] *ASC Research Center, TU Wien, Austria*

The Lattice Boltzmann Method (LBM) is a mesoscopic method in between microscopic particle simulation models and the macroscopic Navier-Stokes fluid simulation. On a regular lattice, q particle distribution functions exist for each lattice node. These distributions correspond to different spatial directions. Fluid packets stream along lattice links and collide at lattice nodes, according to the chosen velocity set $DdQq$ in d spatial dimensions. The method is well-suited for parallelization. The multi-relaxation-time (MRT) collision operator provides good numerical stability. It is based on a transformation of particle distributions to moment space, a relaxation of moments to equilibrium and a back-transformation. The two transformations can be reformulated as general matrix-matrix multiplication (GEMM) for which soft- and hardware acceleration is studied.
Recent developments in artificial intelligence have led to the emergence of various types of hardware accelerators in particular for GEMM. Tensor Cores (TCs) are NVIDIA's special purpose matrix-multiplication cores. Tensor processors can generally perform $\mathbf{D} = \mathbf{A} \cdot \mathbf{B} + \mathbf{C}$ of hardware-defined matrix-sizes in a single cycle. Efficient matrix-multiplication of large matrices is achieved by splitting matrices depending on matrix-core hardware-sizes. The BLAS (basic linear algebra subroutines) software library provides software optimization for linear algebra routines, including GEMM. NVIDIA's implementation named cuBLAS, can deploy TCs. In the present study, the application of GEMM routines and TCs to the MRT operator of LBM based on the BLAS library is analyzed.

Execution times are analyzed for code sections of the MRT-operator, based on the benchmark problem of a lid-driven cavity (LDC), taken on NVIDIA's A100 GPU, an enterprise-grade GPU of the Ampere micro-architecture, provided by the VSC 5 of the Austrian Scientific Cluster (ASC). Different domain sizes (total lattice sizes) are analyzed to compare problem sizes and study scaling effects. The results are compared for single (FP32) and double precision (FP64) floating point data. CuBLAS API routines do not support a programmatic activation of TCs [1]. The activation of TCs is governed by internal heuristics and optimization strategies and depends on the problem sizes. This guarantees best run-times for any problem size.
The activation of TCs for different domain sizes is monitored. A comparison with naive CUDA kernels shows the performance advantage of TCs, depending on the problem size.
The activity of TCs is analyzed in the cuBLAS GEMM routines. In addition, the timing results of NVIDIA A100 GPU were compared to timings executed on the consumer-grade GPU NVIDIA RTX 2080, which includes only TCs for single precision (FP32) computation.

References

[1] https://developer.nvidia.com/cublas

GPU algorithm for efficient runtime detection of coalescence and breakup events in phase-field multiphase flows

Lea Enzenberger[a], Diego Perissutti[a,b], Domenico Zaza[a,b], Alessio Roccon[a,b], and Alfredo Soldati[a,b]

[a] *Institute of Fluid Mechanics and Heat Transfer, Vienna University of Technology (TU Wien), Austria*
[b] *Polytechnic Department of Engineering and Architecture, University of Udine (UNIUD), Italy*

An integrated GPU-oriented computational framework is introduced for large-scale, interface-resolved simulations of multiphase turbulence, combining a high-performance flow solver with runtime droplet analysis capabilities. The core solver, MHIT36 [1], performs direct numerical simulation (DNS) of the incompressible Navier–Stokes equations coupled with a phase-field method to capture interfacial dynamics. Simulations are conducted in a triply periodic domain representative of homogeneous isotropic turbulence. Transport of the phase-field variable is described using the Accurate Conservative Diffuse Interface (ACDI) [2] formulation. From a computational perspective, MHIT36 is designed for modern GPU-accelerated architectures. The code employs a two-dimensional domain decomposition with MPI parallelism, using the cuDecomp library [3] for pencil transpositions and halo exchanges, and cuFFT together with OpenACC directives to offload compute-intensive kernels to GPUs. This strategy delivers excellent scalability while retaining a modular and extensible code structure. Building on this solver, we introduce GALENE36, a fully GPU-accelerated and parallel module for runtime droplet identification, tracking, and event detection in diffuse-interface simulations. Droplet identification relies on connected-component labeling (CCL) applied to a thresholded phase field, a well-established approach in image analysis but still challenging to deploy efficiently in large-scale, three-dimensional, distributed GPU simulations. GALENE36 implements a GPU-optimized CCL strategy tailored to structured Cartesian grids and domain-decomposed data, minimizing synchronization and communication overhead while preserving scalability. Temporal tracking is performed using a voxel-overlap criterion between labeled droplets at consecutive time steps, enabling robust association of parent and child structures. This overlap-based approach allows reliable detection and classification of breakup and coalescence events, while naturally accounting for the diffuse-interface representation inherent to phase-field methods. Together, MHIT36 and GALENE36 form an integrated computational application that couples high-fidelity multiphase DNS with native, runtime droplet statistics, avoiding costly post-processing and reducing the storage demands.

References

[1] Roccon, A., Enzenberger, L., Zaza, D. and Soldati, A., MHIT36: A phase-field code for GPU simulations of multiphase homogeneous isotropic turbulence, Comput. Phys. Commun. **316**, 109804 (2025).

[2] Jain, S.S., Accurate conservative phase-field method for simulation of two-phase flows, J. Comput. Phys. **469**, 111529 (2022).

[3] Romero, J., Costa, P., Fatica, M., Distributed-memory simulations of turbulent flows on modern GPU systems using an adaptive pencil decomposition library, PASC Proceedings, pp. 1-11 (2022).

Numerical analysis of electrocoagulation using computational fluid dynamics for sustainable water treatment

Amina Tahreen

Department of Chemical Engineering and Sustainability, International Islamic University Malaysia, Malaysia

Electrocoagulation (EC) is an effective and environmentally sustainable water treatment technology for the removal of contaminants. Despite its demonstrated efficiency at laboratory scale, the design and optimization of EC reactors remain challenging due to the strong coupling between hydrodynamics, electric fields, electrochemical reactions, and species transport [1]. Experimental investigation of these interactions is often time-consuming and resource-intensive, motivating the development of complex numerical models to support reactor analysis and scale-up [2]. In this study, a three-dimensional transient Computational Fluid Dynamics (CFD) model of a batch EC reactor with a working volume of 250 mL was developed. The model resolves electric potential distribution, and species transport within the reactor domain.

The computational domain was discretized using progressively refined meshes ranging from 6,139 to 46,737 finite elements, corresponding to approximately 10,030–69,989 degrees of freedom (DoF). To evaluate computational scalability, simulations were executed using single-core and dual-core parallel configurations. Runtime measurements show that simulation times increased with mesh resolution, ranging from approximately 1 s for the coarsest mesh to 4 s for the finest mesh using a single core. Parallel execution with two cores provided moderate reductions in runtime for larger meshes, achieving speedups of up to 1.33 times, corresponding to a parallel efficiency.

The CFD results reveal non-uniform electric field distributions and localized regions of enhanced transport near the electrode surfaces [3]. Increasing current density resulted in stronger electric potential gradients and increased predicted aluminum release rates. Although simulations in this study were performed at laboratory scale, the numerical formulation was structured for parallel execution, allowing future extension to finer meshes, longer transient simulations, and larger reactor geometries. Such extensions would benefit from high-performance computing resources to reduce computation time and enable broader parametric exploration. Overall, this work demonstrates the role of CFD as a scalable digital tool for EC reactor analysis and supports its integration into sustainable water treatment research and design.

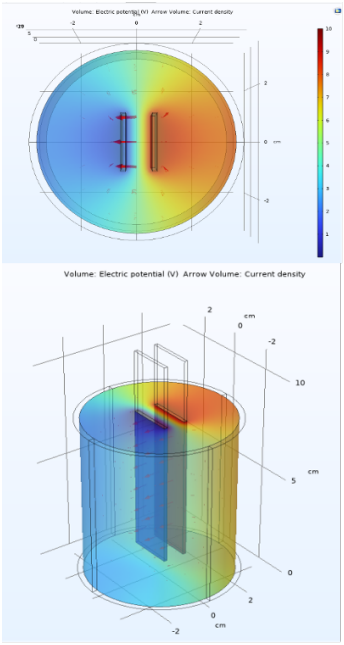


Fig. 1: Electric potential of simulated EC reactor. [3]

References

[1] Tahreen, A., Jami, M.S., Ali, F., Yasin, N.M.F.M., and Ngabura, M., Pollution **7**, 617 (2021).

[2] Tahreen, A., Jami, M.S., and Ali, F., J. Water Process Eng. **37**, 101440 (2020).

[3] Tahreen, A., Jami, M.S., Iwata, M., and Ali, F., PSETC, 1-5 (2025).

HPC-Enabled AI-Agent-Driven Digital Twin Framework for Real-Time Exergoeconomic Optimization of PV-Supported Thermal Energy Systems

Antonija Rajic

Faculty of Electrical Engineering and Information Technology, TU Wien, Austria

Introduction: Digital twins and model predictive control (MPC) have recently emerged as powerful approaches for improving the efficiency and operational performance of modern building energy systems [2]. However, most existing digital twin implementations primarily focus on minimizing operational energy costs and are typically executed on conventional computing infrastructure. The integration of thermodynamic quality indicators such as exergy into real-time control frameworks remains relatively unexplored, while the computational complexity of predictive multi-energy system optimization is rarely addressed using high-performance computing (HPC) environments [3].

The overall architecture of the proposed framework is illustrated in Fig. 1. The physical energy system consisting of a photovoltaic generator, heat pump, thermal storage unit, and building thermal mass is coupled with a digital twin model that receives sensor measurements and forecast data. An AI-driven nonlinear MPC controller determines optimal control actions, while computationally intensive predictive simulations and scenario evaluations are executed on HPC infrastructure to ensure real-time feasibility of the optimization process [3]. Exergy destruction within system components is quantified to capture irreversible thermodynamic losses and incorporated into the optimization objective together with electricity costs and thermal comfort constraints [1].

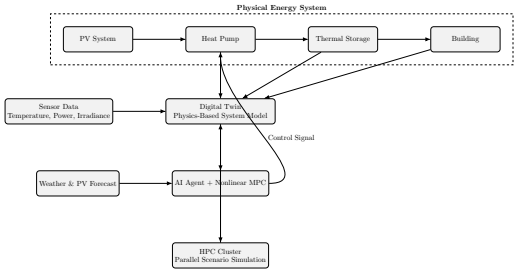


Fig. 1: HPC-enabled digital twin for PV-supported thermal systems.

Computational Framework: Depending on the prediction horizon and the number of forecast scenarios, a single MPC optimization cycle may require hundreds to thousands of dynamic simulations. When considering N prediction steps, S forecast scenarios, and C system components, the computational complexity scales approximately as $\mathcal{O}(N \cdot S \cdot C)$. To ensure real-time feasibility, the Digital Twin framework distributes simulation and optimization tasks across HPC infrastructure. An AI-based orchestration layer distributes scenario evaluations and predictive simulations across distributed HPC compute nodes, significantly accelerating the optimization process. The predictive optimization process is executed periodically within a receding-horizon control scheme, where new measurements and forecast updates trigger a new optimization cycle.

References

[1] G. Tsatsaronis, *Progress in Energy and Combustion Science*, **19**, 227 (1993).

[2] A. Aghazadeh Ardebili et al., *Energy Informatics*, **7**, (2024).

[3] J. Smith et al., *Applied Energy*, 300, 117395 (2023).

High-Fidelity Flow Simulations empowered by HPC

Muhammad Mizan and Bernhard Semlitsch

Institute of Energy Systems and Thermodynamics, TU Wien, Austria

High-Performance Computing (HPC) has become key for analysing complex fluid dynamic problems in various scientific, engineering, and industrial research fields. The non-linear nature of the governing equations, i.e. the Navier-Stokes equations, as the interplay of transport phenomena, demands high-fidelity models and high resolution in large computational domains to obtain meaningful results, which can only be delivered in reasonable timeframes using HPC. Moreover, the file size of the results demands pre- and post-processing on powerful machines with graphical user interfaces. We describe our workflow, illustrated with examples, for performing large-scale industrial flow simulations on the Austrian Scientific Computing (ASC) systems.

In the fluid flow machinery research group, we investigate flow phenomena occurring in turbomachinery, such as Pelton turbines. Pelton turbines are used to balance fluctuations in electricity demand. The potential energy of water stored in high-altitude reservoirs is converted into kinetic energy in the form of water jets at the nozzles, transformed into mechanical energy at the runner and, finally, into electrical energy by the generator. The efficiency-critical process involves the high-velocity water jet transferring kinetic energy into rotational energy by impacting the runner buckets. Perturbations generated upstream of the nozzles deform the water jet [1], which causes losses. However, flow disturbances can be mitigated during acceleration or by favourably placing the nozzle closing body holders. Fig. 1 illustrates the water jet surface when interacting with a normal-oriented surface, where the impact of nozzle convergence is examined. Such fundamental studies require significant computational resources, since all fluid dynamic scales must be considered.

Our typical workflow for performing high-fidelity flow simulations on the ASC environment is illustrated in Fig. 2. The geometry representation is prepared on workstations using computer-aided design software, while the numerical mesh is generated on ASC access nodes with graphical user interfaces and the required memory capacity. The simulation settings are provided via control files, which are executed on the computing cluster. The results are reconstructed, analysed, and visualised on the ASC access node.

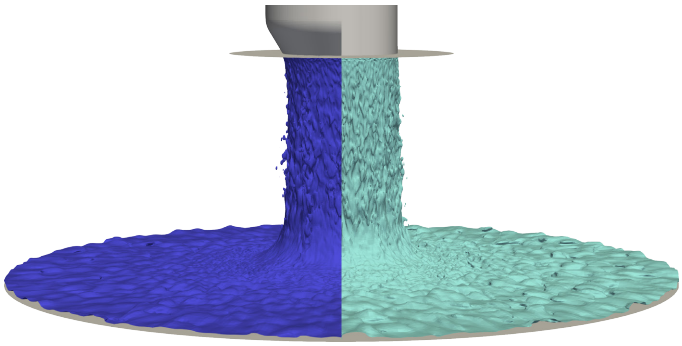


Fig. 1: The surface of simulated water jets impacting a flat plate is shown with different nozzle convergences.

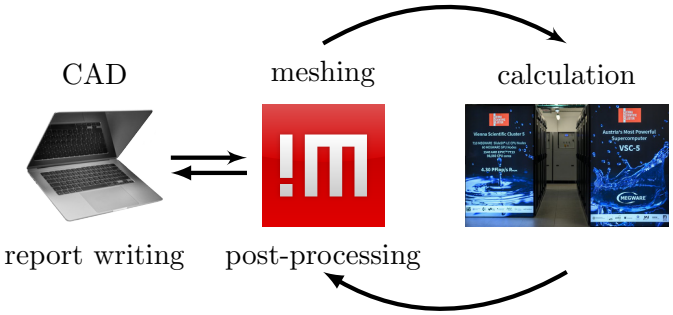


Fig. 2: The workflow for detailed flow simulation on the ASC environment is illustrated.

References

[1] Semlitsch, B., International Journal of Multiphase Flow **174**, 104786 (2024).

Poster:

EuroCC Austria: The Austrian Competence Centre for Supercomputing

Bettina Benesch, Anna Remizova, and Andreas Lindner

Advanced Computing Austria ACA GmbH, Karlsplatz 13, A-1040 Wien, Austria

The Austrian Competence Centre for Supercomputing—EuroCC Austria—is the one-stop shop for access to Austrian and European high-performance computing (HPC) systems and expertise. As HPC is a key enabler of scientific discovery and data-driven innovation, EuroCC Austria plays a central role in this landscape as an active member of a European network of 36 national competence centres.

Core services: The competence centre provides access to Austrian HPC infrastructure, assistance in applying for access to EuroHPC supercomputers (**Fig. 1**), personalised HPC onboarding sessions, and best-practice proofs of concept on Austrian and European HPC systems. The centre is prominent for its successful history of delivering regular high-quality HPC training courses that have been visited by students and professionals from all over Europe. Furthermore, it offers connections to leading experts across the European HPC ecosystem as well as to AI Factories.

Highlights: Since the launch of the EuroCC initiative in 2020, numerous companies have been successfully onboarded to Austrian and European supercomputers, enabling them to efficiently develop their ideas in areas such as data analytics and AI-driven applications. Among the companies that successfully utilised HPC with the help of EuroCC Austria, one, AlongRoute, used Graph Neural Networks to optimise shipping routes in the Mediterranean. Their forecasts, based on the spatiotemporal dynamics of sea state, winds, and currents, achieve higher accuracy than existing solutions. Another company, Beat Shaper, developed an AI-driven system – trained exclusively on licensed data – to compose and refine music through an intuitive interface, from simple prompts that generate initial melodies to advanced, fine-grained parameter controls.

Fig. 1: European map highlighting countries with EuroHPC supercomputers. These systems are accessible through various access calls by the EuroHPC Joint Undertaking to users all over Europe.

Strengthening the European HPC ecosystem: Entering the third phase of the EuroCC project in April 2026, EuroCC Austria is building on its well-established foundation, with an increased focus on classical HPC topics. AI-related services will be provided by the Austrian AI Factory (AI:AT). A key priority in this phase is to deepen integration within the network of EuroHPC facilities and projects. This involves competence centres of other countries, centres of excellence, AI Factories, and further initiatives.

The name EuroCC Austria will be phased out over the coming year. The national competence centre will continue its activities and services under the umbrella of Advanced Computing Austria (ACA).

Poster:

Integration of Advanced Computing service and e-learning platform Merlin

Emir Imamagić, Jurica Špoljar, Daniel Vrčić, and Zvonko Martinović

University of Zagreb, University Computing Centre, Croatia

University Computing Center (SRCE) provides infrastructure and services to the Croatian scientific and academic community. With that task on hands integration of multiple services is key in simplifying and enhancing user experience. After connecting the Advanced Computing service with Croatian Research Information System—CroRIS, and PUH data storage and management, it is now connected with e-learning platform Merlin. Merlin is a robust e-learning platform maintained by the SRCE to support digital education across Croatian higher education institutions. Merlin is based on the Moodle open-source system, which the SRCE's E-learning Center has further developed and adapted to the needs of users, and it is presently the most modern e-learning platform. The Merlin virtual learning environment consists of the Merlin learning platform, the webinar software, e-portfolio system, and it is connected to the ISVU system (Information System of Higher Education Institutions) and DABAR system—Digital Academic Archives and Repositories [1]. The SRCE Advanced Computing service is available to users for the following purposes: research projects financed from public sources and registered in CroRIS; preparation of bachelor's, master's and final specialist theses and PhD dissertations; conducting regular courses and workshops at public higher education institutions and public scientific institutes in the system of science and higher education [2]. While requesting access for a research project is simple, and all data is filled in from the CroRIS, other two methods require manually entering all the data, and inviting each student/participant manually. This method of registration creates additional work for the Advanced Computing service team, which must verify whether the person is indeed a mentor teacher or course leader or workshop organizer. Today teachers can submit a request for use of the Advanced Computing service through Merlin and easy the process of using the Jupyter platform in the teaching and learning process. The connection between the Advanced Computing service and the Merlin platform was achieved by expanding the user management application with an application programming interface (API) that enables the entry and retrieval of data on requests for regular courses, as well as the entry of new users/students and retrieval of data on existing ones [3]. This method of registration significantly simplifies the process of opening a project in the Advanced Computing service, since data such as the name of the e-course and its duration, as well as the list of students, are already present in the Merlin platform. At this moment users can open a request to use the Jupyter platform, but depending on the usage and requirements of the user, it is possible to extend the link to other resources of the Advanced Computing service. This is another extension of the virtual learning environment provided by SRCE, and it allows higher education institutions, teachers and students to access all the necessary resources for conducting online classes through the Merlin platform interface.

References

[1] https://moodle.srce.hr/

[2] https://www.srce.unizg.hr/en/advanced-computing

[3] Imamagić E. and Martinović Z., Srce novosti **103**, 8 (2025).

Poster:

HPC Vega: Supporting Services

Teo Prica[a], Samo Lorenčič[a], and Dejan Lesjak[b]

[a] *IZUM - Institute of Information Science, Slovenia*
[b] *JSI - Jožef Stefan Institute, Slovenia*

HPC Vega is the first launched petascale EuroHPC Joint Undertaking (JU) supercomputer and has entered its fifth year of operation. It provides the necessary infrastructure for the Slovenian scientific community, projects within the EuroHPC JU share of Vega, and industry.

Modern HPC systems require user-friendly tools that simplify access while maintaining transparency and control. We will present the Open OnDemand instance and the User Portal deployed on HPC Vega [1]. Open OnDemand is a web service that enables graphical access to HPC resources in order to manage files, submit and monitor jobs, track their status in the job queue, and inspect job outputs [1]. Complementing this service, the HPC Vega User Portal provides a central overview of user projects and key system metrics [2]. Designed to support efficient planning and informed use of HPC resources, the portal offers clear and up-to-date insights into project status and resource utilization. Built on the Grafana data visualization platform and using the Slurm accounting database as a data source, it provides transparent monitoring for all HPC Vega users [2].

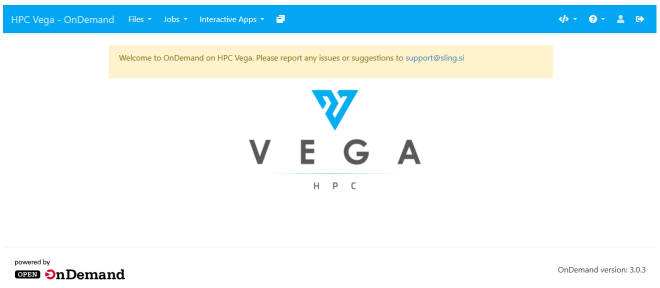


Fig. 1: Open OnDemand User Dashboard.

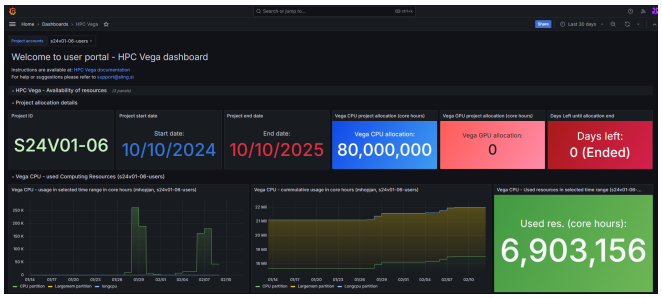


Fig. 2: User Portal based on Grafana.

In addition, the EuroHPC Federation Platform (EFP) project was created as a result of the EuroHPC JU objective to provide a federated platform for end-user workflows and a secure ecosystem of HPC infrastructure within the European Union [3]. In the first phase, users will be provided with critical functions such as user authentication and authorization infrastructure (AAI), an interactive user interface, workload scheduling, federated resource allocation, monitoring and management, information about their projects, system maintenance, software catalog provided by the European Environment for Scientific Software Installations (EESSI), and technical support [3]. Later in the project, the expansion of the set of functionalities and additional support services will be made available to end users.

References

[1] EuroHPC Vega Open OnDemand: https://ondemand.vega.izum.si/

[2] EuroHPC Vega User Portal: https://portal.vega.izum.si

[3] EuroHPC Federation Platform (EFP): https://my-eurohpc.eu/

Poster:

EPICURE: Unlocking European-level HPC Support

Žiga Zebec[a], Samo Miklavc[a], Darin Lah[a], Teo Prica[a], Alja Prah[b], Sebastien Strban[b], and Dejan Lesjak[b]

[a] *IZUM - Institute of Information Science, Slovenia*
[b] *JSI - Jožef Stefan Institute, Slovenia*

The EPICURE [1] project is funded by the EuroHPC Joint Undertaking (JU) and provides specialized technical support services to researchers granted access to EuroHPC JU supercomputers throughout the continent. The project provides several services, including code enablement and scaling, performance analysis and benchmarking, code refactoring, and code optimization [1]. To this end, the project will draw on the experience and knowledge of all partners in high-performance computing (HPC) operations and support, using training activities and hackathons to share knowledge [1].

EPICURE also includes close collaboration with the different Competence Centers and Centers of Excellence. The project consortium includes 16 partners from 14 different European countries, including the Institute of Information Science (IZUM) and the Jožef Stefan Institute (JSI) from Slovenia. The project adopts a specialized support model for HPC end-users, strategically focusing on advanced computational challenges [1]. While support tasks—such as secure shell (SSH) key provisioning, login assistance, and basic job submission workflows—are expected to be managed by users or first-level support, EPICURE delivers expert guidance in higher-value domains. These include customized software installation, performance-aware workflow design, and in-depth code and resource optimizations tailored to maximize computational efficiency in HPC infrastructure [2].

Fig. 1: EPICURE: HPC Support.

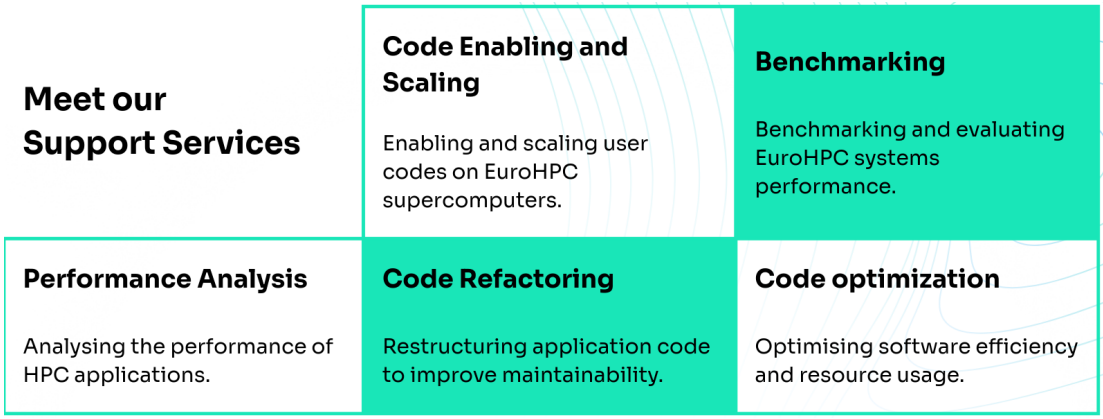


Fig. 2: EPICURE: Meet our Support Services.

References

[1] EPICURE: https://epicure-hpc.eu/

[2] EuroHPC Vega User Documentation: https://doc.vega.izum.si/

Poster:

FFplus: Bridging SMEs with State-of-the-Art HPC and Generative AI

Tina Marc

Arctur d.o.o. and FFplus Consortium

FFplus builds on the methods and accomplishments of the Fortissimo project series, which have left an invaluable legacy in the European HPC and business landscapes. Starting with Fortissimo (2013-2016), Fortissimo 2 (2015-2018) and FF4EuroHPC (2020-2023), more than 130 experiments involving 330 partners have been executed, resulting in 120 success stories. These success stories have encouraged actors across the European industrial landscape to implement new digitalization technologies in manufacturing, and to develop new products and services that bolster the EU economy.

A central objective of FFplus is to provide SMEs with six open calls with more than €24 million in Financial Support, funding Business Experiments—helping SMEs integrate HPC into real industrial challenges—and Innovation Studies, supporting start-ups and SMEs developing generative AI technologies, including large language models (LLMs). Furthermore, selected sub-projects could get the access to EuroHPC's state-of-the-art supercomputing infrastructure, allowing them to exploit world-class European systems for simulation, data analytics, and large-scale AI model development. By coupling funding with access to leading HPC centres, FFplus ensures that SMEs can transform ambitious ideas into scalable, market-ready solutions.

Fig. 1: Foundation Model for European Geospatial Mapping. Copyrights: GEODETICCA VISION, FFplus.

The project has already demonstrated exceptional impact and demand. The first Business Experiments call attracted 126 proposals, resulting in 19 funded sub-projects, the first Innovation Studies call received 62 proposals, with 18 projects selected. The second open call for Business Experiments set a new record: more than 400 proposals were submitted, and in February 2026, also the second Open Call for Innovation Studies also set a new record, reaching the maximum submission limit of 250 proposals in less than 24 hours, clearly demonstrating the extraordinary interest of European SMEs in scaling generative AI using EuroHPC infrastructure. Building on the Fortissimo legacy, FFplus is accelerating Europe's digital transformation. By bridging SMEs with Europe's most advanced supercomputing resources, FFplus helps European SMEs to strengthen industrial resilience, foster AI-driven innovation, and position European industry at the forefront of global competitiveness.

References

[1] Fortissimo, https://www.fortissimo-project.eu/ (2026)

[2] FF4EuroHPC, https://www.ff4eurohpc.eu/ (2026)

[3] FFplus, https://www.ffplus-project.eu/ (2026)

Poster:

NOUS: Advancing Europe's Sovereign Cloud through HPC, Edge Computing and Data Spaces

Tristan Pahor and Tina Marc

Arctur d.o.o. and NOUS Consortium

Europe's cloud market remains heavily dependent on non-European hyperscale providers, which currently dominate more than 70% of the sector. At the same time, the rapid growth of data-intensive applications, AI workloads, IoT ecosystems and emerging Data Spaces creates urgent challenges: fragmented infrastructures, limited interoperability, insufficient integration of HPC resources, and concerns over data sovereignty, security and regulatory compliance. The NOUS project addresses these structural gaps by proposing a new paradigm for a federated European cloud ecosystem.

NOUS aims to develop a comprehensive technological blueprint for interconnected European cloud services. Rather than building yet another standalone cloud platform, NOUS designs an open, interoperable architecture that seamlessly connects High-Performance Computing (HPC), quantum computing, cloud, edge and IoT environments with emerging European Data Spaces.

The project tackles three core challenges: (1) enabling trusted and sovereign data exchange across sectors, (2) integrating HPC and advanced computing into mainstream cloud services, and (3) supporting the Edge-to-Fog-to-Cloud continuum for real-time, data-intensive applications. Its architecture combines IaaS/PaaS/MLaaS services with federated learning frameworks, AI-driven workload orchestration and low-energy Distributed Ledger Technologies (DLT) to guarantee data provenance, lifecycle transparency and secure cataloguing. Semantic data lakes and standards-based interoperability modules facilitate convergence across Mobility, Energy and Green Deal Data Spaces.

Through industrially driven use cases—such as climate modelling, smart mobility systems, energy optimisation and advanced materials research, NOUS validates how a sovereign, HPC-enabled cloud infrastructure can deliver scalable, secure and high-performance digital services.

By bridging advanced computing infrastructures with interoperable Data Spaces, NOUS supports the creation of a Single European Data Market, strengthens digital sovereignty and enhances Europe's global competitiveness in next-generation cloud services.

References

[1] https://nous-project.eu/ (2026)

[2] https://dl.acm.org/doi/10.1145/3685651.3686660 (2026)

Poster:

Privacy Preserving Generative AI and Model Sanitization

Bernd Saurugger[a], Robert Harb[b], Jakub Pekár[c], and Heimo Müller[b]

[a] *TU Wien*
[b] *Medical University of Graz*
[c] *Masaryk University*

Digital pathology generates large whole slide images (WSIs), with individual files often exceeding several gigabytes and research cohorts scaling to tens or hundreds of terabytes. These data volumes, combined with compute-intensive deep learning workflows, necessitate access to powerful high-performance computing (HPC) infrastructures for efficient model training and experimentation. We use the VSC-5 and MUSICA clusters to advance generative modeling in computational pathology, demonstrating scalable distributed training of diffusion models on datasets with varied tissue morphologies. Our work is funded by the European Union through the projects RI-SCALE and OSCARS (Grant Agreement Numbers 10188168 and 101129751).

Our application focuses on diffusion-based image synthesis for histopathology patches, enabling data augmentation for tasks like tumor classification and survival prediction, while meeting strict privacy requirements, through model sanitization via Differential Privacy (DP) and evaluation against membership inference attacks. Such augmentation is essential as real pathology data is limited by class imbalances, scanner variability, and privacy constraints. Generated patches realistically mimic tissue textures and colors. These models, with hundreds of millions of parameters, require substantial GPU memory usage and parallelization to achieve stable training with large effective batch sizes. We train on WSI-derived 256×256 patches, processing millions of samples.

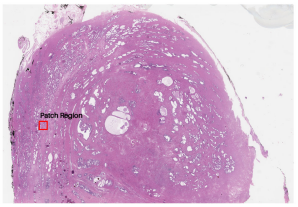


Fig. 1: WSI overview. Mid-zoom region (red)

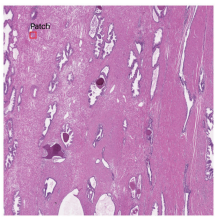


Fig. 3: Mid-zoom. Patch location (red).

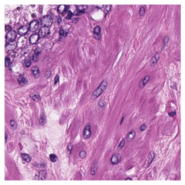


Fig. 2: Patch extracted from the WSI.

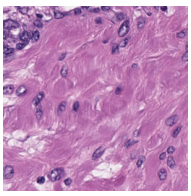


Fig. 4: Model sample. Cosine similarity: 0.89.

To bridge application needs with infrastructure capabilities, we optimized our workflows for MUSICA's GPU partition, enabling memory-efficient patch loading and communication overlap during training. The nodes of the GPU partition of MUSICA (zen4_0768_h100x4) have 2× AMD EPYC 9654 CPUs (192 cores total), 768 GB DDR5 RAM, 4× NVIDIA H100 SXM5 GPUs (94 GB memory per GPU), 7.68 TB local NVMe storage, and 4× NDR200 InfiniBand links enabling efficient multi-node communication. Our distributed training implementation on MUSICA utilizes PyTorch's Distributed Data Parallel (DDP) framework with the NCCL communication backend, where we launch exactly one training process per GPU to ensure optimal resource utilization. This approach enables seamless scaling from single-node configurations utilizing all 4 GPUs per node up to multi-node deployments spanning 4 nodes with a total of 16 H100 GPUs. Scaling experiments on representative diffusion model training workloads for large digital pathology datasets clearly demonstrate efficient performance, achieving approximately 75% parallel efficiency when scaling from 1 to 16 GPUs across multiple nodes, primarily limited by all-reduce communication overhead and batch size scaling effects. These results underscore MUSICA's effectiveness for data-intensive AI workloads in biomedical research.

Poster:

HPC meets EOSC: Emerging Tasks, Opportunities and Concerns

Marie Czuray, Katharina Flicker, Andreas Rauber, and Bernd Saurugger

Austrian Scientific Computing (ASC)

The European Open Science Cloud (EOSC) Federation is transforming how compute and data-intensive research communities access and combine resources across institutional, national, and disciplinary boundaries. This has direct consequences for the way we design and operate high-performance computing (HPC) infrastructures. While EuroHPC systems and national supercomputing centres continue to provide high-performance compute resources, the EOSC Federation is establishing a network that connects workflows, data repositories and services. These are becoming increasingly important in defining how users experience HPC. In Austria, these developments are framed by the EOSC Support Office Austria[4]. This contribution describes how the EOSC Federation affects HPC from the perspectives of infrastructure providers and advanced user communities. It explores the emerging tasks and responsibilities of HPC centres and the opportunities this evolution offers operations teams and users.

To get a first impression of the emerging federation, we can take a look at the EOSC EU Node, National Nodes such as SURF in the Netherlands or ICSC in Bologna, or Thematic Nodes such as CERN or BBMRI-ERIC. These demonstrate federating capabilities, such as seamless discovery and access to data, Jupyter Notebooks that distribute compute across several sites and integrate results locally. HPC centers are expected to publish services and resources into a federation, to support a common access and authentication infrastructure (AAI), and to participate in shared catalogues, accounting, and policy frameworks. This creates new tasks for HPC centers as outlined in the Federation Handbook [1], listing mandatory and optional capabilities, roles (e.g. Operations Manager, Security Officer) and governance structures. Federating capabilities (frequently supported by Open Source implementations, e.g. [2]) include interoperable AAI [3], Resource Catalogues and Registry Services, Service Monitoring, Accounting or Application Deployment Management. A Node may also offer (domain-)specific higher-level services, ranging from access to compute resources, Jupyter Notebooks, and data repositories to specific analysis pipelines, simulation software, or visualization services.

The EOSC Federation supports science by making a broader range of services available to the research community. It also allows better resource utilization and specialization in specific services, which are then made available to a larger community. This integrates HPC more tightly into end-to-end research workflows. For operators, the emerging EOSC Federation model provides a common vocabulary and set of building blocks for structuring services, documenting service maturity and arguing for sustainable funding.

References

[1] EOSC Association: EOSC Federation Handbook Version 2 (Jan 2026). DOI: 10.5281/zenodo.18454649

[2] EOSC EU Node Open Source Code: https://open-science-cloud.ec.europa.eu/about/eosc-eu-node-open-source-code

[3] C. Kanellopoulos et al., EOSC AAI Architecture 2025 (March 2025) DOI: 10.5281/zenodo.15388270

[4] https://eosc-austria.at/

Poster:

MUSICA - From Scratch To A Fully Functional Server Room @UIBK

Martin Thaler

HPC Team, Central IT Service, University of Innsbruck, Austria

Preparation And Planning: It all began in November 2021 with a site visit to the VSC (now ASC) infrastructure at the Arsenal in Vienna, accompanied by Ernst Haunschmid. At the time, the VSC5 server room was still under construction. A half-day tour through the facility's complex corridors and stairways provided us with an in-depth understanding of both the indoor and outdoor technical installations. Drawing on these insights, we began planning the MUSICA server room in Innsbruck. The University ultimately commissioned IKB (Innsbrucker Kommunalbetriebe) to execute the project. As this was the HPC team's first large-scale infrastructure project, the final interior design (illustrated in Fig. 1) represents a significant milestone in our department's history.

The Completed Server Room: The University of Innsbruck and its ASC partners now operate a state-of-the-art facility. Key safety features include Residual Current Monitoring (RCM) at all outlets and strategically placed leakage sensors beneath the raised floor. Furthermore, the water cooling circuits are physically isolated from the primary data center area to mitigate potential risks. The server room has a modular design and can be expanded in the future.

Relocation And Installation Of The HPC System: In September 2025, the MUSICA INN HPC system was successfully relocated from Vienna (Arsenal) to its permanent home. Because the new server room is situated off-campus, tasks that were once routine occasionally presented logistical and organizational challenges. However, we benefited from the exceptional and unbureaucratic support provided by our partner IKB and the University's Building and Infrastructure (GI) department. As shown in Fig. 2, the HPC system MUSICA INN has been successfully installed in Innsbruck.

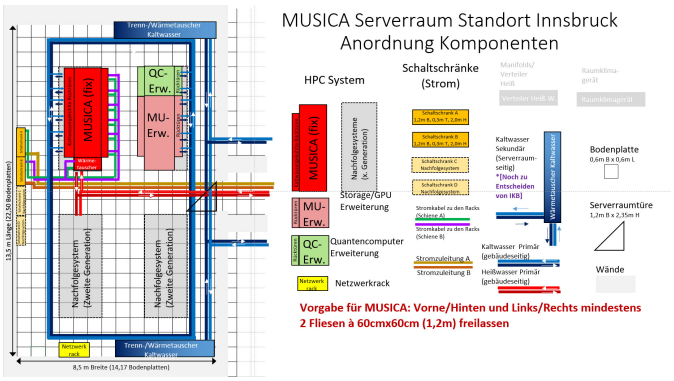


Fig. 1: Final draft for interior design.

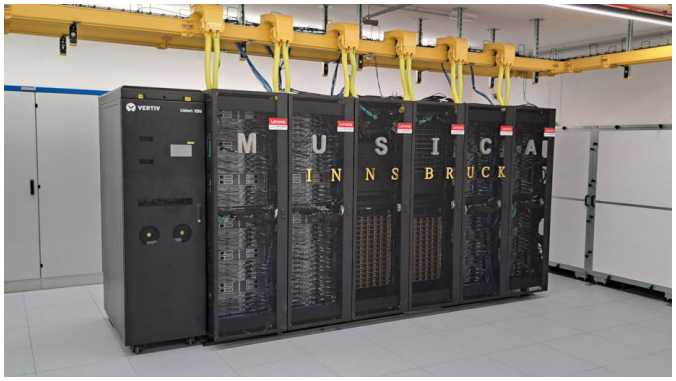


Fig. 2: HPC System MUSICA Innsbruck

References

[1] https://docs.asc.ac.at/systems/musica.html

[2] https://www.uibk.ac.at/en/zid/systeme/hpc-systeme

[3] https://www.ikb.at/internet/rechenzentren

Poster:

Quantum Computing for Scientific Computing

Ezhilmathi Krishnasamy[a,b,c], Janez Povh[b], Xing Cai[d], Leon Kos[a], and Pascal Bouvry[c]

[a] *University of Ljubljana, Slovenia*
[b] *Rudolfovo – Science and Technology Centre Novo Mesto, Slovenia*
[c] *University of Luxembourg, Luxembourg*
[d] *Simula Research Laboratory & University of Oslo, Norway*

Addressing the climate crisis through innovative energy solutions is crucial, with fusion energy being a promising approach. The Particle-In-Cell (PIC) method, based on kinetic modeling[1], is commonly used for both simplified and complex simulations. A key challenge of PIC is solving the system of equations, often managed by iterative solvers like Krylov. While modern CPUs and GPUs can expedite this process, significant instabilities persist, necessitating a unified programming model for various GPU vendors and compute clusters. Additionally, quantum computing represents a promising avenue for solving these equations and may play a vital role in future scientific computing advancements.

Introduction: Classical computers have historically been used for mathematical modeling in scientific computing to solve complex problems through various processes. Advances in technology, particularly in electronics, have led to faster processors and supercomputing resources for efficiently solving partial differential equations with numerical methods.

However, classical processors face three significant limitations: the memory wall, the Instruction-Level Parallelism (ILP) wall, and the power wall[2]. While advancements in electronics may help address these issues in the future, exploring alternative methodologies is still essential despite the exponential growth in computing power.

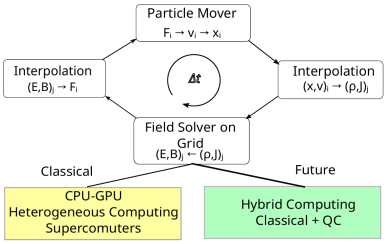


Fig. 1: Schematic overview of a hybrid computing solving a system of equations in PIC-based plasma modelling.

Methods: In light of these considerations, exploring alternative computing methodologies is essential, with quantum computing emerging as a leading candidate for enhancing computational efficiency. Quantum computers offer distinct advantages over their classical counterparts, including energy efficiency and significantly faster computation capabilities.

For instance, factoring a large number with n bits using a classical computer necessitates exponential time–approximately $\exp(n^{1/3})$–whereas a quantum computer employing Shor's algorithm can accomplish the same task in polynomial time, approximately $O(n^2 \log n)$[3]. Furthermore, our ongoing work will investigate and harness quantum computing techniques to overcome current limitations in simulation and modeling related to fusion energy development. We will consider various numerical techniques and their approaches, especially in solving systems of equations (e.g., in PIC, as shown in Fig. 1), ultimately contributing to effective and sustainable solutions for the climate crisis.

Acknowledgements: This work was supported by the Slovenian Research and Innovation Agency (ARIS) under the Early-Stage Researchers funding scheme, grant No. ARIS-RZK-2025/54 and additionally supported by N2-0335 HEXAPIC - Heterogeneous EXAscale Particle-In-Cell code.

References

[1] Dawson, J.M., Rev. Mod. Phys. 55, 403 (1983).

[2] Hennessy, J.L., and Patterson, D.A., Computer Architecture: A Quantitative Approach, Elsevier (2011).

[3] Shor, P.W., SIAM Review 41, 303 (1999).

Poster:

A Domain-Aware Controller for Managed LLM Inference on Shared HPC Infrastructure in Digital Humanities

Michael Otto[a], Lukas Waldhofer[a], David Fleischhacker[a], Max Resch[b], and Florian Atzenhofer-Baumgartner[a]

[a] *Department of Digital Humanities, University of Graz, Austria*
[b] *Center for Cultures and Technologies of Collecting, University for Continuing Education Krems, Austria*

Motivation: DH research increasingly relies on LLM workflows—corpus annotation, NER, knowledge extraction—that differ from chatbot scenarios [1, 2]. DH workloads mix long-running pipelines with interactive queries; both synchronous and asynchronous modes are needed, as is resource management across users with uneven demand. Existing HPC LLM services focus on general-purpose serving; questions of federated identity and institutional quotas remain open. This poster describes a controller for managed LLM inference developed for dhinfra.at.

Controller: A Python/FastAPI controller handles model deployment, routing, and accounting. Features include LiteLLM/vLLM backends (OpenAI-compatible, yellow in **Fig. 1**), scoped API keys with rate limiting, health monitoring with failover, Redis/Celery queues for batch jobs, SLURM integration, and user voting for model rotation ("fixed" vs. "seasonal"). Always-on API serving is thus bridged with HPC job submission.

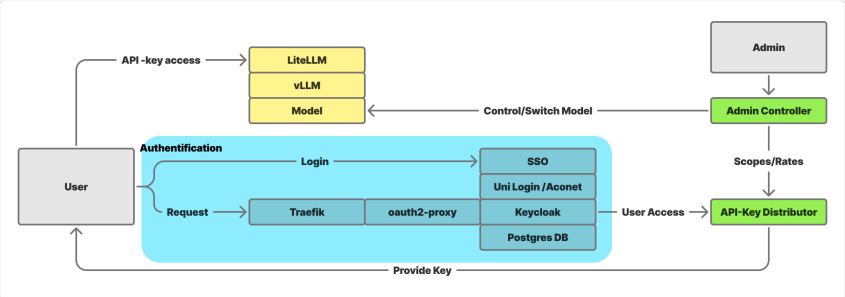


Fig. 1: Controller workflow for managed LLM inference with federated, ACOnet-based authentication. Green: admin components; blue: authentication flow; yellow: inference backends.

Access and Governance: Authentik brokers ACOnet logins; Traefik and oauth2-proxy enforce edge authentication on Request (blue in **Fig. 1**). SSO identities are coupled with institutional quotas and per-user rate limits—not available in API-key-only approaches. The focus is on the orchestration layer: resource allocation, quota enforcement, and scheduling on shared HPC hardware.

Discussion: Open questions include: (1) federated identity with fine-grained quotas, (2) scheduling that mixes always-on inference with SLURM batch jobs, (3) model caching on heterogeneous GPUs. The controller is an open alternative to commercial platforms; institutional control and data sovereignty are preserved. Similar requirements may apply to other HPC deployments.

References

[1] Marwaha, R., Zhou, Q., Day, K., Dabholkar, A., and Kindratenko, V., Frameworks for Large Language Model Serving in HPC Environments. *SC Workshops '25*, St Louis, MO, USA (2025).

[2] Sada, M. F. et al., Serving LLMs in HPC Clusters: A Comparative Study of Qualcomm Cloud AI 100 Ultra and NVIDIA Data Center GPUs. arXiv:2507.00418 (2025).

Poster:

High-Performance Computing for AI-Based Insider Threat Analytics: An Experimental Study

Maizura Ibrahim[a], Dejan Lesjak[b], Nur Fatini Abd Ghani[a], Mohamad Safuan Sulaiman[a], and Andrej Filipčič[b]

[a] *Malaysian Nuclear Agency, Malaysia*
[b] *Jožef Stefan Institute, Slovenia*

Insider threats remain one of the most challenging security risks in critical infrastructure due to their dynamic, behavioral, and temporal nature. Addressing such threats requires continuous assessment of personnel trustworthiness and reliability using heterogeneous big data sources and interpretable analytical models. This paper presents an experimental study that deploys an AI-based insider-threat analytics workflow, grounded in the Trustworthiness Metrics Value (TMV) model, on an HPC infrastructure at the Jožef Stefan Institute, Slovenia. Due to the complexity of the algorithm and large-scale data requirements, HPC infrastructure is essential for execution. This work builds on our previously published framework by implementing and benchmarking its insider-threat analytics components [1].

The TMV model [2] estimates the probability of a malicious insider by aggregating two complementary dimensions of personnel assessment: organizational reliability metrics and social-media-derived trustworthiness metrics. Organizational data are processed to compute time-aware reliability indicators. In parallel, social media data are analysed through text preprocessing and feature-computation techniques, including sentiment polarity, emoticon mining, and trust-mapping lexicon-based scoring using existing LLMs.

Figure 1 depicts the resulting AI-based insider threat analytics workflow deployed on SiGNET cluster. Experiments were executed to examine behavior under varying parameter configurations and batch evaluation scenarios. Rather than emphasizing predictive-model optimization, this study focuses on practical considerations for executing a time-aware insider-threat analytics pipeline on supercomputing resources, including workload structuring, data orchestration, reproducibility, and scalability. These experiments demonstrate that HPC is instrumental in supporting continuous and interpretable insider-threat analytics, particularly when trustworthiness and reliability assessments must be evaluated continuously over time.

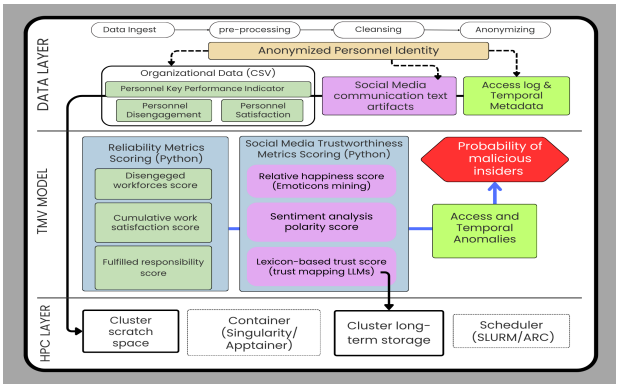


Fig. 1: AI-based insider threat analytics workflow on SiGNET cluster.

References

[1] Ibrahim, M. et al., Big data analytics nuclear security framework, IOP Conf. Ser.: Mater. Sci. Eng. **1106**, 012026 (2021).

[2] Ibrahim, M. et al., Predicting Insider Threats in Nuclear Security: A New Technique for Quantifying Personnel Trustworthiness Using Social Media Data, Nuclear Malaysia Research and Development Seminar 2024, Bangi (Malaysia), August 27–29 (2024).

Poster:

Training Alternative Large Scale Representations on Current High-Performance Computers

Till Kahlke[a], Sebastian Salwig[a], Florian Hirschberger[b], Dennis Forster[c], and Jörg Lücke[a]

[a] *Artificial Intelligence Lab, University of Innsbruck, Austria*
[b] *Machine Learning, Carl von Ossietzky University Oldenburg, Germany*
[c] *Data Analytics and AI, Frankfurt University of Applied Sciences, Germany*

Large scale representations of data densities are central to modern generative Artificial Intelligence (AI) systems. In this context, almost all such scalable approaches rely on deep neural networks (DNNs), with generative adversarial nets (GANs), normalizing flows (NFs) and diffusion models (DMs) as prominent examples. However, alternative data density representations have long been investigated, and approaches based on Gaussian mixture models (GMMs) represent well-known examples, with advantages, e.g., in terms of interpretability and robustness. Compared to DNN based representations, conventional GMM based approaches suffer from unfavorable scaling with model and dataset sizes, however. Their practical applicability to large scale settings has, therefore, been severely limited by their high computational demand. But the computational challenge of training large scale GMMs can be addressed using novel variational approximations and mixtures of factor analyzers (MFAs; [1]) as a restricted form of GMMs [2]. As a result, GMM based approaches can achieve sublinear scaling with model and dataset size [2, 3]. However, the specific approximations introduce implementation challenges that have to be addressed in order to realize an algorithm that not only achieves a theoretical reduction in computational complexity but also executes efficiently on HPC hardware.

In this work, we report technical insights on how recent sublinear learning algorithms for GMMs can efficiently be executed on current hardware. Our approach leverages sparse data structures used for the variational optimization, shared memory parallelization, flexible memory access, specific matrix inversion approaches, and other dedicated analytical and algorithmic solutions (which are not directly available in standard AI software frameworks). In our studies, we observed that latency bounds are the primary performance bottleneck, due to irregular memory access patterns. Nevertheless, runtime speed-ups of orders of magnitudes can be achieved compared to conventional GMM training on large scale applications. Using a single AMD Genoa EPYC 9554 CPU, we observed that GMMs with more than 10 B parameters trained on approximately 100 M images can be optimized in less than nine hours [2]. These scales were previously attainable almost exclusively by DNNs and the training time contrasts with the days or weeks of training on many GPU nodes required for DNN based representations of similar sizes.

References

[1] Ghahramani, Hinton. The EM algorithm for mixtures of factor analyzers. *Technical Report CRG-TR-96-1, University of Toronto* (1996).

[2] Salwig*, Kahlke*, Hirschberger, Forster, Lücke. Sublinear Variational Optimization of Gaussian Mixture Models with Millions to Billions of Parameters. *arXiv:*2501.12299 (2025). *joint first authorship.

[3] Hirschberger*, Forster*, Lücke. A Variational EM Acceleration for Efficient Clustering at Very Large Scales. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**(12):9787–9801 (2022). *joint first authorship.

Poster:

Open-source framework for parametric study of hydrofoil profiles and motivation for using Physics-Informed Neural Networks (PINN).

Aleksander Grm[a,b] **and Nikola Vukašinovič**[b]

[a] *FPP, University of Ljubljana, Slovenia*
[b] *LECAD@FS, University of Ljubljana, Slovenia*

This research presents a robust, automated, open-source computational framework for generating high-throughput hydrodynamic datasets for Physics Informed Neural Network (PINN) applications. Recognising the significant computational overhead of conventional Computational Fluid Dynamics (CFD) in parametric design spaces, this study proposes an integrated approach using GMSH for algorithmic mesh generation and OpenFOAM for high-fidelity fluid flow simulations.

The framework is used to conduct an extensive parametric investigation on an HPC system for three distinct hydrofoil geometries: the NACA 0012, NACA 2412, and NACA 4412 profiles. Simulations are performed across a comprehensive operational envelope, covering angles of attack (α) from $-15°$ to $+15°$ and Reynolds numbers (Re) from 10^4 to 10^7. The resulting structured database contains spatial distributions of velocity and pressure fields, providing the empirical foundation for a deep learning architecture developed within the PyTorch ecosystem. To ensure physical consistency, the PINN architecture enforces the governing Navier–Stokes equations by incorporating the conservation law residuals directly into the composite loss function.

By automating the transition from geometric definition to numerical solution, the framework provides the necessary ground truth for training and validating PINNs. The accuracy of the surrogate model is evaluated by comparing the L^2 relative error and pressure coefficient (C_p) distributions with OpenFOAM steady-state results. This approach aims to accelerate the prediction of hydrodynamic performance in foil-assisted vessel analysis while ensuring that the surrogate model strictly adheres to governing physical laws. The proposed methodology offers a scalable and accessible pathway for developing rapid-response surrogate models in maritime engineering.

References

[1] Bonnet et al., Airfrans: High fidelity computational fluid dynamics dataset for approximating reynolds-averaged navier–stokes solutions, *Advances in Neural Information Processing Systems*, **35**:23463–23478, 2022.

[2] Cuomo et al., Scientific machine learning through physics–informed neural networks: Where we are and what's next, *Journal of Scientific Computing*, **92**()3):88, 2022.

[3] Hakeem et al., Enhancing computational fluid dynamics simulations with machine learning: Techniques, challenges, and future prospects, *Annual Methodological Archive Research Review*, **3**(5):201–219, 2025.

Poster:

VIVID-DTE – Verification-oriented Interactive Visualisation and Decision Support for the EUROfusion Digital Twin Environment

Leon Kos and Matic Brank

LeCAD laboratory, Mech.Eng. University of Ljubljana, Slovenia

The VIVID-DTE project develops advanced visual analytics and operator-oriented decision-support tools for EUROfusion's Digital Twin Environment (DTE). The integrated DTE is used for pre/post-processing, solvers running under HPC workflows with model validation, verification (V&V), and operational insight. VIVID-DTE implements an interactive, geometry-aware, and workflow-driven visualisation framework that integrates simulation results, experimental measurements, and synthetic diagnostics into a unified, reproducible environment. DTE activities rely on tightly coupled modelling chains spanning plasma physics, plasma-facing components, and plant-level systems. With increasing model complexity and dataset size, static figures and offline post-processing introduce latency and reduce cross-disciplinary interpretability. To address these limitations, VIVID-DTE establishes an interactive visual analytics stack directly integrated with DTE workflows and IMAS-based data structures. **System Architecture:** The framework extends SALOME with interactive 3D and temporal visualisation and IMAS-MUSCLE3-based communication for multi-scale coupling and in-situ co-processing. The modular architecture comprises of: **Data Layer:** IMAS-aligned adapters supporting HDF5, and CAD/mesh ingestion with study-level provenance tracking. Study files are preserved with geometry, model parameters, and results. **Compute & In-situ Layer:** Direct processing, single node MPI for parallel visualisation, or HPC batch compute of physics codes, progressive data reduction for multi-resolution in-situ streaming, and GPU-accelerated remote rendering enable interactive exploration of large datasets. **Visual Analytics Layer:** Linked CAD 3D geometry views, time-synchronised plots, residual/error maps, uncertainty encodings, and synthetic-versus-experimental overlays are rendered within a unified study context.

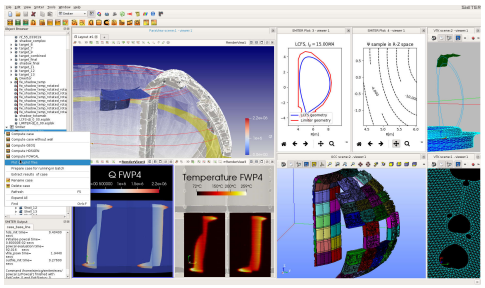


Fig. 1: SMITER [1]/SALOME DTE.

Workflow & UI Layer: A PySide6/Qt-Designer-based interface enables no-code composition of dashboards and workflows inspired by PDS-WF and SOLPS-GUI patterns. Custom widgets exchange typed signals and slots, and complete workflows persist as XML/.ui files with parameter sets. Users construct visualisation and analysis pipelines via drag-and-drop widgets representing data sources (see Fig. 2), analytics modules, and visual panes. Entire dashboards are replayable and versioned, forming persistent IMAS-aligned studies. **Decision-Support Layer:** Operator dashboards provide KPIs, operating limits, alerts, and scenario replay, supporting traceable and explainable decisions.

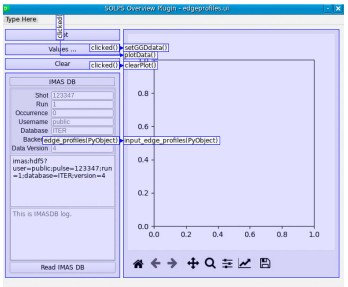


Fig. 2: Workflow designer.

VIVID-DTE accelerates model convergence and enhances plant-wide understanding through digital twin visualisation views. Reproducible UI dashboards and shareable HDF5-based study files support durable knowledge transfer across sites. The modular and IMAS-aligned architecture ensures interoperability.

References

[1] Kos,L., Pitts, R.A., Simič, G., Brank,M., Anand, H., and Arter,W., Fusion Eng.Des. **146**, 1796 (2019).

Poster:

Heterogeneous Exascale Particle-in-Cell

Stefan Costea[a], David Lajevec[b], Miha Radež[a], Jernej Kovačič[a], Matic Brank[a], Leon Bogdanović[a], Ivona Vasileska[a], and Leon Kos[a]

[a] *Faculty of Mechanical Engineering, University of Ljubljana, Slovenia*
[b] *Faculty of Mathematics and Physics, University of Ljubljana, Slovenia*

The Heterogeneous Exascale Particle-in-Cell (HEXAPIC) simulation code is an ongoing development aimed at enabling high-performance plasma physics simulations [1] on exascale computing systems. HEXAPIC decomposes the Particle-in-Cell (PIC) workflow into modular components and uses MPI for distributed-memory parallelism to reduce communication volume and data motion in large-scale runs. Target use cases include plasma–material interaction in scrape-off-layer and divertor regions of tokamak devices, as well as plasma sources for propulsion and materials processing.

We developed a functional CPU backend covering the full PIC time step. We present strong- and weak-scaling results for multi-node setups and discuss performance-limiting factors. The electrostatic field solver was upgraded by integrating the HYPRE library [2] solver for the discretized electrostatic Poisson operator.

Next step of development is heterogeneous refactoring which is based on alpaka3 [3] library. Alpaka3 provides a performance-portable abstraction for hierarchical parallelism across CPUs and accelerators. As illustrated in Fig. 1, the same HEXAPIC kernels can be compiled against multiple backends (OpenMP, CUDA, HIP, SYCL/oneAPI) by selecting the target device at build/run time, avoiding duplicated implementations for each vendor API. Alpaka addresses shared-memory parallelism within a node and therefore connects naturally with HEXAPIC's MPI layer for distributed-memory scaling across nodes.

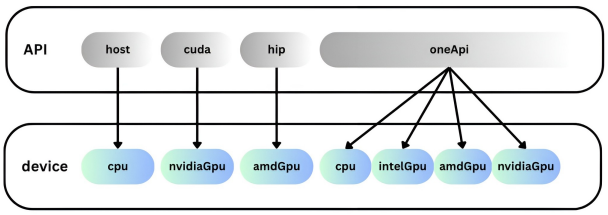


Fig. 1: Alpaka's unified kernel interface maps to host and multiple accelerator backends, enabling seamless heterogeneous execution.

We also implemented a CI/CD pipeline to validate both software and physics. Continuous integration builds all supported backends, runs unit and regression tests, and performs static checks; physics regression tests verify conservation diagnostics and reference benchmark cases using tolerance-based acceptance criteria.

Acknowledgments
The authors acknowledge the project N2-0335 was financially supported by the Slovenian Research and Innovation Agency (ARIS) as well as by the ARIS research core funding P2-0405.

References

[1] Birdsall, C.K. and Langdon, A.B., Plasma Physics via Computer Simulation (1st ed.) (1991).

[2] Falgout, R.D., and Yang, U.M., Lect. Notes Comput. Sci. **2331**, 632 (2002).

[3] Matthes, A., Widera, R., Zenker, E., Worpitz, B., Huebl, A., and Bussmann, M., High Performance Computing, 496 (2017).

Poster:

LEONARDO data centric general purpose partition at AURELEO

Luis Casillas-Trujillo and Claudia Blaas-Schenner

ASC Research Center, TU Wien, Austria

The LEONARDO supercomputer located at CINECA in Bologna is still one of the most powerful computing systems in Europe and the world. The LEONARDO project is part of a collaborative effort, with Austria as a partner in the consortium. Due to this partnership, Austria has a dedicated share of the supercomputer, and Austrian researchers can access LEONARDO's resources through the AURELEO calls. Besides LEONARDO's GPU partition, called the Booster partition, which perhaps LEONARDO is mostly known for, the data centric general purpose partition (DCGP) offers extremely low latency and high data throughput to provide the highest AI and HPC application performance and scalability. The DCGP partition is based on BullSequana X2140 three-node CPU Blade and is equipped with two Intel Sapphire Rapids CPUs, each with 56 cores. It uses NVIDIA Mellanox HDR 200Gb/s InfiniBand connectivity, with smart in-network computing acceleration. The DCGP started pre-production in January 2024 and reaching full production in February 2024. AURELEO's DCGP call is currently in its second iteration and has already proven to be an important avenue to provide Austrian scientists with computational resources to conduct state of the art research and push the boundary of simulation with more challenging and complex simulations. The AURELEO DCGP call program supports a broad spectrum of scientific fields, with awarded projects spanning areas such as condensed matter physics, computer science, fluid dynamics with topics ranging in atomistic simulation of complex materials, polymer processing and algorithm development among others. Many of these projects focus on the development of artificial intelligence (AI) methods and tools, which are being applied to address challenges in these diverse fields. The AURELEO DCGP call [1] opens once a year and it has a simple, straightforward application process. It also comes with support through the High Level Support Team (HLST), which assists users to efficiently run their applications and maximize the usage of the allocated resources.

References

[1] https://asc.ac.at/access/aureleo-austrian-users-at-leonardo-supercomputer/

Poster:

Spreading the Word—ASC Outreach

Atul Singh[a], Diego Medeiros dalla Costa[a], and Siegfried Höfinger[a,b]

[a] *ASC Research Center: TU Wien, Operngasse 11, A-1040 Vienna, Austria*
[b] *Department of Physics: Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931-1295, USA*

Academic outreach events are often designed to foster collaboration, exchange of ideas, and to assist in career development of early stage researchers. However, the question remains whether these goals are really met in practice. Drawing on experience made in three such outreach events over the past year, an IT Vendor Meeting at TU Wien a bilateral research networking initiative organized by the Indian Embassy and Austrian partners [1], and a Post-Doc Day [2] focused on advanced stage researchers, this talk reflects on what such gatherings enable, and where they fall short in addressing real-life challenges. Moreover, insights from participation of MUSICA/ASC in these events is briefly discussed offering ideas about various ways different HPC units can help to maximize value for users—either from industry or academia. It considers how industry engagement is framed within academia, how international collaborations are initiated at an early stage, and how networking formats influence the willingness of researchers to share their every-day challenges, ambitions, and uncertainties. By juxtaposing these events, the talk also invites discussion on how academic and research-focused meetings can evolve beyond commonly used formats toward more meaningful, transparent, and mutually beneficial interactions—especially for researchers navigating complex career and funding landscapes.

The EUMaster4HPC is a new initiative for students interested in earning a Masters degree in HPC. It has a strong focus on European collaboration and is in continuous growth ever since its inauguration in 2022. Master theses can be carried out at various sites that have submitted suitable HPC projects. In 2025 two candidates have been hosted at the ASC Research Center. Such collaborative projects not only are of great value for European outreach in the HPC community, but also enable insight into current trends and expectations by the upcoming next generation of specialists in supercomputing.
Another ASC outreach activity is currently concerned with providing an interactive view of ASC compute servers and corresponding infrastructure. The intention here is to come up with a 3D model of the HPC cluster that helps people visually understand how the machine works as one powerful system. People who don't have a deep understanding of the topic can explore its architecture in an interactive and comprehensive way to better grasp how supercomputers look and function. The visual 3D model shall be integrated into the upcoming ASC website so that visitors can explore underlying hardware in a self-organized exploratory process (combination of individual components sketched in Fig. 1).

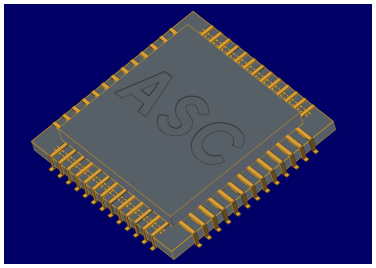


Fig. 1: Sample artwork illustrating component-wise design of the visual 3D model engineered for interactive exploration by the end user (i.e., visitor of the ASC website).

References

[1] https://www.facebook.com/IndiaInAustria/posts/-calling-all-diaspora-researchers-scientists-embassy-of-india-vienna-is-pleased-/1215312903959099/
[2] https://cc.lbg.ac.at/careerday/

Keynote Talk:

Beyond Isolated Quantum Computing Paradigms: Hybridization and Supercomputing

Bojan Žunkovič

University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia
Rudolfovo, Science and Technology Centre, Novo Mesto, Slovenia

Quantum computing has multiple universal formulations. Gate-based circuits, adiabatic and annealing approaches, and measurement-based quantum computing were traditionally developed and studied as distinct paradigms. Increasingly, however, we observe deep and practical connections between them, driven by mid-circuit measurements, hybrid classical–quantum feedback loops, and variational optimization strategies. These emerging links reshape not only algorithm design but also the role of high-performance computing in quantum technologies.

Rather than viewing quantum processors as isolated accelerators, we should understand them as components within a broader computational ecosystem that includes large-scale classical supercomputing. In the NISQ and early fault-tolerant eras, the primary obstacles are limited coherence, restricted circuit depth, trainability issues, and resource overheads. We will argue that overcoming these challenges will not come from a single quantum paradigm in isolation. Instead, progress will emerge from combining different models of quantum computation and tightly integrating them with advanced classical simulation and optimization techniques available on modern supercomputers.

In this talk, we will present three major models of quantum computation and illustrate, through recent work [1,2], how integrating distinct paradigms creates new opportunities for algorithm design and classical–quantum cooperation. In particular, we will show that merging ideas of variational and adiabatic quantum computation can overcome outstanding obstacles in both. This perspective suggests that significant progress can be made through hybridization across different quantum and classical computational paradigms.

References

[1] Žunkovič, B., Torta, P., Pecci, G., Lami, G., and Collura, M. Phys. Rev. Lett., **134(13)**, 130601 (2025).

[3] Žunkovič, B., Ballarin, M., Wright, L. and Lubasch, M., arXiv:2602.17612v1 (2026)

Quantum and Simulated Annealing-Based Iterative Algorithms for QUBO Relaxations of the Sparsest k-Subgraph Problem

Omkar Bihani[a], Roman Kužel[a], Dunja Pucher[b], and Janez Povh[a,c]

[a] *Rudolfovo - Science and Technology Centre Novo Mesto, Novo Mesto, Slovenia*
[b] *Alpen Adria Universitaet Klagenfurt, Klagenfurt, Austria*
[c] *Faculty of Mechanical Engineering, University of Ljubljana, Ljubljana, Slovenia*

In this talk, we present three QUBO (Quadratic Unconstrained Binary Optimization) relaxations for the sparsest k-subgraph (SkS) problem: a quadratic penalty relaxation, a Lagrangian relaxation, and an augmented Lagrangian relaxation, as introduced in [1,2]. The effectiveness of these approaches strongly depends on the choice of the penalty parameters. We present some original theoretical results characterizing the parameter values for which the QUBO relaxations are exact. For practical implementation, we propose three iterative algorithms: Quadratic Penalty Iterative Algorithm (QPIA), Lagrangian Relaxation Iterative Algorithm (LRIA), and Augmented Lagrangian Iterative Algorithm (ALIA), designed to address the computational challenges of solving SkS for larger instances. These algorithms dynamically adjust quadratic penalty and Lagrangian parameters and use the D-Wave quantum annealing and Simulated Annealing (SA) solvers in each iteration.

We provide extensive numerical experiments which (i) validate the exactness of the three relaxations using the exact QUBO solver BiqBin [3], which is parallelized and run on HPC, and (ii) empirically validate the effectiveness of the proposed iterative algorithms on various graph datasets, including Erdős–Rényi (ER) graphs, Bipartite graphs, and D-Wave topology graphs [1].

References

[1] Omkar Bihani, Roman Kužel, Janez Povh, Dunja Pucher: Quantum and Simulated Annealing-Based Iterative Algorithms for QUBO Relaxations of the Sparsest-Subgraph Problem, arXiv:2509.08544 (2025)

[2] Bihani, O., Povh, J., & Pucher, D. On Qubo Relaxations to the Sparsest k-Subgraph Problem. In: DROBNE, Samo (ed.), et al. SOR '25: proceedings of the 18th International Symposium on Operational Research in Slovenia: Bled, Slovenia, September 24-26 (2025)

[3] Gusmeroli, N., Hrga, T., Lužar, B., Povh, J., Siebenhofer, M., & Wiegele, A. BiqBin: a parallel branch-and-bound solver for binary quadratic problems with linear constraints. ACM Transactions on Mathematical Software (TOMS), 48(2), 1-31 (2022)

Quantum computer integration in multi-site HPC infrastructure

Peter Kandolf

University of Innsbruck - FSP Scientific Computing

In the course of the FFG project *Quantum Accelerated Computing Infrastructure* (QACI), the University of Innsbruck acquired a quantum computer (QC). The QC system is integrated into the local high performance computing (HPC) infrastructure, effectively creating a hybrid quantum and high performance computing (QC/HPC) infrastructure. The procured QC is not an *experimental device*, but rather designed for stability, and we expect high availability. Therefore, it can be operated as a workload-specific accelerator, a quantum processing unit (QPU), for the cluster.

The aim is to create a setup that allows for a unified workflow experience between HPC and QC. The HPC infrastructure can perform pre- and post processing while the quantum computer, or QPU, functions as an accelerator for specialized workloads. In a nutshell, we submit a *quantum circuit* to the QPU and receive the result. What happens in between is handled by to the QPU. This setup is similar to graphics processing units (GPUs) but differs in some key aspects.

In order to allow for a unified scheduling with Slurm, the possibility to gather some basic statistics, and increase security while still ensure a seamless integration, we inserted a proxy between the cluster and the QC, see Fig. 1. The proxy handles basic user authentication via a JSON Web Token (JWT), transport encryption with HTTPS, and extracts the number of *shots*, *qubits*, and *circuits*. Furthermore, we combine them with the user and job-id to get some basic statistics.

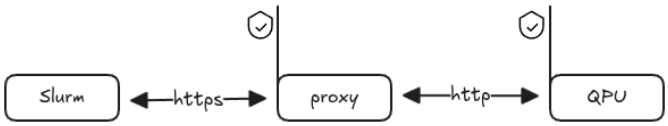


Fig. 1: Overview of the setup. On the left we see our HPC cluster as Slurm, in the middle a proxy to apply some basic security features, and on the right our QPU.

Apart from actually acquiring a system, the QACI project also aims to provide easy access for (Austrian) researchers and research groups to production quantum resources, to advance the research by exploring applications of quantum algorithms and to optimize current quantum systems. A second focus is set on teaching, to make sure a generation of quantum-aware researcher is trained to fully utilize the growing number of quantum resources.

The talk provides a detailed background of this project, describing the existing setup and the integration of the *trapped-ion* quantum system into the local HPC infrastructure. Furthermore, we will share some details about the applications that were already run successfully during the project, as well as the possibilities to access the resources for your own research.

References

[1] Anich G., Frisch A., Schwärzler H., Kandolf P., and Kumar P., in-preparation, 1 (2026).

jz-tree: Lightning fast neighbor search and friends-of-friends with dual tree traversal in JAX and CUDA

Jens Stücker

University of Vienna, Department of Astrophysics, Vienna, Austria

K-nearest neighbour search (kNN) and friends-of-friends clustering (FOF) are fundamental algorithms in many scientific disciplines, including astrophysics, machine learning, and data mining. In my talk I will present JZ-TREE, a novel implementation of these algorithms using dual tree traversal in JAX and CUDA.

I will demonstrate how a hierarchical spatial tree structure in Morton (or "Z") order with a fixed depth and a flexible number of children per node is very well suited for GPU computing. I will discuss how dual tree traversal algorithms can be implemented in custom CUDA kernels that may be invoked from jax's foreign function interface. Good memory coalescence is achieved by ensuring that the children of each node are stored in a continuous segment of memory that may be read collaboratively by a thread block. Finally, I will discuss how the presented algorithms can be scaled efficiently to extremely large data sets in multi-GPU and multi-node environments. While communication adds a small overhead, performance remains primarily compute bound. The examples in Fig. 1 and Fig. 2 show kNN and FOF performance for a uniform random distribution of points on the domain $[0, 1]$ in 3 dimensions.

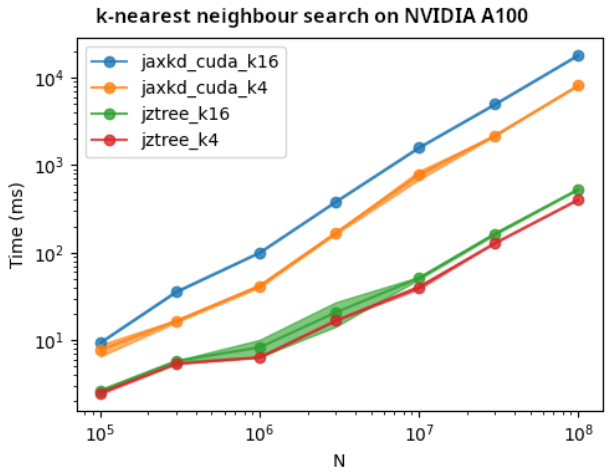


Fig. 1: Performance of JZ-TREE's kNN versus JAXKD (with CUDA) [1] – the previous go-to kNN implementation in JAX. Performance was measured on a single NVIDIA A100 as a function of the number of points. 'k4' and 'k16' indicate different numbers of returned neighbours. Our implementation achieves speed ups of order 10 for large problem sizes.

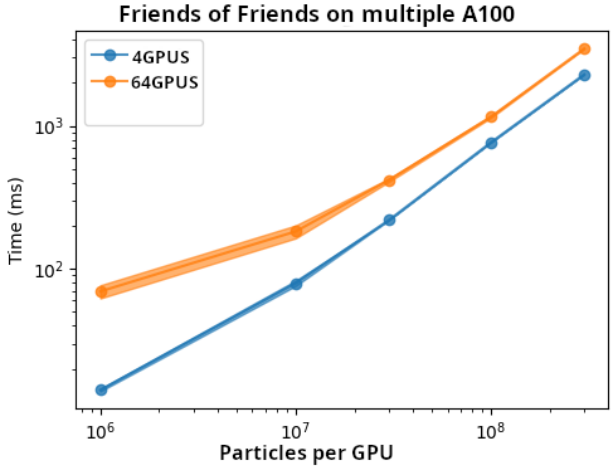


Fig. 2: Scaling of the FOF implementation of JZ-TREE for a distributed setup with 4 and 64 NVIDIA A100 GPUs. The linking length is 20% of the mean interparticle separation $\Delta x = (V/N)^{1/3}$. The x-axis shows the number of particles **per** GPU so that e.g. the last data point of 64 GPUs shows the performance for $2 \cdot 10^{10}$ particles.

References

[1] https://github.com/dodgebc/jaxkd/tree/main

Performance-Portable Particle-in-Cell with Multigrid Solvers on Heterogeneous CPU–GPU Node

Ivona Vasileska, Pavel Tomšič, and Leon Kos

Faculty of Mechanical Engineering, University of Ljubljana, Slovenia

Particle-in-Cell (PIC) codes for plasma simulations continue to be an essential tool, but their performance is challenged by irregular memory access, atomic scatter operations, and strong coupling between particle and field calculations. Modern High-Performance Computing systems include heterogeneous CPU-GPU nodes, making it essential to adopt programming models that offer both high performance and long-term portability.

In this paper, we introduce a performance-portable electrostatic PIC code using SYCL and Kokkos. Both versions have the same algorithmic structure and support: (i) particle pusher and field solver, (ii) atomic charge deposition, (iii) Monte Carlo collision (MCC) operators, and (iv) a scalable field solver using the Hypre library. For the field solver, we integrated the PFMG multigrid solver from the Hypre library and compared it with a baseline conjugate gradient (CD) approach. This work is an extension of the previous portability studies for PIC kernels [1, 2].

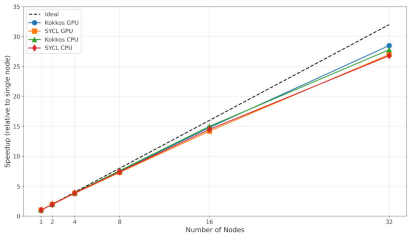


Fig. 1: Total runtime per PIC time step

The code follows a structure-of-arrays particle layout that aims to improve memory merge on GPUs and vectorization on CPUs. The most compute-intensive kernels are expressed as data-parallel loops in SYCL and Kokkos, while deposition uses portable atomic updates. MCC operators are implemented with per-particle random number generation and branch-heavy scattering, representing a realistic workload for low-temperature plasma simulations.

Experiments on a heterogeneous CPU–GPU platform show that both SYCL and Kokkos achieve comparable GPU performance, with differences below 10% for the full PIC step. As shown in Fig. 1, the GPU implementation offers a speedup over the multi-core CPU calculations. At the phase level, the dominant operations in the whole code runtime are the particle pusher and charge deposition (∼60–70%), with the field solver becoming more critical for large grid sizes, as presented in Fig. 2. By replacing the CG solver with the Hypre PFMG solver, a large decrease in the field solver's portion and the overall speedup for larger grid sizes is achieved.

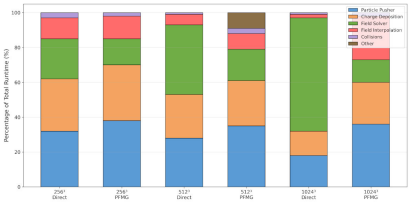


Fig. 2: Phase-level breakdown of GPU runtime

The study demonstrates that portable programming models can achieve near-equivalent GPU performance for PIC workloads, while scalable multigrid solvers are essential to maintain efficiency as grid resolution increases. The combination of SYCL/Kokkos with Hypre provides a maintainable path for PIC codes targeting future heterogeneous HPC systems.

References

[1] Vasileska, I., Tomšič, P., Kos, L., and Bogdanovič, L., "Unveiling performance insights and portability achievements between CUDA and SYCL for particle-in-cell codes on different GPU architectures," in *47th MIPRO*, 1115–1120 (2024).

[2] Vasileska, I., Tomšič, P., and Kos, L., "Accelerating sheath Particle-in-Cell simulations with StarPU," in *48th MIPRO*, 1200–1205 (2025).

LAMMPS Molecular Dynamics Simulations of Laser-induced Periodic Surface Structure Formation: Removal of desorbed atoms between Laser Shots

Matthias Weber and Wolfgang Husinsky

Institut für Angewandte Physik, TU Wien, Austria

Laser-Induced Periodic Surface Structures (LIPSS) emerge from complex, multi-physical laser–matter interactions involving ultrafast electronic excitation, thermal relaxation, and hydrodynamic surface reorganization. Understanding and predicting their formation requires simulation approaches that combine physical fidelity with large spatial and temporal scales, posing significant computational challenges.

This work presents simulation framework for LIPSS formation based on Molecular Dynamics (MD), designed explicitly for execution on modern HPC systems. The framework integrates atomistic MD simulations with continuum-scale electromagnetic field calculations to capture both thermal and non-thermal mechanisms of surface pattern formation. Key models include a Two-Temperature Model (TTM) coupled to MD for electron–lattice energy exchange, as well as Coulomb mechanisms to address ultrafast, ionization-driven material removal.

The implementation is built around LAMMPS and extended through custom Python routines interfaced via the LAMMPS Python API. These routines enable flexible surface morphology generation, precise atom removal between laser pulses, and efficient handling of large atom counts beyond the capabilities of native MD tools. This is done by embedding two coupled python routines into the input script, reading in a reference surface file from Ovito to examine the relative position of every atom and remove it with respect to the from laser-shot to laser-shot varying target surface. Parallelization is achieved using MPI(MPI4Py), which takes advantage of the domain decomposition of LAMMPS through the Python interface, while large-scale simulations are executed in SLURM-based workflows to efficiently use the HPC clusters with GPUs (*in most cases 8 nodes with 2 GPUs for each node; GPU enhanced Coulomb and EAM potentials in the LAMMPS simulation*) on VSC5. This configuration typically result in a 3-5 time improvement of the LAMMPS running time. To handle the substantial data volumes generated, the workflow incorporates optimized I/O strategies, automated post-processing with OVITO in headless mode, and iterative coupling with COMSOL Multiphysics for recalculating laser-induced electromagnetic fields on evolving surface geometries. The entire pipeline is fully automated, enabling systematic simulations and parametric studies across different physical models and parameters.
The presented framework demonstrates how custom MD extensions and tightly coupled multi-software workflows can be efficiently deployed on HPC infrastructures to study complex laser–matter interaction phenomena.

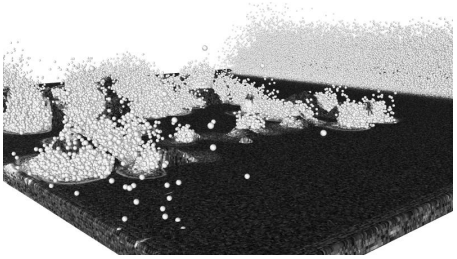


Fig. 1: Excitation of Atoms through incident Laserpulse—Visualized with Ovito

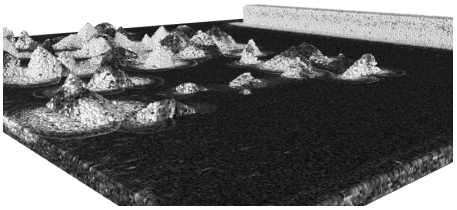


Fig. 2: Target after Atom Removal with Python Routine—Visualized with Ovito

Time-Series Forecasting and Alert Classification for Proactive IT Infrastructure Monitoring

Márk Dénes and András Schweighardt

DAM Invisible Technology Zrt., Hungary

Motivation: Modern IT infrastructure relies on continuous metric monitoring to maintain service availability. Our organization monitors over 3,000 hosts around the clock using the Icinga monitoring system. When metric values breach predefined thresholds, alerts are forwarded to the operations team as incident tickets. However, static threshold-based alerting produces a large volume of false positive alerts caused by transient fluctuations, scheduled tasks, and undocumented recurring system behaviors. Analysis of our production data reveals that approximately 95% of generated alerts are false positives, resulting in significant operational overhead and a fundamentally reactive approach to incident management.

Research Questions: We investigate three questions central to transforming IT operations from reactive to proactive: (1) Can diverse infrastructure metrics—CPU load, memory utilization, disk usage, swap space, network latency, process counts—be reliably forecast one hour ahead? (2) Can threshold-triggered alerts be classified as true or false positives? (3) Can these capabilities be combined to predict incidents before they occur?

Metric Forecasting: We employ the Informer architecture [1], a transformer-based model for efficient long-sequence time-series forecasting. To address the heterogeneity of monitored hosts spanning Linux and Windows platforms, we define a unified nine-feature metric schema and construct a pre-trained base model on the shared metric space of the full host fleet. We evaluate both the general pre-trained model and host-specific fine-tuned variants. Key challenges include irregular metric characteristics—from near-constant disk utilization to chaotic CPU load patterns—missing data, and domain-specific threshold semantics where metrics follow either "higher is worse" (e.g. CPU load) or "lower is worse" (e.g. free memory) conventions.

Alert Classification: We construct a labeled dataset by matching Icinga alerts with incident management records. The resulting dataset of over 41,000 alert windows across 338 hosts exhibits extreme class imbalance with a true positive ratio of only 5%. We design a classification model based on feature extraction from the multi-variate metric time series and investigate strategies to handle the severe imbalance while maintaining reliable detection of genuine alerts across all metric types.

Results and Outlook: Our forecasting model captures metric trends one hour ahead, providing early warning of approaching threshold violations. The alert classifier achieves 87.5% accuracy on held-out test data despite extreme class imbalance. Future work focuses on combining both models into a unified proactive alerting pipeline, scaling model capacity with additional training data, and exploring multi-host prediction to leverage cross-host correlations.

References

[1] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W., Proc. AAAI Conf. on Artificial Intelligence **35(12)**, 11106 (2021). arXiv:2012.07436.

Supporting file intensive AI Workloads on High Performance Computing

István Tamás PhD[a,b], Mihály Terjék[a,b], and Erzsébet Horváth[a,b]

[a] *Digital Government Development and Project Management Ltd.*
[b] *National Competence Center (NCC) - Hungary*

Utilizing High-Performance Computing (HPC) resources for various Artificial Intelligence (AI) workloads is becoming increasingly popular. However, AI-related work differs significantly from traditional HPC usage (e.g., using SMP, MPI, or multi-GPU parallelization for molecular dynamics versus training deep learning models for image recognition or image classification).

One of the key differences is the amount of raw data required for these workloads. Traditional computations usually require a different approach, and while these workloads can also create files with large file sizes (checkpoint files, trajectories, etc.) the problem with AI workflows arises when tremendous amount of training, validation and test data (e.g., images in the case of convolution) must be moved to storage (smaller in size but higher in file numbers). To improve the accuracy of these models, we may also need to incorporate cross-validation methods (e.g., K-fold) into our workflow. Using this method means that the training and validation data need to be resampled multiple times, without overlap. This approach again adds overhead to the data preparation phase and strains the HPC storage infrastructure (mainly due to smaller files and a large number of images, which can cause problems with file I/O).

The HPC cluster in Hungary (Komondor) is receiving an increasing number of AI projects from Academia and industry. The initial experience with AI projects showed us that we need to implement different techniques to support them, especially in the industry, where quick results are key.

We tested HDF5 (Hierarchical Data Format), LMDB (Lightning Memory-Mapped Database), Fuse (Filesystem in Userspace), and the simplest way to organize our test data: folders on our file system. We developed an image preparation pipeline using Python to load all data into HDF5 and LMDB files and tested several neural network models for comparison. For the tests, we trained AlexNet, DenseNet121, DenseNet201, ResNet151, InceptionResNet_v2, ConvNeXt, and ViT_16 (visual transformer) on our dataset and benchmarked their efficiency and precision. All testing and benchmarking are carried out with the PyTorch Python library on A100 GPUs. Around 382k images were compressed into 22 files (5-fold cross-validation), resulting in an overall file size of only 32 GB. This technique enabled easier data management and file transfer to the cluster.

We found that LMDB works well with our cluster, and we are continuously developing these pipelines for our users. Academic users are usually more familiar with various technologies (e.g., HDF5 or LMDB). Still, our goal is to minimize friction for industrial partners, who may not have time to develop these workflows. As the Hungarian NCC, one of our main goals is to open the HPC infrastructure to industrial partners as much as possible, and for this, we need to develop effective ways to make HPC usage easier, with an increasing demand for these workloads in mind.

Furthermore, after optimizing storage usage, the next step is to implement effective multi-GPU and multi-node GPU usage for AI workloads on our cluster using DDP (Distributed Data Parallel) and other approaches. As mentioned earlier, AI workflows require a fundamentally different approach to HPC, and optimizing storage utilization is one of the first key steps.

Innovative Expansion of HPC Infrastructure for Scalable AI Inference Using MACx GPUs

Tomi Ilijaš, Tristan Pahor, and Tomislav Šubić

Arctur d.o.o., Nova Gorica, Slovenia

The rapid growth of AI inference workloads is placing increasing pressure on high-performance computing (HPC) infrastructures originally optimized for scientific simulations and model training. At the same time, the strong dependence of many AI platforms on CUDA-based GPU ecosystems introduces cost, supply-chain, and vendor-lock challenges for HPC operators. This work presents an approach developed at Arctur for expanding an existing HPC infrastructure with more than 100 MACx accelerators in order to build a scalable and cost-efficient AI inference platform.

MACx GPUs are a general-purpose accelerator architecture developed by an international consortium of companies from Europe and Asia led by Torriatte Labs (Japan), and distributed in Europe by Toraks d.o.o. The platform is designed as an alternative to CUDA-centric accelerator ecosystems and supports common GPGPU workloads and multiple precision modes, including FP64, FP32, BF16, FP16, and INT8. The accelerators are compatible with widely used AI frameworks such as PyTorch and TensorFlow and integrate with open-source toolchains.

We describe the integration of MACx accelerators into an existing HPC cluster through a hybrid scheduling architecture that enables both traditional HPC workloads and AI inference pipelines to coexist on shared infrastructure. The deployment emphasizes modular cluster design and heterogeneous resource scheduling to maximize utilization.

Initial results indicate that inference clusters based on non-CUDA accelerators can achieve competitive throughput while improving resource utilization and reducing reliance on proprietary GPU ecosystems. The presented architecture illustrates a practical pathway toward more open, vendor-diverse AI infrastructure, which is particularly relevant for European HPC initiatives pursuing technological sovereignty and sustainable AI deployment.

References

[1] Garcia Lopez, P., Barcelona Pons, D., Copik, M., et al., AI Factories: It's Time to Rethink the Cloud-HPC Divide, arXiv: 2509.12849 (2025)

[2] Turisini, M., Amati, G., Cestari, M., LEONARDO: A Pan-European Pre-Exascale Supercomputer for HPC and AI Applications, arXiv: 2307.16885 (2023).

[3] Nikolic, S., Filipovic L., Ilijas T., Vukotic M.: FIT4HPC? - Accelerating digital transformation by supercomputing opportunities, J Supercomput **81**, 1069 (2025)

HPC-Enabled LLM Fine-Tuning and Machine Translation of Legal Texts on LEO5

Aleksandr Trklja

University of Innsbruck

Recent advances in Legal NLP have accelerated the adoption of transformer-based models and Large Language Models (LLMs) in legal text processing and linguistic analysis (Trklja and McAuliffe, 2018, Zhong et al., 2020; Katz et al., 2023; Lai et al., 2023). However, despite this growth, research on fine-tuned legal LLMs remains limited, as large-scale domain adaptation requires computational resources rarely available in legal and linguistic research. This project addresses this gap by developing an HPC-enabled workflow for fine-tuning and machine translation using open-weight transformer models on parallel European Court of Justice (ECJ) corpora. The work was made possible through access to the LEO5 supercomputing infrastructure at the University of Innsbruck, which provides the GPU capacity necessary for training domain-adapted models on large multilingual datasets.

The pipeline includes automated XML cleaning, sentence segmentation and alignment of metadata-rich judicial documents, followed by parameter-efficient fine-tuning (LoRA/PEFT) for English–German legal translation and subsequent machine translation experiments. Model performance is evaluated using established MT metrics (BLEU, chrF, COMET) (Papineni et al., 2002; Blagec et al., 2022). GPU-enabled LEO5 nodes were essential for managing the memory footprint of transformer architectures and ensuring stable optimisation within available VRAM.

The study demonstrates how HPC infrastructure enables domain-specific LLM development and high-quality machine translation for complex multilingual legal corpora, illustrating the expanding role of AI workloads from the humanities within high-performance computing environments.

References

[1] Blagec, K., Dorffner, G., Moradi, M., Ott, S., & Samwald, M. (2022). A Global Analysis of Metrics Used for Measuring Performance in Natural Language Processing. arXiv:2204.11574.

[2] Katz, D. M., Hartung, D., Gerlach, L., Jana, A., & Bommarito, M. (2023). Natural Language Processing in the Legal Domain. arXiv:2302.12039.

[3] Lai, J., Gan, W., Wu, J., Qi, Z., & Yu, P. S. (2023). Large Language Models in Law: A Survey. arXiv:2312.03718.

[4] Papineni, K., Roukos, S., Ward, T., & Zhu, W-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In ACL 2002 Proceedings, 311 318.

[5] Trklja, A., & McAuliffe, K. (2018). The European Union Case Law Corpus (EUCLCORP). In Proceedings, 217–226.

[6] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How Does NLP Benefit the Legal System: A Summary of Legal Artificial Intelligence. arXiv:2004.12158.

Index of presenting authors

List of ASHPC26 participants

Adam McCartney	adam.mccartney@tuwien.ac.at	ASC Research Center, TU Wien
Aleks Polak	apolak@nvidia.com	NVIDIA
Aleksander Grm	aleksander.grm@fpp.uni-lj.si	University of Ljubljana, Faculty of Mechanical Engineering
Alexander Ostermann	alexander.ostermann@uibk.ac.at	University of Innsbruck, Forschungsschwerpunkt Scientific Computing
Ali Meral	ali.meral@tuwien.ac.at	ASC Research Center, TU Wien
Alja Prah	alja.prah@ijs.si	Jožef Stefan Institute
Alois Schlögl	alois.schloegl@ist.ac.at	Institute of Science and Technology Austria, Scientific Computing
Andras Dorn	dorn.bandi@damit.hu	DAM Invisible Technology
András Dorn	dorn.andras@damit.hu	DAM Invisible Technology
András Schweighardt	schweighardt.andras@damit.hu	DAM Invisible Technology
Andre Heidekrueger	andre.heidekrueger@amd.com	Advanced Micro Devices GmbH
Andreas Lindner	andreas.lindner@advanced-computing.at	Advanced Computing Austria ACA GmbH
Andreas Rauber	rauber@ifs.tuwien.ac.at	ASC Research Center, TU Wien
Andreas Ravazzolo-Mehrle	amehrle@mathworks.com	MathWorks GmbH
Andrej Filipcic	andrej.filipcic@ijs.si	Jožef Stefan Institute
Antonija Rajic	e11821923@student.tuwien.ac.at	TU Wien
Anže Hubman	anze.hubman@ki.si	National Institute of Chemistry, Slovenia, Theory Department
Atul Singh	atul.singh@tuwien.ac.at	ASC Research Center, TU Wien
Bernd Saurugger	bernd.saurugger@tuwien.ac.at	ASC Research Center, TU Wien
Bettina Benesch	bettina.benesch@advanced-computing.at	Advanced Computing Austria ACA GmbH
Bettina Lindner	blindner@mathworks.com	MathWorks GmbH
Bojan Zunkovic	bojan.zunkovic@fri.uni-lj.si	University of Ljubljana and Rudolfovo
Branimir Kolarek	branimir.kolarek@irb.hr	Ruđer Bošković Institute, Centre for Informatics and Computing
Christian Kracher-Fischer	christian.kracher@univie.ac.at	University of Vienna, Zentraler Informatikdienst
Christina Müllner	christina.muellner@tuwien.ac.at	ASC Research Center, TU Wien
Christopher Huggins	christopher.huggins@vastdata.com	VAST Data
Claudia Blaas-Schenner	claudia.blaas-schenner@tuwien.ac.at	ASC Research Center, TU Wien
Darin Lah	darin.lah@izum.si	Institute of Information Science (IZUM)

David Lajevec	david.lajevec@gmail.com	University of Ljubljana, Faculty of Mechanical Engineering
Davor Davidović	davor.davidovic@irb.hr	Ruđer Bošković Institute, Centre for Informatics and Computing
Dejan Lesjak	dejan.lesjak@ijs.si	Jožef Stefan Institute
Diego Medeiros Dalla Costa	diego.dalla.costa@tuwien.ac.at	ASC Research Center, TU Wien
Dieter Kranzlmüller	Dieter.Kranzlmueller@lrz.de	Leibniz Supercomputing Centre (LRZ)
Dieter Kvasnicka	dieter.kvasnicka@tuwien.ac.at	ASC Research Center, TU Wien
Dominic McKendry	dominic.mckendry@thinkparq.com	ThinkParQ GmbH
Eduard Reiter	eduard.reiter@uibk.ac.at	University of Innsbruck, Forschungsschwerpunkt Scientific Computing
Elias Wimmer	elias.wimmer@tuwien.ac.at	ASC Research Center, TU Wien
Emmanuel Kasprzyk	emmanuel.kasprzyk@tuwien.ac.at	TU Wien, Service Unit of High Performance Computing / Campus IT
Ernst Haunschmid	ernst.haunschmid@tuwien.ac.at	ASC Research Center, TU Wien
Ezhilmathi Krishnasamy	ezhilmathi.krishnasamy@gmail.com	University of Ljubljana, Faculty of Mechanical Engineering
Filip Kocina	filip.kocina@tuwien.ac.at	ASC Research Center, TU Wien
Florian Atzenhofer-Baumgartner	florian.atzenhofer-baumgartner@uni-graz.at	University of Graz
Florian Goldenberg	florian.goldenberg@tuwien.ac.at	ASC Research Center, TU Wien
Florian Jäger	florian.jaeger@tuwien.ac.at	TU Wien, Service Unit of High Performance Computing / Campus IT
Florian Klauser	florian.klauser@ijs.si	Jožef Stefan Institute
Franci Merzel	franci.merzel@ki.si	National Institute of Chemistry, Slovenia, Theory Department
Gašper Jug	gasper.jug@ijs.si	Jožef Stefan Institute
Georg Heiler	georg.kf.heiler@gmail.com	Complexity Science Hub Vienna (CSH) and Austrian Supply Chain Intelligence Institute (ASCII)
Gerhard Hummer	gerhard.hummer@biophys.mpg.de	Max Planck Institute of Biophysics
Gregor Molan	gregor.molan@comtrade.com	CT Management d.o.o., Comtrade
Gundolf Haase	gundolf.haase@uni-graz.at	University of Graz, Mathematics and Scientific Computing
Herbert Störi	herbert.stoeri@tuwien.ac.at	ASC Research Center, TU Wien
Hernan Picatto	hernan.picatto@ascii.ac.at	Austrian Supply Chain Intelligence Institute (ASCII)
Inga Lorenz	inga.lorenz@vastdata.com	VAST Data
Istvan Tamas	istvan.tamas@dkf.hu	Digital Government Development and Project Management Ltd., HPC Technological and Service Office

Iulia Rinea	iulia.rinea@ai-at.eu	Advanced Computing Austria ACA GmbH
Iuliia Wimmer	julia.wimmer@tuwien.ac.at	ASC Research Center, TU Wien
Ivan Vialov	ivan.vialov@tuwien.ac.at	ASC Research Center, TU Wien
Ivona Vasileska	ivona.vasileska@fs.uni-lj.si	University of Ljubljana, Faculty of Mechanical Engineering
James McKevitt	jamesm20@univie.ac.at	University College London, Mullard Space Science Laboratory
Jan Javoršek	jona.javorsek@ijs.si	Jožef Stefan Institute
Jan Zabloudil	jan.zabloudil@tuwien.ac.at	ASC Research Center, TU Wien
Janez Povh	janez.povh@rudolfovo.eu	Rudolfovo – Science and technology center Novo Mesto
Jelena Joksimović	jelena.joksimovic@rudolfovo.eu	Rudolfovo – Science and technology center Novo Mesto
Jens Oliver Stücker	jens.stuecker@univie.ac.at	Universitätssternwarte Wien
Jesper Larsson Träff	traff@par.tuwien.ac.at	TU Wien, Faculty of Informatics
Johannes ROSINA	johannes.rosina@jku.at	Johannes Kepler Universität Linz, Institute for Theoretical Physics
Jonas Sieberer	jonas.sieberer@plus.ac.at	Universität Salzburg, FB Informatik
Jure Borišek	jure.borisek@ki.si	National Institute of Chemistry, Slovenia, Theory Department
Jurica Špoljar	jspoljar@srce.hr	University of Zagreb, University Computing Centre (SRCE)
Katrin Muck	katrin.muck@tuwien.ac.at	ASC Research Center, TU Wien
Krisztian Koronics	koronics.krisztian@damit.hu	DAM Invisible Technology
Latha Venkataraman	Latha.Venkataraman@ist.ac.at	Institute of Science and Technology Austria and Columbia University
Lea Enzenberger	lea.enzenberger@tuwien.ac.at	TU Wien
Leon Kos	leon.kos@lecad.fs.uni-lj.si	University of Ljubljana, LECAD Lab. Mech. Eng.
Leon Schwarzäugl	leon.schwarzaeugl@imba.oeaw.ac.at	IMBA – Institut für Molekulare Biotechnologie GmbH
Lukas Winkler	l.winkler@univie.ac.at	University of Vienna, Department of Astrophysics
Maizura Ibrahim	maizura@nm.gov.my	Malaysian Nuclear Agency, Information Technology Centre
Malgorzata Goiser	malgorzata.goiser@tuwien.ac.at	ASC Research Center, TU Wien
Manuel Hofbauer	manuel.hofbauer@ait.ac.at	AIT Austrian Institute of Technology GmbH
Marie Czuray	marie.czuray@tuwien.ac.at	ASC Research Center, TU Wien
Marija Bijelic	marija.bijelic@advanced-computing.at	Advanced Computing Austria ACA GmbH
Mark Antal	antal.mark@damit.hu	DAM Invisible Technology
Mark Márk Dénes	denes.mark@damit.hu	DAM Invisible Technology
Mark Dokter	mark.dokter@advanced-computing.at	Advanced Computing Austria ACA GmbH

Marko Ferme	marko.ferme@um.si	University of Maribor
Marko Tkalcic	marko.tkalcic@famnit.upr.si	University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies
Markus Hickel	markus.hickel@tuwien.ac.at	ASC Research Center, TU Wien
Markus Oppel	markus.oppel@univie.ac.at	University of Vienna, Institute of Theoretical Chemistry
Markus Stöhr	markus.stoehr@tuwien.ac.at	ASC Research Center, TU Wien
Markus Widmer	markus.widmer@hpe.com	Hewlett Packard Enterprise
Martin Thaler	martin.thaler@uibk.ac.at	University of Innsbruck, Zentraler Informatikdienst
Martin Žnidaršič	martin.znidarsic@ijs.si	Jožef Stefan Institute
Maša Lah	masa.lah@ki.si	National Institute of Chemistry, Slovenia, Laboratory for Molecular Modeling
Matteo Ciardi	matteo.ciardi@tuwien.ac.at	TU Wien, Institute of Theoretical Physics (ITP)
Matthias Weber	e11939547@student.tuwien.ac.at	TU Wien
Max Hodapp	maxludwig.hodapp@mcl.at	Materials Center Leoben
Max Resch	max.resch@donau-uni.ac.at	University for Continuing Education Krems, Department for Arts and Cultural Studies
Michael Iro	michael.iro@ai-at.eu	ASC Research Center, TU Wien
Michael Neumayer	michael.neumayer@univie.ac.at	University of Vienna, CUBE – LiSC
Michael Otto	michael.otto@uni-graz.at	University of Graz, Department of Digital Humanities
Mladen Borovič	mladen.borovic@um.si	University of Maribor, Laboratory for Heterogeneous Computer Systems
Moritz Siegel	moritz.siegel@tuwien.ac.at	ASC Research Center, TU Wien
Muhammad Mizan	muhammad.mizan@tuwien.ac.at	TU Wien
Nataša Miletić	natasa.miletic@famnit.upr.si	University of Primorska, FAMNIT, Data Science
Orlenys Natali Troconis	o.troconis@cineca.it	CINECA, High Performance Computing
Pavel Tomšič	pavel.tomsic@fs.uni-lj.si	University of Ljubljana
Peter Kandolf	peter.kandolf@uibk.ac.at	University of Innsbruck, Forschungsschwerpunkt Scientific Computing
Philipp Gschwandtner	philipp.gschwandtner@uibk.ac.at	University of Innsbruck, Department of Computer Science
Relindis Rott	Relindis.Rott@v2c2.at	Virtual Vehicle Research GmbH
Robert Elsässer	elsa@cs.sbg.ac.at	Universität Salzburg, FB Informatik
Ruben Laso	ruben.laso.rodriguez@univie.ac.at	University of Vienna, Faculty of Computer Science

Samo Lorenčič	samo.lorencic@izum.si	Institute of Information Science (IZUM), HPC
Samo Miklavc	samo.miklavc@izum.si	Institute of Information Science (IZUM), HPC
Sanaz Sattari	sanaz.sattari@tuwien.ac.at	ASC Research Center, TU Wien
Sascha Hunold	sascha.hunold@tuwien.ac.at	TU Wien
Sayed Maudodi	sayed.maudodi@amd.com	Advanced Micro Devices GmbH
Sebastian Kalcher	sebastiank@nvidia.com	NVIDIA
Sebastian Sitkiewicz	sebastian.sitkiewicz@pwr.edu.pl	Wroclaw Centre for Networking and Supercomputing
Sebastien Strban	sebastien.strban@ijs.si	Jožef Stefan Institute
Séverine Habert	shabert@nvidia.com	NVIDIA
Siegfried Höfinger	siegfried.hoefinger@tuwien.ac.at	ASC Research Center, TU Wien
Silvio Heinze	silvio.heinze@oeaw.ac.at	Austrian Academy of Sciences, Institute for Urban and Regional Research
Simon Panyella Pedersen	simon.pedersen@tuwien.ac.at	TU Wien, Institute of Theoretical Physics (ITP)
Srinath Bulusu	srinath.bulusu@tuwien.ac.at	TU Wien, Institute of Theoretical Physics (ITP)
Stephan Köstlbacher	stephan.koestlbacher@aithyra.at	AITHYRA GmbH – Research Institute for Biomedical Artificial Intelligence of the Austrian Academy of Sciences
Stefan Schulz	schulzstefan@eviden.com	Bull GmbH, Sales HPC, AI & Quantum Central Europe
Sylvia Ploeckinger	sylvia.ploeckinger@univie.ac.at	University of Vienna, Department of Astrophysics
Teo Prica	teo.prica@izum.si	Institute of Information Science (IZUM), HPC
Thanayut Seethongchuen	thanayut.seethongchuen@tuwien.ac.at	ASC Research Center, TU Wien
Thomas Haschka	thomas.haschka@tuwien.ac.at	TU Wien, Service Unit of High Performance Computing / Campus IT
Thomas Rattei	thomas.rattei@univie.ac.at	University of Vienna, Centre for Microbiology and Environmental Systems Science
Till Kahlke	till.kahlke@uibk.ac.at	University of Innsbruck, Department of Computer Science
Tina Marc	tina@arctur.si	Arctur d.o.o.
Tobias Pfennig	tobias.pfennig@megware.com	MEGWARE Computer Vertrieb und Service GmbH, High Performance Computing (HPC)
Tomaž Šuštar	tomaz.sustar@arnes.si	Academic and Research Network of Slovenia (ARNES)
Tristan Pahor	tristan.pahor@arctur.si	Arctur d.o.o.
Ümit Seren	uemit.seren@gmi.oeaw.ac.at	Vienna BioCenter, Scientific Computing
Valentin Hirschbrich	valentin.hirschbrich@tuwien.ac.at	ASC Research Center, TU Wien

Victoria Döller	victoria.doeller@tuwien.ac.at	ASC Research Center, TU Wien
Yin Wang	Yin.Wang@uibk.ac.at	University of Innsbruck, Theoretical Chemistry
Zdravko Krpić	zdravko.krpic@ferit.hr	University of Osijek
Žiga Zebec	ziga.zebec@izum.si	Institute of Information Science (IZUM)
Zoé Lloret	zoe.lloret@univie.ac.at	University of Vienna, Climate Dynamics and Modeling

16 participants preferred not to be listed.

DOI: https://doi.org/10.25365/phaidra.765

ISBN: 978-3-200-10998-8

Published by:
EuroCC Austria
c/o Universität Wien
Universitätsring 1
1010 Vienna, Austria
https://eurocc-austria.at/

Edited by:
Ivan Vialov, ASC Research Center, TU Wien, 2026

Layout:
Irene Reichl and Claudia Blaas-Schenner, ASC Research Center, TU Wien, 2016