ASHPC25
Austrian-Slovenian
HPC Meeting 2025

ASHPC25

Austrian-Slovenian HPC Meeting 2025

Rimske Terme, Slovenia
19–22 May 2025

ashpc.eu

Austrian-Slovenian HPC Meeting 2025 – ASHPC25

Rimske Toplice, 19–22 May 2025

https://ashpc.eu

Welcome to ASHPC25

The organizing and program committees are delighted to welcome you to the Austrian-Slovenian HPC Meeting, ASHPC25, held this year in the beautiful surroundings of Rimske Terme, Slovenia.

Historically, ASHPC (and former AHPC) meetings have been a place to bring together those who provide computing resources to academic research and those who use these resources. The topics of the submitted abstracts demonstrate that this goal has been achieved again this year.

The application areas range from weather and climate research to physics, chemistry, material science, and biology. The keynotes by Sam Hatfield, Alessandro Laio, and Jan Šuntajs will provide valuable insights and explain why HPC is crucial for their respective research fields. Additionally, we have received a significant number of submissions in the area of Artificial Intelligence and Machine Learning (AI/ML). These submissions cover both the applications of AI/ML and advancements in benchmarking and tooling to improve algorithms for running AI. The keynotes by Dan Alistarh and Erwin Laure will offer further details on the recent developments in this area. We are also pleased to have submissions from the operational side of various HPC research organizations, which offer a rich mix of experiences about how to operate HPC systems, the challenges faced, and how these challenges are addressed. We are also pleased to welcome international contributions from countries including Croatia, Czechia, Germany, Hungary, Italy, Poland, Slovakia, and the United Kingdom.

In total, 60 abstracts were submitted and while we had more submissions for talks than available speaker slots, we were able to accept approximately 60% of them. To address this, we have introduced a 1-minute lightning talk for each poster contribution, ensuring a platform for all valuable contributions. The selection process was rigorous, with each contribution being evaluated by at least two reviewers. We hope that your time here at Rimske Terme will be enriching, offering opportunities to forge new connections, strengthen existing ones, and spark innovative ideas for the future of high-performance computing.

We would like to thank all the participants, sponsors, and collaborators for making this event possible.

On behalf of the program and organizing committees, welcome to ASHPC25 – let's make it a memorable and fruitful experience.

Sponsors

Platinum sponsor

Silver sponsor

Main financial support for ASHPC25 is provided by the Austrian HPC Association.

Schedule

Monday, May 19, 2025

Start		Title
14:00	Get Together & Coffee	
15:00	Central European NCCs Workgroup Meeting – CE-NCCs-WG (for NCC / EuroCC 2 staff only)	
17:00	EuroHPC Ecosystem: How to get Access for HPC and AI (open to all)	
17:00	Philipp Gschwandtner	EuroHPC Systems and Access: From your Workstation to 10.000 CPUs and GPUs
17:20	Žiga Zebec	AI Factory: A Comparative Guide to Access AI Factory's Computing Power
17:40	Claudia Blaas-Schenner Philipp Gschwandtner Žiga Zebec	Q & A – How to get Access for HPC and AI
18:00	Registration	
19:00	Dinner	

Tuesday, May 20, 2025

Start		Title
09:00	**Welcome by the Program and Organizing Chairs**	
09:15	Erwin Laure	HPC & AI - Competition or Collaboration? (Keynote)
10:00	**Plenary: Block 1**	
10:00	Aleš Čep	Slovenian Genomic Data Infrastructure
10:15	Julian Mangott	`atropy`: a dynamical low-rank solver for the chemical master equation
10:30	Ratko Pilipović	Sparse matrix operations on RISC-V vector architecture
10:45	**Coffee Break**	
11:15	**Plenary: Block 2**	
11:15	Sebastian Sitkiewicz	HPC Info - Enhanced Resource Tracking for SLURM Users and Administrators
11.30	Florian Goldenberg	Strong Scaling Days: Testing massively parallel applications on VSC
11.45	Matjaž Pančur	Flexible and compact high density modular data center
12:00	Mark Dokter	Blurring the lines between compilation and runtime optimization in the DAPHNE open source software framework
12:15	Ruben Laso	Unveiling the Energy Cost of Parallel Performance Portability in C++
12:30	**Lunch**	
14:00	Sam Hatfield	Exascale Numerical Weather Prediction at ECMWF (Keynote)
14:45	**Plenary: Block 3**	
14:45	Blaž Gasparini	High clouds, high grid spacing to the rescue?
15:00	Maximilian Meindl	Using machine learning to distinguish km-scale climate models and observations on a regional scale
15:15	Mostafa Kiani Shahvandi	Strategies for faster deep learning-based modeling of global barystatic processes
15:30	**Lightning Talks: Poster presentations**	
	Ivona Vasileska	Performance Evaluation of Parallel Approaches for 1D PIC Simulations on GPUs
	Luis Casillas-Trujillo	AURELEO: Austrian Users at LEONARDO supercomputer
	Wiktor Nastał	Integrating Linux with HTTP: Secure and Automated Workflows
	Michael Blaschek	FLEXWEB - A flexible particle dispersion model web interface
	Karina Pešatová	Empowering Women in High Performance Computing: Activities of the Central European Chapter of Women in HPC
	Karina Pešatová	Understanding and Addressing Evolving Training Needs in High-Performance Computing: Insights from the 2024 Training Services Survey
	Andrej Mihelič	Treating atomic photoionization with a modified time-dependent surface flux method
	Martin Ljubič	Influence of membrane composition on the signaling of the NKG2A/CD94/HLA-E complex investigated by all-atom simulations
	Teo Prica	Optimization of Small Language Models (SLMs)
	Tomas Karasek	NCC Czechia: Sucess Stories
	Aiko Voigt	ICON @ VSC: strong scaling tests with a global km-scale model of the atmosphere
	Adam McCartney	VSC's Software Stack Envisioned
	Štefan Costea	Heterogeneous Exascale Particle-in-Cell
	Zoé Lloret	Kilometer-scale Climate Modeling of TRAPPIST-1e Using ICON-Sapphire: Advancing Exoplanet Habitability Studies

Start		Title
	Emir Imamagić	Support Systems for National Advanced Computing Service
	Victoria Bringmann	Is the IPMI Exporter a Reliable Tool for Power Monitoring?
	Neli Sedej	Predicting rates of conformational change of proteins from projected molecular dynamics simulations
	Tina Črnigoj Marc	FFplus: Driving SME and Startup Innovation by unleasing the potential of HPC and Generative AI
	Tina Črnigoj Marc	EXCELLERAT CoE: The European Centre of Excellence for Engineering Applications
	Tomas Kozubek	European Master for HPC study programme
	Lukas Winkler	Scaling Differentiable Simulations in Cosmology to Multiple GPUs
16:00	**Coffee Break/Poster Exhibition**	
17:00	**Roundtable**	Achieving Gender Balance in HPC: Retention and Representation from the Central European Perspective
18:00	**VSC User Meeting**	
19:15	**Dinner**	

Wednesday, May 21, 2025

Start		Title
09:00	Dan Alistarh	Compressing AI Models at GPT Scale (Keynote)
09:45	**Plenary: Block 4**	
09:45	Matevž Jug	Learning macroscopic equations of motion from particle-based simulations of a fluid
10:00	Majid Salimi Beni	Optimizing Distributed Deep Learning Training by Tuning NCCL#
10:15	Ioannis Vardas	ncclsee: A Lightweight Profiling Tool for NCCL
10:30	Siegfried Hoefinger Simeon Harrison	HPC for Hybrid Threat Resilience: Insights from the HYBRIS Project
10:45	**Coffee Break**	
11:15	**Plenary: Block 5**	
11:15	Uroš Lotrič	Go wrapper for CUDA
11:30	Stefano Elefante	Benchmarking A40, L40S, H100 GPUs
11:45	Domen Verber	Accelerating Differential Evolution for High-Performance Computing: Leveraging Modern GPU Architectures and Mixed-Precision Arithmetic
12:00	Davor Davidović	Benchmarking the scalability and communication of the CholeskyQR2-IM algorithm on national and European HPC resources
12:15	Domen Vreš	GaMS meets Vega and Leonardo: Training Slovene LLMs
12:30	**Lunch**	
14:00	Jan Šuntajs	High performance computing at the boundaries of quantum chaos (Keynote)
14:45	**Plenary: Block 6**	
14:45	Mátyás Koniorczyk	QUBO and quantum annealing: applications and perspectives
15:00	Janez Povh	Hybrid Classical-Quantum Exact Solver for the QUBO Problem
15:15	Philipp Gschwandtner	Bootcamp Performanceoriented Softwarengineering — Experiences from working with real-world problems and developers
15:30	**Coffee Break**	
16:00	**Plenary: Block 7**	
16:00	Florian Goldenberg	MUSICA: Current situation and developments
16:15	Jan Zabloudil	How to play MUSICA
16:30	Markus Hickel	When space runs out: A new central storage system for VSC
16:45	**AI Factories**	
16:45	Orlenys Troconis	CINECA infrastructure, from AI factory to quantum computers
17:00	Markus Stöhr	AI Factory Austria AI:AT
17:15	Jan Jona Javoršek	SLAIF: Slovenian AI Factory (with a new EuroHPC AI optimized system)
17:30		Roundtable AI Factories
19:15	**Dinner**	

Start		Title
14:00	**Parallel session: Container Forum**	
14:00	Alja Prah	EPICURE and Containers, activities in EuroHPC
14:45	Barbara Krasovec	Confidential containers in multi-tenant HPC environments
15:00	Wiktor Nastał	Integrating Linux with HTTP: Secure and Automated Workflows
15:15		Container Forum
16:00	**Parallel session: Container Forum**	
16:00		PC/AI dev-ops and administrator birds of feather meeting

Thursday, May 22, 2025

Start		Title
09:00	Alessandro Laio	When data are big in the "wrong "direction: identifying compact and informative distance measures in high-dimensional feature spaces (Keynote)
09:45	**Plenary: Block 8**	
09:45	Žiga Zebec	AI-zyme: machine learning enhanced protein design
10:00	Draško Tomić	Cell simulation: the ultimate way for developing better drugs
10:15	Alexander Genest	H2O adsorption at Co3O4 (111) surface from a DFT perspective
10:30	Maša Lah	Open-boundary molecular dynamics of ultrasound using supramolecular water models
10:45	**Coffee Break**	
11:15	**Plenary: Block 9**	
11:15	András Dorn	Enhancing AI Models in Local Language for IT Service Management
11:30	Jernej Stare	Exploring the origin of enzyme catalysis by simulation methods
11:45	Sergii Khmelevskyi	Efficient theoretically guided search for functional metallic thermoelectrics
12:00	Siegfried Hoefinger	Research Software Engineers (RSE) at the VSC Research Centre
12:15	**Closing**	
12:30	**Lunch**	

Contents

Keynote Talk:

HPC & AI - Competition or Collaboration

Erwin Laure

Max Planck Computing and Data Facility, Garching, Germany

Since Generative AI has become mainstream through Large Language Models like e.g. employed in ChatGPT or DeepSeek, AI is increasingly considered as a potential tool in scientific workflows. While classical AI is in mainstream use in image based research (e.g. for analyzing brain scans) for many years, other domains are still in the explorative phase. But this is changing at an enormous speed as e.g. exemplified by the recent announcement of ECMWF to use AI in their weather forecast[1]. Yet, how far AI can replace classical simulations, is still subject to ongoing debates.

At the same time, AI has a profound impact on HPC hardware industry. Double precision, typically employed in scientific simulations, is not needed for AI and chip manufacturers start to reduce double precision capabilities in favour of low precision units. This is not surprising, given an AI market that is several orders of magnitude larger than the HPC one.

In this talk we review some of the impact AI has made in scientific computing, using examples from practical AI use within the Max Planck Society. We also review the impact, AI has on hardware industry and how this affects classical scientific computing. Whatever the future will bring: AI has come to stay and while it is a competition to classical HPC in some respect, those being able to effectively exploit AI capabilities will likely have a competitive advantage.

[1] https://www.ecmwf.int/en/about/media-centre/news/2025/ecmwfs-ai-forecasts-become-operational

Slovenian Genomic Data Infrastructure

Aleš Čep, Marko Ferme, and Milan Ojsteršek

Faculty of Electrical Engineering and Computer Science, University Of Maribor, Slovenia

High-Performance Computing (HPC) systems are critical for solving computationally demanding problems across various scientific disciplines. In bioinformatics, vast amounts of data are generated through sequencing technologies, requiring complex analyses such as genome assembly and variant calling. HPC provides the computational power and storage capacity needed to handle these intensive workloads. This presentation highlights the infrastructure of a national node built to enable secure and interoperable genomics data analysis. It facilitates collaboration across research and medical institutions throughout Europe [1].

The national node leverages HPC infrastructure to serve as a central hub for phenotypic and genomic data contributed by institutional nodes. It provides researchers and medical professionals with a unified platform to access all relevant data while ensuring a secure and robust environment for processing and analysis. To guarantee flexibility, scalability, and isolation between components, all services and workflows within the node are containerized using Docker. The infrastructure is built around three key functionalities:

- **Storage:** Securely stores genomics data files, such as BAM and CRAM files, in an encrypted format to ensure data confidentiality. Access to non-public data requires authentication and authorization. Additionally, downloading raw genomic data files is subject to approval by a Data Access Committee.

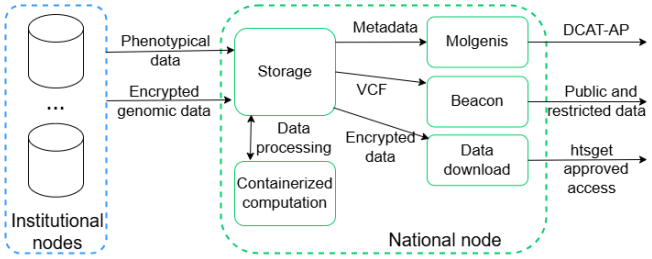


Fig. 1: Overview of the national node infrastructure.

- **Data Access Tools:** Includes tools like Beacon, a data discovery tool that enables researchers to search genomic and biomedical datasets, and Molgenis, a platform for managing metadata and ensuring efficient data exploration and management.
- **Computation Environment:** Provides a framework for processing genomics data. It supports bioinformatics workflows such as genome assembly, variant calling, and other analyses.

Institutional nodes are equipped with small HPC nodes, while the national node leverages both HPC Vega and the ARNES data infrastructure. It provides a scalable and secure infrastructure for managing and analyzing genomic data in a secure processing environment. By integrating advanced storage solutions, robust security measures, and computational frameworks, the national node facilitates efficient data discovery, processing, and analysis. This infrastructure fosters collaboration among research and medical institutions, driving advancements in genomics and bioinformatics across Europe.

References

[1] European Genomic Data Infrastructure, "The European Genomic Data Infrastructure (GDI) Project," `https://gdi.onemilliongenomes.eu/`, accessed January 20, 2025.

`atropy`: a dynamical low-rank solver for the chemical master equation

Stefan Brunner[a], Lukas Einkemmer[a], Julian Mangott[a], and Martina Prugger[b]

[a] *Institut für Mathematik, Universität Innsbruck, Austria*
[b] *Max-Planck-Institut für Plasmaphysik, Germany*

Understanding the reactions in the biological cell is crucial for the development of new drugs and therapies. Numerical simulations became an indispensable tool for studying the reactions networks in the cell, but for an accurate description, stochastic effects have to be taken into account. This requires the solution of the chemical master equation (CME), which however suffers from the *curse of dimensionality*: the solution of the CME is a high-dimensional probability distribution and the memory requirements grow exponentially with the number of chemical species. This renders the simulation of such networks infeasible on even the largest available supercomputers. Therefore, time-dependent model order reduction techniques such as the dynamical low-rank (DLR) approximation are needed to tame the complexity.

In this talk, we present `atropy`, a software which computes the DLR approximation for the CME efficiently with binary tree tensor networks (`https://atropy.gitlab.io/`). The main idea of the DLR approximation is to separate the reaction network hierarchically into smaller partitions. Each partition is described by a set of low-dimensional basis functions, which depend on all species in their partition. This approach treats all reaction pathways inside of a partition exactly. An approximation is only performed if a reaction crosses the partition boundary. The approximation error is controlled by the so-called *rank*, i.e., the total number of basis functions for each partition. Our approach allows to keep tightly coupled species together in the same partition without introducing any error. The hierarchical decomposition overcomes the curse of dimensionality and simulations, which were previously intractable on supercomputers, can now be performed on common computing devises such as laptops.

The DLR approximation is related to the Hartree–Fock approximation in quantum mechanics. In the latter approach, single-orbital basis functions, which only depend on the coordinates of a single electron, are combined to obtain an approximation for the high-dimensional wave function. In [1], this idea has been directly applied to the CME. The problem with this method is that each of the low-rank factors is only allowed to depend on a single species. However, in many biological networks, chemical species are tightly connected by reaction pathways. It is doubtful that species can be considered independently, while obtaining an accurate approximation with a small rank.

We demonstrate by a 20-dimensional reaction cascade (i.e. a reaction network which consists of twenty different chemical species) the effectiveness of our approach: solving the full CME for this problem requires approximately 10^{31} MB of main memory, whereas for the low-rank approximation with tree tensor networks only 2.3 MB are needed. Our approach is, in contrast to Monte Carlo methods, completely noise-free and therefore resolves fine details of the probability distribution.

References

[1] Jahnke, T., and Huisinga, W., Bulletin of Mathematical Biology **70**, 2283 (2008).

Sparse matrix operations on RISC-V vector architecture

Andrej Sušnik[a], Uroš Lotrič[a], Josip Knezović[b], Ratko Pilipović[a], and Davor Sluga[a]

[a] *Faculty of Computer and Information Science, University of Ljubljana, Slovenia*
[b] *Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia*

Matrix operations with sparse matrices play an important role in many areas, such as scientific simulations, machine learning, and graph analysis. Unlike operations with regular dense matrices, working with sparse matrices can result in performance issues due to non-local and irregular memory access patterns. To address these challenges, different compressed matrix formats are employed [1]. These formats aim to reduce memory usage, enhance computational efficiency, or both. The choice of format can significantly affect the performance of sparse matrix operations.

The challenges related to sparse matrix operations differ depending on the computer architecture in use, whether it is Central Processing Units (CPUs), Graphics Processing Units (GPUs), or Field Programmable Gate Arrays (FPGAs). Recently, there has been increasing interest in vector architectures for next-generation high-performance computing. Unlike traditional SIMD (Single Instruction, Multiple Data) extensions, vector architectures support longer vector lengths, leading to a higher level of parallelism and increased throughput. A notable example is the Vitruvius+ vector coprocessor [2] designed by the European Processor Initiative (EPI) and built on the RISC-V architecture. This coprocessor can handle vectors of up to 256 double-precision elements, enhancing the performance of the main CPU in suitable applications.

In our research, we evaluate the performance of the Vitruvius+ vector coprocessor, specifically regarding sparse matrix operations. We utilize sparse matrix-vector multiplication (SpMV) as our benchmark. Initially, we explore the effectiveness of various compressed matrix representations in optimizing computations on Vitruvius+. Following this, we analyze how different key architectural elements of vector processors impact the efficiency of SpMV, highlighting the opportunities for performance optimizations.

Acknowledgements: This research was partially supported by Slovenian Research and Innovation Agency under Grant BI-HR/23-24-009 (Bilateral Collaboration Project).

References

[1] Mohammed, T. and Mehmood, R., 2022. Performance Enhancement Strategies for Sparse Matrix-Vector Multiplication (SpMV) and Iterative Linear Solvers. arXiv preprint arXiv:2212.07490.

[2] Minervini, O. Palomar, O. Unsal, E. Reggiani, J. Quiroga, J. Marimon, C. Rojas, R. Figueras, A. Ruiz, A. Gonzalez, J. Mendoza, I. Vargas, C. Hernandez, J. Cabre, L. Khoirunisya, M. Bouhali, J. Pavon, F. Moll, M. Olivieri, M. Kovac, M. Kovac, L. Dragic, M. Valero, A. Cristal, Vitruvius+: An area-efficient risc-v decoupled vector coprocessor for high performance computing applications, ACM Trans. Archit. Code Optim. 20 (2) (Mar. 2023). doi:10.1145/3575861.

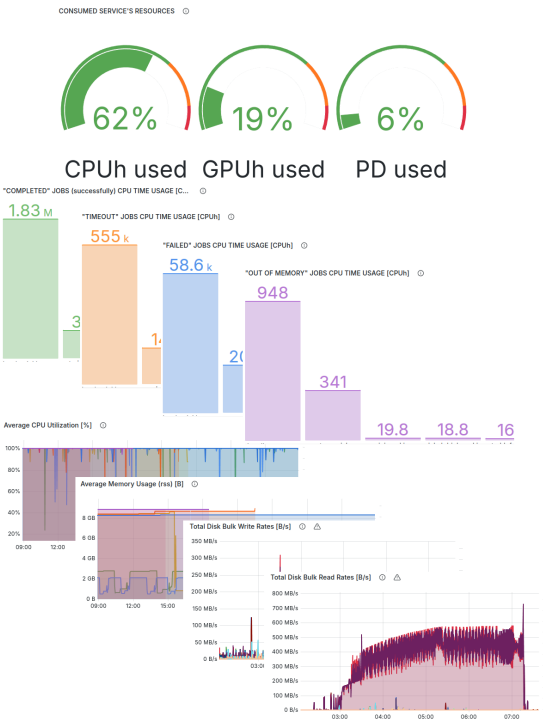
HPC Info: Enhanced Resource Tracking for SLURM Users and Administrators

Sebastian Sitkiewicz

sebastian.sitkiewicz@pwr.edu.pl
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 stat 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 or failed jobs by each user of their grant, whereas for the HPC administrators if 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 22.01.2025)

[3] Grafana software, https://grafana.com (accessed on 22.01.2025)

Strong Scaling Days: Testing massively parallel applications on VSC

Florian Goldenberg and Siegfried Höfinger

VSC Research Center, TU Wien, Austria

On most HPC systems, it is not usual to get a significant part of all nodes for one single user or for one submitted job. To provide this opportunity, we used the downtime during and after major maintenance to allow for such very large jobs. The event was open under the title "Strong Scaling Days" for interested user groups that could ascertain to be able to saturate a high number of nodes.

A total of five research groups accepted the opportunity, ranging from meteorology to physics to mathematics and computer science. The aim of all groups was to test the scalability of large and complex problems in their respective fields. In the most general sense, scalability is the ability to handle more work in the same timeframe as the size of the cluster grows. Translated to software, this usually refers to parallelisation efficiency, meaning the ratio between actual speedup and the theoretical maximum speedup based on increasing CPU core number. In case of the performed runs, the groups kept the problem size constant and increased the used cores to test this relation.

All groups managed to get the intended code running, sometimes after some needed adaptation to the large job size. Five of the groups ran their test on the CPU partition of VSC-5, while one of them additionally ran it on the CPU only VSC-4. GPUs were not involved at all.

For a medium number of nodes, up to around 100, all teams saw a strong scaling effect, with the time to solve the problem decreasing with the number of nodes used. However, only one team managed to get that effect up to 200 nodes, while all other saw the scaling stall or even reverse. No test scaled well above 200 nodes.

Additionally, some groups encountered problems with the MPI based communications or with the used application software. In several cases, adapting the version or parameters of MPI or certain code optimisations could solve the problem, but not in all cases. This also indicated a general problem with either the MPI implementations or the fabric of our clusters, as scaling always broke at 100 to 200 nodes, while in theory it should stay constant for the entire cluster.

The specific results and problems of the five strong scaling tests will be shown in more detail in our presentation.

Flexible and compact high density modular data center

Matjaž Pančur and Iztok Lebar Bajec

University of Ljubljana, Faculty of Computer and Information Science

The Faculty of Computer and Information Science (UL FRI) has a longstanding tradition of artificial intelligence (AI) research and development. This has notably expanded with the advent of advanced deep neural network technologies and generative AI. Presently, the majority of our laboratories is engaging in research and development activities that are directly or indirectly related to AI. This progression necessitates ongoing investment in increasingly more powerful and interconnected GPUs, consequently resulting in a significant rise in energy consumption.

The existing UL FRI data center was constructed approximately 15 years ago, coinciding with our relocation to a new facility. It was conceived as a conventional communication/business IT facility with Tier III capabilities. Designed with standard 60x100 cm racks, a maximum IT load capacity of 7 kW per rack, a 40 kW cap, N+N power supzply, and air cooling provided by two computer room air conditioners (CRACs) via raised floors, but also with a lack of any nowadays standard features such as hot/cold aisle containment. Due to this the existing data center is characterized by inefficiencies in energy usage, reflected in a Power Usage Effectiveness (PUE) of approximately 2. Over the past decade the number of UL FRI faculty and researchers has nearly doubled. Increasing hardware demands, primarily driven by GPU systems with power requirements beyond 3.5 kW and up to 14 kW per individual server, have outstripped the existing data center's cooling and electrical capabilities for redundancy and Tier III compliance. Operating as a research data center without Service Level Agreements (SLAs) or high availability (HA) requirements, we have utilized the center's full non-redundant capacity of 80 kW IT load. Due to the requirement for continuous operation of both CRACs to maintain appropriate data center temperatures and a lack of cooling redundancy, we have developed a controlled shutdown procedure for servers in the event of CRAC unit failure.

To address these challenges, we are undertaking the construction of a new, state-of-the-art, flexible, high-density modular data center on our facility's rooftop terrace. This turnkey "design and build" project will emphasize the implementation of green technologies with low carbon emissions, focus on energy efficiency (minimal PUE and Total Cost of Ownership (TCO)), and align with future technological trends such as direct-to-chip (D2C) liquid cooling, while also accommodating for the expansion of air-cooled compute. All components of the system will adhere to the highest energy standards, with provisions for heat recuperation and reuse. Given the uncertainty regarding future investments at UL FRI over the next decade (air-cooled vs D2C cooled systems), an essential requirement for the new facility is maximal flexibility while preserving low PUE and TCO.

The specifications for the new data center include support for up to 400 kW of IT workload with capability to cater for traditional air-cooled servers as well as D2C liquid cooled racks. Due to space restrictions, as a design choice, rear door heat exchangers (RDHx, [1]) will be used to allow up to 40 kW of air-cooled IT load per rack. For D2C liquid cooled systems we plan to support at least 150 kW per rack.

References

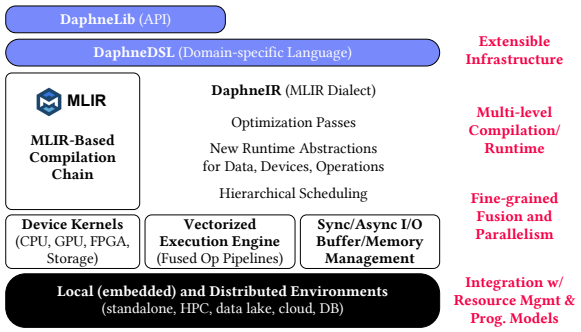
[1] Simon, V.S., et al., Energy Analysis of Rear Door Heat Exchangers in Data Centers With Spatial Workload Distribution, in Proceedings of IPAC2023. doi: 10.1115/IPACK2023-112078.

Blurring the lines between compilation and runtime optimization in the DAPHNE open source software framework

Mark Dokter[a,b,c]

[a] *EuroCC Austria*
[b] *Know Center Research GmbH*
[c] *Advanced Computing Austria GmbH*

Data-parallel computation frameworks like Apache Spark or ML systems like PyTorch are used in modern data-driven applications to leverage heterogeneous data collections to find interesting patterns or to build machine learning pipelines for accurate predictions. A key observation is that these new systems share many techniques with traditional high-performance computing (HPC), and the architecture of underlying HW clusters converges. Yet, the programming paradigms, cluster resource management, as well as data formats and representations differ substantially across data management, HPC, and ML software stacks.



There is a trend though, toward complex data analysis pipelines that combine these different systems. Examples are workflows of distributed data pre-processing, tuned HPC libraries, and dedicated ML systems, but also HPC applications that leverage ML models for more cost-effective simulation. Major obstacles are (1) limited development productivity for integrated analysis pipelines due to different programming models, (2) unnecessary data movement and under-utilization due to separate, statically provisioned clusters, and (3) lack of a common system infrastructure. For these reasons, **DAPHNE**'s ("Integrated **D**ata **A**nalysis **P**ipelines for large-scale data management, **H**PC, and machi**NE** learning") overall objective is the definition of an open and extensible systems infrastructure for integrated data analysis pipelines. In the past four years, a consortium, funded under grant 957407 of the EU Horizon 2020 program, has been developing this software system as an open source project. In this talk we look at the many features that make up the DAPHNE system infrastructure from its language support through the compiler to the runtime system.

References

[1] Zamuda, A. and Dokter, M. International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom), (pp. 1-8) (2024).

[2] Vontzalidis, A., Psomadakis, S., Bitsakos, C., Dokter, M., Innerebner, K., Damme, P., Boehm M., Ciorba F., Eleliemy A., Karakostas V., Zamuda A. and Tsoumakos, D. European Conference on Parallel Processing (EuroPar), (pp. 242-246) (2023).

[3] Damme P. et al. 12th Conference on Innovative Data Systems Research (CIDR), 12 pgs. (2022).

Unveiling the Energy Cost of Parallel Performance Portability in Cpp

Ruben Laso[a], Siegfried Benkner[a], and Sascha Hunold[b]

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

Performance Portability: Modern data centers and consumer-grade systems are equipped with a wide variety of devices from different vendors: multi- and many-core CPUs, GPUs, FPGAs, etc. Therefore, performance portability has become increasingly important in research and industry. However, the energy consumption of performance-portable code is often overlooked as it is expected to scale linearly with execution time. We aim to measure the impact of using performance-portable frameworks on energy consumption.

Methodology: We take performance-portable and platform-specific implementations of several benchmarks and compare their execution time, energy consumption, and energy-delay product (EDP). We use data from the RAPL, NVML, and IPMI interfaces to measure energy consumption. We also analyze the impact of other factors such as compilers, performance portability frameworks, and CPU and GPU frequencies.

Preliminary results: Using the CloverLeaf mini-app [1], we compare the native CUDA implementation to the performance-portable version using parallel execution policies of the Cpp's standard library (STL). Figure 1 shows the results on a system with a GPU NVIDIA Tesla P100 (maximum frequency of 1480 MHz). The following compiler infrastructures are evaluated: CUDA 12.6, NVIDIA HPC SDK 24.11, and AdaptiveCpp 24.10 [2]. First, the performance-portable implementations of the benchmark are (unexpectedly) up to 3.7% faster than the CUDA version. Second, lowering the GPU frequency to 70% of its maximum (1139 MHz) reduces performance by 3.5 % for the CUDA and AdaptiveCPP versions, but only by 0.7% for the NVIDIA HPC SDK. Last, the performance-portable versions save up to 10.8% in the EDP compared to the CUDA implementation.

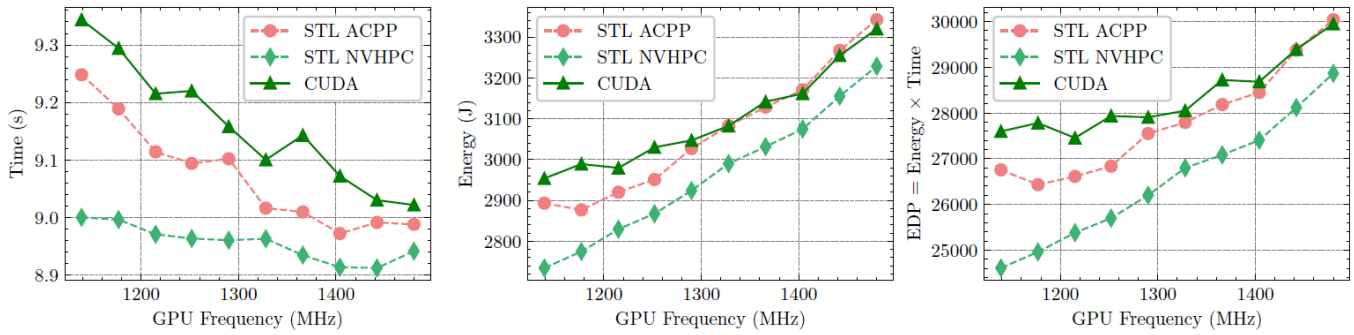


Fig. 1: Execution time, energy consumption, and energy-delay product of the CloverLeaf mini-app.

References

[1] Lin, W., Deakin, T., and McIntosh-Smith, S., "CloverLeaf" (2024). `https://github.com/UoB-HPC/CloverLeaf`.

[2] Alpay, A., and Heuveline, V., "AdaptiveCpp Stdpar: Cpp Standard Parallelism Integrated Into a SYCL Compiler," Proceedings of the 12th International Workshop on OpenCL and SYCL (2024).

Keynote Talk:

Exascale Numerical Weather Prediction at ECMWF

Sam Hatfield

European Centre for Medium-Range Weather Forecasts

For many decades now, the European Centre for Medium-Range Weather Forecasts (ECMWF) has spearheaded developments in global numerical weather prediction. The continued growth in forecast skill over the past few decades is due in large part to increases in the grid resolution of ECMWF's Earth-system model, the Integrated Forecasting System (IFS). This increase has gone hand-in-hand with developments in high-performance computing, with new generations of supercomputer permitting higher model resolutions and complexity. Weather forecast skill is especially sensitive to the resolution of the atmospheric component, for which resolutions are approaching the so-called "storm-resolving" level, which indicates a grid spacing of lower than 10 kilometres. A step change in the fidelity of global atmospheric simulations is expected as the model resolution approaches this "kilometre scale", in particular for the representation of extreme weather events.

However, recent developments in supercomputing present a barrier as we push towards these kilometre-scale simulations. The impetus for this new class of forecast system comes from the Destination Earth project, whose goals are to develop a series of Earth-system digital twins to aid in the prediction and mitigation of extreme weather events under a changing climate. In order to run this new class of Earth-system simulation efficiently, one must make effective use of accelerators, namely GPUs, and large communication networks.

This talk will give an overview of activities at ECMWF towards the goal of running kilometre-scale Earth-system simulations on pre-exascale and exascale supercomputers. The talk will present lessons learned from earlier experiments on supercomputers such as Summit. I will concentrate in particular on the spectral transform library ecTrans which the IFS atmospheric component crucially depends on, and which neatly contains several key computation and communication paradigms. I will also explore the opportunities of the new breed of data-driven models, which are led by ECMWF's AIFS machine learning model. These models rival traditional "physics-based" models such as the IFS, and are extremely cheap at inference time. The training of these models is a high-performance computing problem in its own right.

High clouds, high grid spacing to the rescue?

Blaž Gasparini[a], Rachel Atlas[b], Aiko Voigt[a], Martina Krämer[c,d], and Peter Blossey[e]

[a] *Department of Meteorology and Geophysics, University of Vienna, Austria*
[b] *CNRS-Laboratoire de Météorologie Dynamique, LMD, France*
[c] *Institute for Atmospheric Physics, University of Mainz, Germany*
[d] *IEK-7, Forschungszentrum Jülich, Jülich, Germany*
[e] *Department of Atmospheric and Climate Science, University of Washington, USA*

Climate models used for projections typically have a horizontal grid spacing of about 100 km and a vertical spacing of 0.5 to 1 km in the upper troposphere. They have been the backbone of climate science for decades, but they have intrinsic limitations due to their coarse grid spacing. Most notably, thunderstorms, the fundamental building block of tropical climate, occur at subgrid scales and cannot be resolved by a climate model's dynamical core. Instead, imperfect and often empirical parameterizations are used, leading to significant uncertainty in climate projections due to varying representations of thunderstorms (Fig. 1a).

Thunderstorms are the major source of tropical precipitation and inject large amounts of ice crystals high in the atmosphere, forming persistent anvil clouds, the most frequent and radiatively important cloud type in the tropics. Increasing the model grid spacing to O(1 km) is thought to be sufficient to directly resolve deep convective updrafts, which are the main source of high clouds in the tropics. This allows for a more physical simulation of the anvil cloud lifecycle, as shown by km-scale model output that is often barely distinguishable from satellite observations (e.g., Fig. 1c).

However, removing the need to parameterize one process exposes biases in micro-scale processes that remain parameterized, such as cloud formation and dissipation (known as cloud microphysics). While advanced microphysical schemes are considered expensive and are rarely implemented in global km-scale models or for weather forecasting, we show that most microphysical model bias can be removed by inexpensive modifications to the code. Our work reveals a simple, numerically inexpensive recipe that substantially improves simulations of tropical cirrus in the System for Atmospheric Modeling (SAM) model and can be applied to other atmospheric models.

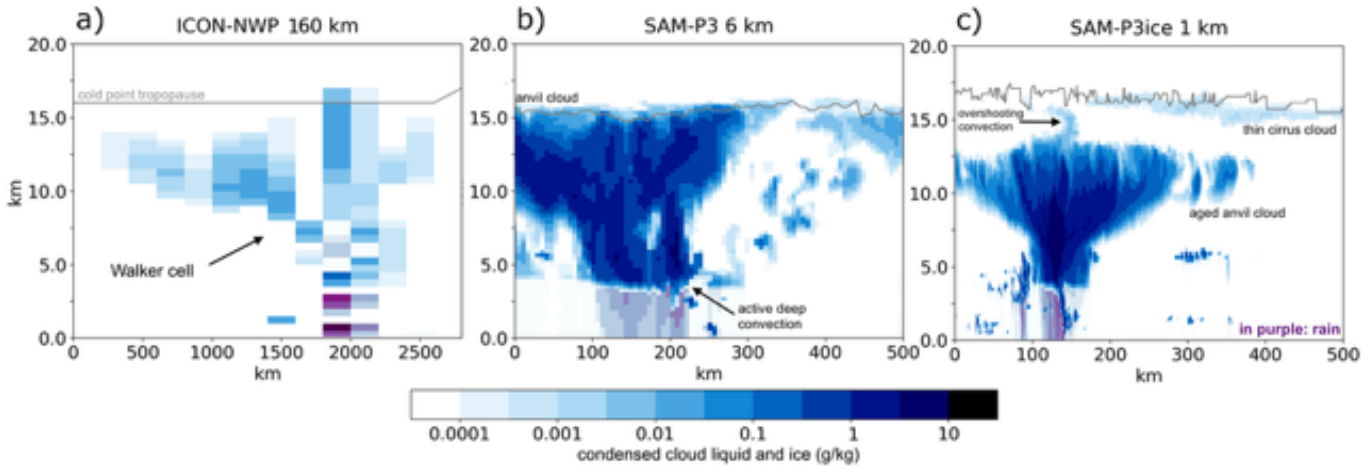


Fig. 1: Snapshots of cross-sections from the tropical Pacific simulated by (a) the ICON-NWP climate model, (b) a standard version of the SAM climate model, and (c) SAM with an improved microphysical scheme. The grey line indicates the tropopause. Blue indicates cloud condensate, purple indicates rainfall.

Using machine learning to distinguish km-scale climate models and observations on a regional scale

Maximilian Meindl[a], Aiko Voigt[a], and Lukas Brunner[b]

[a] *Department of Meteorology and Geophysics, University of Vienna, Austria*
[b] *Research Unit for Sustainability and Climate Risks, University of Hamburg, Germany*

The use of machine learning (ML) for climate science has attracted considerable attention within the last few years. A number of recent studies have used ML to extract information from global climate data (e.g. regional downscaling), predict future states of the climate system and evaluate models against observations. In particular, *Brunner and Sippel* [1] showed that low-resolution global climate models and observations can reliably be distinguished based on the global distribution of daily temperature, even after removing the mean model bias. ML is thus able to isolate fundamental differences between models and observations even in the presence of substantial internal variability. This raises the questions of whether ML can also distinguish between model and observational data on a regional scale, whether ML is as successful for km-scale models as for coarse-resolution models, and whether more complex bias correction methods reduce the success of ML.

To answer these questions, we use daily temperature fields over Austria, a topographically very complex domain. As training data, we use 200 different, randomly drawn days from each of the 13 ÖKS15 bias-corrected EURO-CORDEX models with an output resolution of 1km, resulting in 2600 samples labeled "model" which are matched by the same number of random days labeled "observation" from the SPARTACUS observation dataset. We use the binary classification approach to distinguish between the two classes of models versus observations. A logistic regression classifier is trained to determine the probability that a daily temperature field belongs to one of the two classes. In order to evaluate the ML algorithm subsequently, all days from the out-of-sample 10-year period 2005-2014 are used as test data.

The ML algorithm succeeds in correctly identifying the overwhelming majority of the test data for the setup used, resulting in an accuracy of 99%. The results remain consistent even when a different sample of 2x2600 random training days is used. In contrast to more complex classifiers, such as a convolutional neural network (CNN), the learned coefficients from the logistic regression allow insights into the spatial patterns that are crucial for distinguishing between models and observations. While the performance of climate models is typically evaluated on climatological timescales, our results highlight that such classifiers can be used to identify patterns of structural model biases. Our method hence offers a computationally efficient approach for model evaluation, especially when handling km-scale climate model data on a regional domain.

References

[1] Brunner, L. and Sippel S., Environmental Data Science **2**, e22 (2023).

Strategies for faster deep learning-based modeling of global barystatic processes

Mostafa Kiani Shahvandi and Aiko Voigt

Department of Meteorology and Geophysics, University of Vienna, Austria

Barystatic processes represent the continental-ocean mass redistribution due to melting of polar ice sheets (Greenland and Antarctica) and global glaciers, and variations in terrestrial water storage. Modeling and prediction of barystatic processes is important since they influence global (hydro)climate and have implications for, e.g., sea-level rise in coastal areas. Recently, deep learning algorithms have been used to model and predict barystatic processes [1,2,3]. However, this task remains challenging because of the spatio-temporal nature of the data, which impedes robust uncertainty quantification.

Here we suggest two simple high-performance computing strategies for the fast modeling and prediction of barystatic processes. Since as noted the data of interest are spatio-temporal, they can be considered as a time series of images. The first strategy is inspired by so-called region proposal networks. In this approach a high-resolution image is broken down into sub-images with smaller size (for instance partitioning an image with size 720×1440 into 4 sub-images with size 180×360). These sub-images are then fused using a neural network. The strategy increases the number of samples to the algorithm, but decreases the number of parameters (as the algorithm is dependent on the dimension of the input images). Therefore, since the computational complexity of the model is proportional to the number of parameters, the strategy might facilitate training and prediction of the deep learning model. By testing the strategy on a yearly dataset spanning 1900 to the present that contains the mass redistribution in all the four domains of barystatic processes (Greenland, Antarctica, global glaciers, and terrestrial water storage), an overall $\sim$12% improvement is achieved with respect to the computational time of the benchmark (benchmark is when the images are not partitioned).

The second strategy concerns the problem of uncertainty quantification. For this purpose, we propose to use so-called deep ensembles. In this approach, several models are trained at the same time and subsequently used to compute the mean and standard deviation across the predictions of individual ensemble members. The advantages of using deep ensembles are manifold: they are asymptotically convergent to the Bayesian deep learning (which is the most rigorous probabilistic framework for the analysis of data), and the individual ensembles can be trained in parallel. Applying the strategy to the aforementioned dataset results in an additional $\sim$15% improvement with respect to the benchmark, in addition to providing uncertainty estimates.

In summary, the two strategies together reduce the computational time by $\sim$27% and allow for the quantification of uncertainties. This suggests that they are useful for analyzing data on barystatic processes.

References

[1] Kiani Shahvandi, M., Adhikari, S., Dumberry, M., Modiri, S., Heinkelmann, R., Schuh, H., Mishra, S., Soja, B., Nat. Geosci. **17**, 705 (2024).

[2] Kiani Shahvandi, M., Noir, J., Mishra, S., Soja, B., Geophys. Res. Lett. **51**, e2024GL111148 (2024).

[3] Kiani Shahvandi, M., Adhikari, S., Dumberry, M., Mishra, S., Soja, B., Proc. Natl. Acad. Sci. U.S.A. **121**, e2406930121 (2024).

Poster:

Performance Evaluation of Parallel Approaches for 1D PIC Simulations on GPUs

Ivona Vasileska, Pavel Tomšič, and Leon Bogdanović

Faculty of Mechanical Engineering, University of Ljubljana, Slovenia

Particle-in-Cell (PIC) [1] simulations are essential for plasma modelling, especially in plasma sheaths. Their performance depends on parallelisation strategies and hardware utilisation. This study evaluates the performance of three parallel approaches — MPI, MPI+OpenACC and MPI+OpenMP — on GPUs for 1D PIC plasma sheath simulations.

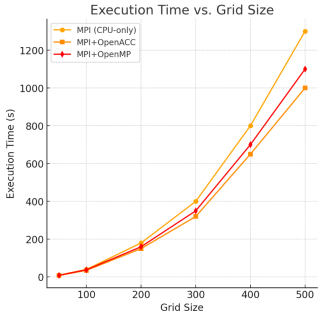


Fig. 1: Execution time across different grid sizes

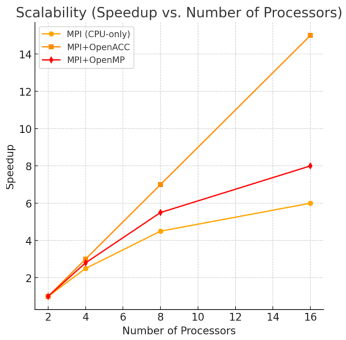


Fig. 2: Scalability analysis

The MPI model supports distributed memory parallelism for large-scale simulations, but lacks thread-level parallelism within nodes. To address this, we investigate hybrid models with GPU acceleration. MPI+OpenACC provides high-level abstraction for porting kernels to GPUs and enables efficient resource utilisation with minimal code changes, as shown in previous studies [2]. In contrast, MPI+OpenMP supports shared memory parallelism and has been adapted for GPU offloading in heterogeneous systems.

Three key metrics are used in the performance benchmarks: execution time, scalability and resource utilisation. The simulations are run on an HPC VEGA cluster with NVIDIA GPUs, comparing execution times for different grid sizes. The MPI+OpenACC implementation achieves superior performance on GPU-dominant tasks due to the efficient execution of the kernel and optimisation of memory transfer. Meanwhile, MPI+OpenMP shows competitive performance on smaller problem sizes, but struggles with memory bandwidth limitations on larger datasets. Scaling tests show that MPI+OpenACC scales effectively as the number of GPUs increases, while the CPU-only MPI implementation is constrained by communication overhead between nodes and does not benefit from GPU acceleration. MPI+OpenMP shows moderate scalability but is limited by sub-optimal GPU utilisation. The resource utilisation metrics show that MPI+OpenACC achieves higher GPU utilisation and lower energy consumption compared to the other approaches, which is consistent with the results of previous analyses of plasma simulations with OpenACC. These results emphasise its suitability for energy-efficient simulations. This study highlights the efficiency of MPI+OpenACC for large-scale 1D plasma sheath simulations on GPUs and suggests extending these results to higher-dimensional models and investigating asynchronous execution for further performance improvements.

References

[1] Tskhakaya, D. et al., "The Particle-In-Cell Method", Contrib. Plasma Phys. **47**, No. 8-9, 563–594 (2007).

[2] Wienke, S. et al., "OpenACC — First Experiences with Real-World Applications", Euro-Par 2012 Parallel Processing, Lecture Notes in Computer Science, pp. 859–870.

Poster:

AURELEO: Austrian Users at LEONARDO supercomputer

Luis Casillas-Trujillo, Ivan Vialov, and Claudia Blaas-Schenner

VSC Research Center, TU Wien, Austria

The LEONARDO supercomputer located at CINECA in Bologna stands as one of the most powerful computing systems, it ranks 9th in the world and 4th in Europe in the latest published list of the Top 500 supercomputers in November 2024. This immense computational power is part of a collaborative effort, with Austria as a partner in the consortium that built the LEONARDO machine. Due to this partnership, Austria has a dedicated share of the supercomputer, and Austrian researchers can access LEONARDO resources through AURELEO calls. These calls provide Austrian researchers with access to two distinct partitions of the LEONARDO supercomputer: the Booster (GPU partition) and the DCGP (CPU partition). The Booster is equipped with four Nvidia Ampere GPUs per node, designed for high-performance parallel processing, while the DCGP features two Intel Sapphire Rapids CPUs per node, optimized for large-scale, data-intensive computations. AURELEO has provided Austrian scientists the computational resources to conduct state-of-the-art research and push the boundary of simulation with more challenging and complex simulations enabled by the power of the LEONARDO cluster. The AURELEO program supports a broad spectrum of scientific fields, with awarded projects spanning areas such as physics, computer science, bioinformatics, with topics ranging in atomistic simulation of complex materials, sustainability, drug discovery, and atmospheric modeling. 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. In addition to granting access to LEONARDO's computational resources, the AURELEO program offers support through its High Level Support Team (HLST). This team helps users to run their applications efficiently and maximize the use of the allocated resources. In this work, we give an introduction to the AURELEO program, including the various calls for project submissions, the application process, and an overview of the different research topics explored by the awarded projects. By supporting diverse and groundbreaking research via providing world-class computational resources, the AURELEO program fosters innovation and contributes to scientific advancements across numerous fields.

Poster, Container Forum:

Integrating Linux with HTTP: Secure and Automated Workflows

Wiktor Nastał

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

This presentation explores innovative methods to enhance Linux system workflows by integrating web-based authentication and automation. The proposed solutions address challenges in security, user experience, and workflow efficiency, leveraging modern authentication protocols and automation tools.

First, the concept of context-aware SSH access is introduced, where the authentication method selected for a user depends on contextual factors such as geolocation. This approach allows for dynamic adaptation of the authentication process, ensuring that the most appropriate and secure method is chosen based on the user's environment and circumstances. Additionally, the integration of OAuth-based authentication is explored, enabling SSH logins to be authenticated through the OAuth 2.0 standard, enhancing both security and flexibility in the login process.

Second, the presentation highlights automation within the user space using tools such as oidc-agent [1], MyToken [2], and HashiCorp Vault. These technologies enable secure and streamlined processes for managing secrets, such as access keys for S3. By integrating these tools, the entire process becomes fully automatable, ensuring a seamless and secure workflow across services.

Finally, the integration of Vault-driven automation is presented as a key solution for managing secrets and automating workflows across working nodes. This approach ensures robust access management, reduces manual intervention, and enhances security for multi-node operations. It strengthens security by ensuring that sensitive information is never exposed in plaintext and that access is granted only when necessary, following predefined policies. As a result, Vault-driven automation safeguards critical resources without introducing unnecessary complexity.

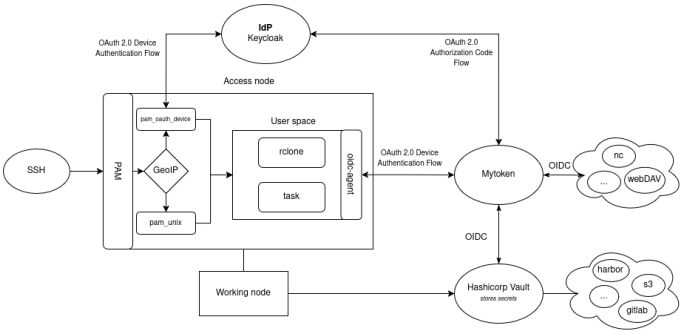


Fig. 1: Proposed system workflow diagram.

By combining these concepts, the presentation demonstrates how Linux environments can securely leverage web protocols to create automated, secure, and user-centric workflows that meet modern computational and operational demands.

References

[1] Zachmann, G., OIDC-Agent: Managing OpenID Connect Tokens on the Command Line (SKILL 2018-Studierendenkonferenz Informatik) (2018).

[2] Zachmann, G., Mytoken-OpenID Connect Tokens for Long-term Authorization (Master's Thesis, Karlsruher Institut für Technologie (KIT)) (2021).

Poster:

FLEXWEB - A flexible particle dispersion model web interface

Michael Blaschek, Marina Dütsch, Lucie Bakels, and Andreas Stohl

University of Vienna: Department of Meteorology and Geophysics

Flexpart (FLEXible PARTicle dispersion model)[1] is a numerical model that simulates the dispersion of gases and aerosols in the atmosphere. In order for Flexpart to be used, it must be installed and run on a (super)computer. However, this is associated with obstacles, as not all scientists have access to a supercomputer and there are often technical problems during installation or execution. In this project, we therefore want to develop a Flexpart Web Service (FLEXWEB) in which Flexpart can be controlled via a web user interface (fastapi, python) and run on a super computer. We will show first results and details on the implementation of such a workflow for a potential operational service on Vienna Scientific Cluster (VSC) and a local cluster. We will further demonstrate that the service can be containerized (docker and singularity) and run in a Kubernetes cluster on Microsoft Azure and make the results easily available to end users. We hope to start a discussion on "How to a Flexpart service or any other service can be deployed on any EuroHPC system operationally?". In this way, we hope to simplify access to Flexpart for scientists worldwide.

References

[1] Bakels, L., Tatsii, D., Tipka, A., Thompson, R., Dütsch, M., Blaschek, M., Seibert, P., Baier, K., Bucci, S., Cassiani, M., Eckhardt, S., Groot Zwaaftink, C., Henne, S., Kaufmann, P., Lechner, V., Maurer, C., Mulder, M. D., Pisso, I., Plach, A., Subramanian, R., Vojta, M., and Stohl, A.: FLEXPART version 11: Improved accuracy, efficiency, and flexibility, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2024-1713, 2024.

Poster:

Empowering Women in High Performance Computing: Activities of the Central European Chapter of Women in HPC

Karina Pešatová

VSB - Technical University of Ostrava, IT4Innovations, Ostrava, Czech Republic

The International Organization of Women in High-Performance Computing (WHPC) [1] is committed to advancing diversity and inclusion in the High-Performance Computing (HPC) field. In 2024, the Central European Chapter of WHPC [2] was launched as part of this global initiative. The chapter aims to promote gender equity in the Central European region through community-building, advocacy, and professional development opportunities. Its founding members represent the Czech Republic, Austria, Poland, and Slovenia, with a vision to expand and engage with neighboring countries, including Slovakia and Hungary.

In 2023, the representation of women in ICT roles in the Central European countries showed significant variation when compared to the EU average of 19.6%. Slovenia exceeded the EU average, with women making up 22.8% of ICT specialists, showcasing progress toward gender diversity. Poland was close to the EU average, with 18.9% of its ICT workforce being female, indicating a balanced representation relative to other countries. Austria fell slightly below the EU average at 18.3%, reflecting a moderate gender gap. Czechia, however, had one of the lowest shares in the region, with only 12.4% of its ICT specialists being women, highlighting the need for targeted initiatives to promote gender inclusion in ICT. [3] Even the EU average, however, remains far from gender balance, underscoring the importance of systemic efforts to achieve equity in the ICT sector across Europe. This poster showcases the chapter's initiatives focused on achieving a positive change in the HPC field:

- Outreach and Networking: Engaging students, early-career professionals, and experienced HPC practitioners through events like networking meetups, panel discussions, and career-building workshops.
- Education and Training: Organizing technical training sessions, mentoring programs, and webinars tailored to empower women in HPC with critical skills and knowledge.
- Collaboration: Partnering with academic institutions, research centers, and HPC service providers to support diverse talent pipelines and equitable practices.
- Success Stories: Showcasing the contributions of women in Central Europe's HPC community and creating role models for the next generation.

Through these efforts, we aim to create an inclusive environment where women can thrive and contribute to advancing HPC research and innovation. This poster invites the community to join us in amplifying these efforts by collaborating on initiatives and sharing best practices to overcome common challenges.

References

[1] https://womeninhpc.org

[2] https://www.linkedin.com/company/central-european-women-in-hpc/

[3] https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20240524-2

Poster:

Understanding and Addressing Evolving Training Needs in High Performance Computing: Insights from the 2024 Training Services Survey

Lucie Kavka and Karina Pešatová

VSB - Technical University of Ostrava, IT4Innovations, Ostrava, Czech Republic

The National Competence Centre in High Performance Computing of the Czech Republic (NCC Czechia) conducts a biennial Training Services Survey to assess the evolving training needs of its stakeholders in academia, industry, and the public sector. As a part of its commitment to providing high-quality education in High Performance Computing (HPC), High Performance Data Analytics (HPDA), Artificial Intelligence (AI), Visualization and Virtual Reality (VVR), and Quantum Computing (QC), the 2024 survey gathered 105 responses. This marked an increase in participation from the 2022 survey and provided valuable insights into the training requirements of the community.

The survey revealed that HPC and AI remain the most in-demand domains, with significant interest in specialized subdomains such as Parallel Programming, GPU-Accelerated Computing, Neural Network Architectures, and Machine Learning algorithm selection. While there was a marked increase in advanced expertise in HPC among respondents between 2022 and 2024, Quantum Computing emerged as a domain where most participants identified themselves as beginners. These findings suggest an opportunity to address knowledge gaps through foundational training programs in Quantum Computing.

The format and delivery of training were also key areas of investigation. Results indicated a growing preference for flexible and accessible learning opportunities, particularly online and self-paced courses, reflecting a broader shift influenced by the pandemic. While in-person and hybrid training formats remain relevant, participants expressed a strong inclination towards shorter sessions, with one-day training being the most convenient. Multi-day events, such as seasonal schools, were considered less feasible due to time constraints, underscoring the importance of concise and efficient learning experiences.

These findings are shaping NCC Czechia's training strategies. The focus is on expanding flexible formats, prioritizing shorter, high-impact sessions, and addressing emerging domains like Quantum Computing. The survey findings not only guide NCC Czechia's offerings but also serve as a valuable resource for other training providers aiming to meet the evolving needs of the HPC community.

The poster will present an overview of the survey's methodology, key findings, and actionable recommendations for tailoring training programs to the dynamic landscape of HPC-related fields. This aligns with NCC Czechia's mission to support the professional growth of its stakeholders and enhance expertise in cutting-edge technologies.

References

[1] https://www.eurocc-czechia.cz/wp-content/uploads/2024/12/Traininig_Services_Survey_Analysis_2024.pdf

Poster:

Treating atomic photoionization with a modified time-dependent surface flux method

Andrej Mihelič[a], Martin Horvat[b,c], Alicia Palacios[d], and Johannes Feist[d]

[a] *Jožef Stefan Institute, Ljubljana, Slovenia*
[b] *Faculty of Mathematics and Physics, University of Ljubljana, Slovenia*
[c] *Lek Pharmaceuticals, Ljubljana, Slovenia*
[d] *Universidad Autónoma de Madrid, Spain*

Theoretical treatment of excitation and ionization of quantum systems with short, intense light pulses with wavelengths from the extreme ultraviolet or the x-ray spectral region generally poses a great computational challenge. The time-dependent surface flux (t-SURFF) method allows the calculation of angle-dependent and total ionization probabilities for laser-driven atomic and molecular targets using relatively small simulation volumes [1–2]. It is based on tracking the photoelectron probability current (flux) through a surface enclosing the ionization centre. While a smaller volume substantially lowers the computational cost, the burden of describing photoionization is transferred to the calculation of the time-dependent final-state (channel) wave functions, whose form must be know at all times. In practice, the final-state continuum wave function is replaced by a simpler, approximate wave function, which results in oscillatory artefacts in the extracted ionization amplitudes and probabilities. In the original method, these artefacts are removed by a temporal average of the extracted ionization amplitudes after the end of the laser pulse. We present a variant of the t-SURFF method in which the temporal average is replaced by an average over the extraction radius (the radius of the sphere enclosing the atom). The method has been tested for the cases of laser-driven hydrogen and helium atoms, and has been found to work reliably for photon energies from the extreme ultraviolet spectral region. The advantage of the modified method is that no extra temporal propagation is required to calculate the average. Compared to the original t-SURFF method, the modified method introduces no additional computational overhead since the order of the spatial average may be interchanged with integration over time (evolution). An efficient implementation—in which the spatially averaged matrix elements between the basis states can be calculated only once—is thus possible. Our implementation takes advantage of both distributed (MPI) and shared-memory (OpenMP) parallelism, and uses a highly-efficient Krylov-Arnoldi-type propagation scheme [3] applicable in the case of non-Hermitian problems.

References

[1] L. Tao, A. Scrinzi, New J. Phys. 14, 013021 (2012).

[2] A. Zielinski, V. Pramod Majety, A. Scrinzi, Phys. Rev. A 93, 023406 (2016).

[3] A. I. Kuleff, J. Breidbach, L. S. Cederbaum, J. Checm. Phys. 123, 04411 (2005).

Poster:

Influence of membrane composition on the signaling of the NKG2A/CD94/HLA-E complex investigated by all-atom simulations

Martin Ljubič[a,b], Jure Borišek[a], and Andrej Perdih[a]

[a] *National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia*
[b] *Faculty of Pharmacy, University of Ljubljana, Aškerčeva 7, 1000 Ljubljana, Slovenia*

Understanding the impact of membrane composition on the dynamics and function of transmembrane proteins is essential for uncovering cellular signaling pathways and identifying novel strategies to modulate signaling processes. Despite its significance, the comprehensive study of lipid roles in shaping the conformation and functionality of large protein systems remains in its early stages. In this study, we utilized all-atom molecular dynamics simulations using the Amber20 PMEMD software package on the HPC Vega system to investigate how varying membrane compositions influence the conformational dynamics of the NKG2A/CD94/HLA-E immune receptor complex [1]. This receptor complex is a critical negative regulator of natural killer cell cytotoxic activity, mediating signals via the ITIM regions in the intracellular domain of NKG2A upon ligand binding [2].

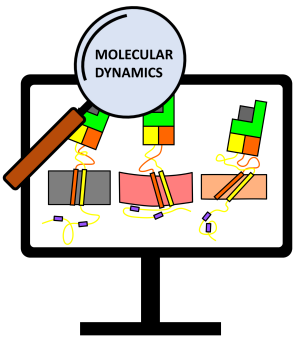


Fig. 1: Schematic representation of the studied models.

Each MD simulation, encompassing over 500 000 atoms, was carried out on four GPUs in parallel to obtain 2 μs trajectories in 4 weeks. Our results reveal distinct variations in the behavior of the immune receptor complex across five representative membrane compositions: POPC, POPA, DPPC, DLPC, and a mixed POPC/cholesterol system. Notably, these differences are most evident in the intracellular domain, affecting mobility, tyrosine exposure, and interdomain communication. For example, the POPA membrane, characterized by a high negative surface charge density, increased lipid interactions and significantly reduced the exposure of the NKG2A ITIM regions to water molecules, potentially inhibiting signal transduction. In contrast, the DPPC membrane, with its elevated transition temperature and gel-like properties, induced curvature effects that modified the exposure of one ITIM region. The DLPC membrane, with its reduced thickness, caused a pronounced tilt in the transmembrane domain, altering the linker protrusion angle and disrupting the hydrogen bonding network in the extracellular domain. Beyond these domain-specific effects, global changes in protein behavior were observed. For instance, POPA exhibited lower overall flexibility, and correlation analyses indicated disrupted communication between protein domains in several models.

These findings highlight the critical role of membrane composition in shaping transmembrane protein dynamics. By demonstrating how receptor behavior varies across different lipid environments, our study provides a foundation for exploring lipid-based interventions to modulate receptor signaling. This work advances our understanding of the molecular mechanisms underlying signal transduction and offers new avenues for developing therapeutic approaches that exploit membrane composition to regulate immune receptor functions.

References

[1] Ljubič M., Perdih A., Borišek J., J. Chem. Inf. Model. **64**, 9374 (2024).

[2] Ljubič, M., Prašnikar, E., Perdih, A., Borišek, J., J. Chem. Inf. Model. **63**, 3486 (2023).

Poster:

Optimization of Small Language Models (SLMs)

Teo Prica[a,b], Vili Podgorelec[b], and Aleš Zamuda[b]

[a] *IZUM - Institute of Information Science, Slovenia*
[b] *UM - University of Maribor, Slovenia*

A Language Model (LM) is a machine learning model designed to perform various useful tasks within Natural Language Processing (NLP) and other tasks within modern Artificial Intelligence (AI). They are trained on vast amounts of data from a broad range of sources. In general, they are divided into Small Language Models (SLMs) and Large Language Models (LLMs) which differ in the number of parameters they contain. They come in different sizes and parameters, which range from millions in SLMs to hundreds of billions in LLMs, determining a models capacity and complexity. [1,2]

Large Language Models (LLMs) are powerful and capable of handling complex tasks such as summarization, recognition, translation, text generating, code programming, and other forms of content, understanding of language, and beyond. In comparison with SLMs they consume more computational resources, including power, making them less energy-efficient and unsuitable for many tasks. [1,2]

Small Language Models (SLMs) are lightweight and quite efficient models, but may not achieve the same accuracy and performance, and tasks such as LLMs. Their limitations and disadvantages are more evident when applied to complex tasks, like performance, underfitting, security vulnerabilities, scalability issues, and more. They may be used to perform tasks such as virtual assistants, basic real-time analysis, and text generation. As the focus is on sustainability in AI, they are designed for environments with limited computational resources. [1,2,3]

SLMs are easier to train, maintain, optimize, fine-tune, and deploy. While applying optimization techniques, and enhancing their functionality, flexibility, efficiency, size, speed, and modularity, they are designed to be easily adapted and integrated within workflows, and tailored solutions. The process is fairly straightforward, as there is already a large set of open-source Pre-trained Language Models (PLMs) available such as BERT (Distil, Medium, Small, and Tiny), TinyLlama, TinyGPT, and beyond. [1,2,3]

The common optimization techniques may be applied to the PLMs without compromising their performance, such as knowledge distillation, pruning, quantization, attention mechanisms, layer locking, parameter sharing, learning rate schedules, early stopping, gradient clipping, regularization techniques like dropout, L1, and L2, hyperparameter optimization with evolutionary optimization, Random Search, Grid Search, Bayesian Optimization, and beyond. [1,2]

References

[1] Podgorelec, V., Lahovnik, T., Vrbaničič, G., Kako ukrotiti velik jezikovni model nad lokalnim korpusom, OTS 2024 **27**, 1-14 (2024)

[2] Zamuda A., Lloret, E., Optimizing Data-Driven Models for Summarization as Parallel Tasks. Journal of Computational Science, **42** (2020)

[3] Zamuda, A., Dugonik, J., Lloret, E., Comparing evolved extractive text summary scores of bidirectional encoder representations from transformers. 11th IcETRAN papers in the IEEE Xplore, 241-246 (2024)

Poster:

NCC Czechia: Sucess Stories

Tomas Karasek and Katerina Beranova

IT4Innovations, VSB – Technical University of Ostrava, Ostrava, Czech Republic

The National Competence Centre for HPC in the Czech Republic (NCC Czechia) was established in 2020 as one of the 31 NCCs across Europe under the auspices of the EuroCC project. Since 2023, its activities have continued under the EuroCC 2 project.

NCC Czechia provides services, access to the knowledge and support for the use of High-Performance Computing (HPC) and associated technologies such as High-Performance Data Analytics (HPDA), Artificial Intelligence (AI) and Quantum Computing (QC) for all stakeholders from the academia, industry, and public institutions not only in the Czech Republic but also Europewide.

We aim to increase the uptake of HPC technologies by stakeholders in the Czech Republic, increase their awareness and preparedness and improve their digital skills related to HPC and associated technologies. In this poster, activities of the NCC Czechia, consisting of dissemination and communication activities [1], training activities [2], workshop organisation, as well as collaboration with the industry, academia and public sector, and collaboration with other NCCs, will be presented. The poster will present highlights of those activities with a focus on the success stories of collaboration with industry, academia and public administration.

The main goal of our cooperation with Bellmer Czech s.r.o. was to guide their experts on how to change their existing Computational Fluid Dynamics Simulations (CFD) so they could utilise HPC infrastructure.

With the Police of the Czech Republic, our NCC focuses on providing expertise in data analytics, algorithms, and methods that could be used to predict crime.

The Institute of Plasma Physics of the Czech Academy of Science focuses on experimental work developing new materials. In the collaboration highlighted in this poster, we provided consultation on how numerical modelling and simulation, HPC and AI could be used to speed up the analysis of data obtained from experiments and how to get deeper insights into studied problems.

References

[1] https://www.eurocc-czechia.cz/en/home/

Poster:

ICON @ VSC: strong scaling tests with a global km-scale model of the atmosphere

Aiko Voigt

Department of Meteorology and Geophysics, University of Vienna

In October 2024, the "VSC Strong Scaling Days" allowed me to perform strong scaling tests with the global ICON-Sapphire [1] model at the Vienna Scientific Cluster VSC5. The task given to the model was to simulate a period of 6 hours and 30 seconds in a prescribed SST setup using only the atmospheric model component and at a horizontal grid spacing of 5 km. In this setup, the model uses 20,971,520 grid cells per level and 90 model levels, resulting in a total of 1,887,436,800 grid cells. The time step is 30 s.

I tested the scaling for 20, 40, 80, 120, and 140 compute nodes, where each VSC5 node consists of 128 AMD Zen3 cores, as well as for pure MPI versus hybrid MPI-OpenMP parallelization. I also evaluated the impact of writing 3-hourly output and running the model in a natively compiled version versus in a singularity container (both using the Intel compiler suite and Intel MPI).

The results are shown in the figure as time to solution, with the blue line showing the time to solution for perfect strong scaling (based on 20 nodes and pure MPI). Both figures show that while the model does not scale perfectly, it shows good strong scaling up to at least 140 nodes or 17,920 cores. The figures also show that at 140 nodes one can expect to achieve 12.7 simulated days per clock day (SDPD). Note also that OpenMP has no strong impact on the simulations and that the container version performs as well as the native version.

While the tests demonstrate that ICON scales well, they also illustrate the need for GPUs to run such kilometer-scale simulations. Assuming perfect strong scaling, 1 SYPD would require approximately 5,200 VSC5 CPU nodes. This is larger than than the entire VSC5 cluster of 710 nodes.

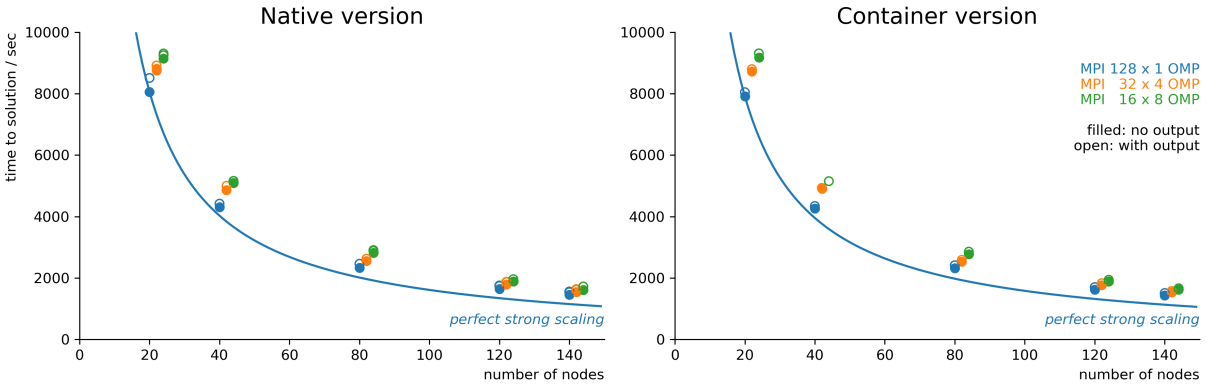


Fig. 1: Time to solution for the native (left) and the container (right) version of ICON as a function of VSC5 nodes.

References

[1] Hohenegger, C., Korn, P., Linardakis, L. et al., Geoscientific Model Development, **16**, 779 (2023)

Poster:

VSC's Software Stack Envisioned

Luis Casillas-Trujillo, Filip Kocina, Adam McCartney, and Moritz Siegel

VSC Research Center, TU Wien, Austria

The *Software and Modules* working group at *VSC* is developing a new software stack. The working group is a cooperation between members of the *Platform*, *Support* and *Sysadmin* teams. The goal is to pool knowledge from these areas and produce a user-focused software layer. In the early part of 2025 the *Software and Modules* working group created two prototypes of a future software stack on *MUSICA*. We compare the two approaches and show how they will inform aspects of the eventual implementation of the new software stack at *VSC*. The first approach is closely integrated with the software and compatibility layers provided by *EESSI*. The second approach is less tightly coupled to the *EESSI* ecosystem, but uses some of the same tools (*EasyBuild* and *lmod*). We will present our findings from this comparison.

The software stack at *VSC* was initially managed by hand. As the number of software installations available grew with time, the package manager *Spack* was introduced. While this improved the situation with respect to dependency management, a number of issues remained like poor performance of *Spack* due to the underlying file system hardware; no underlying compatibility layer to reliably build against a standard set of libraries; pollution of the global module namespace; no update mechanism. To address these issues, the team assessed a number of additional and alternative tools.

The existing distributed file system (gpfs) used in *VSC* is not optimized for software distribution. Software typically has a high amount of metadata, and many small files need to be transferred when a piece of software runs. In order to address this issue The CernVM File System (*cvmfs*) was invented with the specific purpose of distributing scientific software to various types of host in a fast, scalable, and reliable way. This is done using HTTP as a transport mechanism and leveraging several layers of web caches. This approach is used in a number of projects in addition to *CERN*, most notably by Compute Canada and the European Environment for Scientific Software Installations (*EESSI*).

The Compute Canada and *EESSI* software stacks both use *cvmfs* as a means to provide the file system layer within the contexts of their respective stacks. The next layer on top of that is the compatibility layer, which aims to solve the problems associated with portability across operating system upgrades. Such a compatibility layer must provide a set of core libraries and utilities, which can then be used to bootstrap the rest of the software. Both *guix* and *nix* offer valid technical solutions to this problem, and *nix* was initially used during the prototyping phase at Compute Canada. The project eventually settled on the more lightweight approach of using a *gentoo* prefix as a way to provide the compatibility layer. This was also the approach adopted by the *EESSI* project.

One stated goal of the *EESSI* project is to provide a shared stack of scientific software installations that can be used by a wide variety of clients of all sizes. Although it does not explicitly aim to be a tool for building custom software stacks, it does provide a means for extending what is provided using *EasyBuild* and a number of specially configured locations that enable user, site-wide and project-specific installations of custom software (any software not directly provided by *EESSI*).

Finally, *lmod* was chosen as a possible candidate to create a more structured overview of the modules available. Specifically, we looked at it's facilities for creating collections of modules that can be loaded as profiles.

Poster:

Heterogeneous Exascale Particle-in-Cell

Štefan Costea, Miha Radež, Jernej Kovačič, Matic Brank, Leon Bogdanović, Ivona Vasileska, and Leon Kos

Faculty of Mechanical Engineering, 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. Designed to leverage modern heterogeneous architectures, HEXAPIC is intended to efficiently run on a wide range of computational devices, including CPUs, GPUs, and accelerators [2]. The code's modular architecture decomposes the Particle-in-Cell (PIC) algorithm into independent components, allowing dynamic allocation to the most suitable hardware based on computational strengths and simulation needs. An overview and dependency schematics is shown in Figure 1.

A key feature of HEXAPIC is its use of distributed memory programming with MPI (Message Passing Interface), ensuring scalability across large, heterogeneous systems. This approach is essential for optimizing performance in exascale environments, allowing HEXAPIC to harness diverse computational resources effectively. Current development focuses on improving the allocation of tasks, load balancing, and minimizing data movement and communication overhead, all crucial for maintaining high performance in large-scale simulations. Parallel I/O capabilities are also being optimized to manage the vast datasets typical of exascale simulations, addressing potential bottlenecks and improving data throughput.

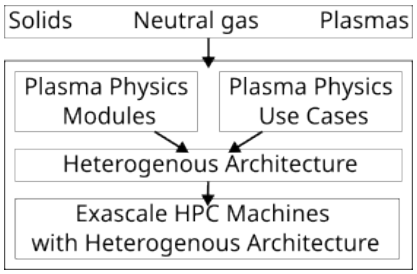


Fig. 1: HEXAPIC overview and dependencies.

HEXAPIC will incorporate advanced solvers and libraries such as PETSc, Hypre, and AMReX to handle both linear and non-linear plasma physics problems. These libraries support adaptive grid refinement and ensure that the code can scale across a variety of physical and computational domains.

The ongoing development of HEXAPIC targets key plasma simulation applications, including the scrape-off-layer and divertor regions of tokamak fusion reactors, material deposition from plasma sources, and ion thrusters for space propulsion. These use cases drive the refinement of the code, with a focus on balancing floating-point performance, energy efficiency and simulation accuracy.

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] StarPU: A Unified Runtime System for Heterogeneous Multicore Architectures. https://starpu.gitlabpages.inria.fr/.

Poster:

Kilometer-scale Climate Modeling of TRAPPIST-1e Using ICON-Sapphire: Advancing Exoplanet Habitability Studies

Zoé Lloret and Aiko Voigt

Department of Meteorology and Geophysics, University of Vienna, Austria

Recent advances in kilometer-scale modeling and exascale computing have made it possible to simulate Earth's climate with unprecedented detail. Alongside this, breakthroughs from the James Webb Space Telescope (JWST) present new opportunities for characterizing Earth-like exoplanets and helped create a growing catalog of over 5,000 confirmed exoplanets. In this work, we focus on one such exoplanet, TRAPPIST-1e, a rocky planet slightly smaller than Earth orbiting in the habitable zone of an ultra-cool dwarf star 40 light-years away from our solar system.

Until now, prior climate simulations of exoplanets have relied on coarse-resolution models (grid spacings >100 km), which require explicit parameterizations for convection and clouds, introducing large uncertainties. We carry out the first global climate simulations of TRAPPIST-1e's atmosphere at 5 km horizontal resolution using ICON-Sapphire, a kilometer-scale model previously applied only to Earth's climate. The model is specifically adapted to Trappist-1e, accounting for its size, rotation rate, surface properties, and stellar radiation, offering a groundbreaking look at Earth-like exoplanet atmospheres.

We aim to explore whether ICON-Sapphire can perform stable and physically consistent simulations of Trappist-1e's climate over long timescales. We also seek to understand how high-resolution simulations differ from previous coarse-resolution models, with a particular focus on the representation of clouds and water vapor. Lastly, we will investigate whether the improved resolution enhances our ability to detect and characterize the atmospheres of Earth-like exoplanets through synthetic spectral observations. This could offer critical insights into the detectability of key atmospheric features and the overall feasibility of confirming habitability via remote sensing. Initial simulations will follow the "Hab 1" protocol of the THAI model intercomparison project [1], assuming a static ocean, and a 1-bar N_2-dominated atmosphere with 400 ppm of CO_2. Development and test runs are carried out on the Vienna Scientific Cluster, and production runs will be conducted on the LUMI supercomputer.

This work makes use of new HPC capabilities to pioneer the use of kilometer-scale climate modeling for exoplanet research, bridging advancements in Earth system science and exoplanetary studies. By critically re-evaluating earlier coarse-resolution models and incorporating detailed physical processes, it provides a robust framework to understand the climates of Earth-like exoplanets and their potential habitability.

References

[1] Fauchez T, et al. TRAPPIST-1 Habitable Atmosphere Intercomparison (THAI): motivations and protocol version 1.0, Geoscientific Model Development, 13, 707–716, (2020).

Poster:

Support Systems for National Advanced Computing Service

Emir Imamagić, Jurica Špoljar, Daniel Vrčić, Katarina Zailac, and Martin Belavić

University of Zagreb, University Computing Centre, Croatia

Within the High Performance Computing (HPC) world, the main focus is usually on the issue of machines and data centres. How much *FLOPS* can be delivered, how much electrical power the machines use, etc. But what is often taken for granted are support systems such as user registration, management, and support, as well as reporting, without which the provision of advanced computing service will not work at all. Those are the crucial, core services without which the machines are just machines.

Acting as the hosting entity for two national advanced computing resources – supercomputer Supek and Vrančić The University of Zagreb, University Computing Centre (SRCE) provides a full package of advanced computing service [1], including core services for the entire Croatian scientific and academic community. Therefore, to satisfy a range of specific needs, a new and custom user registration and management application was developed. Web application *computing.srce.hr* [2] is the focal point for users of SRCE Advanced Computing service and it provides an interface for project registration, individual user registration and personalized accounting information.

The application *computing.srce.hr* [3] is developed using *Django/Python* in the back-end while the front-end is written in *ReactJS*. Users access the application using their institutional credentials, either through the national AAI system AAI@EduHr or through the EduGAIN which enables access for international users. Furthermore, the application is integrated with the Croatian Research Information System (CroRIS), allowing applicants to automatically pull information about their research projects and published papers. We also plan to integrate the application with SRCE's e-learning Merlin to enable easier access to advanced computing for the purpose of workshops and classes incorporated into institutional curricula.

Besides the user management application, one of the most important components is the user support system. SRCE uses open source software osTicket which enables tracking of users' requests and distributing workload among SRCE's Advanced Computing Service operation team.

The last component of support systems is the reporting system. SRCE has provided annual reports for its HPC resources since 2021 (with historical reports available from 2019). A transparent and precise reporting system helps HPC centers demands for future investments, but it primarily provides an insight into how institutions, projects and users are utilizing resources and justification of investment for both: the service provider and funding body. In 2025 work has been done on the inclusion of the Advance Computing Service reporting into the SRCE Dashboard, an overall service reporting system.

In this talk, we will demonstrate the user experience from all of the above core service components: requesting access, gaining accounts, getting help through the support ticketing system, and seeing both personal and annual reports.

References

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

[2] https://computing.srce.hr

[3] https://www.srce.unizg.hr/napredno-racunanje/izvjestaj2023/

Poster:

Is the IPMI Exporter a Reliable Tool for Power Monitoring?

Victoria Bringmann, Waleed Khalid, and Alois Schlögl

Scientific Computing, Institute for Science and Technology Austria (ISTA), Klosterneuburg, Austria

Power supply is a major limitation in addressing the growing need for computational resources. While aiming at a scalable and sustainable operation of the HPC cluster at ISTA, we identified the need for a reliable power monitoring tool and collection of meaningful data. For the past months, we tested the IPMI Exporter [1] within our existing Prometheus [2] / Grafana [3] monitoring infrastructure.

Our experience shows that this monitoring system is easy to set up and the resulting graphs are both compelling and insightful (Fig. 1). Grafana provides a wide variety of options to configure visualizations, such that power data can be presented on a node-, rack-, serverroom- and cluster-level.

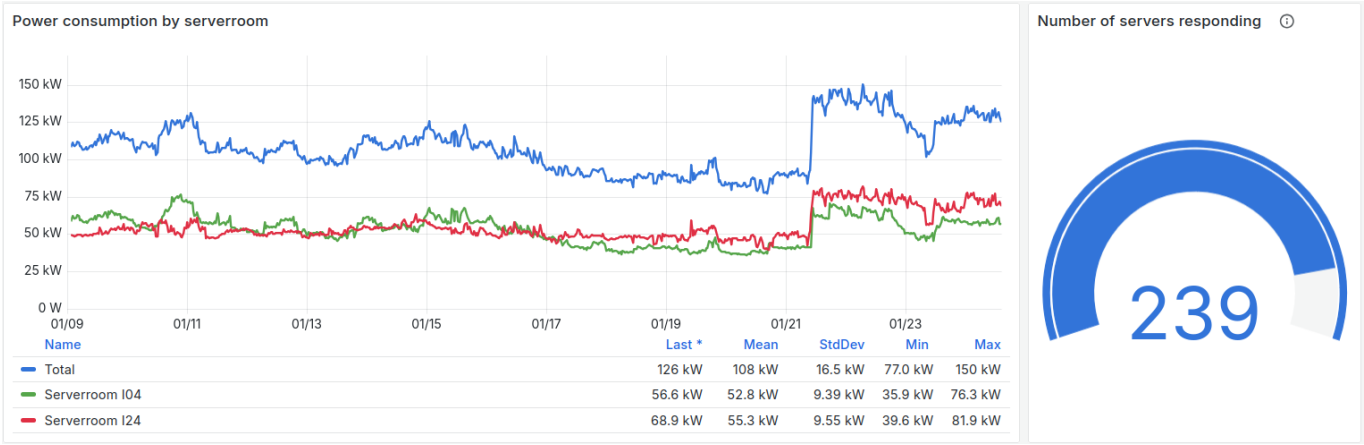


Fig. 1: Example of a Grafana Power Dashboard.

However, it comes with several limitations, especially in the context of heterogeneous cluster environments: IPMI sensors in older servers do not deliver power readings at all. Servers under very high or full load on the other hand tend to suspend system services. The lack of proper documentation and inconsistent hardware implementations across different vendors raises uncertainty in an accurate measurement of the actual power consumption. Another aspect is that Grafana reduces the data resolution with wider time frames selected by the user, which might cause missing or misleading information on power usage.

Depending on the needs, we can conclude that the IPMI Exporter is a useful tool to gain a preliminary understanding of the power consumption of a compute cluster. However, as a complementary measure, we plan to use dedicated power analyzing devices for accurate measurements of the power consumption.

References

[1] Hoffmann, C. *et al.* (2024). *IPMI Exporter*, Version 1.9.0. `https://github.com/prometheus-community/ipmi_exporter`

[2] Prometheus (2023). *Prometheus*, Version 2.48.0. `https://prometheus.io/docs/introduction/overview/`

[3] Grafana Labs (2023). *Grafana*, Version 10.2.2. `https://grafana.com/docs/grafana/latest/`

Poster:

Predicting rates of conformational change of proteins from projected molecular dynamics simulations

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

[a] *Theory Department, National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia*
[b] *Faculty of Mathematics and Physics, University of Ljubljana, Jadranska 19, 1000 Ljubljana, Slovenia*

Understanding the long-time scale behavior of biomolecular systems is of fundamental importance in molecular biophysics. Protein transitions between conformational states associated with structural rearrangements reveal mechanisms of their biological function. For describing the long-time dynamics of such systems one has to identify a few slow collective variables (CVs) while treating the remaining fast variables as thermal noise. This enables us to simplify the dynamics and treat it as diffusion in a free-energy landscape spanned by slow CVs [1, 2].

In this work, we use coarse-grained description of protein dynamics based on the Langevin equation assuming it is Markovian. We design an algorithm for parametrization of multidimensional Langevin equations in which we assume coordinate-dependent diffusion coefficients. Parametrization is based on the principal component analysis of the equilibrium atomistic simulations of protein end-state conformations.

Various combinations of principal components are used to contruct reduced (coarse-grained) subspace in which we perform stochastic dynamics. By finding the appropriate subspace spanned by a given set of principal components with the minimal rate of transition between conformational states we are able to determine optimal CV or a physical reaction coordinate, which captures essential characteristics of conformational change.

As an example we demonstrate the approach on a protein adenosine kinase (AdK) [3]. We performed extensive molecular dynamics simulations using NAMD software on the HPC cluster ARC of the National Institute of Chemistry.

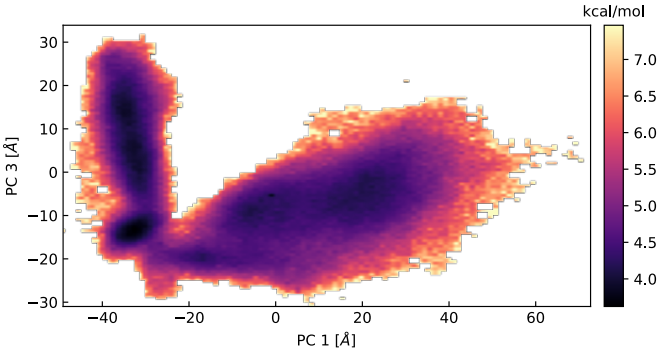


Fig. 1: Projected potential energy surface on a subspace of 2 principal components.

References

[1] Micheletti, C., Bussi, G. and Laio, A., J. Chem. Phys.**129**, 074105 (2008).

[2] Girardier, D.D., Vroylandt, H., Bonella, S., Pietrucci, F., J. Chem. Phys. **159**, 164111 (2023).

[3] Sedej, N., Hubman, A. and Merzel, F. *submitted* (2025).

Poster:

FFplus: Driving SME and Startup Innovation by Unleashing the Potential of HPC and Generative AI

Tina Črnigoj Marc[a] **and FFplus Consortium**[b]

[a] *Arctur d.o.o., Slovenia*
[b] *FFplus Consortium*

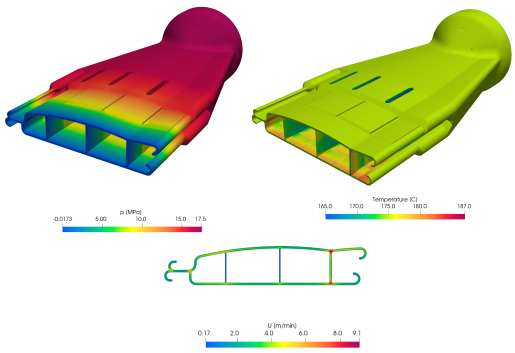
High-Performance Computing (HPC) and Artificial Intelligence (AI) are transformative technologies that can enhance industrial competitiveness, offering significant economic and societal benefits. The FFplus project aims to support SMEs by addressing technical challenges, facilitating access to EuroHPC JU resources, and assisting with business development and outreach.

FFplus will help SMEs through a series of open calls to fund business experiments and innovation studies that showcase the benefits of HPC and generative AI. In total, six open calls will be issued, inviting European SMEs to submit proposals for experiments demonstrating the value of these technologies. Successful applicants will receive funding to carry out their projects.

To encourage adoption, success stories will highlight the business benefits SMEs achieve by integrating HPC and generative AI. The First FFplus Open Call for Business Experiments was Successful: 19 sub-projects were selected for funding, consisting of a total of 43 organisations from 15 countries; meanwhile, 18 Innovation Studies sub-projects were selected for funding, involving a total of 36 organisations and 16 other organisations from 14 countries. The second open call will be scheduled for Business Experiments in late Q2-2025 and Innovation Studies in late Q3-2025.

Innovation Studies will focus on developing large language models (LLMs) for various sectors, while Business Experiments will help SMEs with no prior HPC experience adopt these technologies to solve specific business challenges.

FFplus builds on the successes of the Fortissimo project series, which has left a lasting impact on the European HPC and business landscape. These projects executed over 130 experiments with 330 partners and delivered 120 success stories. These success stories have inspired European industries to adopt digitalization technologies in manufacturing (Industry 4.0) and create innovative products that strengthen the EU economy. FFplus continues this legacy, empowering SMEs to leverage cutting-edge technologies and drive Europe's digital transformation.



References

[1] FFplus, `https://www.ffplus-project.eu/`

Poster:

EXCELLERAT CoE: The European Centre of Excellence for Engineering Applications

Tina Črnigoj Marc[a] and EXCELLERAT P2 Consortium Partners[b]

[a] *Arctur d.o.o., Slovenia*
[b] *EXCELLERAT P2 CoE*

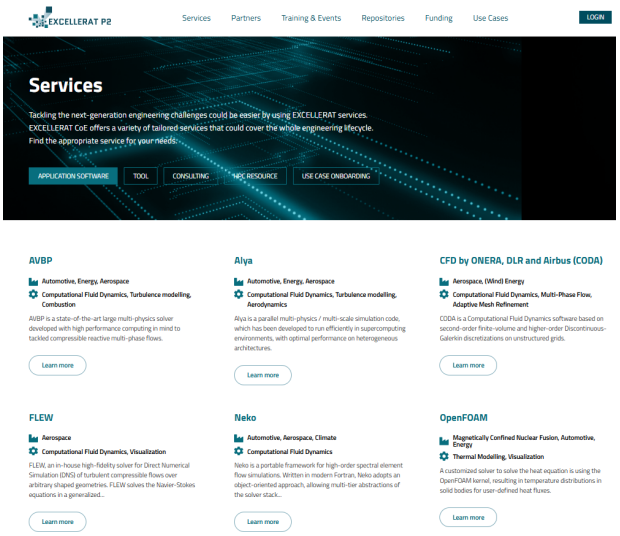
The EXCELLERAT Centre of Excellence (CoE) is a central point for engineering stakeholders seeking expertise in High-Performance Computing (HPC). EXCELLERAT partners focus on data management, analytics, visualization, simulation-driven design, and co-design for Exascale computing, addressing complex engineering challenges and driving innovation in the development phase.

EXCELLERAT aims to advance HPC in engineering, addressing challenges in aerospace, automotive, energy, and manufacturing. Through the EXCELLERAT Service Portal, it provides tools to optimize simulations, integrate data analytics and visualization, and apply AI and Machine Learning to engineering solutions.

The EXCELLERAT Service Portal serves as a central hub for HPC tools, services, and expertise, driving user engagement and generating revenue. The goal is to build long-term industry collaborations, secure partnerships, and create a self-sustaining platform to commercialize research outputs, offering valuable support to the HPC and engineering community.

Key services of the EXCELLERAT Service Portal include:

- HPC-Driven Engineering Applications: Specialized software and use cases that improve simulation scalability, accuracy, and efficiency for industrial applications.
- Workflow Optimization: Tools for streamlining pre-processing, simulation, and post-processing in extreme-scale HPC environments.
- Technology Integration: Using AI/ML to enhance predictive capabilities, reduce computational costs, and provide access to cutting-edge HPC infrastructure.
- Collaborative Ecosystem: Connecting European researchers, industries, and HPC providers to facilitate technology transfer.



- Knowledge and Training: Tailored training programs, consulting services, and resources to build expertise and support HPC adoption in industry.

By offering these services, EXCELLERAT highlights its role in driving HPC innovation in engineering, fostering collaboration across Europe, and ensuring a sustainable future for both research and industry.

References

[1] EXCELLERAT P2., `https://www.excellerat.eu/`

[1] EXCELLERAT P2., `https://services.excellerat.eu/`

Poster:

European Master for HPC study programme

Claudia Blaas-Schenner[a] and Tomas Kozubek[b]

[a] *VSC Research Center, TU Wien, Austria*
[b] *IT4Innovations, VSB – Technical University of Ostrava, Czech Republic*

Launched in January 2022 with a substantial budget of €7 million, the EUMaster4HPC project is driving progress in High-Performance Computing (HPC) education across Europe. Funded by the EuroHPC Joint Undertaking, the project plays a key role in harmonising HPC education by developing a common European Master's programme [1]. The programme has successfully developed and co-designed a comprehensive 120 ECTS curriculum tailored to meet industry needs and offering four key specialisations: Numerical and Data Specialist for the Science Domain, Performance Analyst and Advisor, System Developer and Support, and System Architect. These areas are complemented by transversal skills that are essential in the modern computing landscape.

To enhance global accessibility and foster international cooperation, EUMaster4HPC has established more than 30 dual degree agreements, promoting an enriched educational experience through diverse academic environments. The EUMaster4HPC programme has attracted significant interest, drawing in a broad range of talented applicants from more than 40 different countries. **Now, the call for applications for the fourth cohort is open, and new students are expected to begin their studies in the autumn of 2025/2026.**

Supporting mobility and inclusivity, the project has awarded more than 100 students in the first three cohorts with mobility grants and tuition waivers. The commitment of the initiative to practical learning is evidenced by the acceptance of more than 60 internships with partners from research and supercomputing centres and industry, providing students with real-world applications.

The Moodle platform developed by the project is the cornerstone of HPC education, offering a variety of teaching resources and newly developed MOOCs. These online courses cover important topics such as Massive Parallel Programming on GPUs, Introduction to Quantum Computing, and Energy-Aware Parallel Computing, which are crucial for students aiming to excel in the HPC field. In addition to these educational advances, EUMaster4HPC has facilitated important academic activities, including workshops, challenges, and summer schools. **The next summer school will take place from 14-25 July 2025 at the Middle East Technical University, Ankara, Turkey [2].** This event will cover applications related to data science, modelling and simulation, chemistry, aerospace, materials science, physics, and more.

The future of the EUMaster4HPC project appears bright, with plans to expand its network of partners and awarding universities. The program is committed to continually updating its curriculum to align with the changing needs of the industry and to solidify its position as a cornerstone of European HPC education.

Acknowledgement: The project has received funding from the European High-Performance Computing Joint Undertaking under grant agreement No. 101051997.

References

[1] https://eumaster4hpc.eu

[2] https://eumaster4hpc.eu/summer-school-2025

Poster:

Scaling Differentiable Simulations in Cosmology to Multiple GPUs

Lukas Winkler[a], **Florian List**[b], **and Oliver Hahn**[a,b]

[a] *Department of Astrophysics, University of Vienna, Türkenschanzstraße 17, 1180 Vienna, AT*
[b] *Department of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, AT*

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

Recent libraries such as Jax enable the implementation of physical models that run efficiently on GPUs while also being automatically differentiable, allowing gradient information to propagate through the entire simulation. Building on this, the Disco-DJ framework [1] (**DI**fferentiable **S**imulations for **CO**smology – **D**one with **J**ax), 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].

While Disco-DJ allows running 3D N-body with 512^3 particles on a 1024^3 PM grid in about 0.4 s per timestep on an Nvidia A100 GPU (40 GB RAM), memory constraints prevent scaling to higher resolutions on a single GPU. The vast fields of view covered by recent surveys demand extremely large simulation box volumes where accurately resolving small non-linear scales requires particle resolutions in the range of 1024^3 to 2048^3 – even for fast, approximate methods such as PM simulations.

To scale Disco-DJ to these resolutions, we adapt the codebase to introduce sharding of all major arrays and ensure that the Jax compiler applies this consistently across all calculations. To achieve this, we develop custom SPMD lowering rules to perform distributed FFTs (`np.fft.rfftn`) without collecting all data onto a single device, along with modified versions of `jax.lax.scatter_add` and `jax.lax.gather` that compute density fields and interpolate accelerations back from the grid while communicating particles that might have left the local subset of the box.

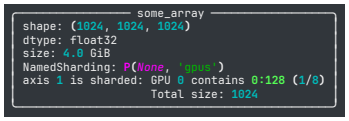


Fig. 1: Debugging output of an array sharded along multiple GPUs

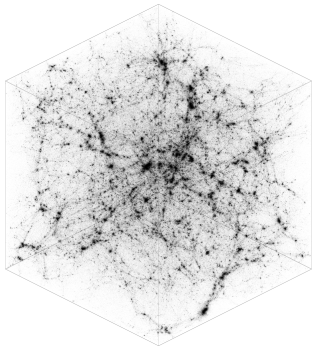


Fig. 2: The cosmic web

References

[1] List, F. et al. (in preparation)

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

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

Roundtable:

Achieving Gender Balance in HPC: Retention and Representation from the Central European Perspective

Karina Pešatová

VSB - Technical University of Ostrava, IT4Innovations, Ostrava, Czech Republic

While awareness of gender imbalance in High Performance Computing (HPC) has grown, systemic change remains slow, especially in the research domain. In Central Europe, women continue to be underrepresented in HPC research teams and leadership roles, and retention through career transitions—from early-stage researcher to senior expert—remains a critical challenge.

This round table invites researchers from across the HPC community to come together for a candid, evidence-informed discussion about gender equity in our field. We will begin with a brief presentation of regional data illustrating the current landscape of women's employment in HPC research institutions, including findings from our member organizations. These insights aim to ground our conversation in lived realities and shared challenges.

Rather than offering a top-down set of solutions, this session seeks to engage participants in a collaborative dialogue around bottom-up strategies that can be initiated within research teams, labs, and departments. Topics may include inclusive supervision and mentoring practices, team culture, institutional support gaps, and the role of peer networks. We especially welcome contributions that reflect practical experiences—what has worked, what hasn't, and what we can try next.

The session aims to empower researchers at all career stages to contribute to shaping more inclusive and resilient research environments in HPC.

WHPC
WOMEN IN HIGH PERFORMANCE COMPUTING

References

[1] https://womeninhpc.org

[2] https://www.linkedin.com/company/central-european-women-in-hpc/

Keynote Talk:

Compressing AI Models at GPT Scale

Dan Alistarh

Institute of Science and Technology, TU Wien, Austria

A key barrier to the wide deployment of highly-accurate machine learning models, whether for language or vision, is their high computational and memory overhead. Although we possess the mathematical tools for highly-accurate compression of such models, these elegant techniques require second-order information about the model's loss function, which is hard to even approximate efficiently at the scale of billion-parameter models. In this talk, I will describe our work on bridging this computational divide, which enables the accurate second-order pruning and quantization of models at truly massive scale. Compressed using our techniques, models with billions and even trillions of parameters can be executed efficiently on GPUs or even CPUs, with significant speedups, and negligible accuracy loss.

Learning macroscopic equations of motion from particle-based simulations of a fluid

Matevž Jug[a,b], Daniel Svenšek[a,b], Tilen Potisk[a,b], and Matej Praprotnik[a,b]

[a] *Laboratory for Molecular Modeling, National Institute of Chemistry, Slovenia*
[b] *Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, Slovenia*

Describing the dynamics of a material in terms of partial differential equations provides a systematic framework for predicting its behavior on large spatio-temporal scales, which are suitable for engineering applications. However, for novel complex materials, only a particle-based description is usually available. Since simulations of these particle-based models can generate vast amounts of data, the discovery of dynamic equations through data-driven means is becoming increasingly popular.

The method that will be presented combines a weak formulation of the unknown equations, sparse regression (specifically, the SINDy [1] framework), and a new model selection measure that balances stability and accuracy to discover equations governing continuum-level dynamics directly from particle-based simulations [2]. From simulations of a simple fluid, modeled using dissipative particle dynamics, our method successfully extracts the mass continuity equation and the Navier-Stokes equation, the latter also containing the correct equation of state. Due to its high robustness to noise, such an approach can be readily applied to simulations or experimental data of more complex materials, where macroscopic dynamic equations are not, or are only partially, known.

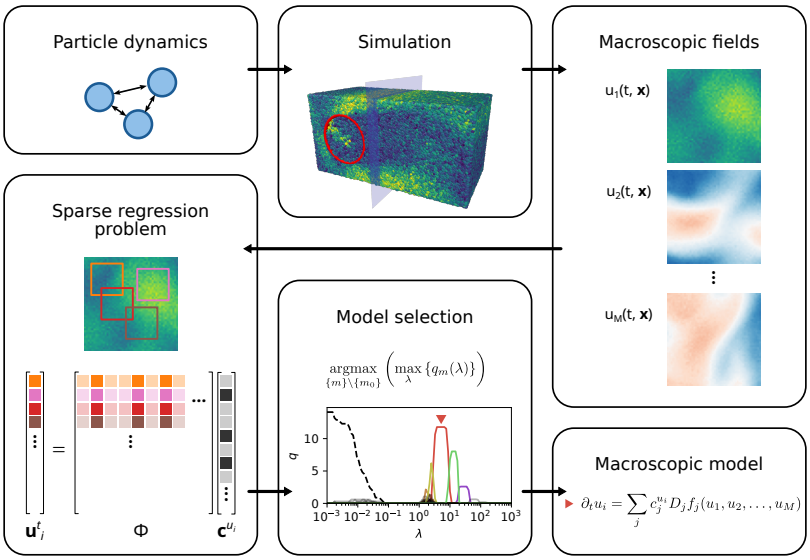


Fig. 1: Overview of data acquisition and our learning framework [2].

References

[1] Brunton, S. L., Proctor, J. L., Kutz, J. N., Proc. Natl. Acad. Sci. USA **113**, 3932 (2016).

[2] Jug, M., Svenšek, D., Potisk, T., Praprotnik, M., Comput. Methods Appl. Mech. Eng. **432**, 117379 (2024).

Optimizing Distributed Deep Learning Training by Tuning NCCL

Majid Salimi Beni[a], Ruben Laso[b], Biagio Cosenza[c], Siegfried Benkner[b], and Sascha Hunold[a]

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

Distributed Deep earning is essential for training large-scale neural networks when the entire data set or model cannot fit into a single machine. The communication layer of such a deep learning framework is responsible for synchronizing model updates and exchanging gradients between nodes, and the communication operations in that layer must be efficient. The NVIDIA Collective Communications Library (NCCL) is a widely used back-end for communication in GPU-accelerated clusters. Similar to the Message Passing Interface (MPI), NCCL's efficiency depends on its parameter configuration [1, 2], including the choice of communication algorithms, buffer sizes, and network types.

NCCL Parameter Tuning: We propose a two-step *offline tuner* to optimize the NCCL parameter configuration for multi-GPU clusters. First, we *profile* the training of the models to determine the most relevant message sizes. Second, we employ a Bayesian optimizer to find an *efficient parameter configuration.*

Experimental Results: Figure 1 compares the performance of two deep learning models (`Bert` and `NasNetMobile`) on 2 nodes of the Leonardo supercomputer, using TensorFlow and Horovod. On top, we compare the bandwidth obtained after tuning the collectives for the most frequently used message size of each model. The tuned configurations improved the bandwidths of the respective NCCL operations in the microbenchmarks by 2.26 and 21.02 times. On the bottom, we show an improvement in training performance of 12% and 13% for `Bert` and `NasNetMobile`, respectively, when using the tuned configuration. Our experiments highlight the significant performance gains achievable through optimizing NCCL in distributed deep learning training.

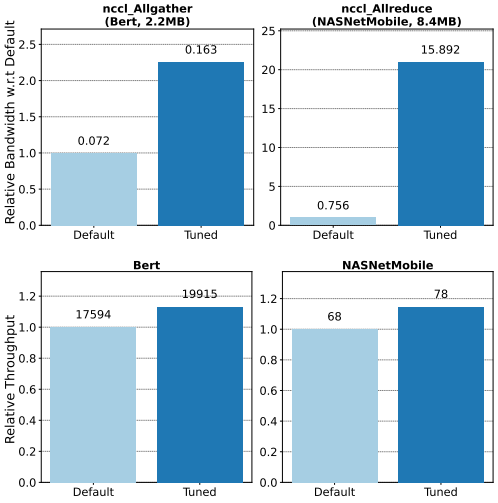


Figure 1: Default vs. Tuned NCCL collectives on 2×4 NVIDIA A100 GPUs. The raw values are shown on top of the bars: Bandwidth in GB/s, and Throughput in samples/s. Higher is better.

References

[1] De Sensi, D., *et al.* "Exploring GPU-to-GPU Communication: Insights into Supercomputer Interconnects," SC (2024).

[2] Salimi Beni, M., Cosenza, B., and Hunold, S., "MPI Collective Algorithm Selection in the Presence of Process Arrival Patterns," CLUSTER (2024).

ncclsee: A Lightweight Profiling Tool for NCCL

Ioannis Vardas[a], Ruben Laso Rodriguez[b], and Majid Salimi Beni[a]

[a] *TU Wien: Parallel Computing Research Group*
[b] *Universität Wien: Scientific Computing Research Group*

To achieve scalable and efficient distributed deep learning, optimized GPU communication is paramount. We introduce `ncclsee`, a lightweight profiler plugin built using version 2 of NVIDIA's Collective Communication Library (NCCL) [1] profiling interface and the NVIDIA CUDA Profiling Tools Interface (CUPTI) [3]. `ncclsee` captures communication patterns in real time, offering insights into GPU communication performance. By focusing on simplicity and efficiency, `ncclsee` enables users to pinpoint and alleviate bottlenecks in distributed workloads, making it useful for debugging and optimizing large-scale AI training workflows.

NCCL is the de facto library for GPU communication with NVIDIA GPUS in deep learning frameworks such as PyTorch and TensorFlow. It delivers high performance by leveraging advanced technologies, including RDMA (Remote Direct Memory Access) and GPUDirect, which enable direct GPU-to-GPU data transfers, over interconnects such as PCIe, Infiniband and NVLink, with minimal CPU involvement. To optimize collective operations like AllReduce, Broadcast, and AllGather, NCCL automatically selects algorithms such as the Ring or Tree, depending on message size, topology, and buffer size characteristics. Once the parameters are selected, NCCL creates a CUDA kernel to perform the collective operation on the GPUs.

Profiling NCCL behavior at scale remains challenging, tools like NVIDIA Nsight can produce highly detailed traces, but the volume of data generated makes them impractical to analyze for large GPU clusters. `ncclsee` tackles this problem by offering summary information on NCCL operations. It leverages NCCL's event callbacks, including start and stop events as well as proxy progress activity, to accurately track asynchronous operations. Because `stopEvent` indicates only that a collective has been enqueued rather than completed, `ncclsee` utilizes CUPTI to measure the time of the corresponding CUDA kernel that performs the NCCL operation on the GPUs. The main challenge for developing `ncclsee` was creating an efficient interface that associates NCCL's profiling API events to the correct CUPTI's event.

`ncclsee` captures summary data for each NCCL operation based on the buffer size, presenting information in a concise format (e.g., Operation: `ncclAllReduce`, Buffer range: 128-4096 Bytes, Calls: 52, Time: 770 ms), allowing users to quickly identify communication patterns and potential bottlenecks.

Ease of use: Integrating `ncclsee` into existing workflows is straightforward. After compiling `ncclsee` producing the `libnccl-profiler.so`, users simply have to set the `NCCL_PROFILER_PLUGIN` environment variable to point to the path of `libnccl-profiler.so`. Once enabled, `ncclsee` records metrics for applications that use NCCL directly or through frameworks that depend on it, such as PyTorch and TensorFlow. `ncclsee` is actively under development, with a functional version already available on GitHub [3]

References

[1] NVIDIA Collective Communication Library (NCCL), `https://developer.nvidia.com/nccl`.

[2] NVIDIA CUDA Profiling Tools Interface (CUPTI), `https://developer.nvidia.com/cupti`.

[3] ncclsee: A lightweight profiling tool for NCCL, `https://github.com/variemai/ncclsee`.

HPC for Hybrid Threat Resilience: Insights from the HYBRIS Project

Simeon Harrison[a], Florian Goldenberg[a], Markus Hickel[a], Siegfried Höfinger[a], Sanaz Sattari[a], Markus Stöhr[b], and Jan Zabloudil[b]

[a] *VSC Research Center, TU Wien, Austria*
[b] *BOKU University, Austria*

Hybrid threats demand sophisticated computational solutions to address their complexity, necessitating the integration of high-performance computing (HPC) frameworks to enable rapid and scalable processing. The HYBRIS project explores the development and deployment of AI-driven methodologies for hybrid threat resilience, leveraging cutting-edge HPC infrastructure to facilitate advanced data processing, model training, and inference. The project focussed on detecting threats deriving from the spread of misinformation through the internet and social media, equipping government decision makers with an early warning tool to oppose misinformation with a timely counter narrative.

This presentation outlines the key findings and methodologies applied during the HYBRIS pilot project, conducted on the Vienna Scientific Cluster (VSC). The project evaluated the scalability and efficiency of large-scale AI workflows, including pre-training, fine-tuning, and inference of transformer-based models on heterogeneous datasets. Using the GDELT Medium Dataset (1.9 GB), the experiments achieved near-linear scaling, reducing model training time from 225 hours on a single GPU to under one hour on eight nodes, which 2 GPUs each. This was achieved with distributed training frameworks and libraries such as Hugging Face Accelerate and DeepSpeed.

Central to the success of the project was the optimization of resource allocation, leveraging techniques like parameter sharding and mixed-precision training. Profiling tools, including Linaro/Forge, provided detailed insights into computational bottlenecks, enabling targeted optimizations. The results demonstrated that systematic use of HPC tools not only accelerates computational workflows but also provides actionable insights into optimizing resource utilization. In addition to the technical achievements, this pilot highlights the practical implications of integrating distributed HPC workflows into AI-driven applications, such as hybrid threat detection and resilience. It serves as a foundation for future advancements in AI-enabled HPC workflows, emphasizing the critical role of collaboration between research institutions and HPC facilities.

Apart from sounding the technical feasibility, the HYBRIS project put great emphasis on ethical and legal considerations that would inevitably be put in place as guard rails to ensure any future implementations would adhere to national and EU laws and regulations, most notably the EU AI Act. HYBRIS did not just thrust HPC to the forefront of today's defence capabilities, but also highlighted the tension between providing national security, ensuring technical feasibility and safe guarding civil rights and liberties.

References

[1] https://www.kiras.at/en/financed-proposals/detail/hybris/

Go wrapper for CUDA

Timotej Kroflič, Davor Sluga, and Uroš Lotrič

Faculty of Computer and Information Science, University of Ljubljana

Programming in CUDA, a framework for general-purpose computing on graphics processing units developed by Nvidia, forces developers to use C/C++ [1] as the main programming language. Nevertheless, Go is often employed in courses on distributed systems and parallel programming due to its powerful support for multi-threading. We aimed to make CUDA accessible in Go, as Nvidia does not offer native libraries for this purpose. Go is a statically typed, compiled language that resembles C syntactically yet incorporates features such as memory safety, garbage collection, structural typing, and built-in concurrency.

We find retaining kernel programming in C acceptable due to the syntactical similarity between Go and C. As a result, the project involved developing a Go wrapper library and a tool called CudaGo[2] similar to *mexcuda* [2], which provides much the same functionality to Matlab. The tool facilitates the execution of CUDA kernels from pre-written C/C++ code while the wrapper manages GPU memory, handles synchronization between the host and device, and more. The wrapper library relies on the CUDA Driver API. CudaGo is a Go-based tool that processes CUDA kernels written in C/C++ and generates Go wrappers in a new package. Compilation of the kernels is performed using NVRTC, Nvidia's runtime compilation library. We provided support for most of the CUDA Driver API. We left out some more exotic functionalities, which will be added in the future.

With this approach, developers can write CUDA kernels in C/C++ and invoke them directly from Go. However, key differences between Go and C required careful handling. For example, the default `int` type in Go is a 64-bit number, while in C, it is (usually) a 32-bit number. Failure to account for such discrepancies would result in incorrect outcomes.

We performed benchmarks on a CUDA implementation of a simple image processing algorithm. We were interested in the execution times of Go-launched kernels compared to C/C++. Results reveal that kernel execution times are unaffected by the tool, and the overhead of invoking kernels from Go is negligible. The average execution time of kernel invocation in C/C++ was 15 µs, while in Go, it was 23 µs. The average difference was thus 7 µs.

References

[1] John R Nickolls, Ian Buck, Michael Garland, K. Skadron, Scalable parallel programming with CUDA, ACM SIGGRAPH 2008 Classes, DOI: 10.1145/1401132.1401152 (2008)

[2] Matlab MexCuda, Running mex functions containing CUDA code

[2]The library and the tool are available on GitHub: `https://github.com/InternatBlackhole/cudago`

Benchmarking A40, L40S, H100 GPUs

Stefano Elefante, Waleed Khalid, and Alois Schlögl

Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria

Vendor-independent performances tests are important in assessing the benefit of different platforms. Here the performance of NVIDIA GPU cards, specifically the A40, L40S, and H100 models is being evaluated. The test utilized a standard code for matrix multiplication such as *matrixMulCUBLAS.cpp* which relies on *cuBLAS* subroutine available in the library CUDA 12.2.2 and the program *tf32TensorCoreGemm.cu* which is designed to exploit NVIDIA Tensor Cores technology [1]. These codes perform rectangular matrix multiplication of various sizes, with the latter also including matrix addition. Tests were conducted on the ISTA local cluster across different machines, each equipped with A40, L40S or H100 GPU cards type with FP32 precision. In order to quantify the accuracy of the results, average and standard deviation for 95% confidence interval were computed.

Table 1: Matrix size

[redacted]	Standard code (A*B)	NVIDIA Tensor Cores code (A*B+C)
1	(1280,960)*(960,640)	(2048,1024)*(1024,2048)+(2048,2048)
2	(2560,1920)*(1920,1280)	(4096,2048)*(2048,4096)+(4096,4096)
3	(3840,2880)*(2880,1920)	(8192,4096)*(4096,8129)+(8192,8192)
4	(5120,3840)*(3840,2560)	(16384,8192)*(8192,16384)+(16384,16384)
5	(6400,4800)*(4800,3200)	(32768,16384)*(16384,32768)+(32768,32768)

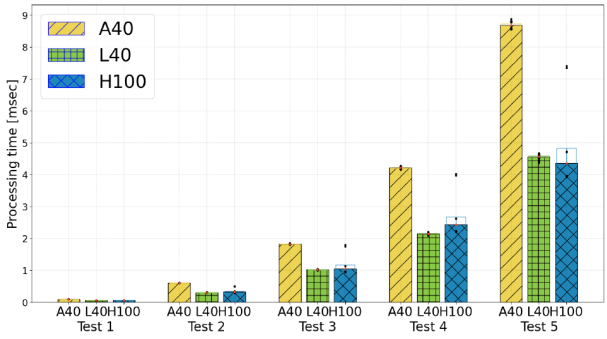


Fig. 1: Test using standard code

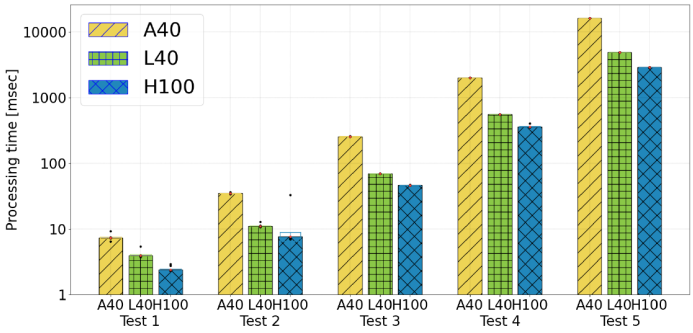


Fig. 2: Test using NIVIDIA Tensor Cores code (log-scale)

The analysis reveals that when a standard matrix multiplication code is utilized, the L40S and H100 GPUs exhibit similar performance levels, whereas the A40 shows longer execution times (Fig. 1). Conversely, when a code designed to exploit NVIDIA Tensor Cores technology is employed, the H100 outperforms both the L40S and A40 (Fig. 2). Furthermore, the H100 features more VRAM, allowing it to handle larger matrices compared to the L40S and A40. These results do not conflict with the benchmark tests provided by the vendor.

References

[1] cuda-samples v12.5 `https://github.com/nvidia/cuda-samples/`

Accelerating Differential Evolution for High-Performance Computing: Leveraging Modern GPU Architectures and Mixed-Precision Arithmetic

Domen Verber

Institute of Informatics, Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia

The rapid evolution of High-Performance Computing (HPC) has significantly enhanced the capabilities of optimization algorithms, particularly Differential Evolution (DE). DE is a stochastic, population-based optimization method renowned for its simplicity and effectiveness across various complex, multidimensional problems. However, its computational demands, especially in large-scale applications, have historically limited its efficiency. The advent of modern GPU architectures, specifically designed to accelerate AI and HPC workloads, offers a promising avenue to overcome these limitations.

Modern GPUs like NVIDIA's Ampere and Hopper architectures are equipped with specialized hardware units like Tensor Cores, optimized for mixed-precision arithmetic operations. These units enable the execution of low-precision formats (e.g., FP16, BF16) alongside standard single-precision (FP32) and double-precision (FP64) computations. This mixed-precision capability allows for significant matrix and vector operations speedups, which are integral to DE processes such as mutation, crossover, and fitness evaluation. By leveraging mixed-precision arithmetic, DE can achieve faster computation times and reduced memory usage, facilitating the handling of larger populations and higher-dimensional search spaces.

Implementing DE on these advanced GPU platforms involves several considerations to harness their capabilities fully. Parallelization strategies must be carefully designed to align with the GPU's architecture, ensuring efficient utilization of its massive parallelism. This includes optimizing memory access patterns to take advantage of high-bandwidth memory and minimizing data transfer between the CPU and GPU to reduce latency. Also, managing the trade-offs between precision and performance is crucial; while lower precision can accelerate computations, it may also introduce numerical instability or reduce solution accuracy if not appropriately handled. Techniques such as adaptive precision scaling and error compensation methods can mitigate these risks, maintaining the robustness of the DE algorithm.

Integrating DE with modern GPU architectures opens avenues for real-time and large-scale applications. For instance, in engineering design optimization, DE can be used to explore vast design spaces more efficiently, leading to innovative solutions that were previously computationally prohibitive. In machine learning, DE can assist in hyperparameter tuning, benefiting from the GPU's ability to handle large datasets and complex models. Furthermore, the energy efficiency of GPUs, enhanced by mixed-precision computing, aligns with the growing emphasis on sustainable computing practices in HPC.

References

[1] Janssen, D. M., Pullan, W., and Liew, A. W-C., "GPU Based Differential Evolution: New Insights and Comparative Study", (2024). [Online]. Available: https://arxiv.org/abs/2405.16551.

[2] NVIDIA Corporation, "Train With Mixed Precision," NVIDIA Developer Documentation, (2022). [Online]. Available: https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html

Benchmarking the scalability and communication of the CholeskyQR2-IM algorithm on national and European HPC resources

Nenad Mijić[a], Abhiram Kaushik Badrinarayanan[b,c], and Davor Davidović[c]

[a] *University Computing Centre, University of Zagreb, Croatia*
[b] *University of Jyväskylä, Finland*
[c] *Centre for Informatics and Computing, Rudjer Bošković Institute, Croatia*

In this work, we benchmark and analyze the performance of our novel algorithm, **CholeskyQR2-IM** , on both national and EuroHPC computing resources. The algorithm efficiently computes the QR factorization of extremely ill-conditioned matrices on distributed multi-GPU systems [1]. It builds on the communication-avoiding CholeskyQR algorithm and includes a block Gram-Schmidt variant to improve numerical stability. The main application of this algorithm is the ChASE library [2], a high-performance library designed to solve large-scale Hermitian eigenproblems on distributed hybrid CPU-GPU systems. The code is highly parallelizable since it can be cast completely in terms of memory-optimized BLAS-3 operation and is based only on the *allreduce* communication routine.

The benchmarks were performed on the national supercomputing resource **Supek**, located at the University Computing Centre (Srce), University of Zagreb, and on the EuroHPC supercomputer **Leonardo**, operated by CINECA in Italy. Both strong and weak scalability benchmarks were performed to analyze how the choice of communication library affects the overall performance. The analysis compared two of the most commonly used communication libraries: MPICH/OpenMPI and Nvidia NCCL. Additionally, we evaluated their performance when custom data types were used in reduction operations.

To perform this analysis, numerous runs with different configuration options had to be prepared and submitted. To streamline this process, we used the JUBE benchmarking environment developed by the Jülich Supercomputing Centre.

The results on SUPEK show a significant speedup when using Nvidia NCCL compared to CRAY MPICH, especially for up to 12 nodes with 4 GPUs per node. However, communication still accounts for a significant portion of the total execution time. While NCCL shows superior performance, its lack of support for custom data types leads to unnecessary communication overhead, especially when only the lower triangular matrix needs to be transferred between GPU memories.

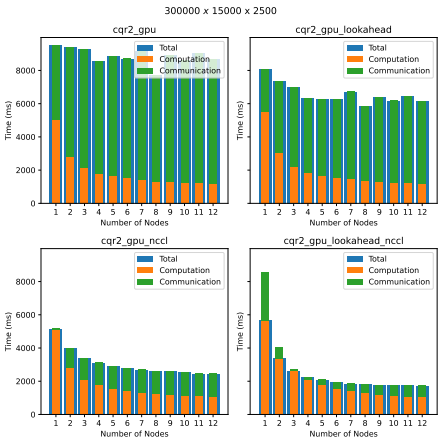


Fig. 1: Strong scaling. MPI vs. NCCL versions on dense matrix $300k \times 15k$.

References

[1] Mijić, N., Kaushik, A., Davidović, D., ArXiv, (2024).

[2] Wu, X., Davidović, D., Achilles, S., Di Napoli, E., Proceedings of the Platform for Advanced Scientific Computing Conference, 12 (2022).

GaMS meets Vega and Leonardo: Training Slovene LLMs

Domen Vreš, Iztok Lebar Bajec, and Marko Robnik-Šikonja

University of Ljubljana, Faculty of Computer and Information Science

Large language models (LLMs) are becoming an essential part of modern automated processes. In this work, we present the development of Slovene LLM GaMS-1B [1] on Vega HPC cluster and initial findings on the development of GaMS-9B on Leonardo HPC cluster. In both cases we used the NVIDIA NeMo framework, which provides support for several parallelisms and allows for easy scaling.

GaMS-1B: GaMS-1B has 1.3 B parameters and was trained on Vega. We used tensor parallelism (TP) to split the model across **4** GPUs (within a single node) and data parallelism (DP) across **16** nodes. Our experience showed that **16** was the optimum value when cluster occupancy and inter-node communication limitations were considered. Using sequence length 2048 and micro–batch size 8, vRAM consumption was 32 GB per GPU (out of 40 GB available). Training the model for 4 epochs (in each epoch we used 28 billion tokens) took **110** hours (run time). Out of these, the model was effectively trained for approximately **64** hours, while **46** hours were lost on job hangs caused by inter-node communication limitations. Time lost on job restarts is excluded.

Inter-node communication limitations: Due to network instabilities on both Vega and Leonardo, NCCL regularly crashes during training, causing jobs to hang. As these events are unpredictable training requires constant monitoring. Additionally, on Vega hung jobs need to be canceled manually. This results in loss of training time and resource allocation budget. By inspecting training logs, we observed that certain nodes cause hangs more regularly than others. By excluding those, we improved the stability of our training jobs.

GaMS-9B: Compared to GaMS-1B training GaMS-9B increases vRAM consumption heavily. This is due to the larger number of parameters and an increased sequence length. To train on Vega, activation checkpointing has to be enabled, which slows down training considerably. On Leonardo this is not required due to more vRAM per GPU (64 GB), but we still have to use TP 8, which splits the model across 2 nodes. The maximum DP value that we were able to use on Leonardo before experiencing job hangs was **16** (32 nodes). Additionally, we test the training on one DGX H100 and A100 node on our internal Frida cluster.

Table 1: Comparison of training on Vega, Leonardo, one DGX-H100, and one DGX-A100 node on the UL FRI Frida cluster.

Cluster	Vega	Leonardo	Frida A100	Frida H100
GPU type	A100	A100	A100	H100
GPU vRAM (GB)	40	64	40	80
# GPUs per node	4	4	8	8
700 M tokens TP8 DP1	38.5 h	29.0 h	27.7 h	10.7 h
1 B tokens with TP8 DP*	6.9 h	2.6 h	39.6 h	15.4 h

The training speed comparison between systems is shown in Table 1. We measure the training time on each by training the 9B model on a corpus with 700 million tokens using TP 8 and DP 1. We then compute the training speed on 1 B tokens using the following DP ranks: 8 for Vega, 16 for Leonardo, and 1 for Frida (a single node of each type); these scores are a proxy of achievable times using the available resources. For the final models, we expect to use approx. 150 B tokens.

References

[1] Vreš, D., Božič, M., Potočnik, A., Martinčič, T., and Robnik-Šikonja, M., Generative Model for Less-Resourced Language with 1 Billion Parameters, Proceedings of Language Technologies and Digital Humanities Conference **JT-DH-2024**, 485 (2024).

Keynote Talk:

High performance computing at the boundaries of quantum chaos

Jan Šuntajs [a,b]

[a] *Department of Theoretical Physics, J. Stefan Institute, SI-1000 Ljubljana, Slovenia*
[b] *Faculty of Mechanical Engineering, University of Ljubljana, SI-1000 Ljubljana, Slovenia*

This talk highlights the importance of high-performance computing (HPC) in addressing some of the fundamental challenges in nonequilibrium quantum physics, particularly in understanding ergodicity-breaking transitions (EBTs) in isolated interacting quantum systems. These transitions delineate ergodic systems (also referred to as quantum chaotic), which equilibrate over time, from nonergodic ones, which retain memory of their initial conditions indefinitely. As such, nonergodic systems hold promise for applications in quantum computing and memory devices, thus making their identification and characterization an area of intense interest for theoreticians and experimentalists alike. The difference between ergodic and nonergodic systems is schematically shown in **Fig. 1.**

We use large-scale numerical calculations to probe the boundary between ergodic and nonergodic systems [1,2,3]. Our methodology relies on full and partial exact diagonalization of the studied Hamiltonian matrices, requiring extensive memory and computational resources that grow exponentially with the physical system size. These challenges are further compounded by the need to perform disorder averaging across 100-10000 samples (depending on the studied system size), which necessitates both the scalability and flexibility of HPC environments. To make effective use of those, we employ advanced SLURM features, including job arrays and job dependencies, which allow for an efficient management of extensive computational workflows. Additionally, we are currently focusing on the implementation of more efficient diagonalization algorithms to push the boundaries of numerically feasible system sizes.

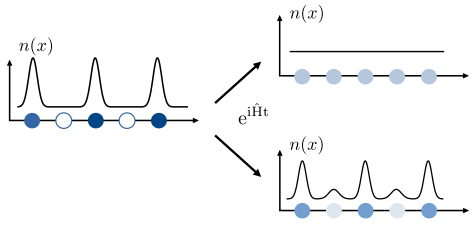


Fig. 1: A schematic of the difference between an ergodic (above right) and nonergodic (below right) quantum system following a time evolution from nonequilibrium state.

The results of our studies have challenged longstanding paradigms in ergodicity-breaking transitions [1], which spurred new research avenues and led to a flurry of new activity in the field. In our subsequent studies, we introduced a toy model of an EBT [2], providing a tractable framework in which analytical predictions of a transition can be readily corroborated numerically. To a large extent, these breakthroughs can be attributed to the use of HPC systems, as they continue to provide the computational backbone necessary for uncovering complex quantum phenomena. Our experiences highlight practical HPC strategies that can benefit other computationally intensive disciplines.

References

[1] Šuntajs, J., Bonča, J., Prosen, T. and Vidmar, L., Phys. Rev. E **102**, 064214 (2020).

[2] Šuntajs, J. and Vidmar, L., Phys. Rev. Lett. **129**, 060602 (2022).

[3] Šuntajs, J., Prosen, T. and Vidmar, L., Phys. Rev. B **107**, 064205 (2023).

Container Forum:

EPICURE and Containers, activities in EuroHPC

Alja Prah

Josef Stefan Institute, Ljubljana, Slovenia

One of the key initiatives supported by the EuroHPC Joint Undertaking is EPICURE—a European project designed to establish and operate a distributed, yet coordinated, HPC application support service [1,2]. This effort responds to a growing need: while local HPC system administrators often face increasing user demand, they frequently lack the time or capacity to address complex, high-level support issues. EPICURE aims to bridge this gap by providing targeted, in-depth support that helps researchers across Europe fully exploit the capabilities of EuroHPC systems.

EPICURE is structured around six work packages (WPs), with the technical core comprising WP2 (Code porting, enabling, and scaling), WP3 (Code optimization), and WP4 (Technical coordination). WP4 does internal coordination, while WP3 handles advanced code transformations, including accelerator porting. WP2, on the other hand, focuses on running performance analysis, offering best-practice guides, and providing optimized builds or containerized versions of scientific applications.

Fig. 1: EPICURE project logo.

In this context, container technologies have emerged as a possible solution for portability, reproducibility, and streamlined deployment. By packaging applications together with their dependencies, containers simplify cross-platform compatibility and significantly reduce the setup burden for users and administrators alike. EPICURE promotes the use of containerized workflows, contributing pre-built containers, build recipes, and guides to support adoption across different EuroHPC sites.

To broaden awareness and knowledge-sharing, EPICURE also hosted a dedicated webinar on containers, showcasing real-world use cases and demonstrating the potential of containers in HPC environments. While container creation in the context of EPICURE currently remains largely case-specific, there is a desire to move toward a more coordinated approach, perhaps in collaboration with the EuroHPC Container Forum, to establish a sustainable and community-driven ecosystem for container development and reuse.

This talk will present an overview of EPICURE's activities, with a particular focus on its container-related efforts. It will highlight both the technical challenges and the collaborative strategies being pursued to ensure that container technologies become a robust, user-friendly component of the EuroHPC landscape.

References

[1] EuroHPC JU. (7 Feb 2024). EPICURE: A new R&I project is launched by the EuroHPC JU.

[2] EPICURE - Unlocking European-level HPC support. Accessed April 10, 2025.

Container Forum:

Confidential containers in multi-tenant HPC environments

Barbara Krašovec[a,b] **and Dejan Lesjak**[a]

[a] *Jozef Stefan Institute*
[b] *EGI CSIRT*

High-Performance Computing (HPC) environments are undergoing significant change. Once characterised by their closed and rigid configurations, these systems are now evolving to become more agile and flexible in response to the diverse and growing demands of the scientific, industrial, and AI communities. This evolution is characterised by a convergence of traditional HPC with cloud and containerised solutions, such as adoption of technologies like OpenStack and Kubernetes, which enable the creation of scalable and adaptable infrastructures.
As a result of all these changes, the types of workloads running on HPC systems are evolving as well, leading to an increase in containerized environments and workloads with large data requirements. API access is becoming the standard method for seamless interaction with HPC resources. Along with these advances, the importance of security is growing. To protect services, HPC environments need to implement security measures at multiple levels, including protecting data in transit and at rest, and securing execution environments. Another challenge is the protection of data-in-use. This raises the question: can hardware-based Trusted Execution Environments (TEEs)[1,2] be used effectively in the HPC context? And how can we ensure the secure and isolated use of containers in the first place?

In this presentation, we will present some strategies for ensuring the isolation and security of containerised workloads. This includes encrypting data at rest and in transit, using user namespaces, implementing cgroups for resource management, using Trusted Execution Environments (TEEs), and employing sandboxing techniques to limit the attack surface. In addition, secure container deployment includes using secure remote builders, signing and encrypting containers, configuring seccomp profiles, and applying granular permissions using Linux kernel capabilities.

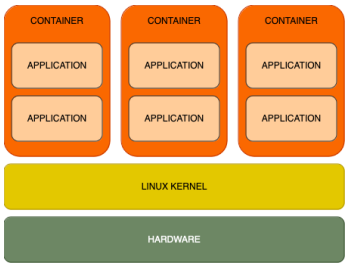


Fig. 1: Containerised application.

We will also discuss the possibility of using TEEs [1], assessing their suitability for HPC design, understanding their purpose [3,4], how to enable batch job execution within a secure enclave, and how to further enhance data-in-use security on CPU-GPU systems. Maintaining a secure multi-tenant environment also requires compliance with security standards, active security monitoring and incident response.

References

[1] Akram, A. et al., SoK: Limitations of Confidential Computing via TEEs for High-Performance Compute Systems, IEEE International Symposium on Secure and Private Execution Environement Design, (2022).

[2] Bertani, A. et al., Confidential Computing: A Security Overview and Future Research, Proceedings of the 8th Italian Conference on Cyber Security, ITASEC (2024).

[3] Goepireddy Reddy, Gopireddy: Confidential Computing: The Key to Secure Data Collaboration in the Cloud, Journal of Scientific and Engineering research, 10(6):271-276 (2023).

[4] Wang, Q., Oswald, D.: Confidential Computing on Heterogeneous CPU-GPU Systems: Survey and Future Directions, arXiv/2408.11601 (2024).

QUBO and quantum annealing: applications and perspectives

Mátyás Koniorczyk[a], Krzysztof Domino[b], and Péter Naszvadi[a,c]

[a] *HUN-REN Wigner Research Centre for Physics, Budapest, Hungary*
[b] *Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, Poland*
[c] *Faculty of Informatics, Loránd Eötvös University, Budapest, Hungary*

We examine various aspects of solving quadratic binary unconstrained optimization (QUBO, aka MAX-CUT or Ising) problems using quantum annealing, quantum-classical hybrid methods, and physics-based heuristics. QUBO solvers have a relevant potential in solving a plethora of discrete optimization problems, hence, they attract a lot of research interest. This is further boosted by the connection with quantum computing, and Ising models which introduces a perspective of physics in this optimization problem. Meanwhile noisy intermediate quantum (NISQ) devices are available already, and are close to become useful, hence one can expect that they will find their place in HPC centres soon. Hence, the question arises how they will be used and what will they be good for.

In one of our previous works [1], we aimed at solving a very practical scheduling problem originating from railway operations research. We addressed this using quantum hardware only, which was instructive because it revealed a number of practical considerations of using a quantum annealer as well as a number of alternative new, physics-motivated heuristics which we will discuss. One of the conclusions is that using purely NISQ quantum annealers limits the problem size. Hence there can exists small yet hard problem instances, while hybrid quantum-classical algorithms offer better practical applicability at the moment.

In the search for small but hard QUBO problems we made an excursion to code theory: the question of Hamming packings in mixed spaces [2]. We have found a way to decompose nontrivial problems of this class into the range where NISQ annealers can be already useful. We present promising preliminary results of this kind.

As for the other research direction, the use of hybrid solvers, our recent work [3] fits to the popular trend of solving practical problems with proprietary closed-source solvers. Through our experience we discuss the practical benefits, and the research methodological limitations of this approach. These practical benefits will be extended, and the methodological limitations will be eliminated by the possible development of more open and possibly special hybrid quantum-classical solvers. In the mean time, both [1] and [3] illustrate the promising perspective of physics-motivated heuristics very suitable for HPC environments; we give a brief overview of these.

References

[1] Domino, K., Koniorczyk, M., Krawiec, K., Jałowiecki, K., Deffner, S., Gardas, B. Quantum Annealing in the NISQ Era: Railway Conflict Management. Entropy 2023, 25, 191

[2] Naszvadi P., Koniorczyk M., https://arxiv.org/abs/2310.01883 (2023)

[3] Śmierzchalski, T., Pawłowski, J., Przybysz A., Pawela Ł., Puchała Z., Koniorczyk M., Gardas B., Deffner S., Domino K., Sci Rep **14**, 21809 (2024).

Hybrid Classical-Quantum Exact Solver for the QUBO Problem

Omkar Bihani[a], Aljaž Krpan[a,b], Roman Kužel[a], and Janez Povh[a,c]

[a] *Rudolfovo - Science and Technology Centre Novo mesto, Slovenia.*
[b] *Faculty of Mathematics and Physics, University of Ljubljana, Slovenia*
[c] *Faculty of Mechanical Engineering, University of Ljubljana,, Slovenia*

The Quadratic Unconstrained Binary Optimization(QUBO) problem is a fundamental NP-hard optimization problem with significant applications in fields such as network design, machine learning, and statistical physics. Exact solutions to QUBO often require substantial computational effort, making it a suitable candidate for hybrid computing approaches. In this study, we extend the capabilities of Biqbin [1], a high-performance computing (HPC) solver for exact solutions, by incorporating quantum computing for lower bound computation. Specifically, we replace the default Goemans-Williamson (GW) heuristic used in Biqbin with the D-Wave quantum solver, creating a hybrid branch-and-bound (B&B) framework.

We solve the MaxCut problems on smaller graph instances by converting them to QUBO. This approach was implemented and tested in both serial and parallel versions of Biqbin on smaller graph instances, demonstrating its feasibility in an HPC environment. We assess the impact of using a quantum solver for lower bounds in comparison to the classical GW heuristic. While the results highlight the robustness and efficiency of the GW heuristic for the tested instances, they also provide insights into the integration of quantum solvers within classical optimization frameworks.

References

[1] Gusmeroli, N., Hrga, T., Lužar, B., Povh, J., Siebenhofer, M., & Wiegele, A. ACM Trans. Math. Softw. **48**, 15 (2022).

Bootcamp Performanceoriented Softwarengineering — Experiences from working with real-world problems and developers

Philipp Gschwandtner

Research Center HPC and Department of Computer Science, University of Innsbruck, Austria

Highly efficient and parallel software is mandatory today, whether in industry or in academia, due to the increase in parallelism in modern hardware and ever-increasing problem sizes and complexities. Practical experience shows that a lack of parallelism and software optimizations are usually compensated for by increased investments in hardware and thus by overprovisioning, which may not only lead to direct economic cost disadvantages, but often also to further losses such as increased energy consumption or low efficiency. However, the development of fast and efficient software is often tedious and requires a deep understanding of programming models, performance, and hardware. In this talk we present our experience with the FFG-funded Bootcamp Performanceoriented Softwarengineering, a qualification measure in which 17 employees from 10 companies were trained to become "digital professionals" in topics including parallel programming, HPC, accelerator computing, code quality, and productivity. These digital professionals deal comprehensively with in-house IT projects and were trained to improve their understanding of the interaction between software and hardware and to increase the productivity of both developers and their software.

We will provide insights into the current state of the industry with regard to parallel programming and HPC use cases, along with selected examples of performance improvements we were able to achieve within just a few weeks of training and development. One example use case originates from the processing of 3D model data in the area of cabinet making and furniture industry. The software was already written in C++ for performance reasons, but lacked any parallelization due to the use of non-thread-safe third party libraries to process proprietary data containers. We analyzed the requirements of this use case and identified parallelization potential using OpenMP tasks along with careful pipelining and synchronization in order to ensure thread-safe interaction with the third party library. Preliminary results showed a speedup of 4x to 5x depending on the input data.

Another use case originates from material science, simulating the generation and propagation of x-rays in computer tomography. The software was written in Python for execution on CPUs and while sensible math packages were used, there was little further optimization in the code. We analyzed the requirements of the use case, ported the Python code to make use of GPUs through numba and CUDA and identified several software components which are used for debugging only and hence could be removed from production runs. A benchmark run originally taking approximately 250 seconds on the reference CPU was reduced to just 45 ms of execution time on a GPU, resulting in a speedup of approximately 5000x to 6000x (depending on the input data).

References

[1] `https://www.uibk.ac.at/de/weiterbildung/gesundheit-mint/pos-tirol/` (German).

MUSICA: Current situation and developments

Florian Goldenberg[a], Elias Wimmer[a], Markus Hickel[a], and Martin Thaler[b]

[a] *VSC Research Center, TU Wien, Austria*
[b] *ZID, University of Innsbruck, Austria*

In the MUSICA project (MUlti-SIte Computer Austria), funded by FFG [1] within the Quantum Austria [2] framework and additionally funded by the Austrian Ministry of Education, Science and Research (BMBWF) [3], we got the opportunity to build a truly distributed HPC platform for science and research. In our presentation, we will report on the ongoing progress implementing the systems and the various challenges encountered so far.

Two parts of the system (Vienna and Innsbruck), as well as all components of the management systems, were delivered to the Arsenal server rooms of the VSC Research Center during Summer 2024. Due to the extreme high demands of the compute nodes in both power and cooling, the needed infrastructure was not finished at that time.

Until the completion of the required construction work, the management systems were installed and configured. The new control and provisioning method will deviate from our previous clusters, utilising OpenStack to create the various needed services, as well as to provision the actual bare-metal compute nodes. This was set up with the help of StackHPC, a consulting firm specialised in such applications. This was possible with all three parts (Vienna, Innsbruck, and Linz).

For the Vienna part, the Weka storage system could also be installed in the server room and put into operation. In tests with older adapted test nodes, it performed very well, but the intended final configuration could not be tested yet.

The power supply (1MW and 1.6kA) and the cooling system (hot water cooling with free cooling units) were finished in late Autumn, which enabled the installation of the compute nodes. The nodes were provisioned using the new system and then tested for stability and functionality. Approval tests and performance benchmarks were done in the end of 2024/beginning of 2025, including HPL. Test operation with selected friendly users is expected to start in spring 2025.

The Innsbruck part will be de-installed and transferred to Innsbruck as soon as the datacenter infrastructure is operational, while the Linz part will be delivered directly to Linz as soon as the rooms there are fully functional.

An open issue is still the data transfer between the three sites, which we plan to solve via the Globus utility running on dedicated nodes.

References

[1] `https://projekte.ffg.at/projekt/4496688`

[2] Quantum Austria is financed by the EU Recovery and Resilience Facility `https://commission.europa.eu/business-economy-euro/economic-recovery/recovery-and-resilience-facility_en`

[3] `https://www.bmbwf.gv.at`

How to play MUSICA

Jan Zabloudil[a,b]

[a] *BOKU University*
[b] *VSC Research Center, TU Wien, Austria*

The Vienna Scientific Cluster's newest high-performance computing (HPC) system, **MUSICA**, became operational at the end of last year, offering state-of-the-art computational resources to researchers across a range of disciplines. The presentation is designed to introduce users to the hardware architecture of the system and provide guidance on how to achieve optimal performance for their computational workloads.

MUSICA's hardware configuration is tailored to support both traditional HPC applications and cutting-edge workloads such as artificial intelligence and machine learning. A key focus will be the architecture of MUSICA's GPU nodes, which are equipped with: **Four NVIDIA H100 GPUs with SXM5 technology**, providing ultra-high inter-GPU bandwidth through NVLink, enabling efficient data sharing and rapid computation. This architecture is particularly well-suited for large-scale simulations, deep learning, and other highly parallelized tasks. **Two AMD EPYC 9654 CPUs**, each offering 96 cores, delivering excellent performance for CPU-intensive tasks and ensuring a balanced architecture for hybrid CPU-GPU workloads. The CPUs also provide exceptional memory bandwidth and connectivity to maximize the efficiency of the GPUs.

Complementing these compute resources is the **NDR InfiniBand fabric**, a high-performance interconnect that ensures low-latency, high-bandwidth communication between nodes. This is critical for applications that require frequent and large-scale data exchange in distributed environments. The presentation will also cover strategies for achieving **optimal performance** on MUSICA, with an emphasis on the importance of **process placement** and resource management. Attendees will learn how to map processes and threads to the underlying hardware to reduce data movement overheads, fully exploit the SXM5 interconnect, and take advantage of the AMD CPUs' capabilities. Guidance will also be provided on how to effectively utilize the NDR InfiniBand fabric for multi-node workloads to ensure scalable and efficient performance. By understanding the architecture and employing best practices, researchers can maximize their use of MUSICA's computational power, enabling them to address larger and more complex scientific challenges. This presentation is designed for both experienced HPC users and newcomers, providing actionable insights to help unlock the full potential of MUSICA.

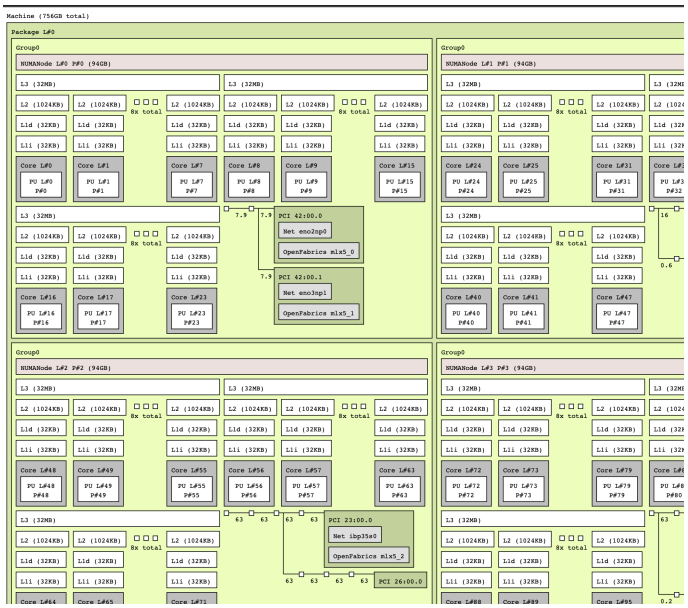


Fig. 1: NUMA node architecture of a MUSICA GPU node

When space runs out: A new central storage system for VSC

Markus Hickel and Florian Goldenberg

VSC Research Center, TU Wien, Austria

Due to the rising demand from data science applications and machine learning workloads, coupled with the ever-growing volume of data being processed and the increasing number of users on our systems, the current storage infrastructure is no longer adequate to meet operational needs. Modern data-driven workflows require storage solutions that can handle large-scale datasets with high performance, reliability, and scalability.

Part of the solution is the already existing Weka filesystem attached to VSC-5 and MUSICA. This 4PB pure NVMe based storage system is intended as extremely fast scratch space to handle demanding I/O workloads requiring high bandwidth and low latency. Its GPU-direct capabilities also makes it ideal for ML/AI workloads.

The main central storage will be replaced by a larger system which is currently being procured. The required capacity is around 40PB, consisting of 10% NVMe drives, the rest being conventional spinning hard-disks. The NVMe pool is intended as an automated fast tier for new or frequently used data, while the HDD pool serves as the slower primary data location.High availability and reliable performance is necessary for the main data storage. All components are redundant and replaceable without service interruption.

Various HPC optimized parallel storage solutions have been considered. These were the Lustre based DDN; ClusterStore from HPE, another Lustre based solution; VDURA, formerly known as panasas; Quobyte, a rather new system from Germany; and a GPFS solution based on the IBM ESS appliance. The system was ordered beginning of April and the decision was made for the GPFS system, which will be provided by EDV-Design. It will consist of 4 ESS units, each extended with several JBODs.

Both systems, Weka and GPFS, will be integrated into the respective high-speed fabric of each cluster, which is HDR Infiniband for VSC-5 and NDR Infiniband for MUSICA, with the possibility to include further upcoming fabrics. The central storage system will be designed to support a variety of access protocols, going beyond traditional POSIX file system functionality. This includes compatibility with S3 for object-based storage, NFS for network file sharing and CIFS.

We will present the structure, setup, and functionality of the VSC Research Center's cluster storage solutions, with a particular focus on the newly installed main data storage system.

CINECA infrastructure, from AI factory to quantum computers

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, Artificial Intelligence (AI) and Quantum Computing (QC). The Tier-0 HPC system **Leonardo** is hosted at the Bologna Technopole in Italy and it is supplied by Eviden. 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 customised Nvidia Ampere A100 GPUs. Ranked 9th on the Top500 list, it delivers 240 petaflops of performance with a total of 4992 compute nodes and more than 100 PB of storage, connected by an InfiniBand HDR fabric.

By 2025, the **LISA** project will enhance Leonardo's capabilities, adding 2.5 exaflops of FP8 performance optimized for AI and machine learning. LISA will also introduce 8-way GPU nodes for greater acceleration in AI workloads. Its non-blocking NDR InfiniBand fabric will be used to improve performance in training large language models (LLMs) by ensuring seamless and efficient data transfer, which is crucial for scaling these models. This will enable faster, more efficient training of LLMs and other AI applications. Looking to the future, the **IT4LIA AI Factory**, arriving in 2026, will provide a system four times more powerful than Leonardo. IT4LIA will accelerate research and industrial innovation in areas like cybersecurity, agritech, and earth sciences. With Leonardo, LISA and IT4LIA, CINECA is building a world-class AI ecosystem, supported by the EuroHPC JU and by National funds, which will have a transformative impact on AI research, industrial applications, and real-world problem-solving.

During 2025, CINECA will also welcome **two quantum computers**, with distinct technologies: a Neutral Atoms Analog Quantum Simulator, with 140 qubits, and a Superconducting Digital Quantum Computer, with 54 qubits. The two quantum computers, which see again a collaboration between EuroHPC JU and National funds, will be available to users through the integration with the classic supercomputer Leonardo.

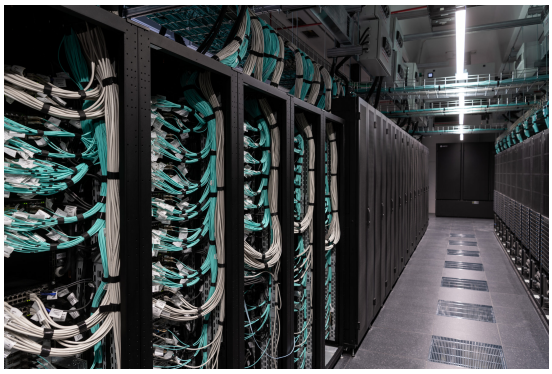


Fig. 1: The supercomputer Leonardo. [1]

Together with Leonardo and LISA, the AI factory services and the Quantum Computers will be available to Italian and European users, with the aim to support the academic and industrial research. I will present an overview of the present and upcoming CINECA infrastructure, showing its potentialities and how to access and exploit the computing and storage resources.

References

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

AI Factory Austria AI:AT

Markus Stöhr[a] and Claudia Blaas-Schenner[b]

[a] *Advanced Computing Austria ACA GmbH and BOKU University, Austria*
[b] *VSC Research Center, TU Wien, Austria*

Funded by the EuroHPC Joint Undertaking and the Federal Ministry for Innovation, Mobility and Infrastructure of the Republic of Austria, the AI Factory Austria (AI:AT) will start its operation in summer 2025. [https://eurohpc-ju.europa.eu/ai-factories]

The **AI Factory Austria (AI:AT)** is a new, large-scale initiative designed to elevate Austria's artificial intelligence (AI) capabilities and accelerate the development and adoption of trustworthy AI solutions in Austria's major industry sectors. Coordinated by Advanced Computing Austria GmbH (ACA) and the AIT Austrian Institute of Technology, it brings together universities, research institutes, and industry partners to establish an AI-optimized supercomputer and a comprehensive AI Factory Hub. As an innovation center and one-stop-shop for AI, AI:AT provides access to cutting-edge supercomputing resources, expert guidance, and collaborative spaces for businesses, researchers, government organizations and innovators. By combining state-of-the-art infrastructure with integrated services, AI:AT accelerates innovation, enhances AI accessibility for businesses of all sizes, and strengthens Austria's position in the European and global AI landscapes. By leveraging Austria's strong research and industrial base, AI:AT supports the national fields of excellence in biotechnology, manufacturing, physics, applied research, and digital governance.

Core Infrastructure and Integration: Based on Austria's strong foundation in High-Performance Computing (HPC), particularly the Vienna Scientific Cluster (VSC), the AI Factory's new supercomputer integrates seamlessly with the Multi-Site Computer Austria (MUSICA) system. Equipped with advanced GPUs, direct water cooling, and high-speed interconnects, it delivers scalable, high-performance computing for AI-intensive applications across diverse fields, from life sciences and manufacturing to materials science. By providing cost-effective, high-capacity resources on shared platforms, the AI Factory enables researchers, SMEs, public organizations and large enterprises to collaborate without capacity constraints, fostering innovation and accelerating AI-driven breakthroughs.

AI Hub and Key Services: The AI Factory Hub is going to be built in Vienna and will serve as a dynamic innovation center, offering a start-up accelerator, operational support, access to trustworthy data, expert consulting for AI-driven solutions, and extensive training programs. With a comprehensive service model spanning AI solution engineering, proof-of-concept development, and regulatory guidance, it fosters cutting-edge advancements while upholding ethical data practices and trustworthy AI principles. A strong emphasis is placed on training and capacity building through hands-on workshops, online courses, internships, and industrial PhD placements — ensuring a continuous pipeline of skilled talent to strengthen Austria's and Europe's AI ecosystems.

Wider Impact and Future Outlook: With a broad range of applications in industry, scientific research, and public administration, the AI Factory is set to drive breakthroughs in, e.g., biotechnology, sustainability, energy and production efficiency, and more effective data-driven administrative processes. By providing an easy to access open one-stop-shop of AI resources, it facilitates the transformation of scientific discoveries and trustworthy AI model developments into real-world solutions. Ultimately, the AI Factory Austria envisions a nationwide AI transformation as a driving force of the transformation in Europe — powered by cutting-edge infrastructure, expert services, and strong partnerships across academia, industry, and government.

SLAIF: Slovenian AI Factory (with a new EuroHPC AI optimized system)

Sašo Džeroski[a,b], Jan Jona Javoršek[a], and Andrej Filipčič[a,c]

[a] *Jožef Stefan Institute, Ljubljana, Slovenia*
[b] *Jožef Stefan Postgraduate School, Ljubljana, Slovenia*
[b] *University in Nova Gorizia, Slovenia*

We present the Slovenian AI Factory (SLAIF) [1], a project enhancing Slovenia's AI capabilities and providing new integrated cutting-edge AI-optimized supercomputer in a brand-new data centre [2].

The system will deliver 10 EFlops in mixed precision for AI applications and 100 PFlops in double precision for HPC tasks. Powered directly by renewable hydroelectric energy, its environmental impact will be reduced. A hybrid HPC-cloud infrastructure will help industrial users with integration of AI capabilities.

The project will support the development of a comprehensive AI eco-system that brings together Slovenia's expertise in AI and HPC, currently spread across many institutions, into a coordinated hub. Bridging the gap between academia and industry, the project will help us provide an AI infrastructure, expertise, and data as a one-stop-shop for businesses, startups, and researchers.

SLAIF will drive AI adoption and innovation across multiple sectors: *the green transition*, leveraging AI to optimize energy systems, environmental monitoring, and smart agriculture, improving sustainability efforts across a range of industries; in *healthcare and biotechnology*, supporting analyses of multi-modal medical data, facilitating personalized medicine, supporting pharmaceutical industry with AI-driven drug discovery tools, accelerating the development of new treatments; *in language technologies* for digital services, media and creative industries, as well as *AI for science*, facilitating, e.g., the design of new materials and building AI models on scientific simulations and analyses. SLAIF will facilitate secure and open access to *AI-relevant datasets* and integrate with European data spaces and data labs.

SLAIF will provide a comprehensive *support framework*: sector-specific AI-development and deployment support, access to pre-trained AI models, data, and tailored cloud-based AI services, real-time AI analytics tools, and domain-specific AI training programs, ensuring smooth integration of AI solutions. On-campus facilities and services, incl. co-working spaces, will further support businesses and researchers. Access to industry-specific case studies, AI tools, and cloud-based resources will reduce entry barriers for businesses of all sizes. SLAIF will support *developing AI talent* with specialized AI training programs and hands-on AI education for students, researchers, and professionals, working with academia and industry leaders.

By integrating world-class AI computing power, industry collaboration, talent development, and data accessibility, SLAIF will offer key support in regional AI infrastructure and integrate with the wider EuroHPC ecosystem. It will boost Slovenia's economic and scientific competitiveness and ensure AI serves the needs of businesses, researchers, and society as a whole, contributing to Europe's broader goal of AI sovereignty and sustainable digital transformation.

References

[1] Džeroski, S., Filipčič, A.. SLAIF. `https://www.slaif.si/`

[2] Arnes. The First ARNES Data Centre of the Future. `https://www.arnes.si/en/the-first-arnes-\data-centre-of-the-future-will-be-located-in-maribor/`

Keynote Talk:

When data are big in the "wrong" direction: identifying compact and informative distance measures in high-dimensional feature spaces

Alessandro Laio[a,b]

[a] *Physics Section, Scuola Internazionale Superiore di Studi Avanzati, Trieste 34136, Italy*
[b] *Condensed Matter and Statistical Physics Section, International Centre for Theoretical Physics, Trieste 34151, Italy*

Real-world data typically contain a large number of features that are often heterogeneous in nature, relevance, and also units of measure. When assessing the similarity between data points, one can build various distance measures using subsets of these features. Finding a small set of features that still retains sufficient information about the dataset is important for the successful application of many machine learning approaches. We introduce an approach that can assess the relative information retained when using two different distance measures, and determine if they are equivalent, independent, or if one is more informative than the other. This test can be used to identify the most informative distance measure out of a pool of candidate. We will discuss applications of this approach to feature selection in molecular modeling, to the analysis of the representations of deep neural networks, and to infer causality in high-dimensional dynamic processes and time series.

AI-zyme: machine learning enhanced protein desing

Žiga Zebec[a], Samo Miklavc[a], and Teo Prica[a]

[a] *IZUM - Institute of Information Science, Slovenia*

One of the most important achievements of humanity is the invention of synthetic, man-made materials with incredible chemical and mechanical characteristics. At the same time, these materials pose a significant risk to our environment in the form of waste. The standard way to recycle these materials is by chemical or mechanical treatment. However, some cutting-edge recycling methods involve enzymes that are highly specific to their substrates. Because most of these materials are man-made, suitable enzymes for their degradation can rarely be found in nature, making enzyme engineering inefficient. To overcome this issue, *de novo* enzymes are generated by means of synthetic biology.

Here we introduce a novel pipeline for *de novo* enzyme design, integrating cutting-edge generative models and structure prediction tools. We utilize the power of diffusion models to generate *de novo* protein structures with or without predefined sequences, using **diffusion models**. Next, we employ a **deep learning–based** protein sequence design method to generate sequences for the structure made in the previous step by the diffusion model. To validate the generated sequences of the *de novo* protein structures, **AlphaFold** is utilized to compare the structure generated from the sequence alone and the initial structure of the *de novo* structure generated by diffusion models. The final comparison is made by superimposing the predicted *de novo* structure and structure generated from sequence by **AlphaFold** (**Chimera-X** or similar tools).

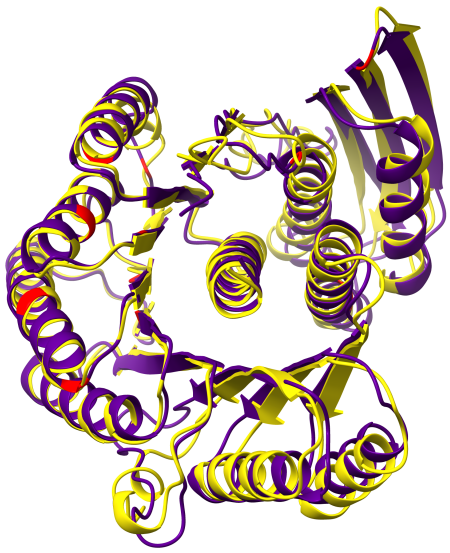


Fig. 1: Superimposed enzyme 9

These tools will be integrated into a pipeline (AI-zyme) using Input/Output (I/O) automatically to generate faster results and provide a better user experience for nonexperts in computer science. This approach provides a powerful methodology for accelerating the discovery and design of new enzymes with wide-ranging applications in biotechnology, medicine, industry, and beyond.

References

[1] Zebec, Ž., Poberžnik, M., and Lobnik, A. Enzymatic Hydrolysis of Textile and Cardboard Waste as a Glucose Source for the Production of Limonene in Escherichia coli. Life, **12(9)**, 1423 (2022)

Cell simulation: the ultimate way to develop better drugs

Draško Tomić

Institut Rudjer Bošković, Croatia

All-atom cell simulation promise to deliver the most effective drugs against many complex diseases such as cancer, neurodegenerative diseases, cardiovascular diseases and infectious diseases. However, this is a computationally extremely demanding task that cannot be accomplished even on the most powerful supercomputers available today [1]. The approximate modeling of cells in the preclinical phase of drug development can be done in different ways, e.g. by coarse-grained simulation [2] or the use of KEGG metabolic pathways [3]. However, details that are crucial for the development of a new cancer drug may be lost. This is one of the main reasons why clinical cancer trials end up ineffective in 90% of the cases. However, with the continuous optimization and increase in scalability of molecular dynamics simulators such as NAMD, Amber, and Gromacs, with the advent of ever more powerful HPC clusters equipped with ultra-fast GPU accelerators, sophisticated quantum algorithms and quantum computers with more and more qubits, we are well on our way to getting closer to this goal. This talk will give a brief overview of the successes achieved so far in simulating large molecular structures and report on the results we have obtained on this topic on the EuroHPC JU supercomputers Vega, LUMI, JURECA and JEDI. It also highlights the main obstacles that currently prevent the simulation of extremely large molecular structures and identifies the components that can benefit most from the introduction of advanced quantum algorithms and quantum coprocessors.

References

[1] Samuel Russell PP, Alaeen S, Pogorelov TV. In-Cell Dynamics: The Next Focus of All-Atom Simulations. J Phys Chem B. 2023 Nov 23;127(46):9863-9872. doi: 10.1021/acs.jpcb.3c05166.

[2] Pivkin, I. V., & Karniadakis, G. E. (2008). Accurate coarse-grained modeling of red blood cells. Physical review letters, 101(11), 118105. https://doi.org/10.1103/PhysRevLett.101.118105

[3] Tomic D, Murgic J, Frobe A, Skala K., Vrljicak A, Medved RB, Kolarek B. et all, V. (2024). Exploring potential therapeutic combinations for castration-sensitive prostate cancer using supercomputers: a proof of concept study. Scientific reports, 14(1), 18824. https://doi.org/10.1038/s41598-024-69880-9

H_2O adsorption at Co_3O_4 (111) surface from a DFT perspective

Alexander Genest, Thomas Haunold, and Günther Rupprechter

Institute of Materials Chemistry, Technische Universität Wien, Getreidemarkt 9/BC, 1060 Vienna, Austria

Co_3O_4 was identified as a versatile catalyst material that allows easy exhaust cleaning without the need for elevated temperatures. However, water molecules, constantly present in exhaust gases, have been determined to hamper the smooth progress of CO oxidation, a major flaw in its functionality[1]. Using Density Functional calculations in a plane wave setting using VASP with PBE including a treatment for vdW forces, we will demonstrate how H_2O molecules adsorb at a model (111) surface of Co_3O_4. Crucial is the dissociation of water to form surface hydroxyl groups using lattice oxygen. This formation of surface hydroxyl species is favorable for up to the conversion of four H_2O molecules, which in turn benefits the adsorption of even more water species. Near ambient pressure XPS experiments traced the fate of a Co_3O_4(111) film under a 5 mbar pressure of water[2].

To compare those XPS experiments, we align the determined shifts with computed ones for model structures. Using representative oxygen centers to compare to experiments was not supportive of the interpretation of the experiments. Using machine-learned force fields adapted to the system water on Co_3O_4(111) was used to produce more typical structures that might resemble the geometry in the experiment. These force fields cut down typical calculation times of DFT computations on CPUs with 32 cores from 1-3 hours to several seconds, such that MD simulations equilibrating the water structure became accessible. For those structures, we calculated XPS shifts for all oxygen centers to get a range of typical shifts.

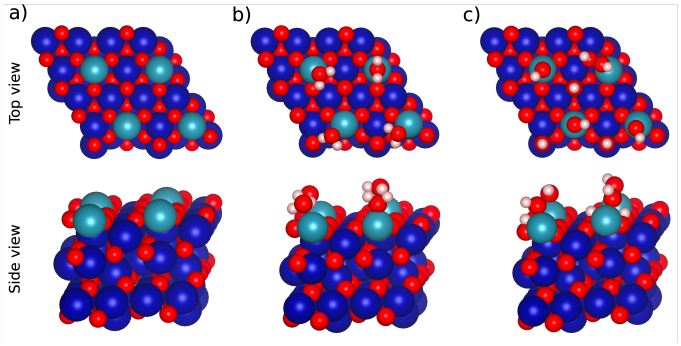


Fig. 1: Adsorption of Water on Co_3O_4(111).

References

[1] Yigit, N.; Genest, A.; Terloev, S.; Möller, J.; Rupprechter, G., J Phys Condens Matter **34** 354001 (2022).

[2] Haunold, T.; Anić, K.; Genest, A.; Rameshan, C.; Roiaz, M.; Li, H.; Wicht, T.; Knudsen, J.; Rupprechter, G., Surf Sci **751** 122618 (2024).

Open-boundary molecular dynamics of ultrasound using supramolecular water models

Maša Lah[a,b], **Nikolaos Ntarakas**[a,b], **Tilen Potisk**[a,b], **Petra Papež**[a], **and Matej Praprotnik**[a,b]

[a] *Laboratory for Molecular Modeling, National Institute of Chemistry, Slovenia*
[b] *Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, Slovenia*

Computational studies have the potential to optimize ultrasound use in diagnostics and therapy, for example by providing optimal frequencies to enhance contrast agent signals or to modulate protein function. Depending on the system size and required detail, sound waves can be modeled using methods ranging from continuum methods to molecular dynamics (MD). Coarse-grained particle-based methods such as the dissipative particle dynamics (DPD), where a single bead represents multiple molecules, can efficiently simulate ultrasound waves at the mesoscale.

Studying sound wave propagation requires careful implementation of boundary conditions. Here, we use open-boundary molecular dynamics (OBMD) to simulate ultrasound in an open system. In open systems, waves dissipate as they propagate, unlike in periodic boxes, where propagation across periodic boundaries can distort wave dynamics. In OBMD, the outer layers of the simulation box act as particle reservoirs, allowing the particles to diffuse in and out of the system freely. The equations of motion for the particles in the bulk remain unchanged, whereas external forces are applied to the outer layers to impose boundary conditions, such as introducing sound waves through a time-dependent sinusoidal pressure perturbation.

We study the acoustic properties of liquid water—speed of sound and ultrasound attenuation—using Martini 3, DPD, and many-body DPD models. This serves as a verification of the OBMD implementation in Mirheo, a high-throughput DPD simulation package.

All models reproduce the damped traveling wave equation in response to a sinusoidal pressure perturbation (Fig. 1), with the speed of sound close to the experimental value. Furthermore, the results of our simulations fully agree with the predictions of attenuation in water and accurately capture both the dispersion and the frequency dependence of the attenuation coefficient.

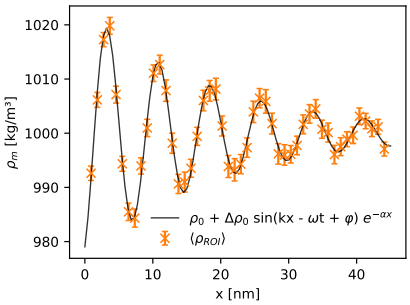


Fig. 1: Density variation of an ultrasound wave throughout the simulation box for a DPD water model.

This method represents an important step toward developing a general virtual ultrasound machine. Despite significant effort, simulating sound of arbitrary frequency remains challenging, as it requires simultaneously resolving both the fast and slow processes that govern its propagation. Such methodology opens up exciting new opportunities for research in biomedical engineering, materials science, and soft matter physics.

References

[1] Lah M., Ntarakas N., Potisk T., Papež P., and Praprotnik M., J. Chem. Phys. **162**, 024103 (2025).

Enhancing AI Models in Local Language for IT Service Management

András Dorn (CEO)

DAM Invisible Technology Plc., 1118 Hungary, Budapest, Budaörsi út 64.

Enhancing automated ticket categorization accuracy in IT service management remains a significant challenge, particularly in language-specific and domain-specific contexts like Hungarian. Existing transformer-based models often struggle with nuanced classification in specialized IT environments, necessitating the development of custom AI-driven solutions.

To address this, domain-specific BERT-based classifiers were developed and trained on the CulturaX dataset using curriculum learning techniques. Initial experiments with publicly available models (e.g., hubert-base-cc and NYTK\PULI-BERT-large) revealed limitations in context comprehension and classification accuracy. To overcome these constraints, proprietary models, IRIS-BERT-base and IRIS-BERT-large, were created, optimized for extended token contexts, and trained using the Hungarian HPC CC Komondor supercomputer.

Experimental results demonstrate significant improvements in classification accuracy and F1 Macro scores compared to traditional statistical approaches and off-the-shelf transformer models. Ongoing research focuses on refining IRIS-BERT-large to further optimize performance. These findings contribute to the field of AI-driven IT service management by showcasing the impact of custom language models on operational efficiency and decision-making accuracy.

The model was designed and developed by DAM Invisible Technology Plc. for the IR:IS system, with computational support from the Hungarian HPC CC Komondor.

Exploring the origins of enzyme catalysis by simulation methods

Jernej Stare, Janez Mavri, Alja Prah, Martina Rajić, Aleš Novotný, and Andrzej J. Kałka

National Institute of Chemistry, Ljubljana, Slovenia

The paramount function of enzymes is their ability to catalyze biomolecular reactions. In comparison to plain aqueous environment, enzymes provide an enormous speedup of a wide variety of reactions involved in life processes. Enzymes are therefore essential for the existence of life in its present form – and yet, no consensus has been achieved so far on the very fundamental question, namely, what is the physical driving force behind enzyme catalysis.

Our group has been investigating the hypothesis that enzyme electrostatics is primarily governed by electrostatic interactions established between the large array of charged/polar moieties existing in an enzyme's scaffold, and its active site, in that electrostatic stabilization is substantially larger in the transition state than in the reactant state, thereby lowering the activation barrier and boosting the kinetics. By using our own technique based on embedding quantum-chemical calculations of a reacting moiety in an environment consisting of point charges [1], we demonstrated a massive barrier lowering for a variety of considered enzymes and their reactions, suggesting that the hypothesis of electrostatics as source of catalysis is valid.

In addition, by using the same methodology, we elucidated rather subtle differences in enzyme catalytic performance caused by genetically driven enzyme point mutations [2]. We confirmed significant increase in activation barrier for several mutated variants of monoamine oxidase A enzyme, which have been known as pathogenic and associated with severe behavioral and mental disorders known as Brunner syndrome. In this way, we provide links between clinical genetics and the underlying physical background and advance the understanding of genetic diseases on a molecular level.

Our electrostatic embedding technique uses the established program package *Gaussian*, requiring atomic coordinates of the reacting moiety together with coordinates and values of point charges representing the enzymatic surroundings. Because hundreds of representative structures are needed to properly account for thermal fluctuations, and because analysis of contribution of individual residues to the catalytic effect can also include a comparable number of residues, a comprehensive treatment of an enzymatic reacion may require over 100,000 individual (independent) quantum chemistry omputations, each involving 32-64 CPU cores, and each taking about 1-5 minutes. This represents a challenging task in terms of CPU resources. For the sake of economy we have been able to either reduce the number of individual tasks down to 10,000 calculations per studied case, or use less demanding but also less accurate semiempirical quantum chemistry approaches. We pursue most of our simulations at the local HPC center at the National institute of Chemistry consisting of approximately 100 nodes and 5,000 CPU cores, but in the future we envisage a more ambitious treatment, which possibly includes the use of massive HPC infrastructure within the SLING framework such as VEGA.

References

[1] A. Prah, E. Frančišković, J. Mavri, J. Stare, *ACS Catal.* **2019**, *9*, 1231.

[2] A. Prah, D. Pregeljc, J. Stare, J. Mavri, *SCI. Rep.* **2022**, *12*, 21889.

Efficient theoretically guided search for functional metallic thermoelectrics.

Sergii Khmelevskyi[a]

VSC Research Center, TU Wien, Operngasse 11, Vienna, Austria

Thermoelectric materials capable of converting waste heat into electricity hold immense potential for addressing global energy challenges and enabling sustainable energy technologies. While traditional research has focused on semiconductors, recent advances reveal that metallic systems can achieve competitive thermoelectric performance [1], offering advantages such as mechanical robustness and high power factors.

Central to our approach is the use of advanced ab initio methods, including the Coherent Potential Approximation (CPA) compositional screening, to predict and optimize the electronic and thermal transport properties of metallic alloys. By targeting interband scattering and density of states (DOS) engineering, we have demonstrated unprecedented thermoelectric properties in Ni-Au alloys [2], achieving power factors exceeding 30 mW/K^2m and a record-high zT of 0.5 for metals. These findings challenge the conventional view of metals as poor thermoelectric candidates. We have explored the thermoelectric behavior of Fe_2VAl semimetallic Heusler alloys, where substitutional Fe disorder leads to Anderson localization transitions with profound impacts on thermoelectric properties. The CPA calculations revealed how tuning disorder in stoichiometric Fe_2VAl could optimize its electronic structure for enhanced performance. This work underscores the role of substitutional disorder and band structure modifications in advancing thermoelectric materials.

The utilization of the Coherent Potential Approximation (CPA) offers several distinct advantages over supercell calculations, particularly when applied to large-scale searches for novel thermoelectric alloys. These benefits are rooted in CPA's efficiency, accuracy, and suitability for modeling disordered systems, which are common in alloy-based thermoelectrics. Standard supercell calculations require constructing large unit cells to approximate disorder, introducing computational overhead and finite-size effects that may not fully capture the true nature of randomness in the alloys. CPA's computational efficiency and ability to generate high-quality data make it an excellent candidate for integration with machine learning algorithms. These models can be trained on CPA-generated data to predict thermoelectric properties across vast compositional spaces.

We will discuss ongoing efforts and prospective to extend these principles to ternary and quaternary alloy systems, incorporating elements from the 3rd to 5th periods of the periodic table. Leveraging automation and artificial intelligence, one can potentially navigate vast compositional spaces, identifying promising candidates with sharp DOS features and enhanced thermoelectric responses.

References

[1] F. Garmroudi, M. Parzer, A. Riss, A. V. Ruban, S. Khmelevskyi, M. Reticcioli, M. Knopf, H. Michor, A. Pustogow, T. Mori, and E. Bauer. Nature Communications **13**, 3599 (2022).

[2] F Garmroudi, M Parzer, A Riss, C Bourgès, S Khmelevskyi, T Mori, E. Bauer, A. Pustogow. Science Advances **9**, eadj1611 (2023).

[3] R. Jha, N. Tsujii, F. Garmroudi, S. Khmelevskyi, E. Bauer, T. Mori. Journal of Materials Chemistry C **12**, 8861 (2024).

Research Software Engineers (RSE) at the VSC Research Centre

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

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

Research Software Engineers (RSEs) are a new class of specialized workforce intended to bridge the gap between domain-specific software developers and HPC units more involved in basic data centre operation. Recently the VSC Research Centre at TUW has started to promote RSEs as a new job role for research fellows joining the team with ambitions in scientific HPC. This will enable better understanding of urgent needs in science and research and can provide a better basis to reach out to the scientific community to co-design new strategies of mutual interest to both R&D as well as infrastructure divisions.
In this contribution we will briefly summarize the current status of RSEs in the VSC Research Centre.

i) SH: Concept and outlook with a recent example [1], ongoing activities and the invitation to join;

ii) AS: An already active RSE reporting first hand experiences from supporting CAE users. This is a frequently requested type of cooperation as often the CAE software suites are used by researchers hailing from a non-computer science background. This eventually makes it difficult for them to progress with their research, unless the associated ecosystems of software suites are well integrated into their individual research workflows. RSEs will work towards assisting such researchers, effectively reducing their barrier to entry, not only to the VSC's vast research computing infrastructure, but also with their collaborations within the "future-steps" of their own research. Here practical tips and tricks are presented from the domains of meshing as well as solving for Ansys Fluent with the focus on HPC systems. This is mainly to address the lack of knowledge in adequately pinning the cores and allocating compute nodes for an efficient Ansys Fluent run on HPC systems, which is commonly not known among the Ansys Fluent community, since the specific support recently shifted behind an Ansys paywall;

iii) DMdC: Another testimony from an already active RSE collaborating with various domain scientists on a daily basis, particularly by enhancing simulation results with augmented reality (AR, see Fig. 1);

iv) IV: A newly established blog describing typical cases of HPC usage to an interested community with the aim of collaboratively growing and sharing R&D;

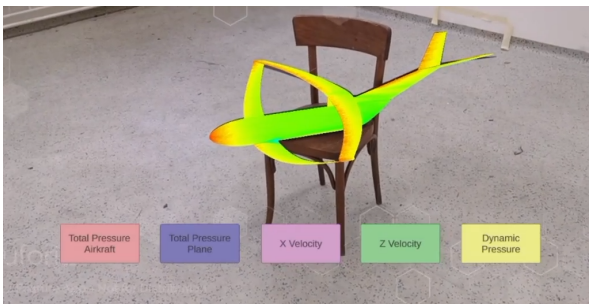


Fig. 1: Wind tunnel results for a double wing plane studied by the SpaceTeam of TUW.

References

[1] Hickel, M., and Höfinger, S., RSE-HPC-2024 (2024).

Index of presenting authors

List of ASHPC25 participants

Alistarh, Dan		Institute of Science and Technology Austria
Beranova, Kateřina	katerina.beranova@vsb.cz	IT4Innovations, VSB - Technical University of Ostrava, NCC Czechia
Blaas-Schenner, Claudia	claudia.blaas-schenner@tuwien.ac.at	VSC Research Center, TU Wien
Blaschek, Michael	michael.blaschek@univie.ac.at	University of Vienna
Bringmann, Victoria	victoria.bringmann@ista.ac.at	Institute for Science and Technology Austria
Buzan, Elena	elena.buzan@upr.si	University of Primorska
Čep, Aleš	ales.cep@um.si	Universtiy Of Maribor
Česnik, Blaž		ARNES
Costea, Stefan	stefan.costea@fs.uni-lj.si	Faculty of Mechanical Engineering, University of Ljubljana, Slovenia
Črnigoj Marc, Tina		Arctur d.o.o.
Davidović, Davor		Rudjer Bošković Institute
Debreczeni, Atila	attila.debreczeni@dkf.hu	NCC Hungary
Dellago, Christoph	christoph.dellago@univie.ac.at	University of Vienna
Denes, Mark		DAM Invisible Technology Zrt.
Dokter, Mark	mark.dokter@advanced-computing.at	EuroCC Austria
Dolenc, Tomi	tomi.dolenc@arnes.si	ARNES
Dorn, Andras	dorn.andras@damit.hu	DAM Invisible Technology Zrt.
Drobnjak, Marko	marko.drobnjak@arnes.si	open science expert
Dugonik, Jani		University of Maribor, Faculty of Electrical Engineering and Computer Science
Elefante, Stefano	stefano.elefante@ist.ac.at	Institute of Science and Technology Austria (ISTA)
Gasparini, Blaž	blaz.gasparini@univie.ac.at	University of Vienna
Genest, Alexander		Institute of Materials Chemistry, TU Wien
Goiser, Malgorzata	malgorzata.goiser@tuwien.ac.at	VSC Research Center, TU Wien
Goldenberg, Florian		TU Wien
Großöhme, Peter	peter.grossoehme@megware.com	MEGWARE
Gschwandtner, Philipp	philipp.gschwandtner@uibk.ac.at	University of Innsbruck
Haase, Gundolf	gundolf.haase@uni-graz.at	University of Graz
Hackl, Benjamin	benjamin.hackl@uni-graz.at	University of Graz
Harisch, Damjan	damjan.harisch@arnes.si	ARNES
Harrison, Simeon	simeon.harrison@tuwien.ac.at	EuroCC Austria & VSC Research Center, TU Wien
Hatfield, Samuel	samuel.hatfield@ecmwf.int	European Centre for Medium-Range Weather Forecasts
Heydemueller, Joerg	jheydemueller@ddn.com	DataDirectNetworks GmbH
Hickel, Markus		VSC Research Center, TU Wien, Austria

Hoefinger, Siegfried	siegfried.hoefinger@tuwien.ac.at	VSC Research Centre
Hoffmann, Zoltan	hoffmann.zoltan@damit.hu	DAM Invisible Technology Zrt.
Horvath, Erzsebet	erzsebet.horvath@dkf.hu	NCC Hungary
Hubman, Anže	anze.hubman@ki.si	PhD student
Hunold, Sascha	sascha.hunold@tuwien.ac.at	TU Wien
Hyryavets, Halyna	halyna.hyryavets@slovakianscc.com	Slovak National Supercomputing Centre (EuroCC Slovakia)
Imamagić, Emir	eimamagi@srce.hr	University of Zagreb, University Computing Centre (SRCE)
Jug, Matevž	matevz.jug@ki.si	National Institute of Chemistry, Slovenia
Kacin, Peter		Sysadmin
Kalcher, Sebastian	sebastiank@nvidia.com	NVIDIA
Karasek, Tomas	tomas.karasek@vsb.cz	IT4Innovations, VSB Technical University of Ostrava
KERO, STEFAN	stefan.kero@eviden.com	Sponsor Partner
Khalid, Waleed	waleed.khalid@ista.ac.at	Institute of Science and Technology, Austria
Khmelevskyi, Sergii	sk@cms.tuwien.ac.at	VSC Research Center, TU Wien
Klauser, Florian	florian.klauser@ijs.si	JSI
Kollarik, Tomas	tomas.kollarik@slovakianscc.com	Slovak National Supercomputing Centre (EuroCC Slovakia)
Koniorczyk, Mátyás	koniorczyk.matyas@wigner.hun-ren.hu	HUN-REN Wigner Research Centre for Physics
Kozubek, Tomas	tomas.kozubek@vsb.cz	IT4Innovations, VSB – Technical University of Ostrava
Kralj, Marko	marko.kralj@izum.si	Institute of Information Science, Maribor, Slovenia
Krašovec, Barbara	barbara.krasovec@ijs.si	IJS
Lah, Maša	masa.lah@ki.si	National Institute of Chemistry
Laio, Alessandro	laio@sissa.it	SISSA
Laso, Ruben		Research Group for Scientific Computing, Faculty of Computer Science, University of Vienna
Laure, Erwin	erwin.laure@mpcdf.mpg.de	MPCDF
Lebar Bajec, Iztok	ilb@fri.uni-lj.si	University of Ljubljana, Faculty of Computer and Information Science
Lesjak, Dejan	dejan.lesjak@ijs.si	Jožef Stefan Institute, Slovenia
Lindner, Andreas	andreas.lindner@advanced-computing.at	Advanced Computing Austria ACA GmbH
Ljubič, Martin	martin.ljubic@ki.si	National Institute of Chemistry, Hajdrihova 19, 1001 Ljubljana, Slovenia
Lloret, Zoé	zoe.lloret@univie.ac.at	University of Vienna
Lotrič, Uroš		University of Ljubljana
Maffi, Mateja	mateja.maffi@fs.uni-lj.si	NCC Slovenia, ULFS
Maudodi, Sayed	sayed.maudodi@amd.com	AMD
McCartney, Adam	adam.mccartney@tuwien.ac.at	VSC
Meindl, Maximilian	maximilian.meindl@univie.ac.at	University of Vienna/Department of Meteorology and Geophysics
Merzel, Franci	franci.merzel@ki.si	Theory Department, National Institute of Chemistry

Mihelic, Andrej	andrej.mihelic@ijs.si	Jozef Stefan Institute, Ljubljana, Slovenia
Muck, Katrin	katrin.muck@tuwien.ac.at	TU Wien, VSC Research Center
Neumayer, Michael	michael.neumayer@univie.ac.at	Universität Wien
Ojsteršek, Milan	milan.ojstersek@um.si	University of Maribor, Faculty of Electrical Engineering and Computer Science
Ojsteršek, Tadej	tadej.ojstersek1@um.si	University of Maribor, Faculty of Electrical Engineering and Computer Science
Ostermann, Alexander	alexander.ostermann@uibk.ac.at	University of Innsbruck
Pančur, Matjaž	matjaz.pancur@fri.uni-lj.si	University of Ljubljana, Faculty of Computer and Information Science
Pellegrini, Bozidara	bozidara.pellegrini@slovakianscc.com	Slovak National Supercomputing Centre (EuroCC Slovakia)
Penas Franqueira, Adela	adelapenasfranqueira@yahoo.es	CESGA
Pešatová, Karina	karina.pesatova@vsb.cz	VSB - Technical University of Ostrava, IT4Innovations
Pfennig, Tobias	tobias.pfennig@megware.com	MEGWARE
Pilipović, Ratko	ratko.pilipovic@fri.uni-lj.si	Faculty of Computer and Information Science, University of Ljubljana
Potisk, Tilen		National Institute of Chemistry
Povh, Janez	janez.povh@rudolfovo.eu	Rudolfovo - Science and technology center Novo mesto
Praprotnik, Matej	praprot@cmm.ki.si	National Insitute of Chemistry
Prica, Teo		IZUM
Rauber, Andreas	rauber@ifs.tuwien.ac.at	TU Wien
Reiter, Eduard	eduard.reiter@uibk.ac.at	Universität Innsbruck
Rosina, Johannes	johannes.rosina@jku.at	Institute for Theoretical Physics, JKU
Salimibeni, Majid	majid.salimibeni@tuwien.ac.at	TU Wien
Sattari, Sanaz	sanaz.sattari@tuwien.ac.at	VSC
Schlögl, Alois	alois.schloegl@ist.ac.at	Austria
Schmidt, Thorsten	thorsten.schmidt@cornelisnetworks.com	Cornelis Networks GmbH
Schneider, Sarah	sarah.schneider@tuwien.ac.at	Technical University Vienna
Sedej, Neli	neli.sedej@ki.si	National Institute of Chemistry
Siegel, Moritz	moritz.siegel@tuwien.ac.at	VSC Research Center
Šikaleska, Violeta	violeta.sikaleska@izum.si	IZUM, Institute of Information Science
Singh, Atul	atul.singh@tuwien.ac.at	VSC Research Center, TU Wien, Austria
Sluga, Davor		University of Ljubljana, Faculty of Computer and information Science
Soylu Yılmaz, Elis	esoylu@ogu.edu.tr	Eskişehir Osmangazi University
Špoljar, Jurica	jspoljar@srce.hr	University of Zagreb University Computing Centre (SRCE)
Stanič, Samo	samo.stanic@ung.si	University of Nova Gorica
Starc, Alenka	alenka.starc@arnes.si	Arnes
Stöhr, Markus	markus.stoehr@tuwien.ac.at	BOKU University
Šuntajs, Jan	jan.suntajs@ijs.si	Institut Jožef Stefan, Department of Theoretical Physics; University of Ljubljana, Faculty of Mechanical Engineering
Terjék, Mihály	mihaly.terjek@dkf.hu	DKF
Thaler, Martin	martin.thaler@uibk.ac.at	University of Innsbruck

Tomić, Draško	drasko@irb.hr	Rudjer Bošković Institute
Tomšič, Pavel	pavel.tomsic@fs.uni-lj.si	University of Ljubljana, Faculty for Mechanical Engineering
Troconis, Orlenys	o.troconis@cineca.it	CINECA
Ujlaki, Gyula	gyula.ujlaki@dkf.hu	NCC Hungary
Valh, Dejan	dejan.valh@izum.si	IZUM
Vardas, Ioannis	vardas@par.tuwien.ac.at	TU Wien
Vasileska, Ivona	ivona.vasileska@fs.uni-lj.si	Faculty of Science, Technology and Medicine, University of Luxembourg, Luxembourg
Verber, Domen		University of Maribor, Faculty of Electrical Engineering and Computer Science
Vialov, Ivan	ivan.vialov@tuwien.ac.at	Vienna Scientific Cluster Research Center
Voigt, Aiko	aiko.voigt@univie.ac.at	University of Vienna
Vreš, Domen	domen.vres@fri.uni-lj.si	University of Ljubljana, Faculty of Computer and Information Science
Wang, Yin	yin.wang@uibk.ac.at	University of Innsbruck
Wimmer, Elias	elias.wimmer@tuwien.ac.at	TU Wien / VSC
Winkler, Lukas	l.winkler@univie.ac.at	University of Vienna
Zabloudil, Jan	jan.zabloudil@tuwien.ac.at	BOKU University
Zebec, Žiga	ziga.zebec@izum.si	IZUM

DOI: https://doi.org/10.25365/phaidra.662

ISBN: 978-3-200-10466-2

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

Edited by:
Damjan Harisch, Academic and Research Network of Slovenia

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