siam 2026

Searchable Abstracts Document

SIAM Conference on Parallel Processing for Scientific Computing (PP26)

March 3–6, 2026

Zuse Institute Berlin and Free University of Berlin, Berlin, Germany

This document was current as of February 19, 2026. Abstracts appear as submitted.

siam

3600 Market Street, 6th Floor
Philadelphia, PA 19104-2688 U.S.
Telephone: 800-447-7426 (U.S. & Canada) +1-215-382-9800 (Worldwide)
meetings@siam.org

IP1

Computational Climate Science (Public Lecture)

Computational climate science, using a variety of approaches over the years, attempts to estimate the structure of the climate systems and its sensitivity to perturbations. In this talk I will provide a brief overview of the approaches employed over the years, and outline new capabilities that are expanding the scope of questions that it can address. I will introduce and describe two grand challenge problems that may be possible to resolve given small computing and (hopefully larger) algorithmic advances in the coming five years. One is to estimate the structural stability of the climate system; the second is to computationally simulate (rather than model) the response of global cloud cover to warming. They are related, as without the former the latter makes little sense.

Bjorn Stevens
Max Planck Institute for Meteorology
bjorn.stevens@mpimet.mpg.de

IP2

In Transit Learning at Exascale: A Streaming ML Workflow for Scientific Simulations

Exascale simulations generate voluminous data, creating significant I/O bottlenecks for downstream analysis. This work presents a streaming workflow that bypasses traditional file systems by directly feeding simulation output into machine learning frameworks. Data is transformed in transit, enabling asynchronous training and continual learning. Using a real-world application, we demonstrate scalable deployment on large scale supercomputer(s) and address challenges in adapting to non-steady processes via experience replay.

Sunita Chandrasekaran
University of Delaware, U.S.
schandra@udel.edu

IP3

Multiscale Computing: A Unique Opportunity for Digital Twins?

The concept of Digital Twins has emerged more than a decade ago and has been promoted still then as a potential solution for many industrial challenges. Still their industrial adoption is limited to high value use cases only. Scalable solutions and workflows to build real-time models as required for most Digital Twin applications are missing. While brute force Machine Learning approaches achieve impressive results, their data and training requirements limit their applicability in many use cases. To overcome this issue we believe that multi-scale approaches, which have been proven successful for many applications, are a key concept to foster broad Digital Twin applications. In this talk we review state-of-the-art industrial multiscale concepts, ranging from multigrid solvers to reduced order modelling and machine learning. We highlight theses along concrete use cases and discuss open challenges requiring further research. Specific focus will be given to obstacles limiting broader industrial adoption today with the goal to spur furth research to overcome the limited scalability of Digital Twins.

Dirk Hartmann
Siemens Digital Industries Software, Germany
hartmann.dirk@siemens.com

IP4

Scaling Vascular Digital Twins: From Millions of Heartbeats to Petabytes of Data

High performance computing has transformed our ability to model complex physical systems. The emergence of vascular digital twinspatient-specific, physics-informed simulations of blood flowextends this frontier into healthcare, requiring scalable solvers for multi-physics models across anatomically detailed 3D domains. These simulations, often based on lattice Boltzmann or Navier-Stokes formulations, must capture millions of cardiac cycles while integrating multimodal data from imaging and wearables. The result is petabyte-scale datasets and continuous, time-resolved computation. I will discuss advances in parallel time integration, communication-avoiding solvers, and data-driven adaptivity that enable sustained performance at exascale. Coupling real-time sensor data with physics-based models introduces new challenges in temporal synchronization, reduced-order modeling, and uncertainty quantification. Together, these developments illustrate how algorithmic innovation and system co-design are reshaping what is computationally possible in medicine and beyond

Amanda Randles
Duke University
amanda.randles@duke.edu

IP5

Parallelism in Sparse and Data-Sparse Direct Solvers

Efficient solutions of large-scale, ill-conditioned and indefinite algebraic equations are ubiquitously needed in numerous computational fields, including multiphysics simulations, machine learning, and data science. Because of their robustness and accuracy, direct solvers are a crucial component in building a scalable solver toolchain. In this talk, we will discuss recent advances of sparse direct solvers along two axes: 1) reducing communication and latency costs in both task- and data-parallel settings, and 2) reducing computational complexity via low-rank techniques such as hierarchical matrix algebra. In addition to algorithmic principles, we also illustrate the key parallelization challenges and best practices to deliver high speed and reliability on modern heterogeneous parallel machines.

Sherry Li
Lawrence Berkeley National Laboratory
xsli@lbl.gov

IP6

Extreme Computing Universals

Extreme in computing means more than large in scale. It can indicate constraints of real-time, low power, low memory capacity or bandwidth per core, or low available concurrency. Some universals in dealing with extremes are: reside high on the memory hierarchy (e.g., by blocking or processing on the fly), reduce synchrony in frequency and/or in span (e.g., by performing extra flops), reduce communication in number and/or volume of messages (e.g., by exploiting extra memory), employ dynamic scheduling and balancing (e.g., by runtime systems based on DAGs), avoid over-resolving with respect to output accuracy re-

quirements (e.g., adapt precision, fidelity, and inner tolerances), reformulate applications before computing (e.g., with smarter bases or discretizations), co-design algorithms with hardware (e.g., specialized heterogeneity in processing, memory, and networking elements), exploit the right to re-order (e.g., linearization vs partitioning, lagged evaluations, colorings), exploit hierarchical or multiple alternative versions of the same system, exploit data sparsity to meet curse of dimensionality with blessing of low rank, take resilience into algorithms, relieving hardware and systems, look over the transoms for optimizations beyond optimized components, code to specialized back-ends while presenting high-level APIs to users, consider science per Joule. Most are classical but have new significance. We provide some illustrations and welcome others.

David E. Keyes
King Abdullah University of Science and Technology (KAUST)
david.keyes@kaust.edu.sa

IP7

AI's Hardware Revolution and the Scientific Computing Opportunity

The rapid success of AI has driven unprecedented investment in massively parallel processor architectures, creating systems with fundamentally different characteristics from traditional HPC platforms. At Cerebras Systems, we replaced much of traditional architecture to focus on strong scaling limited only by what is physically possible to build. The result brings innovations across architecture, algorithms, packaging, and systems-level deployment. IO bandwidth matches computational bandwidth, memory hierarchy is flattened, and the overall architectural paradigm demands new programming models. These systems achieve strong scaling 1000 beyond contemporary exascale platforms for certain scientific applications, but they come with tradeoffs the scientific computing community must understand. This talk explores what becomes possible for scientific applications on these advanced architectures. Strong scaling parallelismpursuing temporal scale, not just spatial scalerequires different and harder techniques than weak scaling approaches. Math kernels must be reimagined for memory organizations with very different work-per-core ratios. We examine applications from PDEs to large-scale simulations, showing where data-intensive algorithms thrive when bandwidth hierarchies flatten. We also confront important tensions: precision arithmetic support, algorithmic flexibility, and whether market-driven innovation serves science missions. As AI deployments scale to unprecedented sizes, the scientific computing community has an opportunity to shape how these platforms evolve. Exploring the envelope of feasibility positions the community to create the synergies between scientific computing and AI that both fields ultimately require.

Michael James
Cerebras Systems, U.S.
michael@cerebras.net

IP8

Brain-Inspired Computing: Opportunities for Neuromorphic Systems in the Future of Computing (Public Lecture)

Neuromorphic computing, a brain-inspired computing technology, provides the opportunity for low-power, intelligent computing systems. Neuromorphic computers have the potential to be used in a variety of scenarios, from the edge to high-performance computing systems. In this talk, I will overview the field of neuromorphic computing, and I will give an introduction to spiking neural networks, as well as some of the most common algorithms used in the field. I will discuss the potential for using neuromorphic systems in real-world applications from scientific data analysis to autonomous vehicles. Finally, I will discuss remaining challenges in the field, as well as opportunities for future applications.

Catherine Schuman
University of Tennessee
Department of Electrical Engineering and Computer Science
cschuman@utk.edu

SP1

SIAG/SC Prize Presentations and 2026 SIAM Activity Group on Supercomputing Early Career Prize Lecture - Making Waves in the Cloud: A Paradigm Shift for Scientific Computing through Compiler Technology

Scientific models are today limited by compute resources, forcing approximations driven by feasibility rather than theory. They consequently miss important physical processes and decision-relevant regional details. Advances in AI-driven supercomputing specialized tensor accelerators, AI compiler stacks, and novel distributed systems offer unprecedented computational power. Yet, scientific applications such as ocean models, often written in Fortran, C++, or Julia and built for traditional HPC, remain largely incompatible with these technologies. This gap hampers performance portability and isolates scientific computing from rapid cloud-based innovation for AI workloads. In this talk, we bridge that gap by transpiling existing programs using the MLIR compiler infrastructure. This process enables advanced optimizations, deployment on AI hardware, and automatic differentiation. In particular, we demonstrate execution of a state of the art Julia-based ocean model (Oceananigans), with ¿277 custom single-node CUDA kernels on thousands of distributed GPUs and Google TPUs. Our results demonstrate that cloud-based hardware and software designed for AI workloads can significantly accelerate simulations, opening a path for scientific programs to benefit from cutting-edge computational advances.

William S. Moses
University of Illinois at Urbana-Champaign, U.S.
wsmoses@illinois.edu

CP1

Parallelizing the Approximate Minimum Degree Ordering Algorithm: Strategies and Evaluation

The approximate minimum degree algorithm is widely used before numerical factorization to reduce fill-in for sparse matrices. While considerable attention has been given to the numerical factorization process, less focus has been placed on parallelizing the approximate minimum degree algorithm itself. In this paper, we explore different parallelization strategies, and introduce a novel parallel framework that leverages multiple elimination on distance-2 independent sets. Our evaluation shows that parallelism within individual elimination steps is limited due to low computational workload and significant memory contention. In contrast, our proposed framework overcomes these challenges by parallelizing the work across elimina-

tion steps. To the best of our knowledge, our implementation is the first scalable shared memory implementation of the approximate minimum degree algorithm. Experimental results show that we achieve up to an 8.30× speedup using 64 threads over the state-of-the-art sequential implementation in SuiteSparse.

Yen-Hsiang Chang
University of California, Berkeley
yenhsiangc@berkeley.edu

Aydin Buluç
Lawrence Berkeley National Laboratory
abuluc@lbl.gov

James W. Demmel
UC Berkeley
demmel@berkeley.edu

CP1

Application Failures and Machine Computational Efficiency

We present a framework for evaluating uptime efficiency of Exascale-class scientific computers when application failure rates are appreciable. This is the situation that confronts current leadership-class scientific computing platforms and large AI training installations. What distinguishes scientific computing platforms is the heterogeneity of their applications. We argue that this diversity requires that failure rates and mean intervals between failures should be specified in terms of usage (e.g. node-hours) rather than time, as is currently customary. We consider the usage loss terms due to failures, to checkpointing, and to restart costs, and update the framework of Daly (2006) allowing users to specify optimal checkpointing usage intervals that minimize such losses. We derive the machine computational efficiency, which specifies the expected fractional resource allocation that is available for scientific computation. We illustrate the methodology using one year of production runtime data from the Frontier supercomputer at Oak Ridge National Laboratory.

Carlo Graziani, Lusch Bethany
Argonne National Laboratory
cgraziani@anl.gov, blusch@anl.gov

Messer Bronson
Oak Ridge National Laboratory
bronson@ornl.gov

CP1

Energy Consumption in Parallel Neural Network Training

The increasing demand for computational resources of training neural networks leads to a concerning growth in energy consumption. While parallelization has enabled upscaling model and dataset sizes and accelerated training, its impact on energy consumption is often overlooked. To close this research gap, we conducted scaling experiments for data-parallel training of two models, ResNet50 and FourCastNet, and evaluated the impact of parallelization parameters, i.e., GPU count, global batch size, and local batch size, on predictive performance, training time, and energy consumption. We show that energy consumption scales approximately linearly with the consumed resources, i.e., GPU hours; however, the respective scaling factor differs substantially between distinct model trainings and hardware, and is systematically influenced by the number of samples and gradient updates per GPU hour. Our results shed light on the complex interplay of scaling up neural network training and can inform future developments towards more sustainable AI research.

Philipp Huber
Karlsruhe Institute for Technology (KIT)
philipp.huber@kit.edu

David Li, Juan Pedro Gutiérrez H. Muriedas, Deifilia Kieckhefen, Achim Streit
Karlsruhe Institute of Technology, SCC
david.li@kit.edu, juan.muriedas@kit.edu, deifilia.to@kit.edu, achim.streit@kit.edu

Markus Götz
Karlsruhe Institute of Technology, SCC
Helmholtz AI
markus.goetz@kit.edu

Charlotte Debus
Karlsruhe Institute of Technology (KIT), SCC
charlotte.debus@kit.edu

CP1

From GPUs to RRAMs: Distributed In-Memory PrimalDual Hybrid Gradient Method for Solving Large-Scale Linear Optimization Problems

The exponential growth of computational workloads is surpassing the capabilities of conventional architectures, which are constrained by fundamental limits. In-memory computing (IMC) with RRAM provides a promising alternative by providing analog computations with significant gains in latency and energy use. However, existing algorithms developed for conventional architectures do not translate to IMC, particularly for constrained optimization problems where frequent matrix reprogramming remains cost prohibitive for IMC applications. Here we present a distributed in-memory primaldual hybrid gradient (PDHG) method, specifically co-designed for arrays of RRAM devices. Our approach minimizes costly write cycles, incorporates robustness against device non-idealities, and leverages a symmetric block-matrix formulation to unify operations across distributed crossbars. We integrate a physics-based simulation framework called MELISO+ to evaluate performance under realistic device conditions. Benchmarking against GPU-accelerated solvers on large-scale linear programs demonstrates that our RRAM-based solver achieves comparable accuracy with up to three orders-of-magnitude reductions in energy consumption and latency. These results demonstrate the first PDHG-based LP solver implemented on RRAMs, showcasing the transformative potential of algorithmhardware co-design for solving large-scale optimization through distributed in-memory computing.

Huynh Vo
School of Industrial Engineering and Management, OSU
Energy Systems and Infrastructure Assessment Division, ANL
lucius.vo@okstate.edu

Md Tawsif Rahman Chowdhury
Department of Electrical and Computer Engineering, WSU
mtawsifrc@wayne.edu

Paritosh P. Ramanan
School of Industrial Engineering and Management
Oklahoma State University
paritoshpr@gatech.edu

Gozde Tutuncuoglu
Department of Electrical and Computer Engineering, WSU
gozde@wayne.edu

Junchi Yang, Feng Qiu
Energy Systems and Infrastructure Assessment Division, ANL
junchi.yang@outlook.com, fqiu@anl.gov

Murat Yildrim
Department of Industrial and Systems Engineering, WSU
murat@wayne.edu

CP2

The Tubal Arnoldi Method for Tensor Function Approximation

This talk presents Krylov subspace methods based on the tensor t-product for approximating quantities related to third-order tensor functions. In particular, it introduces the tensor t-tubal Arnoldi method, which projects high-dimensional tensor problems onto lower-dimensional tensor Krylov subspaces. The proposed method leverages novel algebraic properties of the t-product to efficiently handle large-scale tensor computations and is naturally well-suited for parallel computing architectures, making it attractive for high-performance scientific applications. Applications include the evaluation of parameter-dependent tensor functions and the solution of multidimensional ordinary differential equations. Numerical experiments demonstrate the effectiveness, scalability, and accuracy of the proposed method compared to existing approaches.

Fatima Bouyghf
ENAC Toulouse, OPTIM, MORO
fatimabouyghf3@gmail.com

CP2

Large-Scale Flow and Aeroacoustic Hpc Simulations with Cruna

We present recent advances in the parallelization of CRUNA, a simulation and optimization framework based on the compressible NavierStokes equations with volume penalization in the finite-difference time domain. To enable efficient large-scale computations, CRUNA has been parallelized and benchmarked through scaling tests on the CPU CLX partition of the high-performance computing centre at ZIB, demonstrating excellent performance and scalability. Applications include aeroacoustic simulations of sound generation from flow around porous-coated cylinders at high Reynolds numbers.

Yannick Schubert
TU Berlin
schubert@tnt.tu-berlin.de

CP2

Analysis of Ultra-Weak Discontinuous Galerkin Method for Cahn-Hilliard Equation with Memory

Phase separation in materials with memory effects can be described by the time-fractional CahnHilliard equation. We design and analyze a fully discrete numerical framework for this model, where spatial discretization is carried out using an ultra-weak discontinuous Galerkin approach with flexible flux choices, and the temporal discretization employs a non-uniform L1 scheme tailored for fractional dynamics. The method is rigorously shown to be uniquely solvable and unconditionally stable under a convexconcave splitting formulation. In addition, optimal convergence is established through a priori error analysis. A notable feature of the scheme is that it inherits the mass conservation and energy decay properties of the continuous problem. Numerical tests confirm the theoretical accuracy and further illustrate the effectiveness of the method in capturing coarsening patterns and long-time dynamics of phase separation.

Deeksha Singh
IIT Guwahati
deekshas@iitg.ac.in

CP3

Quantum Machine Learning Applications for Enhancing Computational Fluid Dynamics Simulations

Quantum Machine Learning (QML) is increasingly vital for modeling and optimizing cryogenic systems, particularly those involving liquid hydrogen in aviation, quantum technologies, and large-scale scientific experiments. By leveraging quantum physics-informed AIsuch as Physics-Informed Neural Networks (PINNs), Fourier Neural Operators (FNO), and Geometry-Aware Operator Transformers (GAOTs)QML enables accurate simulation of low- temperature heat and mass transfer processes, critical for cryo-cooler design, storage, and transport systems. In aerospace, QML enhances predictions of thermofluidic behavior under extreme cryogenic conditions, im- proving safety, efficiency, and anomaly detection in hydrogen-based propulsion and storage systems. These models also optimize mesh generation and turbulence modeling in computational fluid dynamics (CFD), reducing computational costs while increasing predictive accuracy. Cryogenics underpins quantum computing, where maintaining coherence in superconducting qubits requires precise thermal control. Quantum combined with Physics-Inspired AI models assists in simulating quantum dynamics and optimizing cryostat performance and error mitigation strategies. By addressing the challenges of high-dimensionality and multi-scale modeling, these methods offer unparalleled efficiency and accuracy, paving the way for breakthroughs in scientific and engineering applications.

Aditya A Sesh
quasi ai
aditya.a.sesh@alumni.tu-berlin.de

CP3

Tiled Execution Intermediate Representation (teir)

We present the Tiled Execution Intermediate Representation (TEIR), a compact IR for high-performance tensor operations that expresses computation as a composition of primitives over subtensors ("tiles"). TEIR separates what is executed from how it is scheduled through two records: (1) TEIR-Primitives and (2) TEIR-Schedule. The TEIR-Primitives record defines three primitives: first-access, main, and last-access. The main primitive encap-

sulates the core computation. Specifically, tiled tensor contractions are implemented using GEMM or BRGEMM (batch-reduce GEMM) primitives. TEIR-Schedule assigns a per-axis execution policy (sequential, shared-memory parallelism, or consumed inside primitives) and specifies axis extents and strides. We demonstrate high performance with the TEIR-enabled Python package etops on x86 and Arm CPUs across multiple ISAs (x86 AVX-512; Arm Neon and SME).

Alexander Breuer
University of California, San Diego
alex.breuer@uni-jena.de

CP3

Fuzzy Computation for Community Analysis in Large and Time-Variant Social Graphs

This study presents a fuzzy computing approach for dynamic community detection, termed Fuzzy Time-variant Community Groups (FTCG), utilizing fuzzy weighting techniques to track evolving network structures over time. The methodology was validated on a small 5-node graph and applied to large-scale datasets, including Amazon product networks, Bitcoin transactions, and Cellular Phone Network data. Two novel link-weighting techniques were introduced to enhance the detection of temporal community changes, while a Fuzzy Modularity measure was proposed to evaluate community quality. The impact of varying threshold values was analyzed, demonstrating how different thresholds influence community detection outcomes. Experimental results confirm the approachs effectiveness in capturing network dynamics, particularly in the Bitcoin and Cellular datasets, proving its robustness in Social Network Analysis (SNA) and its potential for informed decision-making in evolving systems.

Dr. Ubaida Fatima
NED University of Engineering and Technology Karachi
ubaida@neduet.edu.pk

CP4

GPU Parallelization of Adaptive Conservative Time Integration (ACTI) for Multiphysics Simulations

Adaptive Conservative Time Integration (ACTI) is a conservative finite volume scheme that allows for arbitrarily large timesteps to be taken when solving partial differential equations. In this project, we parallelize the ACTI scheme with Compute Unified Device Architecture (CUDA) on GPUs and formulate the scheme as an algorithm to test the scalability of ACTI. We tested the parallelized ACTI algorithm on tracer transport and two-phase flow with heterogeneous permeability fields. We ran multiple simulations in 1D, 2D, and 3D while varying grid sizes and high performance computing architecture. Parallelized ACTI on GPUs reduces computational runtime by orders of magnitude compared to ACTI on CPUs. For example, results show that a 3D, one million cell tracer problem can be solved 84x faster on a single GPU than a single core CPU. Two-phase flow simulations exhibit similar efficiency improvements. These results facilitate the parallelization of ACTI for extreme computational efficiency gains in fluid dynamic simulations.

Brea Swartwood
Stanford University
brea7@stanford.edu

CP4

Accelerating Matrix Multiplication in Multiple Double Arithmetic with Tensor Cores of a NVIDIA A100 Graphics Processing Unit

A multiple double is a sequence of nonoverlapping doubles. Exploiting hardware arithmetic, multiple double arithmetic multiplies the precision. To compensate for the cost overhead of multiple double arithmetic, the tensor cores in the NVIDIA A100 graphics processing unit are applied. Specialized for matrix multiplications, tensor cores only support elementary, noncomposite floating-point operations. The renormalization of results of multiple double arithmetical operations into nonoverlapping doubles cannot be performed by tensor cores, as renormalizations involve branching. The renormalization of multiple doubles into nonoverlapping doubles is relaxed, widening the gaps between the doubles with trailing zero bits. Data staging algorithms arrange the convolutions of low with high doubles into inner products for execution on tensor cores. The renormalizations are handled by the streaming multiprocessors. Experiments demonstrate the correctness and performance.

Jan Verschelde
University of Illinois, Chicago, U.S.
janv@uic.edu

Howard Chen
University of Illinois, Chicago
hchen221@uic.edu

CP4

ACORN-QRE: A Practical Method of Generating Secure One-Time Pads for Use in Encryption

The Additive Congruential Random Number (ACORN) generator gives rise to sequences with long period approximating to uniformity in up to k dimensions (for any value of k). ACORN-QRE (Wikramaratna, REAMC Report-007, 2023, https://eprint.iacr.org/2023/1080) is a straightforward modification which avoids the linearity of ACORN, while preserving the uniformity. This can generate one-time pads that are demonstrably resistant to attack by current computers or by future computing developments (including quantum). The pads, which can use any alphabet, work with a Vernam-type cypher to securely encrypt both files and communications. In this paper, we present performance data for a software implementation of ACORN-QRE with a key of 1079 bits, each bit assigned randomly as 0 or 1. On a standard laptop or desktop computer, this works to securely encrypt binary files of arbitrary size. Encrypted files can be safely shared over any public network or even left for collection on a publicly-accessible web site without fear of their being intercepted and later read by a malicious actor. Only the sender and the intended recipient (in possession of the relevant key, which must be kept secure) are able to decrypt the file. Thus, the problem of securely sharing GBs or even TBs of data is reduced to one of securely sharing a key comprising 1079 random-looking bits. The ACORN-QRE algorithm is patented in UK (GB2591467) and USA (US11,831,751B2); the patents are owned by REAMC Limited.

Roy S. Wikramaratna
REAMC Limited

rwikramaratna@gmail.com

CP5

Multivariate Data Analysis of Earthquake Flagged by Volcanic Eruptions and Tsunami Waves

The tectonic plates are always slowly moving, but they get stuck at their edges due to friction. When the movement on the edge overcomes the friction, a sudden slip of the fault will cause an earthquake that releases energy in waves that travel through the earth's crust, causing the shaking that we feel. Most machine learning methods focus on analyzing the focal depth and fault to determine the earthquake's intensity and destruction. The goal is to apply multivariate data analysis and Python tools to study earthquakes that are flagged with tsunami waves and volcanic eruptions. Creating a data frame with columns time, tsunamis, waves, volcanic eruptions, longitude, and magnitude is an essential step for early predictions. Mathematical and statistical methods are essential for applying data mining, clustering, classification, and decision trees. Longitude and latitude are the x and y axes. Combining three data sets of earthquakes, volcanic eruptions, and tsunami waves will provide a machine learning module to learn the locations, radius, intensity, and, frequently, the three geological phenomena will strike together. The historical data set will range from 2015 BC to the present date. For a wide view of the same locations and how frequently earthquakes combined with volcanic and tsunamis will occur.

Rami A. Aboushadi
Jackson State University
batool757575@gmail.com

CP5

A Load Balancing Library for Algorithm Development and Testing with Amrex

Load balancing is a perpetual challenge for distributed-domain HPC codes. As system architectures, physics models and research requirements increase in complexity, the underlying assumptions of approximation-based load balancing algorithms have become less accurate, reducing the scalability of HPC simulations on modern hardware. Additionally, the diversification of HPC simulation codes to target unique, specific research objectives has generated a need for bespoke load balancing strategies, often prohibitively increasing the time and complexity to find effective solutions. This talk will present the work done thus far to design a load-balancing library for the rapid building, testing and deployment of effective load-balancing algorithms. This effort focuses on algorithms suitable for dynamic, domain-decomposition applications, such as those of the mesh-and-particle code, AMReX. It is designed to allow scientists to explore new strategies through the modification and improvement of existing algorithms without running a large number of at-scale simulations. An overview of the library will be presented, including its ready-to-use testing infrastructure and library of algorithms. Recent success stories in developing the current algorithmic suite will also be presented as well as an overview of future algorithmic targets and their use cases for modern HPC architectures and HPC+AI code coupling paradigms.

Kevin N. Gott
LBNL
kngott@lbl.gov

CP5

Correlation of Performance Portability and Code Complexity in Gpu Parallelization Paradigms for Scientific Computing

In recent years, numerous parallel programming paradigms have emerged for offloading compute-intensive tasks to GPUs. These technologies include AdaptiveCpp (hipSYCL), Kokkos, OpenCL, CUDA, OpenACC, OpenMP, BoostCompute, and ISO C++'s parallel algorithms. Previous research examining these paradigms has highlighted three key aspects: portability, performance, and productivity. These aspects address practical questions such as where the software can run, how much overhead a generalized solution induces compared to a native application, and how easily a developer can prototype a new application. In this work, we will provide a comprehensive comparison of the performance of these paradigms on GPUs in the context of four distinct applications, ranging from simple vector addition and matrix multiplication to GPU-accelerated n-body particle simulations and polyhedral gravity modelling. This cartesian product of paradigm and computational problem is assessed regarding numerical stability, application efficiency as proposed by Pennycook et al. (2019), and code complexity metrics such as Lines of Code and the Halstead complexity. Given their performance and complexity, this investigation creates a reference for the suitability of GPU programming paradigms for application scientists to inspire uncomplicated GPU-accelerated scientific software, as we show that simple, understandable code does not imply less performance.

Jonas Schuhmacher, Robin Brase, Hans-Joachim Bungartz
Technical University of Munich
jonas.schuhmacher@tum.de, robin.brase@tum.de, bungartz@cit.tum.de

CP5

Parallel-in-Time Kalman Smoothing Using Orthogonal Transformations

The talk will present a numerically-stable parallel-in-time linear Kalman smoother. The smoother uses a novel highly-parallel QR factorization for a class of structured sparse matrices for state estimation, and an parallel adaptation of the SelInv selective-inversion algorithm to evaluate the covariance matrices of estimated states. Our implementation of the new algorithm, using the Threading Building Blocks (TBB) library, scales well on both Intel and ARM multi-core servers, achieving speedups of up to 47x on 64 cores. The algorithm performs more arithmetic than sequential smoothers; consequently it is 1.8x to 2.5x slower on a single core. The new algorithm is faster and scales better than the parallel Kalman smoother proposed by Srkk and Garca-Fernndez in 2021.

Shahaf Gargir, Sivan Toledo
Tel Aviv University
shahafgargir@mail.tau.ac.il, stoledo@mail.tau.ac.il

CP6

Fault-Tolerance Systems Unlock Freedom Path-

ways in Colombia

During the sixteenth century, Fault-Tolerance Systems (FTS) were essential for enslaved Colombians to curate/utilize spy networks to secure maroon communities as counter spaces for the exchange and perseverance of indigenous knowledge that led to liberation. This included redundancy/replication and fail-safe algorithms under real-time conditions. Pathways of escape were generated throughout the system to maintain reliability. This FTS system had redundancy: if one of the code keepers became detained, multiple code keepers retained their codes (braids)and had the autonomy to roam and disseminate escape codes. Codes were distributed in a decentralized system with multiple concealed nodes/escape routes. Tributaries (arroyos) were checkpoints where cultural knowledge/escape routes were disseminated as a means to disrupt/dismantle enslavement. The enslaved had their own language as a backup system to distribute escape codes if the braid codes were detected by the Spanish enslavers. This FTS was adaptable/ reconfigured to depict/uncover army strongholds as a means of avoiding them via the mountainous landscape of the cimarones (maroon communities). This system was continually taught/re-taught at tributaries. The application of FTS was successful in establishing pathways to freedom for enslaved Colombians in the sixteenth century (Greensword, 2022; Guillen, 2018;Landers, 2013; Opong-Nyantekyi,2023; Pasham, 2020; Thomas, 2020; Torres & Obseso, 2012).

Dr. Lafrance A. Clarke
Independent STEM Researcher
University of South Florida Alumnus
yetu9100@gmail.com

CP6

Parallel Frame Interpolation: Improving Tomographic Imaging Workflows

Deep-learning-based frame interpolation networks can generate intermediate slices that closely match experimentally acquired tomographic images, enabling a reduction in the number of acquired projections and thus shorter acquisition times. However, the computational cost of synthesizing these slices remains a challenge for large-scale or time-sensitive workflows. We present ongoing efforts to parallelize the interpolation pipeline by splitting images pairs into independent tasks executed concurrently across multiple computing resources. This approach leverages the embarrassingly parallel nature of the problem and can run on multi-GPU systems or HPC clusters. Preliminary observations suggest this strategy can reduce wall-clock time, making deep interpolation more practical for high-resolution tomographic reconstruction. Quantitative benchmarks and scalability analysis will be presented at the conference.

Ahmed El Kerim
ENS Paris Saclay
ahmed.el-kerim@safrangroup.com

Clément Remacha
SafranTech
clement.remacha@safrangroup.com

CP6

Category Theory and Genetic Drift

This presentation introduces a whole new way of looking at genetic drift. Instead of using just probability models (which can be kinda rigid), we use category theory to build a model that actually respects how messy and reversible population changes can be. Groupoids help us capture that reversibility and functors let us track how things evolve across populations. We also bring in group actions, orbits, and fixed points to analyze how allele frequencies get stuck or keep shifting. We show how universal properties can describe drift itself as like the glue that holds everything together. This isn't just 'abstract nonsense' it is a powerful way to rethink 'randomness' in evolution.

Taylor G. Mendes
Spelman College
taylormendes@spelman.edu

CP6

Colibri2: Distributed HPC Software for Building Chemical Reaction Networks

Exploring chemical potential energy surfaces (PESs) is notoriously difficult because they are continuous, high-dimensional spaces where the number of minima and transition states grows exponentially with each additional atom. Colibri2, an open-source software, tackles this challenge by representing PESs as reaction networks (RNs), discrete network representations of chemical states and transformations between them. To make this approach scalable, Colibri2 employs a distributed architecture built on Redis, PostgreSQL, and optimized parallel Python, enabling efficient use of multicore and cluster environments. To reduce unnecessary calculations, Colibri2 performs on-the-fly energy evaluations of molecules, which guides node expansion based on energy differences. This enables dynamic pruning of irrelevant network branches, avoiding costly computations while preserving chemically meaningful pathways. This design supports the generation of networks with billions of nodes, which can then be analyzed using optimized path-finding algorithms such as Dijkstra, A^*, and machine learning techniques to identify feasible reaction pathways and mechanisms. By combining scalable graph algorithms with high-performance computing, Colibri2 provides a general-purpose platform for predictive reaction pathway discovery across complex chemical systems.

Miko M. Stulajter
Predictive Science Inc.
mstulajt@uci.edu

Liliana Garcia, Dmitrij Rappoport
University of California, Irvine
mteutla@uci.edu, d.rappoport@uci.edu

CP7

Mapping Sparse Triangular Solves to GPUs via Fine-grained Domain Decomposition

Solving sparse linear systems typically uses preconditioned iterative methods, but applying preconditioners via sparse triangular solves introduces bottlenecks due to irregular memory accesses and data dependencies. This work leverages fine-grained domain decomposition to adapt triangular solves to the GPU architecture. We develop a fine-grained domain decomposition strategy that generates non-overlapping subdomains, increasing parallelism in the application of preconditioner at the expense of a modest increase in the iteration count for convergence. Each subdomain is assigned to a thread block and is sized such that the subdomain vector fits in the GPU shared memory, eliminating the need for inter-block synchronization

and reducing irregular global memory accesses. When compared to rocSPARSE triangular solves, we achieve a 10.7× speedup for triangular solves and a 3.2× speedup over an ILU0-preconditioned biconjugate gradient stabilized (BiCGSTAB) solver on an AMD Instinct MI210 GPU. Furthermore, our approach delivers a 3.3× geometric mean speedup over the BiCGSTAB + ILU0 implementations from the hypre library on the same GPU.

Atharva Gondhalekar
Virginia Tech
atharva1@vt.edu

Kjetil Haugen
Haugen Labs
kjetil@haugenlabs.com

Thomas Gibson
Advanced Micro Devices
thomas.gibson@amd.com

Wu-chun Feng
Virginia Tech
wfeng@vt.edu

CP7

LAPIS: A Performance Portable, High Productivity Compiler Framework

Portability, performance, and productivity are three critical dimensions for evaluating a programming model or compiler infrastructure. Several modern programming models for computational science focus on performance and portability. On the other end, several machine learning focused programming models focus on portability and productivity. A clear solution that is strong in all three dimensions has yet to emerge. A second related problem arises when use cases from computational science converge with machine learning. The disparate popular frameworks of these fields require programmers to manually integrate codes written in different frameworks. We present LAPIS, an MLIR-based compiler that addresses both challenges. We demonstrate that LAPIS can automatically lower sparse and dense linear algebra kernels from computational science and artificial intelligence use cases. We also show how LAPIS facilitates the integration of codes between PyTorch and Kokkos. Finally, we compare kernel performance with the default MLIR implementations on diverse architectures to demonstrate portability.

Brian Kelley, Sivasankaran Rajamanickam
Sandia National Laboratories
bmkelle@sandia.gov, srajama@sandia.gov

CP7

On Combining Pipelining and S-Step Concepts in Preconditioned Conjugate Gradient Methods

On large-scale parallel computers, global communication becomes a major bottleneck for the Preconditioned Conjugate Gradient (PCG) method, an iterative solver for large sparse linear systems. Scalable PCG variants reduce the number of global synchronization points by a factor of O(s) (s-step methods), or overlap communication with local computations (pipelined methods). The pipelined s-step PCG method P-sPCG$_{\text{mon}}$ combines these two approaches, but introduces significant local computational overhead increasing with step size s. Moreover, it suffers from poor numerical stability. Choosing a suitable basis type for the s-step basis matrices usually strongly improves numerical stability. Thus, we generalize P-sPCG$_{\text{mon}}$, which was designed to only use the monomial basis, to support arbitrary basis types, denoting our new method as P-sPCG. Moreover, we also reformulate the more stable s-step method CA-PCG such that its global communication is overlapped with computations, denoting another new pipelined s-step method as P-CA-PCG. Numerical experiments on 40 real-world test problems show that P-sPCG improves numerical stability compared to P-sPCG$_{\text{mon}}$, while P-CA-PCG is even more stable than P-sPCG. Strong scaling experiments with a synthetic test problem confirm that for larger values of s, P-CA-PCG outperforms P-sPCG and all existing s-step methods except for one, which, however, showed significantly worse numerical stability in our experiments.

Viktoria Mayer
University of Vienna
viktoria.mayer@univie.ac.at

Wilfried N. Gansterer
Department of Computer Science
University of Vienna
wilfried.gansterer@univie.ac.at

CP7

Compiler-supported reduced precision and AoS-SoA transformations for heterogeneous hardware

This study evaluates AoS-to-SoA transformations over reduced-precision data layouts for a particle simulation code on several GPU platforms: We hypothesize that SoA fits particularly well to SIMT, while AoS is the preferred storage format for many Lagrangian codes. Reduced-precision (below IEEE accuracy) is an established tool to address bandwidth constraints, although it remains unclear whether AoS and precision conversions should execute on a CPU or be deployed to a GPU if the compute kernel itself must run on an accelerator. On modern superchips where CPUs and GPUs share (logically) one data space, it is also unclear whether it is advantageous to stream data to the accelerator prior to the calculation, or whether we should let the accelerator transform data on demand, i.e. work in-place logically. We therefore introduce compiler annotations to facilitate such conversions and to give the programmer the option to orchestrate the conversions in combination with GPU offloading. For some of our compute kernels of interest, Nvidias G200 platforms yield a speedup of around 2.6 while AMDs MI300A exhibits more robust performance yet profits less. We assume that our compiler-based techniques are applicable to a wide variety of Lagrangian codes and beyond.

Pawel Radtke
Computer Science, Durham University
pawel.k.radtke@durham.ac.uk

Tobias Weinzierl
Durham University
tobias.weinzierl@durham.ac.uk

CP8

Data Assimilation Framework for Outflow Parameter Estimation in Patient-Specific Cardiovascular Modelling

Patient-specific cardiovascular modelling has emerged as a powerful approach for investigating haemodynamics and enabling real-time, data-informed clinical decision-making.

A critical challenge lies in prescribing physiologically consistent outflow conditions, which strongly influence simulated flow fields. Windkessel (RCR) models are commonly used, yet their parametersproximal resistance (R_p), distal resistance (R_d), and compliance (C)are often uncertain. We present a stabilised finite element framework coupled with ensemble-based data assimilation to estimate these parameters from flow and pressure data. The unsteady incompressible NavierStokes equations are solved with RCR outlet conditions, and an ensemble Kalman-type method is used to minimise the mismatch between simulated and observed outlet flow rates. Unlike adjoint-based approaches, this statistical formulation avoids explicit gradient evaluations, improving robustness and facilitating integration with clinical data. Synthetic datasets generated from forward simulations in a patient-specific aortic model are used for validation. The framework accurately recovers prescribed RCR values and supports sensitivity analysis of how outlet variations affect haemodynamic metrics such as wall shear stress, pressure gradients, and flow distribution. These results highlight the potential of ensemble-based assimilation to enhance the reliability and predictive capability of personalised cardiovascular modelling.

Km Surabhi Km Surabhi
SISSA Trieste
surabhirathor92@gmail.com

CP8

Fine-Grained Energy Measurements for Algorithm Selection in Particle Simulations

In recent decades, the number and performance of HPC systems have risen sharply. Yet, estimates show that emissions from supercomputers are significant, comparable to those of airplanes (in kg CO2 per hour). While the benefits of HPC are undeniable, addressing energy consumption is crucial. Our study focuses on particle simulations, both relevant to and benefiting from HPC. These simulations offer numerous algorithmic choicescontainers (Linked Cells, Verlet Lists, etc.), data layouts (AoS, SoA), parallel traversals, and hyperparameters (cell size factor, vectorization). The combinations yield hundreds of possible configurations, making manual optimal selection impractical. Moreover, the particle phase-space changes over time, often requiring different algorithms at different stages. This motivates AutoPas, a particle simulation library that dynamically selects algorithms to minimize node-level energy use. We highlight AutoPass energy-saving potential and key challenges, such as lack of fine-grained energy measurements, noise, and hardware portability. Solutions and future directions are discussed, with molecular simulation examples from md-flexible, an MD simulator built on AutoPas.

Manish K. Mishra, Hans-Joachim Bungartz
Technical University of Munich
manishk.mishra@tum.de, bungartz@cit.tum.de

CP9

Hybrid Neural Modelling

Hybrid Neural Models combine mechanistic models with neural networks to capture unknown componentsranging from fixed parameters to time-dependent processes or even missing states. This approach offers the flexibility and predictive power of neural networks while retaining the interpretability and analytical structure of mechanistic models. It also forces researchers to confront what prior knowledge and constraints to encode, balancing expressivity with tractability.In this talk, I will outline the methodological opportunities and challenges of hybrid neural modelling, with examples from epidemiology (SIRS models), neuroscience (neurotransmission). A central motivation comes from systems where rates or parameters evolve over time: calcium ion influx modulates vesicle fusion in neurotransmission, while vaccination uptake shapes epidemic dynamics. Estimating such time-varying quantities is difficult when their form and drivers are unknown. Hybrid neural ordinary differential equations address this by embedding mechanistic models within an ODE framework and using neural networks to learn missing dynamics directly from time-series data, without prior assumptions about parameters or rate functions. This enables joint training across datasets, allowing rapid inference for new data and improved forecasting. Finally, I will also show how hybrid neural ODEs can be combined with manifold learning to quantify uncertainty, further broadening their applicability.

Thomas Gaskin
Department of Methodology
London School of Economics and Political Science
t.gaskin@lse.ac.uk

Anastasia Bankowski
Zuse Institute Berlin
bankowski@zib.de

CP9

Multiresolution Analysis Based Convolution in Hybrid Environments

The convolution operator in the multiresolution analysis (MRA) framework considered here enables efficient low-separation-rank representations of operators, addressing the challenge of extending these representations to multiple dimensions in practical applications. Within this framework, operators are evaluated using only one-dimensional integrals. Moreover, the vanishing moment property of the multiwavelet basis employed in the construction yields sparse representations for a broad class of operators. The outcome of this construction is a tree-based structure that maps naturally onto data-flow execution models. We realize this approach on top of the Template Task Graph (TTG) programming model, a distributed, task-based data-flow model that addresses key challenges in modern high-performance computing by improving programmer productivity and enabling performance portability across heterogeneous architectures. This talk will discuss the construction of the convolution operator as a data-flow graph-based algorithm in TTG and how GPUs are utilized to accelerate this essential operator. To the best of our knowledge, this ongoing work represents the first major step toward GPU-accelerated convolution in MRA.

Nilesh Chaturvedi
Institute for Advanced Computational Science
AMS Department, Stony Brook University
nilesh.chaturvedi@stonybrook.edu

Joseph Schuchart
Institute for Advanced Computational Science
Stony Brook University
joseph.schuchart@stonybrook.edu

Robert J. Harrison
Institute for Advanced Computational Science

AMS Department, Stony Brook University
robert.harrison@stonybrook.edu

CP9

A Robust Xg-PINN Approach for Physics-Informed Forward and Inverse PDE Learning in Cancer Detection

Physics-Informed Neural Networks (PINNs) have emerged as a powerful tool for solving forward and inverse problems governed by partial differential equations. It integrates the residuals of physical constraints into the loss function of neural architecture. However, conventional PINNs often suffer from poor convergence, gradient pathologies, and instability, especially in stiff, multi-scale, and data-sparse regimes. Gradient-enhanced PINN addresses local sensitivity by incorporating derivative information, while extended PINN offers improvements via adaptive sampling and residual balancing. Despite these efforts, existing methods remain limited in handling residual accumulation and global error propagation. In this work, we propose the Extended Gradient-Enhanced PINN (Xg-PINN), which integrates gradient-informed loss terms with a spatio-temporal domain decomposition strategy. This allows localized, parallelizable training while enforcing physical continuity across interfaces, thus improving numerical stability and scalability. We evaluate Xg-PINN on benchmark PDEs and apply it to modelling bioheat transfer and cellular transport in breast cancer studies. Results demonstrate enhanced convergence, accuracy, and physical fidelity compared to state-of-the-art PINN variants, marking a significant step toward reliable physics-informed learning for complex scientific problems.

Scindhiya Laxmi
Indian Institute of Technology (Indian School of Mines) Dhanbad, India
scindhiya@iitism.ac.in

CP10

Parallel Locally Optimal Iterative Methods

New iterative methods, which were recently proposed for the computation of approximate low-rank factorizations, inverses, and more, are parallelized. Of particular interest are the locally subspace-optimal variants of those algorithms. All the methods proposed rely on short recurrence update formulae, some of them guaranteeing provably faster convergence than current state-of-the-art alternatives. When these methods are applied to sparse matrices, we deploy non-zero dropping strategies in order to achieve sparse approximate matrix inverses and factorizations with controlled memory consumption. We address how random sketching can be used to further decrease memory traffic, and speed-up computations, while maintaining acceptable accuracy. In this talk, we pay particular attention to partitioned variants of the proposed methods with the intent to set the basis for parallel implementations. Experiments are performed on publicly available matrices to showcase how the newly proposed partitioned schemes perform in comparison to current state-of-the-art methods.

Nicolas Venkovic
NXP Semiconductors
nicolas.venkovic@tum.de

Amir Bouslama, Aryaman Jeendgar, Hartwig Anzt
Technical University of Munich
amir.bouslama@tum.de, aryaman.jeendgar@tum.de, hartwig.anzt@tum.de

CP10

Matrix Series Approximation technique for Solving the Riccati Matrix Differential Equation

The Riccati matrix differential equation (RDE) is of significant importance in modern control theory and practical engineering applications. In this paper, we first prove, based on the properties of controllable and observable systems, that a matrix series is convergent, which is the solution to a continuous algebraic Lyapunov equation. With the aid of the convergent matrix series and the Hermitian negative definite solution of a continuous algebraic Riccati equation, we derive a new expression for the solution of the RDE. The convergence analysis and error analysis of this method are also given. Then, we provide a novel algorithm based on the matrix series approximation technique to solve the RDE. Finally, we show the effectiveness and superiority of the derived algorithm through some numerical experiments.

Ze Zhang
Xiangtan University
Xiangtan University
zezhang@smail.xtu.edu.cn

Jianzhou Liu
Xiangtan University
liujz@xtu.edu.cn

Bastien Vieublé
Chinese Academy of Sciences
bastien.vieuble@amss.ac.cn

Juan Zhang
Xiangtan University
zhangjuan@xtu.edu.cn

CP10

CDFCI: High-Performance Parallel Software for Large-Scale Many-Body Eigenvalue Problems

We present CDFCI, a high-performance shared-memory parallel software package for computing low-lying eigenpairs of large-scale, non-relativistic fermionic Hamiltonians, arising in electronic structure and condensed matter simulations. CDFCI utilises a coordinate-descent-based selected configuration interaction (CI) approach, while leveraging dynamic task scheduling to achieve efficient workload distribution on modern multi-core CPU architectures. Numerical experiments on representative ab-initio electronic Hamiltonians and lattice models demonstrate state-of-the-art numerical accuracy, near-linear strong scaling up to 256 cores, and competitive performance when compared with existing selected CI and DMRG implementations. The software is open-source, well documented, and provides a Python interface facilitating integration into workflows such as PySCF. This is joint work with Zhe Wang, Jianfeng Lu, and Yingzhou Li.

Yuejia Zhang
Fudan University

yuejiazhang21@m.fudan.edu.cn

CP10

A New Pdlp Solver for Highs

The primal-dual hybrid gradient method for LP (PDLP) is an exciting new practical solution technique for very large scale sparse linear programming (LP) problems that can be run entirely on GPUs. HiGHS is the world's best open-source linear optimization software, and the cuPDLP-C implementation of PDLP in HiGHS is being replaced by a new in-house solver HiPDLP. Based on established techniques, and novel enhancements, HiPDLP is written in C++ and CUDA. This talk will discuss the novel enhancements in HiPDLP, the techniques for running it on GPUs using CUDA, and its performance relative to other PDLP implementations.

Yanyu Zhou
University of Edinburgh
s2032537@ed.ac.uk

CP11

Benefits and Challenges of Large-Scale Block-Adaptive Fluid Simulations for Turbulent Flow Using Wavelets.

Current state-of-the-art fluid simulations are limited by available computing resources. The main performance factors can be the grid resolution and efficiency of the chosen physics model. We present a Wavelet Adaptive Block-Based solver for Interactions with Turbulence (WABBIT) for computing 3D flows in simple and complex geometries. The solver dynamically evolves octree grids of uniform blocks. These efficiently balance the refinement of a solution and hence the memory and computational cost with the accuracy requirements for adequate representation of relevant flow characteristics. Distributing the blocks among MPI processes permits an efficient parallelization on large-scale supercomputers, and adaptation uses a wavelet decomposition with lifted bi-orthogonal interpolating wavelets. The flow is solved using the artificial compressibility method (ACM), avoiding solving a Poisson problem. In this contribution, we investigate the performance scaling of the adaptive solver depending on the problem size and block size. Furthermore, the difference in adaptivity and solution accuracy for the ACM in comparison to a classical projection method is investigated. Comparisons are done for the 3D Taylor Green Vortex test-case, where a smooth initial condition develops into decaying turbulence, and for flapping flight of a bumblebee. Results show the challenges for an effective load balancing of block computations and inter-block communication for largely parallel fluid simulations.

Julius Bergmann
Aix-Marseille University
Technical University Berlin
Julius.Bergmann@univ-amu.fr

Thomas Engels
Institut des Sciences du Movement
CNRS & Aix-Marseille Unviersity
thomas.engels@univ-amu.fr

Angela Busse
Institute of Fluid Dynamics and Technical Acoustics
Technical University Berlin
angela.busse@tnt.tu-berlin.de

Kai Schneider
Aix-Marseille University
I2M-CNRS, Centre de Mathematiques et Informatique
kai.schneider@univ-amu.fr

CP11

Shallow Water Simulations with Sycl on Multi-Accelerator Systems

The shallow water equations model fluid flow in regions where the horizontal scale is much larger than the vertical depth. Applications of the shallow water equations include the modelling of tides, tsunamis or atmospheric flows. In this talk, we will discuss a SYCL implementation of a discontinuous Galerkin discretization for the two dimensional shallow water equations with application to coastal ocean simulations. The implementation uses unstructured triangular meshes to represent the computational domain and targets CPUs, GPUs and FPGAs. We will discuss which abstractions and specializations are necessary to support FPGAs alongside CPUs and GPUs without changing the numerical algorithm. Benchmarks on realistic domains show the efficiency on different architectures and scalability to multi-node systems.

Markus Büttner
University of Bayreuth
markus.buettner@uni-bayreuth.de

Christoph Alt
Paderborn University
christoph.alt@uni-paderborn.de

Tobias Kenter
Paderborn Center for Parallel Computing, Germany
kenter@uni-paderborn.de

Vadym Aizinger
University of Bayreuth
vadym.aizinger@uni-bayreuth.de

CP11

Nextgenpb: A Modern Parallel Finite Element Solver for the PoissonBoltzmann Equation

The PoissonBoltzmann equation (PBE) is central to modeling biomolecular electrostatics, yet achieving accurate solutions typically requires costly grid refinement. Although the linear PBE is widely used in computational biology, only a few solvers exist, and most are legacy codes that are difficult to maintain, extend, and adapt to new hardware. To address this, we developed NextGenPB (NGPB), a finite elementbased adaptive solver built on modern numerical and programming practices to ensure portability across current and future CPU-based architectures and to provide a codebase that is accessible to 21st-century developers. NGPB leverages analytical surface corrections to enhance accuracy without refinement and applies efficient boundary conditions via local grid de-refinement, reducing computational demand while preserving solution quality. We validate NGPB on extended analytical benchmarks involving multiple dielectric spheres and on realistic biomolecular systems. Results show improved accuracy-to-cost ratios compared to existing solvers. Thanks to these advances, NGPB achieves state-of-the-art accuracy with strong parallel scalability on distributed-memory systems. More broadly, the core ideasanalytical corrections at discontinuous interfaces combined with adaptive mesh con-

troloffer a general and scalable strategy for solving PDEs with discontinuous coefficients in high-performance computing contexts.

Vincenzo Di Florio
MOX Laboratory, Politecnico di Milano
vincenzo.diflorio@polimi.it

Carlo De Falco
MOX Modeling and Scientific Computing, Dept. of Mathematics
Politecnico di Milano
carlo.defalco@polimi.it

Walter Rocchia, Sergii Siryk, Sergio Decherchi
Istituto Italiano di Tecnologia
walter.rocchia@iit.it, sergii.siryk@iit.it, sergio.decherchi@iit.it

CP11

Adaptive Pfasst for Shallow Water Equations on the Rotating Sphere

Recent approaches have been developed, applied and analyzed to extend parallelization in simulations to the temporal dimension. Such methods are especially demanded for simulations characterized by large temporal domains, such as weather and climate modeling. Shallow Water Equations Environment for Tests (SWEET) is a barotropic solver employed as a mini-application representing a challenging component of weather simulation within the dynamical core. In recent work, the evaluation of Parallel Full Approximation in Space and Time (PFASST) has been shown to provide high-order accurate solutions for up to 16 processors in SWEET[https://doi.org/10.1016/j.jcp.2019.109210]. Further scaling reduced wall-clock time but led to increasing errors. Convergence of PFASST is problem-dependent, and the optimal number of parallel time steps may vary. In this work, we extended SWEET with Dynamic Processes with Psets (DPP) [arXiv:2403.17107v1] which are design principles to support Dynamic Resource Management (DRM) in OpenMPI via extensions to the MPI sessions interface. DPP facilitates the dynamic adaptation of resources along with various parameters to investigate convergence behavior for the Gaussian bump and Galewsky benchmarks. We then developed an adaptivity criterion to perform convergence-informed adaptation of PFASST and evaluate the efficiency to showcase the benefits of DRM for PFASST.

Keerthi Gaddameedi
Technical university of munich
keerthi.gaddameedi@tum.de

Martin Schriber
Université Grenoble Alpes
martin.schreiber@univ-grenoble-alpes.fr

Dominik Huber
TU Munich
domi.huber@tum.de

Hans-Joachim Bungartz
Technical University of Munich
bungartz@cit.tum.de

Jan Fecht
TU Munich
jan.fecht@tum.de

Valentina Schueller
Lund University
valentina.schuller@math.lu.se

Martin Schulz
Technische Universität München
martin.w.j.schulz@tum.de

Tobias Neckel
Technical University Munich
neckel@cit.tum.de

CP12

Gpu-Accelerated Parallel Selected Inversion for Structured Matrices

Selected inversion is essential for applications such as Bayesian inference, electronic structure calculations, and inverse covariance estimation, where computing only specific elements of large sparse matrix inverses significantly reduces computational and memory overhead. We present an efficient implementation of a two-phase parallel algorithm for computing selected elements of the inverse of a sparse symmetric positive definite matrix A, which can be expressed as $A = LL^T$ through sparse Cholesky factorization. Our approach leverages a tile-based structure, focusing on selected dense tiles to optimize computational efficiency and parallelism. While the focus is on arrowhead matrices, the method can be extended to handle general structured matrices. Performance evaluations on a dual-socket 26-core Intel Xeon CPU server demonstrate that MOSAIC outperforms state-of-the-art direct solvers such as Panua-PARDISO, achieving up to 13X speedup on large-scale structured matrices. Additionally, our GPU implementation using an NVIDIA A100 GPU demonstrates substantial acceleration over its CPU counterpart, achieving up to 5X speedup for large, high-bandwidth matrices with high computational intensity. These results establish MOSAIC as a robust and versatile framework for large-scale selected inversion, offering a scalable solution for both many-core CPUs and modern GPU accelerators.

Esmail Abdul Fattah, Hatem Ltaief
King Abdullah University of Science and Technology (KAUST)
esmail.abdulfattah@kaust.edu.sa,
hatem.ltaief@kaust.edu.sa

Haavard Rue
KAUST
haavard.rue@kaust.edu.sa

David E. Keyes
King Abdullah University of Science and Technology (KAUST)
david.keyes@kaust.edu.sa

CP12

On-Device Wavelet Compression for 3D Scientific Data on GPUs

The increasing use of GPUs for scientific simulations often makes data movement the bottleneck. Integrating compression directly on the GPU can alleviate this in two ways: it increases the effective amount of data the GPU can store, reducing the need to transfer intermediate data to the CPU, and when transfers are unavoidable, it reduces the number of bytes moved, improving the effective

bandwidth. In this work, we present a GPU compressor for 3D floating-point data based on the Discrete Wavelet Transform (DWT) coupled with coefficient thresholding and sparse compaction into indexvalue pairs. The implementation processes independent 3D blocks, enabling each thread block to process a 3D tile end-to-end in shared memory. On an NVIDIA A100 and representative 3D datasets, we achieve compression ratios up to an order of magnitude higher than state-of-the-art GPU compressors at comparable reconstruction quality, while maintaining competitive throughput.

Pedro Caires, Aleksandar Ilic, Leonel Sousa
INESC-ID
josepedrocaires@tecnico.ulisboa.pt,
aleksandar.ilic@inesc-id.pt, leonel.sousa@inesc-id.pt

Clement Flint
INESC
@tbd

CP12

Static Load Balancing for Molecular-Continuum Flow Simulations

Molecular-continuum simulations enable a decomposition of the computational domain into molecular dynamics (MD) and computational fluid dynamics (CFD). Using an overlapping domain decomposition, fluid dynamics quantities are exchanged between MD and CFD in coupling cells, which could correspond, e.g., to the cells of the CFD mesh. My group has been contributing to the long-term development of the MD solver ls1 mardyn and developing the macro-micro-coupling tool (MaMiCo) for a decade, which is meant to support the coupling of arbitrary MD and CFD solvers. In my talk, I will discuss recent research and development with particular emphasis on load balancing, incorporated into ls1 mardyn and MaMiCo. Optimal node-level performance is achieved in ls1 mardyn by automated algorithm selection using the AutoPas library, including tuning of parallelization strategies. Load balancing at distributed-memory level is enabled by a balancing strategy, that distributes load at the granularity of coupling cells. This has been realized within MaMiCo and leverages a specialized balancing mechanism in ls1 mardyn/AutoPas. I will demonstrate the feasibility of the approaches we have taken in different settings, for example in coupled molecular-continuum simulations with inhomogeneous density distributions or when leveraging heterogeneous hardware systems.

Philipp Neumann
Deutsches Elektronen-Synchrotron, IT-Dept.
University of Hamburg, HPC & Data Science
philipp.neumann@desy.de

CP12

HPC-friendly GW in exciting: leveraging modern HPC parallelism

exciting is an open-source all-electron density-functional-theory code renowned for its Ha-level precision, with a strong focus on excited-state properties relevant to optoelectronics. Its accuracy stems from the implementation of the linearized augmented plane wave plus local orbital method, regarded as the gold standard for solving the Kohn-Sham equations. Consequently, exciting serves as a reliable benchmark for less accurate methods, i.e., pseudopotential-based approaches, albeit at a higher computational cost. Among its capabilities, exciting implements the GW method, the state-of-the-art approach for computing quasiparticle spectra. However, the previous implementation was poorly adapted to modern HPC systems, with limited parallelism and no GPU support, preventing efficient use of current supercomputers. We present a new HPC-optimized GW implementation featuring a task-based workflow that exploits symmetry operations and parallelism across k/q-points and bands, achieving up to 82% parallel efficiency on 9216 cores. To further accelerate key bottlenecks, GPU support has been added via a hybrid strategy: vendor-optimized libraries handle linear algebra routines, while compute-intensive loops are parallelized using portable OpenMP offload. This approach yields speedups of 410x across different hardware vendors and workloads. Overall, the new implementation enables large-scale, high-precision GW simulations that fully leverage the capabilities of modern HPC systems.

Martí Raya Moreno
Humboldt-Universität zu Berlin
marti.raya.moreno@physik.hu-berlin.de

Ronaldo Rodrigues Pelá
Zuse-Institut Berlin
ronaldo.rodrigues@zib.de

Claudia Draxl
Humboldt-Universität zu Berlin
claudia.draxl@physik.hu-berlin.de

CP13

Analysis of Hepatic Blood Flow Using the NavierStokesDarcy-Forchheimer Penalization Approach and Shear Stress Transport Model: Openfoam Parallel Simulation.

Liver diseases represent a significant global health challenge, contributing to high morbidity and mortality rates. Accurate assessment of hemodynamic parameters, particularly the portal pressure gradient (PPG), is vital for understanding liver function but often necessitates invasive procedures, such as catheterization of the portal vein and inferior vena cava. An alternative is the liver venous pressure gradient (HVPG), which measures the difference between free and wedged liver venous pressures; however, it too is an invasive technique. This study investigates the potential of computational hemodynamics as a non-invasive approach to accurately measure these critical liver hemodynamic values. I present a parallelized simulation of blood flow within the liver, factoring in its complex anatomy. The framework effectively integrates the transient, incompressible Navier-Stokes and Darcy-Forchheimer equations to model the liver's porous medium accurately. By treating the flow in the portal and hepatic veins as free flow while modeling the interior flow of the liver as porous, the method captures the intricate dynamics of fluid movement. I employ the $k-\omega$ Shear-Stress Transport (SST) model to assess turbulence within the blood flow. Using OpenFOAM, I achieve efficient and scalable simulations that yield detailed insights into important blood flow characteristics, such as pressure, velocity, and wall shear stress, enhancing our understanding of liver hemodynamics.

Salman Ahmad
Department of Mathematics,
The Chinese University of Hong Kong

salmanuom206@gmail.com

CP13

GPU-Accelerated ILES of Weakly Compressible Flows Using a GalerkinBoltzmann Formulation

We present a high-order implicit large-eddy simulation (ILES) approach for nearly incompressible flows based on a nodal discontinuous Galerkin (DG) discretization of the continuous Boltzmann equations. The compact, low-dissipative nature of DG is used to confine numerical dissipation within a narrow band of high wavenumbers, mimicking traditional LES behavior without explicit subgrid-scale models. The proposed approach is analyzed within the `libParanumal` Boltzmann solver by utilizing multi-GPU systems. Validation is provided on the TaylorGreen vortex at Reynolds numbers exhibiting a wide range of coherent turbulent scales, and on flow over a sphere to assess the methods ability to capture laminarturbulent transition and coexisting multiscale features. Overall, the results demonstrate the robustness of a high-order DG formulation of the Boltzmann equations for ILES of nearly incompressible flows.

Onur Ata, Atakan Aygun, Ali Karakus
Middle East Technical University
onurata@metu.edu.tr, atakana@metu.edu.tr, akarakus@metu.edu.tr

CP13

GPU Accelerated Discontinuous Galerkin Solutions of GalerkinBoltzmann Formulations for Nearly Incompressible Flows on Moving Meshes

In this talk, we introduce an arbitrary Lagrangian-Eulerian (ALE) formulation of the Galerkin-Boltzmann method, which recovers the NavierStokes equations in the nearly incompressible regime. A perfectly matched layer (PML) is incorporated into the moving domain to damp reflections from the boundaries. The numerical framework is implemented within the libParanumal solver using triangular meshes and is well-suited for multi-GPU systems. Free-stream preservation is achieved by consistently updating the mesh positions with the same temporal order as the solution field. We present moving aerodynamic test cases to demonstrate the accuracy and robustness of the proposed framework.

Atakan Aygun, Onur Ata, Ali Karakus
Middle East Technical University
atakana@metu.edu.tr, onurata@metu.edu.tr, akarakus@metu.edu.tr

CP13

A Nitsche Approach for Coupled Multiphysics Problems

In this talk, we present a mathematical and numerical model for the interaction between free fluid flow and a poroelastic medium. The Brinkmann equation is used to describe the flow within the porous domain, allowing for the inclusion of inertial effects. A thermodynamically consistent framework is proposed for modeling soft tissue perfusion. To approximate the solution, we develop a mixed finite element method based on Nitsches approach and establish the well-posedness of the discrete formulation. Furthermore, we derive a priori error estimates in the energy norm. Numerical experiments are provided to confirm the theoretical convergence rates and to demonstrate the methods capability in accurately capturing the underlying physical behavior.

Aparna Bansal
Indian Institute of Technology Roorkee
a_bansal@ma.iitr.ac.in

Nicolás Barnafi
Pontificia Universidad Católica de Chile
nicolas.barnafi@uc.cl

Dwijendra Narain Pandey
Indian Institute of Technology Roorkee
dwijpfma@iitr.ac.in

Ricardo Ruiz Baier
Monash University
ricardo.ruizbaier@monash.edu

CP14

Parallel High-Resolution Partial Fft-Gmres Algorithms for Subsurface Scattering Problems.

In this paper, we present a hybrid parallel implementation of a novel algorithm combining Partial FFT (PFFT) preconditioning with the GMRES method for large-scale subsurface scattering problems. The solver leverages high-order compact finite-difference discretizations of the 3D Helmholtz equation with spatially varying coefficients. A key innovation is the use of PFFT-based preconditioners derived from lower-order approximations to accelerate convergence. Our implementation efficiently exploits modern heterogeneous architectures through OpenMP multithreading for parallelizing tridiagonal solves across CPU cores, and CUDA streams for overlapping FFT computations and data transfers on GPUs. We analyze the computational complexity and parallel scalability of the method on realistic subsurface models with soil and mine-like targets, demonstrating significant performance improvements over conventional approaches.

Yury A. Gryazin
Idaho State University
gryazin@isu.edu

Xiaoye Sherry Li
Lawrence Berkeley National Laboratory
xsli@lbl.gov

CP14

Evaluating a Real-Time Lossy Array Compression Algorithm for a Lattice Boltzmann Solver

Computer simulations that were previously regarded as CPU-bound become gradually memory-bound as the growth in memory bandwidth cannot keep up with the much higher advancements in raw computing power. This imbalance is quantified with a relative factor of approximately 5.1 per decade since the 1990s, where a rise in memory bandwidth is met with a 5.1-times increase in relative computing power. In practical terms, comparing a NEC SX-4 from 1994 that operated at a balanced arithmetic intensity of 0.125 FLOP/Byte with an Intel Ponte Vecchio accelerator from 2023 that operates at 15.9 FLOP/Byte shows a factor of 127 in the described imbalance. Mixed-precision approaches that were traditionally used to speed up the throughput of calculations on a CPU core level now

provide speedup due to less demanded memory bandwidth. Approaches to compress arrays in a lossless or lossy manner to reduce memory bandwidth were implemented in LLNLs zfp library, although it is not able to process the data in real-time. In this work, a similar approach targeting a real-time lossy array compression (RTLAC) algorithm utilizing known value ranges of variables was developed and applied to a Lattice Boltzmann method for CFD simulations. Here, the main simulation variables are within certain bounds, making them ideal candidates for the RTLAC approach. Accuracy and performance results of the algorithm implemented in the m-AIA solver framework with multiple compression sizes will be presented.

Darjan Krijan
HLRS
University of Stuttgart
darjan.krijan@hlrs.de

Julian Vorspohl
RWTH Aachen University
j.vorspohl@aia.rwth-aachen.de

CP14

Combining Multigrid and Domain Decomposition Methods: An Interface Between Feat and Frosch

We present the development and implementation of an efficient interface between the FEAT Software library and the FROSch Trilinos package for monolithic overlapping Schwarz preconditioners of the generalized Dryja-Smith-Widlund (GDSW) type applied to fluid problems. This work is part of the StroemungsRaum project, financed by the German BMFTR (formerly BMBF) under the SCALEXA program, the German initiative to develop software in the exascale computing era. The main focus of this presentation is the design and integration challenges of coupling FEATFLOW's finite element infrastructure with FROSch's flexible domain decomposition framework from the ShyLU package within Trilinos. The interface enables FEATFLOW to leverage FROSch's advanced overlapping domain decomposition capabilities while maintaining computational efficiency.

Stephan Köhler, Oliver Rheinbach
Technische Universitaet Bergakademie Freiberg
stephan.khler@math.tu-freiberg.de,
oliver.rheinbach@math.tu-freiberg.de

CP14

Investigating Matrix Assembly Strategies for Finite Volume Methods using NeoN

Efficient matrix assembly plays a crucial role for the performance of computational fluid dynamics frameworks. This talk investigates portable performance strategies for matrix assembly of different partial differential equations required by the Finite Volume Method (FVM) using the NeoN DSL

Gregor Olenik
TU Munich, Germany)
Germany
gregor.olenik@tum.de

MS1

Using Llms for Programming Models Compilers' Implementations' Testing

Using LLMs for Programming Models Compilers' Implementations' Testing Abstract: The adoption of Large Language Models (LLMs) for software and testing tasks has grown rapidly since their introduction, but expectations around their capabilities have recently become more grounded. Ensuring the correctness of outputs produced by LLMs is essential to improving their practical value, yet no fully autonomous and comprehensive verification solutions currently exist. Hallucinations pose a significant risk when LLMs are used without careful validation, and the lack of transparency in their reasoning processes complicates trust in their results. To address these challenges while exploring effective applications of LLMs, we propose a dual-LLM framework, combining a generative model with a discriminative model to produce and evaluate large volumes of candidate outputs. We will present evaluations of multiple LLMs with varying parameter sizes, using a set of ten carefully selected metrics to assess both quality and reliability. Our findings will demonstrate that LLMs hold strong potential for automated generation and verification tasks when paired with structured evaluation strategies.

Sunita Chandrasekaran
University of Delaware, U.S.
schandra@udel.edu

MS1

Agentic Workflows for Generating and Optimizing Hpc Code

Developing performant and portable code for high-performance computing (HPC) systems is a complex and time-consuming task that typically requires deep domain expertise. In this talk, I will present a multi-agent framework designed to automate key aspects of HPC code generation using large language models (LLMs). This involves a team of specialized agents, including a Code Writer, Compiler, Tester, and Performance Analyzer, to support iterative code generation, correctness checking, and performance optimization tailored to HPC systems. This work moves us closer to the goal of intelligent, AI-driven software development for high-performance computing environments.

Harishitha Menon
Lawrence Livermore National Laboratory
gopalakrishn1@llnl.gov

MS1

Fine-Tuning Llms with Chathpc for Performance Portability and Modeling

This talk introduces ChatHPC, a process and toolchain for building AI assistants fine-tuned from existing large language models (e.g., Code Llama) to accelerate productivity in high-performance computing (HPC) software development. ChatHPC targets three critical challenges: parallelizing sequential code, optimizing existing parallel implementations, and porting applications across heterogeneous platforms. By leveraging domain-specific fine-tuning on datasets that include HPC benchmarks and real-world applications, ChatHPC enables developers to produce accurate, high-performance, and performance-portable code with significantly less manual effort. The talk also highlights how ChatHPC supports task-based runtimes to enhance portability, as well as its use in performance modelingdemonstrating its potential for heterogeneity in the

HPC environment.

Mohammad Alaul Ha Monil
Oak Ridge National Laboratory
monilm@ornl.gov

MS1

Accelerated Systems for Vector (Semantic) Search

Semantic search has become a critical enabler for applications ranging from scientific discovery to enterprise analytics and next-generation AI systems. Delivering this capability poses classic HPC challengesminimizing search latency and optimizing throughput, while managing cost, storage hierarchies, accelerator utilization, and quality of search. This talk will present our work at NVIDIA on designing high-performance algorithms and architectures for vector search, and discuss how these systems underpin not only LLM use cases such as retrieval-augmented generation (LLM-RAG), but also broader domains where fast, scalable similarity search is essential and becoming increasingly adopted.

Srinivasan Ramesh
NVIDIA
srramesh@nvidia.com

MS2

Using Numerical Profiling to Determine Where Mixed Precision is Usable

Many scientific applications are facing the challenge of determining where they can lower precision in the solvers they use when the prevailing theory indicates that highest available precision is needed. However, scientific intuition can sometimes indicate where precision can be lowered without compromising the quality of solution. A tool that allows exploration of the impact of changing precision on the solution can be of tremendous help in tuning the precision during expensive simulations. RAPTOR is a tool that allows numerical profiling of scientific applications. In this presentation I will describe numerical experiments with RAPTOR to determine where and when can precision be lowered in applications with Flash-X, a multiphysics software for modeling reactive and multiphase flows.

Anshu Dubey
Argonne National Laboratory
adubey@anl.gov

Faveo Hoerold
ETH Zurich
faveo.hoerold@inf.ethz.ch

Ivan Radanov Ivanov
Institute of Science Tokyo
ivanov.i.aa@m.titech.ac.jp

Akash V. Dhruv
Argonne National Laboratory
adhruv@anl.gov

Mohamed Wahib
RIKEN Center for Computational Science
mohamed.attia@riken.jp

Jens Domke
Satoshi MATSUOKA Laboratory
GSIC, Tokyo Institute of Technology
jens.domke+siampp@riken.jp

MS2

Mixed and Variable Precision for a Hyperbolic PDE Engine

We present experiences and findings with implementing and evaluating mixed and variable precision in the hyperbolic PDE engine ExaHyPE. ExaHyPE relies on the high-order ADER-DG algorithm (discontinuous Galerkin with ADER time stepping). Our mixed-precision implementation can compute several core kernels of ADER-DG selectively with higher or lower precision. We also present first experiences with variable precision, allowing for different precision in different parts of the computational domain. Key findings are that mixed precision can make it possible to compute solutions in half precision, though with low accuracy. For mixed precision, we find that the storage precision often dominates accuracy and high-order convergence is only observed with double precision.

Michael Bader
TU Munich
bader@in.tum.de

Marc Marot-Lassauzaie
Technical University of Munich
marc.marot@tum.de

MS2

The Ozaki Scheme for Reliable Matrix Computation

Recent accelerators favor low-precision arithmetic for performance, often at the cost of numerical accuracy.This talk presents the Ozaki scheme, a practical approach for reliable matrix computation that emulates high-precision results using low-precision operations. We outline the basic mechanism of the scheme and demonstrate its effectiveness in matrix multiplication and related numerical linear algebra problems.

Katsuhisa Ozaki
Shibaura Institute of Technology
ozaki@shibaura-it.ac.jp

MS2

Tbd Mike

Artificial intelligence (AI) has driven a huge demand for hardware accelerators and has increasingly shaped hardware design in recent years. Non-AI high-performance computing (HPC) applications have benefited from rapid, iterative hardware improvements, particularly increased memory bandwidth and a growing number of compute units per chip. However, as Moores law approaches its limits, hardware vendors face difficult trade-offs in deciding which functional units to include on future architectures. With the rise of AI workloads, double-precision units are often the first to be reduced or eliminated to make room for additional units optimized for GEMM operations. If double-precision capability is pretty low against the memory bandwidth, sparse linear algebra kernels may no longer remain memory-bound in double precision. In this work, we explore the implications of such hardware configurations by emulating reduced double-precision support for sparse linear algebra, even though current hardware can

still execute these workloads efficiently.

Yu-Hsiang Tsai
TU Munich
yu-hsiang.tsai@tum.de

MS3

Mixed Precision Krylov Subspace Methods with AMG Preconditioning

Modern supercomputer are equipped with increasingly more GPU accelerators which support computation in arithmetics with different precisions. This popularity is driving the design of new algorithms exploiting mixed-precision techniques to reduce runtime, energy and memory consumption. In this talk, we present the results of our work on mixed-precision Krylov subspace methods with algebraic multigrid (AMG) preconditioning for solving large-scale systems of linear equations. The focus is on the evaluation of various mixed-precision algorithms in terms of their accuracy, execution time, energy consumption and scalability. The algorithms are implemented using Ginkgo and Composyx numerical linear algebra libraries and executed on GPUs. We present results showing that some mixed-precision methods are able to achieve the same overall accuracy as their uniform precision counterparts, while being faster, requiring less memory and consuming less energy. We also discuss the problems and difficulties related to designing and implementing these mixed-precision methods.

Ani Anciaux-Sedrakian, Petr Vacek
IFP Energies nouvelles
ani.anciaux-sedrakian@ifpen.fr, petr.vacek@ifpen.fr

Alfredo Buttari
CNRS-IRIT-Université de Toulouse
alfredo.buttari@irit.fr

Emmanuel Agullo
INRIA
emmanuel.agullo@inria.fr

MS3

Scalable s-step Conjugate Gradient with Chebyshev Basis and GaussSeidel Gram Solve

We present a variant of the s-step Conjugate Gradient (CG) method that constructs a Chebyshev-stabilized Krylov basis and performs coefficient updates using Gauss-Seidel iterations. In s-step CG, multiple search directions are generated per outer iteration, reducing global synchronization costs. This leads to a Symmetric Positive-Definite (SPD) normal system (the Gram system) in the projected basis $P^TAP = P^Tb$, whose efficient solution is critical for performance and scalability. We prove that a single Forward GaussSeidel (FGS) iteration suffices to achieve backward error $\mathcal{O}(\varepsilon)$, where ε is machine precision, and that the projection is backward stable in the A-norm in the sense of Rozlonk, Langou, and Thomas. Experiments confirm that FGS converges rapidly, typically solving the Gram system in only a few iterations, even for large step sizes and system dimensions. The Chebyshev basis construction improves spectral conditioning within each Krylov block, while FGS implicitly enforces orthogonalization. Large-scale tests on SPD systems show convergence comparable to classical CG, with substantial reductions in synchronization per outer iteration on modern NVIDIA GPUs. These results establish Chebyshev-stabilized, GaussSeidel-enhanced s-step Krylov solvers as stable, scalable, and bandwidth-efficient for current and next-generation accelerator architectures.

Pasqua D'Ambra
Institute for Applied Computing
IAC-CNR
pasqua.dambra@cnr.it

Massimo Bernaschi
Istituto Applicazioni del Calcolo, CNR, Italy
massimo.bernaschi@cnr.it

Mauro Giovanni Carrozzo
Institute for Applied Computing IAC-CNR
maurogiovanni.carrozzo@cnr.it

Stephen Thomas
HPC and DC-GPU, Advanced Micro Devices, Austin, TX, USA
stephen.thomas2@amd.com

MS3

Stage-Parallel Implicit Runge-Kutta Methods Via Low-Rank Matrix Equation Corrections

Implicit RungeKutta (IRK) methods are highly effective for solving stiff ordinary differential equations (ODEs) but can be computationally expensive for largescale problems due to the need of solving coupled algebraic equations at each step. The key idea here is to reformulate a perturbed stage system in a stable way and to retrieve the exact solution through the solution of a Sylvester matrix equation with a known low-rank structure on the right-hand side. We focus on two major IRK familiessymmetric and collocation schemesand extend the methodology to nonlinear settings via a simplified Newton iteration. A set of numerical experiments, including ODEs derived from spatial discretizations of PDEs, confirm the effectiveness of the proposed approach.

Mariarosa Mazza
Università di Roma
mariarosa.mazza@uniroma2.it

Fabio Durastante
Università di Pisa
fabio.durastante@unipi.it

MS3

Iterative Solvers in the Exascale Era: Revisiting Domain Decomposition and Multigrid

The rapid growth of hardware parallelism, particularly through GPUs, has opened new opportunities for solving large–scale linear systems. Yet, exploiting the full performance of modern architectures remains challenging, as communication bottlenecks limit the scalability of classical iterative solvers at extreme scales. Classically, domain decomposition (DD) and multigrid (MG) have been among the most efficient iterative strategies: DD partitions problems into localized subproblems suited for parallelism, while MG accelerates convergence through hierarchical error reduction. Both, however, face limitations at scale — DD through frequent interface exchanges and potential load imbalance, MG through difficulties with complex geometries, irregular discretizations, or anisotropy, as well as its more dense communication pattern due to its

multiplicative nature. We explore DD and MG methods in the context of exascale computing, focusing on GPU–optimized MG solvers and on developing a more general framework that integrates these complementary strategies. By investigating MG techniques not only as solvers for global problems but also within DD context, we explore pathways to alleviate scaling bottlenecks. We aim to contribute to robust, architecture–aware solvers that combine the strengths of DD and MG, with the long–term goal of enabling more scalable iterative methods for next–generation scientific and engineering applications.

Ivan Prusak, Martin Kronbichler
Ruhr University Bochum
ivan.prusak@rub.de, martin.kronbichler@rub.de

Ivan Pribec
Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities
ivan.pribec@lrz.de

Enes Soydan
Ruhr University Bochum
enes.soydan@ruhr-uni-bochum.de

MS4

Scheduling Algorithm Selection Strategies for Parallel Applications

The increasing computational demands of scientific and data science applications have led to the adoption of heterogeneous high performance computing systems, which rely on hierarchical parallelism to deliver performance. On such systems, effective scheduling and load balancing are critical to maximizing application performance and achieving efficient resource utilization. In particular, the recent adoption of advanced scheduling algorithms in node-level parallelization frameworks, such as OpenMP, introduces the challenge of automatically selecting the most suitable algorithm for a given workload and system. In this work, we investigate the problem of automatic scheduling algorithm selection in applications that use OpenMP. We develop and evaluate two complementary approaches for algorithm selection: one based on expert knowledge and the other leveraging reinforcement learning (RL). Our experiments show that the RL-based approach successfully selects high-performing scheduling algorithms but incurs significant exploration costs, with the choice of reward function emerging as the most influential factor. In contrast, expert-based selection requires less exploration and effectively exploits domain knowledge, at the risk of overlooking the highest-performing scheduling algorithm choice. Looking forward, we aim to extend this methodology to hybrid OpenMP+MPI-based applications to enable coordinated scheduling decisions to efficiently exploit multiple levels of parallelism.

Reto Krummenacher
University of Basel
reto.krummenacher@unibas.ch

Jonas H. Müller Korndörfer
University of Bern
Switzerland
jonas.korndorfer@unibe.ch

Ali Mohammed, Ahmed Eleliemy
HPE HPC/AI EMEA Lab
Switzerland
ali.mohammed@hpe.com, ahmed.eleliemy@hpe.com

Quentin Guilloteau
Inria
France
quentin.guilloteau@inria.fr

Florina M. Ciorba
University of Basel
Switzerland
florina.ciorba@unibas.ch

MS4

Embedding-Based Methods for the Selection of Iterative Linear Solvers and Preconditioners

The solution of sparse linear systems is central to many large-scale scientific computing problems, where the choice of solver and preconditioner strongly correlates with overall execution time and solution quality. However, an ever-expanding number of implementations available across distinct numerical libraries makes choosing the most efficient combination (or sometimes even a stable and convergent one) for a given problem challenging, particularly for non-expert users. This work presents a novel pipeline leveraging embedding and linear modelling techniques and compares it to classic machine learning approaches to the solver-preconditioner selection problem. We then apply the developed model to a selection of Krylov solver implementations and preconditioners from the PETSc framework over different datasets. We then use different metrics to analyze the results, since the raw accuracy value can be misleading, given the imbalance in the data. Beyond raw accuracy, we analyze multiple performance metrics, since imbalanced datasets often render standard accuracy misleading. Results demonstrate that embedding-based selection achieves up to $\sim$15% higher NDCG scores and reduced variance across different metrics compared to the classic black-box approaches, highlighting the method's robustness and applicability in parallel solver selection tasks.

Hayden Liu Weng, Felix Dietrich, Hans-Joachim Bungartz
Technical University of Munich
h.liu@tum.de, felix.dietrich@tum.de, bungartz@tum.de

MS4

Algorithm Selection in High Performance Computing and Its Application to Molecular Dynamics

When developing high-performance scientific programs, numerous algorithmic choices need to be made. These range from simple parametric choices, such as OpenMP chunk size or CUDA thread block dimensions, to more fundamental algorithmic choices, like preconditioners, sparse matrix formats, or neighbour identification algorithms. Often, it is found that there is no single best algorithmic choice, and the best choice may vary depending on hardware and simulation scenario, and can sometimes change during a program. Hence, it is often unrealistic for the program developer to manually optimise for all users' use cases. In this talk, we will review automated algorithmic configuration and selection within High Performance Computing. Whilst not exhaustive, we aim to highlight a range of different motivations, methodologies, and problems within the field. In addition, we will also take a deeper dive into Molecular Dynamics and the particle simulation library AutoPas. AutoPas provides a black box particle container that aims to automatically select and tune the

optimal algorithmic configuration from its internal algorithm library of neighbour identification algorithms, data structures, and parallelisations. We will, in particular, discuss recent works on improvements to its algorithm selection process, emphasising practicality in addition to performance.

Samuel J. Newcome, Hans-Joachim Bungartz
Technical University of Munich
samuel.newcome@tum.de, bungartz@cit.tum.de

MS4

Evolving Large-Scale PDE Solvers: A Grammar-Based Design Framework Using Distributed Evolutionary Algorithms

We present a grammar-based framework that uses *context-free grammars* to infuse domain knowledge and structure into the design of large-scale PDE solvers. A population of candidate solver programs is evolved using *evolutionary algorithms*. The grammar rules guarantee that the solver codes are evolved with algorithmic and syntactic correctness. In contrast to deep learning approaches that rely on differentiable codes to learn optimal solver components, we adopt a complementary strategy: employing *evolutionary algorithms* to construct efficient PDE solvers from algorithmic building blocks. This has a key practical advantageenabling easy and non-intrusive integration with existing solver frameworks such as *hypre*. We demonstrate this approach by generating efficient and scalable multigrid methods. Numerical experiments on scientific applications using *hypre BoomerAMG* highlight the potential of our method for automated solver design.

Dinesh Parthasarathy
University of Erlangen-Nuremberg, Germany
dinesh.parthasarathy@fau.de

Harald Koestler
University Erlangen-Nürnberg
harald.koestler@fau.de

MS5

Systematic Upscaling to Link Kinetics and Moment Equations

Using moment integrals as a restriction operator, we derive a version of Brandt's systematic upscaling to link plasma kinetics with an evolution of the moments. We will interpret the tau-correction here as a numerical closure for our model, which can be compared with ansatze, such as Landau damping closure of Hammett and Perkins.

Matthew G. Knepley
University at Buffalo
Department of Computer Science and Engineering
knepley@buffalo.edu

MS5

The Landau Collision Operator in the Particle Basis

In this talk we will investigate the application of the particle basis Landau collision operator to Vlasov-Poisson-Landau systems in PETSC-PIC to determine the effectiveness of a collision operator in alleviating instabilities in PIC codes alongside particle resampling schemes by Adams et al. The primary systems of concern will be the well studied Landau damping and Two Stream instability test cases with a full split time stepping scheme for the metric bracket and symplectic Poisson bracket with particle resampling to assemble a sufficiently regular particle basis for the particle basis Landau collision integral.

Joseph Pusztay
University at Buffalo
Department of Computer Science and Engineering
josephpu@buffalo.edu

MS6

A Fully-dynamic Approximation Algorithm for Maximum Weight b-Matchings in Graphs

Matching nodes in a graph G = (V, E) is a well-studied algorithmic problem with many applications. The b-matching problem is a generalization that allows to match a node with up to b neighbors. This allows more flexible connectivity patterns whenever vertices may have multiple associations. The algorithm b-suitor [Khan et al., SISC 2016] is able to compute a (1/2)-approximation of a maximum weight b-matching in $O(|E|)$ time. Since real-world graphs often change over time, fast dynamic methods for b-matching optimization are desirable. In this work, we propose Dyn-b-suitor, a dynamic algorithm for the weighted b-matching problem. As a non-trivial extension to the dynamic Suitor algorithm for 1-matchings [Angriman et al., JEA 2022], our approach computes (1/2)-approximate b-matchings by identifying and updating affected vertices without static recomputation. Our proposed algorithm is fully-dynamic, i. e., it supports both edge insertions and deletions, and we prove that it computes the same solution as its static counterpart. In extensive experiments on real-world benchmark graphs and generated instances, our dynamic algorithm yields significant savings compared to the sequential b-suitor, e. g., for batch updates with 10^3 edges with an average acceleration factor of 10^3. When comparing our sequential dynamic algorithm with the parallel (static) b-suitor on a 128-core machine, our dynamic algorithm is still 59 to 10^4 faster.

Fabian Brandt-Tumescheit
Humboldt University of Berlin
fabian.brandt-tumescheit@hu-berlin.de

Henning Meyerhenke
Karlsruhe Institute of Technology (KIT)
meyerhenke@kit.edu

MS6

Algorithms for Dynamic Link Scheduling in Data Centers

Reconfigurable optical networks are emerging as a promising way to improve datacenter performance by adapting the physical connectivity to traffic patterns. A key challenge is to decide on the fly which direct connections between top-of-rack switches to establish. Modern optical technologies offer very high bandwidth and can reconfigure links within microseconds. To exploit this capability, however, we need scheduling algorithms that are both fast and effective. In this talk, I will show how dynamic link scheduling can be viewed through a graph-theoretic lens: finding and maintaining configurations of optical switches that maximize the communication demand that is offloaded to the optical network can be formulated as an edge-coloring problem on graphs. Building on this perspective, I will introduce algorithms that are not only theoretically

grounded, but also efficient in practice. Experiments on real-world traffic traces demonstrate that these algorithms can compute high-quality schedules with low overhead.

Kathrin Hanauer
University of Vienna
kathrin.hanauer@univie.ac.at

MS6

Algorithm Design for Emerging Architectures

To respond to the constantly increasing demand for computational resources, particularly for machine learning tasks, several emerging architectures have been proposed, including tensor cores and processing-in-memory. In recent years, several works have presented computational models and algorithmic techniques for designing and analyzing efficient algorithms that leverage such architectures. In this talk, we will present some of these results with a particular focus on irregular computations.

Francesco Silvestri
University of Padova
silvest1@dei.unipd.it

MS6

Sets in the Language of Linear Algebra: Applications and Acceleration

In programming, sets are often represented by tree data structures while for AI accelerators, tile-based programming models have taken root. This talk bridges the two through a linear algebraic lens, resulting in set implementations suitable for modern AI accelerators. When viewed from a linear algebraic lens, finite sets correspond to sparse vectors. Consider, for example, a set S with k elements $s_k \in \{0, 1, \ldots, 255\}$: S may be represented by a Boolean vector of length 256 whose positions $s_k \in S$ are true. While sparse vectors may be represented by trees, more common representations befitting sparse vectors of limited length are based on raw arrays– with the sparse accumulator (SPA) data structure a most effective choice as it supports constant-time complexity for access and modification. We show how SPA-based sparse vectors decompose into tiles, and how concurrent tile-based updates are resolved to maintain a globally consistent data structure. One application area is given by so-called incremental graph algorithms, where algorithms operate over (unordered) subsets of the full vertex set that decrease in size as the graph algorithm progresses round-by-round. We investigate the performance of such incremental graph algorithms, and, if time permits, 1) gauge the effectiveness of this set representation in other application areas and 2) explore additional tile-based set representations.

Albert Jan N. Yzelman
Computing Systems Lab
Huawei Zürich Research Center
albertjan.yzelman@huawei.com

MS7

Scaling Bayesian Inference Up, Out, and to Gpus

Bayesian posterior predictive inference can be carried out using either Monte Carlo sampling methods (e.g., Hamiltonian Monte Carlo [HMC] or sequential Monte Carlo [SMC]) or variational inference optimization methods (e.g., autodiff variational inference [ADVI] or normalizing flows [NF] based on neural networks). These algorithms face a serialization bottleneck in moving from a random initialization to the bulk of the probability mass (i.e., mixing). The structure of the statistical model determines which forms of parallelization are practical for evaluating the log density and derivatives. HMC allows multiple chains and SMC multiple particles—both may be improved with synchronous or asynchronous parallelism. ADVI is only scalable with approximate covariance methods, whereas normalizing flows scale well onto GPU because of their neural network architecture. I will conclude with a survey of practical tools for scaling models up on single machines, out across machines, and onto GPUs. Whatever the algorithm, we need to come to grips with the future of computing being ever cheaper and more scalable single-instruction multiple data (SIMD) computations with relatively slow and constrained memory.

Bob Carpenter
Flatiron Institute
bcarpenter@flatironinstitute.org

MS7

Large-Scale Bayesian Modeling for Multivariate Spatio-Temporal Gaussian Processes

Multivariate Gaussian processes (GPs) provide a flexible probabilistic framework for modeling complex interdependent phenomena. In high-dimensional settings, which frequently arise in spatio-temporal applications, they pose, however, significant computational challenges. In this talk, we present DALIA, a highly scalable framework for performing Bayesian inference tasks on spatio-temporal multivariate GPs, based on the methodology of integrated nested Laplace approximations. Our approach relies on a sparse inverse covariance matrix formulation of the GP, puts forward a GPU-accelerated block-dense approach over the arising structured sparsity patterns, and introduces a hierarchical, triple-layer, distributed memory parallel scheme. We present weak scaling performances surpassing the state-of-the-art by two orders of magnitude on a model whose parameter space is 8larger and demonstrate strong scaling speedups of three orders of magnitude when running on 496 GH200 superchips on the Alps supercomputer. Applying DALIA to an air pollution study over northern Italy spanning 48 days, where we showcase refined spatial resolutions over the aggregated interdependent pollutant measurements.

Lisa Gaedke-Merzhäuser
Università della Svizzera italiana
lisa.gaedkemerzhauser@kaust.edu.sa

Vincent Maillou
ETH Zurich
vincent.maillou0@gmail.com

Fernando Rodriguez Avellaneda
King Abdullah University of Science and Technology
fernando.rodriguezavellaneda@kaust.edu.sa

Olaf Schenk
Università della Svizzera italiana
Switzerland
olaf.schenk@usi.ch

Alexandros Ziogas
ETH Zurich
alexandros.ziogas@iis.ee.ethz.ch

Håvard Rue, Håvard Rue
King Abdullah University of Science and Technology
haavard.rue@kaust.edu.sa, haavard.rue@kaust.edu.sa

MS7

Extreme scale Bayesian inference and optimal experimental design for wave propagation-based source inversion

We consider Bayesian source inversion problems and associated optimal experimental design governed by wave propagation in very high dimensional parameter space. This class of problems is fundamentally challenging since it does not lend itself to low dimensional approximation. We achieve a real-time Bayesian inversion capability by exploiting the time-shift invariance of the wave propagation problem and resulting parameter-to-observable map, an offline–online decomposition with a modest number of offline (adjoint) wave propagation solutions along with a wave propagation-free online component, and novel parallel algorithms for the online component that map well onto GPU clusters. Capitalizing on this capability, we design novel greedy algorithms for solving sensor selection OED problems with the goal of maximizing expected information gain for both the parameter field as well as posterior predictive quantities of interest. The real-time Bayesian inversion and OED framework is demonstrated for acoustic–gravity wave propagation-based tsunami early warning in several high-risk subduction zones settings. Bayesian inverse problems with up to 1 billion inversion parameters representing spatiotemporal seafloor motion are solved in less than a second on 512 GPUs, and OED sensor selection problems for tsunami early warning with hundreds of acoustic pressure sensors are solved to near optimality in a few hours.

Sreeram R. Venkat
University of Texas Austin
201 E 24th St, Austin, TX 78712
srvenkat@utexas.edu

Stefan Henneking
University of Texas at Austin
stefan@oden.utexas.edu

Alice Gabriel
University of California, San Diego
algabriel@ucsd.edu

Omar Ghattas
University of Texas at Austin
omar@oden.utexas.edu

MS8

Distributed Computing for Physics-Based Data-Driven Reduced Modeling at Scale

High-performance computing (HPC) has enabled the simulation of complex real-world physical processes at unprecedented fidelity. In aerospace propulsion, for example, HPC is used to simulate rotating detonation rocket engines (RDREs) in support of next-generation engine design. However, such simulations can require millions of core hours, rendering them impractical for routine engineering tasks such as design space exploration and risk assessment. Reduced-order models (ROMs) offer a means to overcome this limitation by providing computationally efficient surrogate models that retain essential physical accuracy. We present a distributed-memory algorithm for the fast and scalable construction of predictive, physics-based ROMs from sparse datasets with extremely large state dimensions. The proposed method learns structured ROMs that approximate the underlying dynamical systems governing the data. We demonstrate strong scalability on up to 2,048 cores of the Frontera supercomputer at the Texas Advanced Computing Center. The approach is evaluated on a realistic RDRE problem for which one millisecond of simulated physical time requires approximately one million core hours. Using a training dataset of 2,536 snapshots, each with a state dimension of 76 million, the algorithm constructs a predictive reduced-order model in 13 seconds on 2,048 cores. Distribution Statement A: Approved for Public Release; Distribution is Unlimited. AFRL-2024-1411

Ionut-Gabriel Farcas
Virginia Tech
farcasi@vt.edu

Rayomand Gundevia
Jacobs Engineering Group, Inc.
Edwards Air Force Base
rayomand.gundevia.ctr@afrl.af.mil

Ramakanth Munipalli
Edwards Air Force Base
Air Force Research Laboratory
ramakanth.munipalli@us.af.mil

Karen E. Willcox
UT Austin
kwillcox@oden.utexas.edu

MS8

Real-Time Bayesian Inference at Extreme Scale: A Digital Twin for Tsunami Early Warning Applied to the Cascadia Subduction Zone

We present a digital twin (DT) for tsunami early warning in the Cascadia subduction zone (CSZ). This DT assimilates pressure data from seafloor sensors into an acoustic-gravity wave equation model, solves an inverse problem to infer spatiotemporal seafloor deformation, and forward predicts tsunami wave heights. The entire end-to-end data-to-inference-to-prediction computation is carried out in real time through a Bayesian framework that rigorously accounts for uncertainties. Creating such a DT is challenging due to the enormous size and complexity of both the forward and inverse problems. For example, a discretization of the spatiotemporal seafloor velocity in the CSZ the parameter field to be inferred gives rise to a system with one billion parameters. Using current methods, computing the posterior mean alone would require more than 50 years on 512 GPUs. We exploit the shift invariance of the parameter-to-observable map and devise novel parallel algorithms for fast offline-online decomposition. The offline component requires just one adjoint wave propagation per sensor; the PDE solver is implemented with MFEM and exhibits excellent scalability to 43,520 GPUs on LLNLs El Capitan system. Fast Hessian applications are enabled by an FFT-based algorithm for the resulting block Toeplitz matrices. Using this framework, the Bayesian inverse solution and wave height forecasts are computed in 0.2 seconds, representing a ten-billion-fold speedup over state-of-the-art methods.

Stefan Henneking

University of Texas at Austin
stefan@oden.utexas.edu

Sreeram R. Venkat
University of Texas Austin
201 E 24th St, Austin, TX 78712
srvenkat@utexas.edu

Veselin Dobrev, John Camier, Tzanio Kolev
Lawrence Livermore National Laboratory
dobrev1@llnl.gov, camier1@llnl.gov, kolev1@llnl.gov

Milinda Fernando
Oden Institute, UT Austin
milinda@oden.utexas.edu

Alice-Agnes Gabriel
UC San Diego
algabriel@ucsd.edu

Omar Ghattas
University of Texas at Austin
omar@oden.utexas.edu

MS8

Fused ensembles of dynamic-rupture earthquake simulations to accelerate Bayesian inference

We investigate fused ensemble simulations as a strategy to accelerate large earthquake modeling workflows with varying parameters, such as required for Bayesian parameter inference with uncertainty quantification (UQ). We implemented fused earthquake simulations in the SeisSol software package. By extending the degrees of freedom tensor, we improve the efficiency of the performance-dominant wave-propagation kernels. Depending on the system architecture, we achieve a speedup of up to 4.56x (for lower-order simulations on the NVIDIA Grace-Grace Superchip) and up to 35% savings in node hours for a full Parallel Multi-Level Delayed Acceptance (MLDA) inversion workflow on the Vista supercomputer.

Vikas Kurapati, David Schneller
Technical University of Munich
vikas.kurapati@tum.de, david.schneller@tum.de

Linus Seelinger
Karlsruhe Institute of Technology (KIT)
Scientific Computing Center
mail@linusseelinger.de

Zihua Niu
Ludwig-Maximilians-Universität München
zihua.niu@lmu.de

Alice Gabriel
University of California, San Diego
algabriel@ucsd.edu

Michael Bader
Technical University of Munich
bader@in.tum.de

MS8

Accelerating Analyst Workflows Via Alternating Schwarz-Based Coupling and Model Order Reduction

This talk will describe a novel domain decomposition-based approach for creating adaptive hybrid models with the help of the Schwarz alternating method (SAM). In this approach, the solution on the full domain is obtained via an iterative process in which a sequence of subdomain-local problems are solved, with information propagating between subdomains through transmission boundary conditions (BCs). The models being coupled can be subdomain-local full order models (FOMs) and/or subdomain-local reduced order models (ROMs). We will present some recent extensions of SAM to enable the overlapping and non-overlapping coupling of non-intrusive ROMs constructed via Operator Inference (OpInf). We will show numerical results that demonstrate the SAMs potential to accelerate analyst workflows by simplifying the meshing step of the mod/sim process and by enabling the plug-and-play integration of data-driven models into mod/sim workflow. Time permitting, we will additionally discuss some perspectives towards enabling on-the-fly switching between subdomain-local models of varying fidelities within the SAM framework.

Irina K. Tezaur, Anthony Gruber
Sandia National Laboratories
ikalash@sandia.gov, adgrube@sandia.gov

Ian Moore
Department of Mathematics
Virginia Tech
ianm9123@vt.edu

Eric Parish
Sandia National Lab
eric.parish9@gmail.com; ejparis@sandia.gov

Chris Wentland
Sandia National Laboratories
crwentl@sandia.gov

Cameron Rodriguez
Columbia University
camjohnrod@gmail.com

MS9

Asynchronous and Distributed Column Generation

We study how to rapidly obtain strong dual bounds for large-scale generic mixed integer programs (MIPs) when high-performance, multi-core and distributed computing resources are available. To this end, we focus on column generation methods applied to extended formulations derived from DantzigWolfe decompositions. We present both asynchronous and distributed variants of column generation that relax the strict synchronization of classical approaches, while preserving optimality guarantees. By decoupling the control flow between the master problem and the pricing subproblems, pricing can be solved in parallel on different dual solutions, while the master problem incorporates improving columns as soon as they become available. Computational experiments on large-scale benchmark instances show that the proposed approaches effectively exploit parallel and distributed resources and achieve substantial speedups, particularly on large instances, when compared to state-of-the art solvers and standard parallelizations of classical column generation.

Saverio Basso

USI-SUPSI, CH
saverio.basso@supsi.ch

Alberto Ceselli
University of Milan
alberto.ceselli@unimi.it

MS10

A Composable Abstraction of Hierarchical Methods for Matrix-Vector Product Acceleration

Numerous hierarchical methods have been developed to speed up matrix-vector multiplications involving hierarchically rank-structured matrices, which have low-rank off-diagonal blocks. They have been demonstrated to reduce the computational costs and memory requirements from quadratic to log-linear and even linear complexity while maintaining high accuracy. Examples of these methods include the Fast Multipole Method (FMM) and Hierarchical Matrices, as well as their variants: Hierarchically Semi-Separable (HSS), Hierarchically Off-Diagonal Low-Rank (HODLR), and Block Low-Rank (BLR) matrices, which are their flat counterparts. All of these methods essentially aim to tessellate the matrix into blocks and approximate the low-rank ones using different strategies. In this work, we leverage this common foundation to provide a unifying framework. We propose an abstraction encompassing all these methods which is composable, meaning it is made of small components that can be changed and tailored to fit the specific needs of different applications. Along with this abstraction, we provide a generic hierarchical matrix-vector algorithm that remains unchanged regardless of how the abstraction's components are specialized. This offers an opportunity to reduce the implementation effort required for hierarchical methods. Experimental validation shows that our generic algorithm yields comparable numerical results to those of the original algorithms.

Antoine Gicquel
Inria
antoine.a.gicquel@inria.fr

MS10

Mixed Precision Hodlr Matrices

Hierarchical matrix computations have attracted significant attention in the science and engineering community as exploiting data-sparse structures can significantly reduce the computational complexity of many important kernels. One particularly popular option within this class is the Hierarchical Off-Diagonal Low-Rank (HODLR) format. In this talk, we show that the off-diagonal blocks of HODLR matrices that are approximated by low-rank matrices can be represented in low precision without degenerating the quality of the overall approximation. We also present an adaptive-precision scheme for constructing and storing HODLR matrices, and we prove that the use of mixed precision does not compromise the numerical stability of the resulting HODLR matrix–vector product and LU factorization. Our analyses further give insight on how one must choose the working precision in HODLR matrix computations relative to the approximation error in order to not observe the effects of finite precision. Intuitively, when a HODLR matrix is subject to a high degree of approximation error, subsequent computations can be performed in a lower precision without detriment. A range of numerical experiments is presented to demonstrate the validity of our theoretical results. This is a joint work with Erin Carson and Xinye Chen.

Xiaobo Liu
MPI for Dynamics of Complex Technical Systems, Magdeburg
xliu@mpi-magdeburg.mpg.de

MS10

On a Shrink-and-Expand Technique for Symmetric Block Eigensolvers

In symmetric block eigenvalue algorithms, such as the subspace iteration algorithm, the locally optimal block preconditioned conjugate gradient (LOBPCG) algorithm, the steepest descent (SD) algorithm, and the trace minimization (TraceMIN) algorithm, a large block size is often employed to achieve robustness and rapid convergence. However, using a large block size also increases the computational cost. Traditionally, the block size is typically reduced after convergence of some eigenpairs, known as deflation. In this work, we propose a non-deflation-based, more aggressive technique, where the block size is adjusted dynamically during the algorithm. This technique can be applied to a wide range of block eigensolvers, reducing computational cost without compromising convergence speed. We present three adaptive strategies for adjusting the block size, and apply them to four well-known eigensolvers as examples. Detailed theoretical analysis and numerical experiments are provided to illustrate the efficiency of the proposed technique. In practice, an overall acceleration of 20% to 30% is observed.

Yuxin Ma
Charles University
yuxin.ma@matfyz.cuni.cz

MS10

Accelerating the Solution of Sparse Linear Systems Using a Mixed-Precision Nested Krylov Method

Low-precision computing is becoming increasingly important for extracting maximum performance from todays computing resources. While mixed-precision strategies have been extensively studied for iterative sparse linear solvers, fully exploiting half-precision arithmetic remains challenging. In this talk, we present a novel nested Krylov solver that integrates FGMRES and Richardson iterations within a deeply nested framework. The solver gradually reduces arithmetic precision toward half precision in the innermost iterations. To preserve the effectiveness of computation, the low-precision inner solvers are restricted to a few iterations at a time. However, the nesting structure ensures their frequent application and good convergence. Numerical experiments show that this approach not only integrates half-precision while preserving convergence speed, but also delivers significant performance gains, achieving speedups of up to 2.42× compared to a double-precision implementation. Moreover, compared to standard Krylov solvers, such as CG and BiCGStab, our method achieves speedups exceeding 2×.

Kengo Suzuki
Kyoto University
suzuki.kengo.8m@kyoto-u.ac.jp

MS11

Scaling Training and Inference of Large Language

Models on Gpu-Based Supercomputers

Training and fine-tuning large language models (LLMs) with hundreds of billions to trillions of parameters requires tens of thousands of GPUs, and a highly scalable software stack. In this talk, I will present a novel four-dimensional hybrid parallel algorithm implemented in a highly scalable, portable, open-source framework called AxoNN. We have implemented several performance optimizations in AxoNN, which have resulted in unprecedented scaling and peak flop/s (bf16) for training of GPT-style transformer models on Perlmutter (620.1 Petaflop/s), Frontier (1.381 Exaflop/s) and Alps (1.423 Exaflop/s). As LLMs grow in parameter count and context length, efficient generation with these models requires scaling of inference frameworks beyond a single node. I will also describe our work on YALIS (Yet Another LLM Inference System), a framework developed on top of AxoNN that supports scalable tensor parallelism, optimized kernels, and pluggable attention backends. Using YALIS, we conduct a systematic study of strong scaling for LLM inference on modern GPU supercomputers, including NVIDIA GH200 (Alps) and A100 (Perlmutter) systems.

Abhinav Bhatele
Lawrence Livermore National Laboratory
bhatele@llnl.gov

MS11

Llms Meet Hpc: Profiling Challenges and Opportunities

Large Language Models (LLMs) have emerged as powerful tools for accelerating scientific discovery, but their training and fine-tuning remain computationally demanding and resource intensive. The scale, irregularity, and dynamic behavior of LLM workloads make them particularly well suited for performance analysis through the lens of high-performance computing (HPC) expertise and tools. In this talk, I will present our ongoing effort to apply HPC-style profiling methodologies, such as runtime tracing, FLOPs accounting, and fine-grained memory analysis, to better understand and optimize the performance of LLM training. Specifically, I will discuss how input length variability and sequence packing strategies impact GPU utilization, load balance, and overall throughput during fine-tuning. I will also introduce profiling-aware packing approaches that improve efficiency by aligning data characteristics with system behavior. In addition, I will highlight memory profiling techniques that capture allocation and deallocation patterns across training steps, helping identify inefficiencies and root causes of out-of-memory failures. This work demonstrates how HPC-driven performance engineering can guide the development of more efficient and scalable LLM training pipelines, while also motivating the design of new tools and systems within the HPC community to meet the emerging demands of large-scale AI.

Gokcen Kestor
Pacific Northwest National Laboratory
gokcen.kestor@pnnl.gov

MS11

High Performance Communication Library and Transport for Llm Training at 100K+ Scale

Each successive generation of the LLaMA model has demonstrated substantial growth in both model size and complexity. The largest multimodal mixture-of-experts model within our LLaMA4 series possesses nearly two trillion total parameters, with 288 billion active parameters and 16 experts. To accommodate the computational demands associated with training such a colossal model, we expanded our AI clusters, deploying approximately 100,000 GPUs. GPU-to-GPU communication latency is a critical factor when coordinating such a vast number of GPUs. Even microsecond delays accumulate across thousands of nodes, consequently impacting the time required for training. We engineered the underlying network infrastructure to provide the necessary backbone for high-speed GPU-to-GPU communication, concurrently innovating our communication library stack to enhance overall communication efficiency. In this presentation, we will provide an overview of the network topology deployed within Meta datacenters and introduce a range of communication optimizations and custom features that facilitated LLaMA4 training through cross-layer codesign, encompassing model algorithms, collectives, and extending to the network transport layer.

Min Si
Facebook, U.S.
minsi.atwork@gmail.com

MS11

Communication-Efficient Optimizers and Reasoning with Sparse MoEs

This talk is based on two of our recent publications Lion Cub: Minimizing Communication Overhead in Distributed Lion and Optimal Sparsity of Mixture-of-Experts Language Models for Reasoning Tasks. The former describes how the Lion optimizers sign-based update can be combined with majority voting collectives and selective momentum synchronization to yield 5x speedup for distributed training. The latter shows how the sparsity in MoEs affect the downstream performance of memorization vs. reasoning tasks, depending on the combination of training data vs. evals.

Rio Yokota
Tokyo Institute of Technology
rioyokota@rio.gsic.titech.ac.jp

MS12

Graph Algorithms on Dataflow Architecture

NextSilicon's dataflow architecture (ICA: Intelligent Compute Accelerator) is a non-Von Neuman architecture. As such, it enables efficient running of graph algorithms that don't lend themselves to acceleration on standard CPUs/GPUs. In the talk we'll explain the ICA architecture and give examples of algorithms that can efficiently run on it.

Oded Margalit
NextSilicon
oded.margalit@nextsilicon.com

MS12

Reordering in the Age of Matrix Units

Reorderingthe process of permuting the rows and columns of sparse matrices-is a crucial strategy for accelerating high-performance sparse linear algebra, as it rearranges nonzero entries to fit modern architectural components such as matrix units. Although reordering techniques achieve modest average speedups compared to hardware-specific kernel optimisations, their benefits come without

the cost of modifying or even interfacing with application-specific hardware or software, greatly widening their potential range of applicability. In principle, these techniques could be applied as a pre-emptive measure to treat ill-behaved sparse matrices, regardless of the applicationa blanket preprocessing solution. Reordering problems are, however, notoriously (NP-)hard, computationally treatable only when driven by heuristics. Heuristics and quality metrics are also fundamental to evaluate the quality of reordering without running matrices through all their possible applications. What are, then, good heuristics for the age of matrix units?

Sylos Labini Paolo, Flavio Vella
University of Trento
paolo.syloslabini@student.unibz.it, flavio.vella@unitn.it

MS12

High-Level Synthesis of Multithreaded Accelerators for Irregular Applications and Graph Analytics

Irregular applications such as graph analytics pose challenges to traditional hardware design due to control divergence, memory latency, and load imbalance. We discuss two high-level synthesis (HLS) methodologies that generates efficient multithreaded accelerators from OpenMP-annotated parallel C/C++ programs. Our approach is embodied in two frameworks: SVELTO, a multithreaded architecture template which targets fine-grained parallelism extraction and scheduling from irregular control structures, and SPARTA, a followup scalable design that supports many more constructs of the OpenMP runtime. They enable productive hardware generation for a range of irregular workloads. We demonstrate significant gains in throughput and efficiency across key graph and sparse analytics kernels, outperforming conventional HLS flows.

Antonino Tumeo
Pacific Northwest National Laboratory
Antonino.Tumeo@pnnl.gov

MS12

A Supercomputer on a Chip: Parallel Algorithms on the Cerebras WSE

Sparse large-scale problems are critical for numerous scientific and machine learning domains, but they are notoriously difficult to parallelize. Accelerating these computations is vital, yet scaling them on conventional hardware remains a significant challenge. The novel Cerebras wafer-scale architecture, however, has shown considerable promise for these sparse algorithms due to its massive parallelism and unique memory fabric. Efficient implementation still requires rethinking traditional programming concepts so algorithms can be adapted to this spatial hardware. We will discuss a few successful problem mappings and learnings thereof and cover future research directions and open problems.

Max Zhao
Cerebras
max.zhao@cerebras.net

MS13

Exposing New Recurrences and Parallelism in Key Computations Within the Non-Equilibrium Green's Function Formalism

The non-equilibrium Greens function method is a powerful tool for the study of electronic transport in materials. A numerical challenge in this technique is the solution of Dyson and Keldysh problems. The former consisting of the extraction of the block tridiagonal part of a block tridiagonal matrix, $D = b3diag(T^{-1})$, and the latter is a contraction of the form $K = T^{-1}\Sigma T^{-*}$, where Σ is also block tridiagonal. In this talk, we first recapitulate the well-known Recursive Greens Function (RGF) method, a sequential algorithm for computing D. Then, inspired by RGF we briefly outline a novel formulation of Recursive Keldysh (RK), again a sequential procedure, which gives us K without having to fully form T^{-1}. Finally, we motivate our parallel formulation of RGF, and present numerical results with a Julia implementation running on various heterogeneous supercomputing architectures. The algorithms presented here, although focusing on the block n-diagonal case with $n = 3$, can be easily extended to $n > 3$.

Edoardo A. Di Napoli
Juelich Supercomputing Centre
e.di.napoli@fz-juelich.de

Gustavo Ramirez-Hidalgo
Forschungszentrum Jülich
g.ramirez.hidalgo@fz-juelich.de

MS13

A Parallel Task-Based Algorithm for the Solution of the Non-Equilibrium Green's Function

Quantum transport calculations, such as the NEGF method, tend to be more challenging regarding time and memory consumption for the new circuits. We first aim to accelerate the computation by using Selected Inverse on retarded Green's function (G^r). This method takes advantage of the sparsity of G^r to only compute the inverse of nonzeros 'block', which provides a good trade-off between performance and fill-in effect (arising from factorization). We also investigate the acceleration from parallel computation using the Nested Dissection method on G^r, which will control the fill-in effect by first reordering blocks.

Matthieu Robeyns
CNRS-IRIT
matthieu.robeyns@irit.fr

MS14

Partitioning Trillion Edge Graphs on Edge Devices

Processing large graphs with billions of entities is critical in fields like bioinformatics, high-performance computing, and navigation. Efficient graph partitioning, which divides a graph into subgraphs while minimizing inter-block edges, is essential to graph processing, as it optimizes parallelism and enhances data locality. Traditional in-memory partitioners like METIS and KaHIP offer high-quality partitions but are often infeasible for huge graphs due to large memory overhead. Streaming partitioners reduce memory usage to O(n), where n is the number of nodes, by loading nodes sequentially and assigning them to blocks on-the-fly. This paper introduces StreamCPI, a novel framework that reduces the memory overhead of streaming partitioners through run-length compression of block assignments. StreamCPI enables partitioning of trillion-edge graphs on edge devices. Within this framework, we propose a modification to the LA-vector bit vector for append support,

usable for online run-length compression across streaming applications. Empirical results show that StreamCPI reduces memory usage while maintaining or improving partition quality. Using StreamCPI, the Fennel partitioner partitions a graph with 17 billion nodes and 1.03 trillion edges on a Raspberry Pi, achieving significantly better quality than Hashing, a widely used feasible approach on edge devices. StreamCPI thus advances graph processing by enabling high-quality partitioning on low-cost machines.

Adil Chabra
University of Heidelberg
adil.chhabra@informatik.uni-heidelberg.de

Florian Kurpicz
Karlsruhe Institute of Technology
kurpicz@kit.edu

Christian Schulz
Heidelberg University, Germany
christian.schulz@informatik.uni-heidelberg.de

Dominik Schweisgut
Humboldt University of Berlin
dominik.schweisgut@hu-berlin.de

Daniel Seemaier
Karlsruhe Institute of Technology
daniel.seemaier@kit.edu

MS14

ScaleRunner: A Fast Mpi-Based Random Walk Engine for Multi-Cpu System

Random walks (RWs) on graphs have a plethora of applications, both in theory and practice. One of the currently most important applications is representation learning (RL) finding a suitable embedding of a graph into some low-dimensional geometric space. The demand for fast RW algorithms lead to a variety of RW engines targeting different computing architectures. In this paper, we address multi-CPU systems and aim at improving upon existing random walk engines such as KnightKing when running first- and second-order RW algorithms. To this end, we introduce ScaleRunner, a C++ library with full CMake integration that executes random walks in parallel. Our main acceleration techniques for ScaleRunner are: (i) each random walk is modeled as a task deployed to a thread-pool, balancing the work load on each CPU separately; (ii) integration of the dynamic graph data structure DHB to speed up graph data caching operations; (iii) collective MPI I/O routines to speed up graph input, path output, and postprocessing operations. Our experiments use a variety of popular benchmark graphs to execute RW algorithms commonly used in RL applications. On average, ScaleRunner speeds up first-order RWs by one order of magnitude and second-order RWs by two orders compared to KnightKing.

Henning Meyerhenke
Karlsruhe Institute of Technology
henning.meyerhenke@kit.edu

Florian Willich
Humboldt University of Berlin
f.willich@hu-berlin.de

MS14

Multilevel Anomaly Detection in Large-Scale Directed Graphs.

We consider the problem of detecting anomalous nodes in large-scale directed graphs. Our method leverages graph partitioning principles to perform anomaly detection in a multilevel fashion. During the coarsening phase, nodes and edges that can be confidently classified as inliers are identified and merged, based on statistical tests applied to both structural properties of the graph and to the weights of nodes and edges. Outlier detection at the coarsest level is then carried out using spectral localization on the reduced graph, while the refinement phase ensures that the neighbors of detected outliers are not themselves part of anomalous substructures. The effectiveness of the proposed unsupervised framework is demonstrated though numerical experiment on synthetic graphs and on graphs that simulate real-world financial transaction behaviour.

Dimosthenis Pasadakis, Malik Lechekhab
Institute of Computing
Università della Svizzera italiana
dimosthenis.pasadakis@usi.ch, malik.lechekhab@usi.ch

Olaf Schenk
Università della Svizzera italiana
Switzerland
olaf.schenk@usi.ch

MS14

Space-Time Efficient Compression of All Resistance Distances on a Large Sparse Network

The resistance distances, a robust alternative to the geodesic distances, on a network/graph $G(V, E)$ play a significant role in exploratory network analysis across various application domains [Klein & Randic, 1993]. However, it is challenging to make the distance matrix $Q(G)$ available for a large sparse network, due to the space and time scaling issues: the amount of memory for storing Q scales quadratically with $n = |V|$, the arithmetic cost for computing Q by its exact representations scales cubically with n in the worst case. We present the ResQ method for compressive representations of matrix Q for an undirected graph G, with near-linear memory consumption and significantly reduced execution time. We split Q to a k-nearest-neighbor (knn) matrix and a far-neighbor matrix. The knn matrix is exact or numerically accurate, preserving the manifold structure in this metric space in order to enable or facilitate subsequent data analysis tasks. The far-neighbor matrix is compressed algebraically and losslessly (case A) or with a loss (case B), determined by an internal criterion. To reduce the total cost, ResQ schedules batch operations according to neighbor propagation and accelerates Q construction via solving $b \ll n$ sparse M-subsystems per batch with neighbor-reference prediction. We demonstrate with empirical results on large graphs that ResQ enables resistance-distance analysis that was previously infeasible.

Dimitris Floros
Nicholas School of the Environment
Duke University
dimitrios.floros@duke.edu

Nikos Pitsianis

Department of Electrical & Computer Engineering
Aristotle University
pitsiani@ece.auth.gr

Xiaobai Sun
Department of Computer Science
Duke University
xiaobai@cs.duke.edu

MS16

Scaling Simulation-As-a-Service: Parallel Workflows for Cae and Ai Models in a Cloud-Native Platform

As the demand for high-fidelity engineering simulations increases, so does the complexity of delivering scalable, reproducible, and user-friendly computational workflows. StrmungsRaum is a cloud-native Simulation-as-a-Service (SaaS) platform designed to democratize access to CAE and AI technologies across industrial domains. In this talk, we present the architectural principles and parallel computing strategies underpinning the systems ability to orchestrate both classical CFD/FEA simulations and modern AI-based surrogate models at scale. We outline our approach to handling multitenant workloads in containerized HPC environments, address challenges in reproducibility, data management, and code lifecycle, and show how modular pipelines allow rapid deployment of hybrid CAE-AI applications. Particular focus is given to how the platform leverages parallelism across different layers: from batch simulation scheduling on Kubernetes-managed clusters to real-time inference of ML models in production. By integrating robust solvers with GPU-accelerated inference engines and a flexible workflow engine, StrmungsRaum enables engineering teams to run parameter studies, optimization tasks, and digital twin updates in parallel while preserving traceability, compliance, and cost control. The talk concludes with insights into architectural trade-offs and lessons learned from scaling simulation workflows across multiple industrial projects.

Markus Geveler
IANUS Simulation GmbH, Dortmund, Germany
m.geveler@ianus-simulation.de

MS16

Parallel Scalable Monolithic Two-Level Nonlinear Schwarz Methods for Navier-Stokes Equations with High Reynolds Numbers

To solve nonlinear partial differential equations, the well-known class of linear domain decomposition methods (DDMs) is typically employed as a preconditioner for the Krylov subspace method used to solve the tangential system at each iteration of a Newton-type method. Nonlinear DDMs are an alternative to classical Newton-Krylov-DDMs that apply domain decomposition before linearization, solving nonlinear problems on each subdomain as well as on the global domain. They have been shown to improve the nonlinear convergence behavior of Newtons method, and when augmented with a second level, they hold the potential for excellent scalability. However, their implementation complexity is high, which has limited most research to sequential implementations or parallel implementations that scale to less than a hundred subdomains. We are developing a highly scalable implementation of one- and two-level nonlinear Schwarz methods based on FROSch, a sub-package of the Trilinos project that implements linear one- and multi-level Schwarz DDMs. Our solver is coupled with the fluid-dynamics software FEAT3. In this talk, we present results demonstrating the nonlinear convergence and parallel scalability of our implementation for a selection of nonlinear fluid-dynamics problems from the FEAT3 library, such as the lid-driven cavity problem. We use a monolithic approach and compare the performance of our implementation with that of a standard Newton-Krylov-Schwarz solver.

Kyrill Ho
Department of Mathematics and Computer Science
University of Cologne
kyrill.ho@uni-koeln.de

Martin Lanser
Universitaet zu Koeln
Mathematisches Institut
martin.lanser@uni-koeln.de

Axel Klawonn
Department of Mathematics and Computer Science and Center
for Data Centand Simulation Science, University of Cologne
axel.klawonn@uni-koeln.de

MS16

Locality Matters: A Level-Based Approach to Gustavson's Sparse Matrix Multiplication on Multicore-Cpus

Sparse matrixmatrix multiplication (C = A * B, SpGEMM) is a critical operation in many scientific applications, but has received increasing attention in recent years due to its ubiquity in AI/ML workloads. Gustavson's algorithm is the most widely used method of computing SpGEMM on multicore CPUs, but it incurs irregular accesses to intermediate structures. These access irregularities pose serious performance difficulties for multicore CPUs, which are optimized for regular access patterns. In this talk, I explore Gustavson's SpGEMM algorithm, identify opportunities to improve access locality, and propose a level-based implementation to alleviate the problems arising from irregular accesses.

Dane Lacey
NHR@FAU, Friedrich Alexander-Universität
Erlangen-Nürnberg
Germany
dane.c.lacey@fau.de

MS16

StroemungsRaum - An Overview from a Numerical and Algorithmic Perspective

The aim of the StroemungsRaum project is to extend the scalability of the StrmungsRaum platform, which has been successfully used for years by the SME IANUS SIMULATION as Simulation-as-a-Service in an industrial environment. The simulation core of the StrmungsRaum platform is the open source software FEATFLOW which has been expanded in such a way that highly scalable CFD simulations can be carried out efficiently on future exascale architectures with heterogeneous hardware components so that significantly more complex and finer-resolution CFD simulations in industrial applications become feasible by combining leading methodological approaches for the efficient, parallel solution of CFD problems with a consistent, stringent application of performance engineer-

ing. We exemplarily concentrate on how parallel-in-space simultaneous-in-time Krylov-Multigrid approaches in combination with nonlinear global-in-time Newton/Domain Decomposition methods and special Pressure Schur Complement techniques can be designed which allow a much higher degree of parallelism for the resulting sequences of time-dependent convection-diffusion velocity and pressure Poisson-type subproblems. To exploit current accelerator hardware in lower precision, we apply the concept of prehandling for the discretized systems of equations. We provide numerical results as proof-of-concept and discuss open problems and challenges for incompressible flow problems in industrial applications.

Stefan Turek
TU Dortmund University
Department of Mathematics
stefan.turek.mathematik.tu-dortmund.de

MS17

Asynchronous Federated Spatial Modeling with Low-Rank Gaussian Processes

Spatial datasets in applications such as environmental monitoring and urban analytics are increasingly collected across distributed platforms where privacy and communication constraints prevent centralization. Federated modeling provides a natural paradigm for collaborative inference in such settings, yet spatial dependence across data sources challenges standard federated optimization assumptions. Existing methods either neglect cross-worker dependencies or rely on synchronous updates, which are inefficient in heterogeneous environments. Thus, we present an asynchronous federated framework for spatial modeling based on low-rank Gaussian process approximations. The proposed approach leverages block-structured optimization and incorporates staleness-aware aggregation, gradient correction, and stabilization mechanisms to ensure robust updates. We establish linear convergence with explicit dependence on staleness, offering theoretical insight into the impact of asynchrony. Numerical experiments demonstrate that the asynchronous method matches synchronous performance under balanced resources and significantly improves efficiency under heterogeneous conditions. These results illustrate how combining spatial statistical structure with asynchronous optimization enables scalable, privacy-preserving inference for large distributed spatial systems.

Sameh Abdullah
KAUST
sameh.abdulah@kaust.edu.sa

MS17

Artificially Intelligent Geospatial Systems: A Case Study in Spatial Energetics for Mobile Health Data

This talk will offer perspectives on the significant paradigm shift taking place in data analysis with the advent of AI technologies. This rapidly evolving field offers substantial intellectual space for statistical theory and methods to not only co-exist with other disciplines within computer science and machine learning, but also play a crucial role in advancing data analysis and probabilistic inference at unprecedented scales. The talk will elucidate three ideas that will synthesize into an artificially intelligent inferential system. The first is "amortized Bayesian inference" that considers training and calculating posterior distributions using generative AI. The second is Bayesian transfer learning for scaling Inference to massive datasets. The third is Bayesian predictive stacking that delivers exact simulation-based inference without resorting to expensive iterative methods such as Markov chain Monte Carlo. These ideas will be synthesized to present a case study in spatial energetics. Spatial energetics is broadly referred to as the study of live movement in real time and has been especially relevant in the context of mobile health data using actigraph units embedded in wearable devices. The case study is a part of the University of California Los Angeles Physical Activity and Sustainable Transportation Approaches and is primarily concerned with learning about a subject's metabolic levels as a function of their mobility attributes and other health attributes.

Sudipto Banerjee
Department of Biostatistics
UCLA
sudipto@ucla.edu

MS17

Distributed Multilevel Sequential Monte Carlo Applied to PDE-Based Bayesian Inverse Problems

We present a distributed multilevel sequential Monte Carlo method for Bayesian inverse problems to identify high-dimensional input parameters of models based on partial differential equations. The method is implemented on a distributed multilevel data structure to divide the computational load across multiple processing devices and to control the sampling and discretization errors introduced by Monte Carlo estimation and finite element approximation. By formulating these errors as the minimization target in a knapsack problem, with available computational resources as constraints, we derive error bounds with respect to a given computational budget. Based on this framework, we discuss convergence properties and present experimental results demonstrating the methods performance.

Niklas Baumgarten
Karlsruhe Institute of Technology
niklas.baumgarten@uni-heidelberg.de

MS17

Distributed and Recursive Bayesian Inference for Complex Spatio-Temporal and Big Data Models

The rapid growth of large-scale and streaming datasets in fields such as environmental sciences, risk management, and public policy has introduced substantial computational challenges for statistical modeling. Distributed inference tackles these challenges by partitioning data across machines and recombining results, while recursive inference enables sequential posterior updating as new information becomes available. Together, these approaches have become essential for modern applications. In this work, we propose a unified Bayesian framework for distributed and recursive inference, built upon the integrated nested Laplace approximation (INLA) methodology and implemented in the `R-INLA` software. The framework extends INLA to distributed and recursive settings, enabling efficient posterior computation without redundant processing. It naturally accommodates federated and privacy-preserving applications, where data remain decentralized but inference must be performed jointly, and it supports real-time updating in streaming contexts. In addition, we introduce automated strategies for partitioning both data and models to reduce computational complexity. The effectiveness of the approach is demonstrated through case studies in spatio-temporal modeling, large-scale monitor-

ing, and distributed analysis, underscoring its scalability and adaptability.

Mario Figueira Pereira, David Conesa, Antonio López-Quílez
University of Valencia
mafipe@alumni.uv.es, david.v.conesa@uv.es, antonio.lopez@uv.es

Håvard Rue, Håvard Rue
King Abdullah University of Science and Technology
haavard.rue@kaust.edu.sa, haavard.rue@kaust.edu.sa

MS18

Adaptive and distributed low-rank methods for modern computer architectures

Low-rank methods have seen significant interest lately as a complexity reduction technique for high-dimensional partial differential equations (PDEs). They have been shown to perform well for many problems, ranging from plasma physics to biology. In principle, such methods are well-suited for modern computer architecture such as GPUs as they have a higher arithmetic intensity than traditional PDE solvers. Adaptive low-rank methods (i.e. methods that change the rank) during the simulation are widely used, but they pose some additional challenges on GPU architectures (both beause they usually consume more memory and need to adapt the data structures during the simulation). Moreover, low-rank methods do not lend themselves naturally to be used in a distributed memory context as they have global data dependency patterns. In this talk, we will present our progress in these areas (both algorithmic and implementation) in the context of our dynamical low-rank framework Ensign (https://github.com/leinkemmer/Ensign).

Stefan Brunner
University of Innsbruck
Austria
stefan.brunner@uibk.ac.at

Lukas Einkemmer
University of Innsbruck
lukas.einkemmer@uibk.ac.at

MS18

The Wigner Poisson Equation, a sampling based adaptive rank methods to overcome intractability of scaling.

The Wigner Poisson equation arises in a variety of applications, including transport models for semiconductor physics and hot warm dense matter in inertial confinement fusion, to name two interesting examples. The model itself comes from the BBGKY formalism applied to the N-body quantum system. It represents the leading order term in the reduction, and when applicable, it reduces the model from N^N to N^6 degrees of freedom. The key complicating factor in the quantum model is its non-locality across all of phase space, making the model intractable using traditional parallel approaches. However, unlike the Vlasov equation, which is its counterpart in classical physics, the model appears to have low rank structure to its solutions over a wide range of parameters relevant to modeling key physics in inertial confinement fusion in the warm dense state. This opens up avenues for addressing this six dimensional model. In this talk we introduce a structure preserving sampling based approach to an adaptive rank algorithm in 1D1V and its corresponding extensions in 2D2V and 3D3V. We establish the scaling of the method is O(N) in memory and storage. This makes simulations of the 3D3V model tractable for the first time, as the algorithm overcomes the need for distributed computing of this non-local model. The current method preserves density to machine precision and preserves momentum and energy to the level of the truncation employed in the adaptive rank method.

Andrew J. Christlieb
Michigan State Univerity
Dept. of Comp. Math., Sci & Engr.
christli@msu.edu

Jing-Mei Qiu
University of Delaware
jingqiu@udel.edu

Sining Gong
Michigan State University
gongsini@msu.edu

Nanyi Zheng
University of Delaware
nyzheng@udel.edu

MS18

Multiscale Modeling of Nonlinear Behaviour in Aerospace Composites

Accurate simulation of nonlinear behavior in aerospace composites remains a significant computational challenge due to complex geometries, material heterogeneities, and localized nonlinearities such as delamination. Classical model reduction techniques often rely on assumptions of scale separation or linearity. I will present a method to efficiently solve large-scale nonlinear PDEs in composite materials without scale separation assumptions. Our approach constructs a coarse approximation space by solving local spectral problems in A-harmonic subspaces, yielding high-fidelity localized basis functions that are naturally adapted to material heterogeneity and geometric complexity. The method is applied to several nonlinear phenomena such as large deformation kinematics and interfacial decohesion through cohesive zone models. A two-level restricted additive Schwarz domain decomposition strategy is employed, ensuring parallel scalability on HPC platforms. This nonlinear multiscale solver demonstrates the ability to robustly handle simulations with millions of degrees of freedomin a fraction of the time required by traditional solvers.

Anne Reinarz
Durham University
anne.k.reinarz@durham.ac.uk

Robert Scheichl
Ruprecht-Karls University Heidelberg
r.scheichl@uni-heidelberg.de

MS18

High-Performance Computing for Macroscopic Plasma Simulations in Fusion Energy Research

In this presentation, we describe selected advances in applications of high-performance computing to fusion energy research. We highlight some recent successes together with on-going challenges and opportunities for innovation. Fusion holds great promise as an energy source for large-

scale power generation. The underlying systems, however, are intrinsically multi-scale and highly nonlinear. Consequently, our understanding of and ability to predict the resultant dynamics has benefitted significantly from advances in high-performance computing. We start by discussing magnetohydrodynamic (MHD) simulations, which model magnetically confined fusion plasmas at the system scale. We will focus primarily on stellarator applications, which are the subset of configurations that use symmetry-breaking geometric shaping to confine fusion plasmas. We will highlight recent advances in high-fidelity simulations of stellarators. From there, we will discuss efforts towards developing multi-fidelity models for multi-scale plasma dynamics that are capable of linking the macro- and microscopic dynamics.

Adelle Wright
University of Wisconsin - Madison
adelle.wright@wisc.edu

MS19

Programming GPUs for Performance with Fortran and the OpenMP API

Modern supercomputers are built to run thousands of tasks in parallel, and GPUs essential for accelerating computations. However, programming GPUs efficiently has long been a challenge-especially because different vendors support different programming models like CUDA, HIP, OpenCL, and SYCL. This often forces scientists to write multiple versions of the same code, for Fortran codes even in a different language, just to run on different hardware. For researchers who primarily use Fortran or prefer portable solutions, this complexity can be a major barrier. In this talk, we present OpenMP, a widely-used and portable programming interface that simplifies GPU programming. OpenMP allows you to write code once and run it across different systems without needing to dive into vendor-specific frameworks. This talk will introduce the basics of GPU offloading using OpenMP, including how to manage data and control flow between CPUs and GPUs. It will then explore performance optimization strategies like minimizing data transfers and asynchronous offload. This presentation offers a practical and accessible entry point into GPU computing for scientists from any field-without needing deep expertise in CUDA, HIP, OpenCL, or SYCLand shows how OpenMP can help accelerate your science.

Michael Klemm
OpenMP ARB
Advanced Micro Devices GmbH
michael.klemm@amd.com

MS19

GALAEXI: An Architecture-Agnostic GPU-Acceleration Approach for Legacy Fortran Software

GALAEXI is a high-order discontinuous Galerkin spectral element method (DGSEM) computational fluid dynamics (CFD) code used for the study of compressible, turbulent flows. It is the device-accelerated version of an existing Fortran-based, CPU-only CFD framework. In this talk, the chosen strategy for porting the Fortran-based compute kernels in GALAEXI to CUDA/HIP C++ is covered. Included are discussions on managing device memory, retaining support for CPU computations, strategies for writing and testing kernels and how to handle GPU-to-GPU MPI communication. Specific emphasis is placed on how the strategies covered can be abstracted in an architecture-agnostic way to allow the greatest degree of portability. The information presented will then be distilled into advice for those seeking to perform a similar porting effort.

Spencer Starr
Institute of Aerodynamics and Gas Dynamics
spencer.starr@iag.uni-stuttgart.de

Patrick Kopper, Yannik Feldner, Anna Schwarz, Andrea Beck
University of Stuttgart
Institute of Aerodynamics and Gas Dynamics
patrick.kopper@iag.uni-stuttgart.de,
yannik.feldner@iag.uni-stuttgart.de, schwarz@iag.uni-stuttgart.de, beck@iag.uni-stuttgart.de

MS19

Building Your Numerical Software with AMD ROCm Libraries

Have you ever wondered how to "translate" an algorithm listing from a linear algebra textbook or a publication into a performant piece of GPU code? AMD ROCm libraries, such as rocBLAS, rocSPARSE, rocSOLVER, and others can be leveraged to accomplish this task. In this mini-symposium talk, we build an implementation of Conjugate Gradient (CG) using ROCm libraries. Next, we extend the implementation to Preconditioned Conjugate Gradient (PCG) by setting up and calling a simple preconditioner; limited and relatively simple set of (mostly) rocSPARSE and rocBLAS routines is required for CG and PCG. To further challenge ourselves, we modify and expand the code to Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG), which is a (non-trivial to implement) eigenvalue solver used for finding largest (or smallest) eigenvalues and corresponding eigenvectors of a symmetric eigenvalue system. In the talk, we provide an overview of ROCm linear algebra libraries and associated concepts (such as "handle", "descriptor", and "matrix info"), and show how to use them to build a "real" application that requires implementing calls to multiple libraries in a single code. We provide guidance on the best practices that lead to better GPU performance (such as avoiding unnecessary memory operations), show how to avoid common problems, and comment on common misconceptions and pitfalls.

Katarzyna Swirydowicz
Advanced Micro Devices
kasia.swirydowicz@amd.com

MS19

Accelerating a Large Scale CFD Code Base Using Iso C++: A Viable Way for APU Clusters?

Graphics and Accelerated Processing Units (GPU/APUs) have become increasingly prominent not only in machine learning but also in traditional high-performance computing (HPC) domains, including scientific computing. This paradigm shift in the computing hardware presents a significant challenge for well-established simulation codes in computational science and engineering. In this work, we present the experience and results of GPU/APU porting and tuning efforts of m-AIA, a multiphysics software framework written in ISO C++ with a strong focus on computational fluid dynamics (CFD) applications. The framework integrates multiple numerical solution schemes for diverse

physical models, all operating on a shared hierarchical Cartesian grid. This design enables efficient multiphysics coupling, adaptive mesh refinement, and dynamic load balancing [1]. Over the past years, major parts of m-AIA were ported using the parallel-STL resulting in GPU/APU-ready code working on NVIDIA and AMD based systems. The vendor-specific implementations of the parallel STL provide a rapid and accessible path to porting computational workloads to GPU/APU systems. It is shown that this approach offers a practical strategy for modernizing scientific codes while maintaining flexibility and maintainability. In this talk, lessons learned from this porting endeavor as well as examples of the performance tuning including some simulation results will be presented. [1] https://zenodo.org/records/13350586.

Julian Vorspohl
RWTH Aachen University
j.vorspohl@aia.rwth-aachen.de

Miro Gondrum
Institute of Aerodynamics
RWTH Aachen University
m.gondrum@aia.rwth-aachen.de

Ansgar Niemoeller, Tim Wegmann
RWTH Aachen University
a.niemoeller@aia.rwth-aachen.de,
t.wegmann@aia.rwth-aachen.de

Matthias Meinke
Institute of Aerodynamics
RWTH Aachen University
m.meinke@aia.rwth-aachen.de

Dominik Krug
RWTH Aachen University
d.krug@aia.rwth-aachen.de

MS20

Forward and Backward Error Bounds for a Mixed Precision Preconditioned Conjugate Gradient Method

The preconditioned conjugate gradient (PCG) is very frequently method of choice for approximating the solution of sparse, symmetric positive definite linear algebraic systems. While the use of a preconditioner can significantly improve the rate of convergence, its application usually involves solving two triangular systems; a potential bottleneck in the algorithm. Employing low precision in the application of the preconditioners has the potential of overcoming this computational bottleneck, while at the same time providing a high-accuracy approximation. Multiple mathematically equivalent PCG variants have been developed, but these can exhibit a different behavior in finite precision. In particular, to the best of our knowledge, these variants have not been analyzed in the context of mixed precision under a unified framework. We provide insights into the behavior of distinct PCG variants through backward and forward stability analyses.

Thomas Bake
Charles University
bake@karlin.mff.cuni.cz

MS20

A Practical, Fully Parallel Implementation of the (H-)Tucker Decomposition Via Randomization

We present a new Tucker decomposition technique combining tensor fiber sampling and randomized sketching techniques. In particular, the fiber sampling overcomes the computational costs of forming the tensor matricizations, while the sketching improves the approximation of the tensor matricization column space. These modifications are based on recent results in matrix randomization; in our framework, they decrease the computational time and memory requirements. Under the assumption of an existing low-rank structure, the approximation errors produced by our variant are comparable to those of existing randomized Tucker methods. Starting from our randomized Tucker decomposition, we introduce the same modifications in the Hierarchical Tucker (H-Tucker) decomposition. This method relies on a binary tree structure to factorize tensors, forming and decomposing a tensor matricization for each tree node. Since the number of nodes is usually larger than d, the computational benefits of our randomized H-Tucker variant are even more appreciable. To enhance further computational efficiency, we develop a randomized H-Tucker parallel variant, thanks to the naturally parallelizable structure of the algorithm. The randomized changes reduce the inter-node communication costs and the local memory requirements.

Martina Iannacito
University of Bologna
martina.iannacito@unibo.it

MS20

Extreme-Scale Matrix-Free Multigrid for the Stokes Equations

Mantle convection drives mass and heat transport in Earths interior, linking deep Earth dynamics to surface phenomena such as plate motion, mountain building, and sea level change. Over geologic time scales, mantle rocks behave as a viscous fluid, allowing the dynamics to be modeled as Stokes flow with strongly variable viscosity. Realistic simulations at kilometer-scale resolution require discretizations with trillions (10) of unknowns, well beyond the capabilities of traditional sparse solvers due to memory and scalability limitations. We present advances in matrix-free multigrid methods designed for this extreme regime. Our approach integrates adaptive Galerkin coarsening for robustness under large viscosity contrasts with code-generationdriven kernel optimization. These developments enable scalable solvers with minimal memory overhead, moving toward geophysically realistic mantle convection simulations on next-generation supercomputers.

Nils Kohl
Ludwing-Maximilians-Universität München
nils.kohl@lmu.de

MS20

Comparison of the convergence rates of iterative refinement and GMRES

Both iterative refinement and GMRES can be used to improve the accuracy of inaccurate linear solvers. To a certain extent, the literature features indirect arguments supporting the convergence superiority of GMRES over iterative refinement. However, because these arguments do not use the iterative refinement terminology or address rounding errors and approximations, the connection between GMRES and iterative refinement was never made evident. In

this context, we propose to compare theoretically and experimentally the convergence rates of iterative refinement and GMRES. We showcase why and in which cases GMRES corrects linear solvers faster. We further explain that one can preserve the complexity advantage of iterative refinement while improving its convergence rate.

Bastien Vieublé
Chinese Academy of Sciences
bastien.vieuble@amss.ac.cn

MS21

Scalable Schwarz-Based Methods for the Numerical Solution and Data-Driven Prediction of Generalized Newtonian Blood Flow

Hemodynamics plays a fundamental role in the functioning of the human body, motivating the development of computationally efficient simulation methodologies for predicting blood flow dynamics. In this work, we employ a generalized Newtonian fluid model with shear-dependent viscosity to describe stationary blood flow and solve the resulting nonlinear partial differential equation (PDE) system using scalable domain decomposition techniques. The simulations are carried out with our in-house software library FEDDLib (Finite Element and Domain Decomposition Library), which interfaces with the FROSch (Fast and Robust Overlapping Schwarz) solver package from the Trilinos library. Specifically, we use the iterative solver GMRES (Generalized Minimal Residual) and an overlapping Schwarz preconditioner with a GDSW-type (Generalized Dryja-Smith-Widlund) coarse space, implemented in FROSch, and demonstrate scalability. However, these scalable solvers remain too costly for real-time applications. To move toward online prediction, we introduce a CNN-based surrogate modeling approach that employs an overlapping Schwarz domain decomposition strategy. In this framework, locally trained CNN surrogates are used as inexact subdomain solvers within an alternating Schwarz method. We describe the proposed pipeline and discuss the strengths and weaknesses of this approach.

Axel Klawonn
Department of Mathematics and Computer Science and Center
for Data Centand Simulation Science, University of Cologne
axel.klawonn@uni-koeln.de

Natalie Kubicki
University of Cologne
nkubicki@uni-koeln.de

Martin Lanser
Universitaet zu Koeln
Mathematisches Institut
martin.lanser@uni-koeln.de

Takashi Shimokawabe
Information Technology Center
The University of Tokyo
shimokawabe@cc.u-tokyo.ac.jp

Janine Weber
Universität zu Köln
Department of Mathematics and Computer Science
janine.weber@uni-koeln.de

MS21

Toward Automatic Generation of High Performance Numerical Codes by LLM

Generative AI technology based on Large Language Models (LLMs) has made remarkable progress in recent years, enabling the automatic generation of program code at a high level. However, significant challenges remain in generating numerical code for high-performance computing. To address this, we investigated several lightweight yet effective techniques to enhance generation performance. In this talk, we first demonstrate how nave linear algebra codes for CPUs can be accelerated by automatically applying OpenMP, SIMD instructions, and blocking techniques. We then present results of code generation for GPUs, covering not only CUDA for NVIDIA GPUs, but also HIP for AMD GPUs and SYCL for Intel GPUs.

Daichi Mukunoki
Nagoya University, Japan
mukunoki.daichi.p2@f.mail.nagoya-u.ac.jp

Koki Morita, Hayashi Shun-ichiro, Tetsuya Hoshino
Nagoya University
morita.koki.f8@s.mail.nagoya-u.ac.jp,
hayashi.shunichirou.y4@s.mail.nagoya-u.ac.jp,
hoshino.tetsuya.w9@f.mail.nagoya-u.ac.jp

Takahiro Katagiri
Information Technology Center, Nagoya University
RIKEN R-CCS
katagiri@cc.nagoya-u.ac.jp

MS21

Innovative Supercomputing by Integrations of Simulations/Data/Learning on Large-scale Heterogeneous Systems and Beyond

Supercomputing is evolving rapidly, with the integration of Simulation, Data, and Learning (S+D+L) being essential for realizing Society 5.0, which merges cyber and physical spaces. In 2015, we launched the BDEC project to develop supercomputers and software for integration of (S+D+L). Wisteria/BDEC-01 began operation in May 2021, combining A64FX nodes for simulation (Odyssey) and NVIDIA A100 GPUs for AI/data analytics (Aquarius). We developed h3-Open-BDEC to maximize performance and minimize energy use through adaptive precision, machine-learning-based data-driven-type simulations, and system software for heterogeneous systems. In January 2025, we began operating the Miyabi system with the University of Tsukuba, featuring 1,120 NVIDIA GH200 nodes and 380 Intel Max 9480 sockets. This talk will focus on integration of (S+D+L) using h3-Open-BDEC on Wisteria/BDEC-01 and introduce innovations with Miyabi. h3-Open-BDEC is also used for QC-HPC hybrid computing in the JHPC-quantum project, launched in November 2023 as a five-year initiative. We will also present an overview and future plans of this project.

Kengo Nakajima
Information Technology Center, The University of Tokyo
RIKEN Center for Computational Science (R-CCS)
nakajima@cc.u-tokyo.ac.jp

Shinji Sumimoto
The University of Tokyo

sumimoto@cc.u-tokyo.ac.jp

Takashi Arakawa
CliMTech
arakawa@climtech.jp

Yao Hu
KOIBUCHI Laboratory
National Institute of Informatics
hu@cc.u-tokyo.ac.jp

Kazuya Yamazaki
Information Technology Center, The University of Tokyo
kyamazaki@cc.u-tokyo.ac.jp

Hisashi Yashiro
National Institute for Environmental Studies
yashiro.hisashi@nies.go.jp

MS21

Parallel and Distributed Sequences of very sparse unstructured matrix by dense skinny matrix computation

AI, including LLM and machine learning applications in general, is currently redesigning the architectures and the distributed and parallel programming approaches: from the chip design, the new arithmetic, the new accelerators to new data structures and programming paradigms, for example. Nevertheless, linear algebra and supercomputing expertise are and will be still required. Convergence between those domains is not guaranteed and we need to build bridges and be able to propose interoperability allowing to avoid separate roadmaps and ecosystems. Decade long researches on distributed and parallel computational science, linear algebra and middleware may be adapted and optimized these new application deployments. In this talk we target the important problem of non-symmetrical sparse matrix computation which is, for example, part of some LLM algorithms. We present several experiments on different clusters of multicore nodes at the Flatiron Institute and on the Fugaku supercomputer for sequences of very sparse non-symmetric and unstructured matrix by dense skinny matrix computation.

Serge G. Petiton
University of Lille, and CNRS
serge.petiton@mailfence.com

Maxence Vandromme
RIKEN
maxence.vandromme@riken.jp

Maxence Buisson
Polytech Lille
maxence.buisson@polytech-lille.net

Geraud P. Krawezik
Flatiron Institute
gkrawezik@flatironinstitute.org

Aksel Kurudere
University of Lille
aksel.kurudere@polytech-lille.net

MS22

An In-Depth Examination of Porting Icon's Turbulent Mixing Framework to Kokkos

ICON, originally written in Fortran, is being modernized for better portability by transitioning to C++ and Kokkos. The initial phase focuses on porting physics packages within the Atmosphere, such as the Turbulent Mixing (TMX) component. In ICON, TMX is a flexible framework that enables the integration of various turbulent mixing schemes through object-oriented programming. The complexity of the code, covering both infrastructure and scientific elements, and its reliance on multiple ICON components make it an ideal guinea pig for evaluating the challenges in future porting efforts. In this presentation, I will explore the intricacies of moving to C++/Kokkos. This process includes validating the code to achieve bit-identical results, pioneering the use of the Memory Manager and new math libraries in C++, and analyzing performance metrics. Ultimately, while the success of the trial is visible to all, we aim to reveal the underlying experience: What did the 'guinea pig' endure in this process?

Harshada Balasubramanian
Max-Planck-Institute for Meteorology
harshada.balasubramanian@mpimet.mpg.de

MS22

A Fortran-C++ Interoperable Memory Manager for Icon

C++ has better GPU support compared to Fortran. Therefore, ICON developers have been exploring ways to rewrite subroutines and kernels in C++ to better adapt to modern supercomputer systems. However, ICON consists of millions of lines of code and cannot be rewritten all at once. Naturally, the solution would be a step-by-step rewrite and replacement of the legacy code. Cross-compilation of Fortran and C++ code is nontrivial because there is no standard interface for Fortran-C++ interoperability. Moreover, the bigger challenge is to pass data efficiently and correctly between the two languages. This talk introduces a memory manager framework inspired by ICON, which handles the memory for the user. The user can program in intrinsic Fortran or C++ code to allocate and retrieve data from the memory manager. The memory manager handles the data passing between Fortran and C++. The memory manager handles memory not only on the CPU, but also on multiple GPUs from different vendors.

Yen-Chen Chen, Yen-Chen Chen
TUM
yen-chen.chen@tum.de, yen-chen.chen@tum.de

MS22

Data-Flow Analysis and Visualization for Fortran Applications

A big obstacle to the modernization of legacy Fortran codes is the large scale and scope of many such applications. These modernization efforts can take multiple forms that can coexists within a single project, such as disentangling module components into dedicated libraries, porting Fortran code to a language with better GPU support like C/C++, or introducing new programming models like task-based parallelism. All those transformations require a good understanding of the codebase at various level of granularity, from the dependencies between modules to the call stack of procedures. In order to support those efforts, we present a tool meant to extract useful informa-

tion by constructing structures like dependency graphs or call graphs and performing data-flow analysis. We discuss how this information can be exploited and visualized to explore an extremely large codebase without overwhelming the user.

Gwenolé Lucas
Technical University of Munich
gwenole.lucas@tum.de

MS22

Modernizing Icon Math Libraries: A C++ and Kokkos Implementation

High-resolution weather and climate simulations with ICON depend on a suite of mathematically intensive kernels, challenging modern supercomputers. These operations include interpolation and differential operators, which are central to accurate simulations where efficiency and portability directly affect model resolution and throughput on next-generation systems. The original Fortran implementation, accelerated with OpenACC, achieved GPU support but struggled with portability across architectures and long-term maintainability. This talk introduces a modernization of ICONs mathematical libraries in modern C++ using the Kokkos performance portability framework. The algorithms remain unchanged, but the implementation now delivers optimized performance on CPUs as well as NVIDIA, AMD, and Intel GPUs without hardware-specific code. Crucially, the new C++ routines are called directly from the existing Fortran code via language interoperability, ensuring seamless integration into ICON without disrupting established workflows. The rewrite addresses critical limitations in the legacy codebase while ensuring accuracy and maintainability. Performance benchmarks demonstrate consistent optimization across architectures, showcasing how legacy scientific applications can benefit from modern C++ practices and performance portability frameworks. This modernization enables ICON to efficiently exploit emerging and diverse computing architectures in high-performance computing environments.

Pradipta Samanta
German Climate Computing Centre
samanta@dkrz.de

MS23

Dense Tiling Meets Structured Sparsity: Scalable Algorithms with sTiles

Scalable Bayesian inference for latent Gaussian models is increasingly essential in modern applications spanning spatio-temporal statistics, environmental modeling, and machine learning. These models often lead to large structured sparse precision (inverse of covariance) matrices, where computational bottlenecks, especially in Cholesky factorization and selected inversion, limit scalability. To overcome these challenges on modern architectures, we must rethink how sparsity and structure are represented and exploited. In this talk, we present sTiles, a tile-based framework designed to accelerate sparse matrix computations by transforming localized sparsity patterns, such as arrowhead, banded, or block-clustered forms, into dense tiles. This shift enables static scheduling, improves data locality, and unlocks efficient parallelism on GPUs and manycore CPUs. By leveraging the structure arising in models discretized from stochastic PDEs, sTiles bridges the gap between sparse and dense computations, achieving scalable performance while maintaining exactness. Our results demonstrate substantial speedups over state-of-the-art solvers and establish sTiles as a practical foundation for large-scale Bayesian inference workflows.

Esmail Abdul Fattah, Hatem Ltaief
King Abdullah University of Science and Technology (KAUST)
esmail.abdulfattah@kaust.edu.sa, hatem.ltaief@kaust.edu.sa

Håvard Rue, Håvard Rue
King Abdullah University of Science and Technology
haavard.rue@kaust.edu.sa, haavard.rue@kaust.edu.sa

David E. Keyes
King Abdullah University of Science and Technology (KAUST)
david.keyes@kaust.edu.sa

MS23

Scalable Approximate Selected Inversion Via Ilu and Spectral Corrections for Sparse Systems

We analyze two parallel numerical strategies for computing selected entries of the matrix inverse of large sparse symmetric systems: the selected inverse method and a factorized approximate inverse method. These techniques are aimed at computing via LU factorizations or incomplete LU (ILU) factorizations. The selected inverse approach exploits the LU / ILU factorization to recover the entries of the matrix inverse within the pattern of the (incomplete) LU factorization. In contrast, the factorized approximate inverse method applies a truncated Neumann series expansion to the (incomplete) LU factors, providing an alternative approach at the cost of reduced accuracy. To improve accuracy while keeping sparsity, we introduce low-rank corrections and eigenvector deflation, which provide an alternative to tightening drop tolerances. Numerical illustrations for both parallel and sequential computations demonstrate the effectiveness and robustness of our approach.

Tahamina Akter
TU Braunschweig
Researcher
tahamina.akter@tu-braunschweig.de

MS23

Bespoke Multiresolution Analysis of Graph Signals

We present a framework for discrete multiresolution analysis of graph signals using the samplet transform, a wavelet-like tool originally developed for scattered data in Euclidean spaces. We define samplets on graphs by subdividing the graph into patches, embedding each patch in Euclidean space to construct samplets, and mapping them back to the graph. This approach ensures orthogonality, locality, and vanishing moments with respect to graph polynomial spaces. Compared to classical Haar wavelets, it enables efficient compression of a broader class of graph signals. We illustrate this on signals defined over vertices on embedded manifolds and implement it efficiently using heavy edge clustering and landmark `Isomap` for low-dimensional embeddings. Our results demonstrate robustness, scalability, and sparse representations with controllable approximation error, outperforming traditional Haar wavelets in compression and multiresolution fidelity.

Giacomo Elefante, Gianluca Giacchi

Università della Svizzera italiana (USI)
giacomo.elefante@usi.ch, gianluca.giacchi@usi.ch

Michael Multerer
Università della Svizzera italiana
michael.multerer@usi.ch

Jacopo Quizi
Università della Svizzera italiana (USI)
jacopo.quizi@usi.ch

MS24

ExaEpi: An Amrex-Powered Agent-Based Model for Epidemiology

ExaEpi is a GPU-capable agent-based model (ABM) for epidemiology modeling powered by AMReX. ABMs are valuable because they provide a fundamental and natural description of disease dynamics and are able to capture emergent phenomena better than traditional SEIR models. However, their use in forecasting and control is limited by the difficulty in calibrating and quantifying the uncertainty associated with a large number of parameters. In this talk, I will discuss how ExaEpi can help address these limitations by enabling many large ensembles to run quickly on exascale compute facilities. I will discuss how agent-based algorithms can be implemented using the particle functionality in AMReX, how good parallel scaling and GPU performance can be achieved, and how to deal with the uncertainty associated with random fluctuations due to the nondeterministic execution order of asynchronous GPU threads. These features enable ExaEpi to be used for sensitivity analysis, uncertainty quantification, and surrogate training for diseases like COVID-19.

Andrew Myers
Lawrence Berkeley National Lab
atmyers@lbl.gov

Peter Nugent
Lawrence Berkeley National Laboratory
penugent@LBL.gov

Erin Acquesta
Sandia National Laboratories
eacques@sandia.gov

Jacqueline M. Wentz
University of Colorado Boulder
jacqueline.wentz@colorado.edu

Nathan Keilbart
LLNL
keilbart1@llnl.gov

Debojyoti Ghosh
Lawrence Livermore National Laboratory
Livermore, CA
ghosh5@llnl.gov

Kyle Nguyen
Sandia National Laboratories
Livermore, CA
kcnguye@sandia.gov

Steven Hofmeyr
Lawrence Berkeley National Laboratory
shofmeyr@lbl.gov

Tan Nguyen
LBNL
tannguyen@lbl.gov

MS24

Gempicx – An Open Source Library for Exascale Geometric Particle-in-Cell Methods

We present the open-sourcening of GEMPICX, a GEo-Metric Particle-In-Cell eXascale framework for structure-preserving methods in plasma physics. Using the AM-ReX[Weiqun Zhang et al., AMReX: A Framework for Block-Structured Adaptive Mesh Refinement, Journal of Open Source Software 4.37 (2019)] framework to control GPU and MPI parallelism on a wide variety of architectures, along with structure-preserving methods presented earlier[Michael Kraus et al., GEMPIC: Geometric electromagnetic particle-in-cell methods. Journal of Plasma Physics, 83(4), (2017)], GEMPICX runs on Linux, MacOS and Windows systems, as well as NVIDIA and AMD GPUs, and through the CMake framework provides easy methods to configure, compile and run, aided by several CMake presets ranging from 1D CPU debug to 3D GPU release builds. Currently capabilities include finite difference cartesian and cylindrical coordinate simulations, drift-kinetic, fully kinetic and mixed mode low storage Runge-Kutta methods and finite volume magnetohydrodynamics. Mesh-refinement capabilities are in the works. The talk will go through the general design of the framework, including testing and documentation, and additionally provide performance metrics for some simple simulations run on the RAVEN and VIPER HPC systems at the Max Planck Computing and Data Facility.

Emil Poulsen
Max Planck Institute for Plasma Physics
emilb.poulsen@ipp.mpg.de

MS24

Phase field methods for solid / fluid mechanics problems in energetic and structural materials with Alamo

Phase field (PF) methods provide a versatile framework for modeling phenomena in mechanics, including fracture, microstructure evolution, topology optimization, and solid-fluid interactions. As diffuse interface methods, PF models typically require adaptive mesh refinement (AMR) for computational tractability, yet few implementations leverage block-structured AMR efficiently. We present Alamo, an AMReX-based code for solving phase field equations coupled with implicit elastic solves. Alamo includes multiple phase field models, material laws, and a custom strong-form nonlinear elasticity solver using source terms in place of boundary conditions, enabling accurate resolution of stresses in diffuse-boundary settings. This approach extends to solid-fluid interactions, where viscous Navier-Stokes equations are solved with source terms that reproduce sharp-interface conditions while maintaining stability. Applications illustrate the frameworks flexibility. In particular, PF models capture deflagration and regression in solid rocket composite propellants (e.g., AP/HTPB), predicting burn rates as a function of composition and morphology. Additional examples include phase field fracture, topology optimization, and microstructure evolution. By combining PF methods with block-structured AMR and strong-form solvers, Alamo provides a general, scalable platform for high-resolution simulations of multiscale, multiphysics

problems with diffuse interfaces.

Brandon Runnels
Iowa State University
brunnels@iastate.edu

MS24

Hipace++: a Gpu-Capable Quasi-Static Particle-in-Cell Code with Cache-Blocking and Asynchronous Communications

Plasma accelerators are a promising technology for next-generation compact particle accelerators, offering accelerating gradients orders of magnitude higher than conventional radio-frequency accelerators. Modeling these devices accurately is computationally intensive, and the quasi-static particle-in-cell (PIC) algorithm is a preferred approach due to its efficiency in many scenarios. We present HiPACE++, a performance-portable, three-dimensional, quasi-static PIC code written in C++ and built on the AMReX framework to run efficiently on modern GPUs. By decomposing the full 3D domain into a sequence of 2D transverse slices and carefully managing memory usage, HiPACE++ achieves orders-of-magnitude speedups compared to CPU-only implementations. Each transverse slice is computed on a single GPU, minimizing communication overhead, while longitudinal domain decomposition enables strong scaling up to 2048 GPUs. We discuss several algorithmic and performance optimizations, including a fully asynchronous communication pipeline implemented using non-blocking MPI functions with a state-machine approach to track message status per slice. A custom multigrid solver was developed to handle a two-component inhomogeneous shielded Poisson equation. By leveraging GPU shared memory, the smoother performs four Gauss-Seidel iterations. Additionally, particle deposition is carried out in GPU shared memory tiles, using a linked-list-inspired method to efficiently assign particles to tiles.

Alexander Sinn
Deutsches Elektronen-Synchrotron
alexander.sinn@desy.de

Maxence Thevenet
Deutches Elektronen Synchrotron (DESY), Germany
maxence.thevenet@desy.de

Carlo Benedetti
Lawrence Berkeley National Laboratory
cbenedetti@lbl.gov

Eya Dammak
Deutsches Elektronen Synchrotron (DESY)
eya.dammak@desy.de

Severin Diederichs
CERN
severin.diederichs@cern.ch

Axel Huebl
Lawrence Berkeley National Laboratory (LBNL), USA
axelhuebl@lbl.gov

Xingjian Hui
Deutsches Elektronen Synchrotron (DESY)
xingjian.hui@desy.de

Rémi Lehe
Lawrence Berkeley National Laboratory
ATAP
rlehe@lbl.gov

Andrew Myers
Lawrence Berkeley National Lab
atmyers@lbl.gov

Jean-Luc Vay
Lawrence Berkeley National Laboratory
Berkeley, CA
JLVay@lbl.gov

Weiqun Zhang
Lawrence Berkeley National Laboratory
Center for Computational Sciences and Engineering
WeiqunZhang@lbl.gov

MS25

Parallel Nystrom Approximation: Lower Bounds and Algorithms

The Nystrom approximation is a widely used technique in randomized linear algebra for low-rank matrix approximation, and it has gained significant traction in machine learning due to its ability to accelerate the computational routines during training and inferenceespecially in kernel-based methods. However, applying Nystrom approximation to large datasets often necessitates distributed memory algorithms, where communication overhead becomes a critical bottleneck on modern supercomputing clusters. Despite its growing relevance, distributed memory strategies for Nystrom approximation remain largely unexplored. In this work, we establish communication lower bounds for Nystrom approximation and introduce novel algorithms to achieve these bounds. Our approach leverages both CPU and GPU computation and GPU-aware communication to fully exploit the capabilities of contemporary supercomputing architectures.

Grey Ballard
Wake Forest University
ballard@wfu.edu

MS25

Learning Efficient Sparse Encoding for High-Performance Tensor Decomposition

We present the reinforcement-learned adaptive tensor encoding (ReLATE) framework, a novel learning-augmented method that automatically constructs efficient sparse tensor representations without labeled training samples. ReLATE employs an autonomous agent that discovers optimized tensor encodings through direct interaction with the application environment, leveraging a hybrid model-free and model-based algorithm to learn from both real and imagined actions. Moreover, ReLATE introduces rule-driven action masking and dynamics-informed action filtering mechanisms that ensure functionally correct tensor encoding with bounded execution time, even during early learning stages. By automatically adapting to both irregular tensor shapes and data distributions, ReLATE generates sparse tensor representations that consistently outperform expert-designed formats across diverse sparse tensor data sets, achieving up to 2X speedup compared to the best sparse format

Jee W. Choi
University of Oregon

jeec@uoregon.edu

MS25

High-Performance Block-Sparse Tensor Contractions on Gpus

This talk presents our work on optimizing block-sparse tensor contractions on GPUs by exploiting all nonzero block interactions, where blocks may vary significantly in tensor volume. We analyze how GPU performance changes with block size and block structure, and we search for high-performing execution configurations for a given block size. The configuration space includes input/output block memory layouts and hierarchical GEMM tiling parameters at the thread-block, warp, and thread levels. Our implementation is built on the CUTLASS framework and evaluated using both synthetic benchmarks and real workloads from ITensor. Results show that achieving strong performance requires careful configuration selection and kernel specialization for block-sparse tensor contractions.

Jiajia Li
North Carolina State University
jiajia.li@ncsu.edu

Sri Harshavardh Reddy Deverapalli, Zecheng Li
North Carolina State University
USA
sdevera@ncsu.edu, @tbd

Karl Pierce
The University Of Maryland
karl.m.pierce@gmail.com

Miles Stoudenmire
Flatiron Institute
mstoudenmire@flatironinstitute.org

MS26

Delta-Pic with Neural Forward-Backward Lagrangian Reconstructions

In this talk, we describe a δf particle simulation method where the bulk density is periodically remapped using a Semi-Lagrangian approach, where the backward flow is reconstructed using neural networks. This method is designed to handle plasma regimes where the densities strong deviate from their initial state and may evolve in general profiles.

Victor Fournet
Max Planck Institute for Plasma Physics
victor.fournet@ipp.mpg.de

MS26

Fast Flood Prediction Using Graph Neural Networks: Application to the Tt River Basin

Fast and reliable flood forecasting is needed for operational situations where decisions must be made quickly to reduce damages. Physically-based hydrodynamic models, such as those solving the shallow water equations on unstructured meshes, are well known for their accuracy and ability to describe complex river systems. However, for real-time forecasting, these models face a trade-off: achieving high precision requires a lot of computation, which can be too slow when many forecasts are needed under strict time constraints. In this work, we study if Graph Neural Networks (GNNs) can provide real-time flood predictions that are accurate enough for operational use. Our method uses the same unstructured mesh representations found in hydraulic simulations, which makes it easy to connect with existing workflows and data. We conduct our study on the Tt River basin (Pyrnes-Orientales, France), using a synthetic dataset produced by high-resolution numerical simulations. Our approach adapts the MeshGraphNet architecture to process unstructured meshes as graphs.

Valentin Mercier
INP Toulouse
valentin.mercier@toulouse-inp.fr

MS26

Accelerating the Convergence of Newtons Method Using Fourier Neural Operators

This talk will address a critical challenge in solving nonlinear elliptic partial differential equations (PDEs): the slow convergence of Newtons method when the initial guess is far from the true solution. To overcome this, we use Fourier Neural Operators (FNOs) to generate high-quality initial guesses. FNOs are particularly appealing because they are resolution-invariant, meaning they can be trained on coarse grids and applied to finer ones without retraining, making them computationally efficient for PDE problems. The FNO is trained to predict an approximate solution to the discretized PDE by minimizing a loss function based on the PDE residual, using data generated from numerical simulations or analytical solutions. We show numerically that this approach significantly reduces the number of Newton iterations required for convergence, especially for strongly nonlinear or anisotropic PDEs, where traditional initial guesses (e.g., zero or linear interpolations) perform poorly. Numerical experiments in one and two dimensions validate the approach. For the anisotropic elliptic PDEs, the FNO-initialized Newton's method achieves convergence in fewer iterations compared to baseline methods, with minimal computational overhead from the FNO evaluation. We will also discuss the trade-offs, such as the cost of training the FNO and the need for sufficient training data to generalize across problem variations.

Victor Michel-Dansac
INRIA
victor.michel-dansac@inria.fr

MS26

A Functional Framework for the Optimization of Neural Quantum States

This talk revisits the infinite-dimensional first optimize, then discretize paradigm for optimization in scientific machine learning. The key idea is to design algorithms at the infinite-dimensional level and subsequently discretize them in the tangent space of the neural network ansatz. We illustrate this approach in the context of the variational Monte Carlo method for quantum many-body physics, where neural quantum states have recently emerged as powerful representations of high-dimensional wavefunctions. In this setting, we recover the celebrated stochastic reconfiguration algorithm, interpreting it as a projected Riemannian L2 gradient descent method. We further explore extensions to Riemannian Newton methods, and conclude with algorithmic considerations related to the efficient scalability of these schemes.

Marius Zeinhofer

ETH Zurich
marius.zeinhofer@math.ethz.ch

MS27

A Comparison of Mixed Precision Iterative Refinement Approaches for Least Squares Problems

Various approaches to iterative refinement (IR) for least-squares problems have been proposed in the literature and it may not be clear which approach is suitable for a given problem. We consider three approaches to IR for least-squares problems when two precisions are used and review their theoretical guarantees, known shortcomings, and when the method can be expected to recognize that the correct solution has been found. It is shown that the IR methods exhibit different sensitivities to the conditioning of the problem and the size of the least-squares residual, which should be taken into account when choosing the IR approach. We end with practical advice for choosing the best approach in practice.

Erin Claire Carson
Charles University
Charles University MFF
carson@karlin.mff.cuni.cz

MS27

Mixed-Precision Domain-Decomposition Preconditioning: A Perturbation Analysis and Practical Guidelines

This work is concerned with domain decomposition preconditioners, specifically the Schwarz method with GenEO coarse space, to solve large, sparse symmetric positive definite problems. Through libraries such as HPDDM, these preconditioners have already been efficiently parallelized in their numerical implementations. However, they still require expensive linear algebra operations in each local subdomain. Motivated by the emergence of fast low precision arithmetic in hardware, we aim in this work to speed them up using mixed precision. To do this, we need to identify the sensitivity of these operations to perturbation and propose an actionable criteria for selecting the appropriate precision for each local subdomain. In order to do so, we develop a perturbation theory for our preconditioner to bound the worst-case loss of efficiency of the preconditioner and study the sharpness of this theoretical bound through numerical experiments with the FreeFEM, PETSc4py, and HPDDM libraries. Our findings show that the only important parameter is the maximum of the sizes of the local subdomain perturbations, weighted by the condition number of the local subdomain matrix. Our results therefore suggest that preconditioners can be constructed in mixed precision while effectively controlling the loss of efficiency.

Tom A. Caruso
Sorbonne University
tom.caruso@inria.fr

MS27

Floating-point Autotuning with Customized Precisions: Benchmarking and Analysis

Achieving optimal performance in numerical computations often hinges on aggressively quantizing data and arithmetic to low-precision formats under rounding error analysis to retain numerical accuracy. The PROMISE software provides a unified, task-specific validation platform for automated precision tuning, enabling a balance between computational efficiency and numerical fidelity. In this paper, we advance PROMISE with customized precision formats associated with emulated mathematical functions, which streamlines the exploration of low-precision configurations. Besides, we present our benchmark by applying PROMISE to evaluate a suite of well-known numerical algorithms. Our simulations reveal the potential for performance gains and memory reduction for numerical algorithms with reduced-precision configurations. Through detailed case studies, we provide actionable insights into selecting the optimal precision levels for various algorithms. This work underscores the transformative potential of automated precision tuning in enhancing efficiency while maintaining robustness across numerical applications.

Xinye Chen
Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
Xinye.Chen@lip6.fr

Thibault Hilaire, Fabienne Jézéquel
LIP6, Sorbonne Université
thibault.hilaire@lip6.fr, fabienne.jezequel@lip6.fr

MS27

Mixed-Precision Algorithms for the Sylvester Equation

We consider the solution of the Sylvester equation $AX + XB = C$ in mixed precision. We derive a new iterative refinement scheme to solve perturbed quasi-triangular Sylvester equations; our rounding error analysis provides sufficient conditions for convergence and a bound on the attainable relative residual. We leverage this iterative scheme to solve the general Sylvester equation. The new algorithms compute the Schur decomposition of A and B in low precision, use the low-precision Schur factors to obtain an approximate solution to the quasi-triangular equation, and iteratively refine it to obtain a working-precision solution to the quasi-triangular equation. In order to solve the quasi-triangular equation to working precision, the unitary Schur factors of A and B must be unitary to working precision, but this is not the case if the Schur decomposition is computed in low precision. We propose two approaches to address this: one is based on re-orthonormalization in working precision, and the other on explicit inversion of the almost-unitary factors. The two mixed-precision algorithms thus obtained are tested on various Sylvester and Lyapunov equations from the literature. Our numerical experiments show that the new algorithms are at least as accurate as existing ones. Our cost analysis, on the other hand, suggests that they would typically be faster than mono-precision alternatives if implemented on hardware that natively supports low precision.

Andrii Dmytryshyn
rebro University
School of Science and Technology
andrii.dmytryshyn@oru.se

Massimiliano Fasi
University of Leeds
School of Computing
m.fasi@leeds.ac.uk

Nicholas J. Higham
The University of Manchester

Department of Mathematics
nick.higham@manchester.ac.uk

Xiaobo Liu
MPI for Dynamics of Complex Technical Systems, Magdeburg
xliu@mpi-magdeburg.mpg.de

MS28

Nonlinear Two-Level Schwarz Methods - Avoiding Global Sparse Matrices

Domain decomposition methods (DDMs) are robust and efficient iterative solvers for discretized partial differential equations (PDEs). They follow a divide and conquer paradigm and are well suited for the high concurrency of modern supercomputer architectures. Applying the DDM paradigm to nonlinear PDEs, nonlinear DDMs acting as nonlinear preconditioners can be derived. These methods often show a faster nonlinear convergence and a larger convergence radius then classical Newtons method. This has been shown for many nonlinear problems, for example, Navier-Stokes equations with high Reynolds numbers, nonlinear diffusion, or Allen-Cahn equations. Additionally, nonlinear DDMs show a favorable ratio between computation and communication and offer possibilities to save global sparse matrix computations. Both properties are advantageous if state-of-the-art hardware with high parallelism and many accelerators is used. We discuss all these aspects and show results for the example of nonlinear two-level Schwarz methods and various nonlinear problems.

Martin Lanser
Universitaet zu Koeln
Mathematisches Institut
martin.lanser@uni-koeln.de

Axel Klawonn
Department of Mathematics and Computer Science and Center
for Data Centand Simulation Science, University of Cologne
axel.klawonn@uni-koeln.de

Kyrill Ho
Department of Mathematics and Computer Science
University of Cologne
kyrill.ho@uni-koeln.de

Janine Weber
Universität zu Köln
Department of Mathematics and Computer Science
janine.weber@uni-koeln.de

MS28

Scalable Three-Level Overlapping Schwarz Preconditioners for Fluid Problems

We consider monolithic overlapping Schwarz preconditioners of the generalized Dryja-Smith-Widlund (GDSW) type for the application to fluid problems. This work is part of the StroemungsRaum project, financed by the German BMFTR (formerly BMBF) under the SCALEXA program, the German initiative to develop software in the exascale computing era. This presentation is concerned with the enhancement of the successful FEATFLOW software library by scalable domain decomposition methods based on the FROSch framework in Trilinos with a focus on GDSW with three and more levels. FROSch is part of the ShyLU package within the Trilinos software library and provides a flexible infrastructure for implementing advanced overlapping domain decomposition preconditioners. We show weak and strong parallel scalabilty of the combined FEATFLOW/FROSch software using different monolithic approaches.

Oliver Rheinbach, Stephan Köhler
Technische Universitaet Bergakademie Freiberg
oliver.rheinbach@math.tu-freiberg.de,
stephan.khler@math.tu-freiberg.de

MS28

SpaceTime parallel spectral deferred correction methods for large-scale incompressible flow simulations

The numerical simulation of unsteady incompressible flows governed by the time-dependent NavierStokes equations remains a major challenge in large-scale computational fluid dynamics due to the high cost of resolving both spatial and temporal scales. While spatial parallelization is well established, the inherently sequential nature of time integration often constitutes a fundamental bottleneck to scalability. To overcome this limitation, we develop a spacetime parallel framework that couples spectral deferred correction (SDC) methods for temporal integration with finite element discretizations in space. The approach is first prototyped within pySDC, using FEniCSx in space to implement a stable and accurate finite element formulation tailored for incompressible flows. This prototype enables the rapid exploration of algorithmic properties and demonstrates the potential of the parallel-in-time method for benchmark problems, such as the DFG 2D-3 cylinder flow. Building on these insights, the methodology is transferred to the high-performance CFD solvers FEATFLOW/FEAT3 for large-scale industrial applications. Numerical experiments confirm that the proposed strategy achieves high accuracy, strong scalability, and significant runtime reductions, establishing a clear pathway from algorithmic prototyping to production-ready implementations in industrial CFD.

Robert Speck
Forschungszentrum Juelich GmbH
Juelich Supercomputing Centre
r.speck@fz-juelich.de

Abdelouahed Ouardghi
Forschungszentrum Juelich
Juelich Supercomputing Centre
a.ouardghi@fz-juelich.de

MS29

Modernizing Octopus: A Fortran Codes Journey from CPU to GPU Acceleration

GPU acceleration has become essential for scientific high-performance computing, particularly for large-scale simulations. Octopus is a simulation code written in Fortran and based on time-dependent density functional theory (TDDFT), developed collaboratively at MPSD and MPCDF. It performs real-time electron dynamics simulations from first principles and requires high computational throughput and scalability on modern supercomputers. Octopus has long supported distributed-memory and shared-memory parallelism via MPI and OpenMP. Moreover, GPU acceleration has been introduced, with support

for CUDA, HIP, and OpenCL backends. To ensure portability across GPU vendors, Octopus was ported to run on AMD GPUs using the HIP backend and portable abstraction layers. The code is fully multi-GPU capable, using MPI-based domain decomposition to distribute workloads. Efforts focused on porting performance-critical kernels, optimizing memory layout, and minimizing host-device data transfers. Despite the challenges of adapting a mature Fortran codebase to heterogeneous architectures, the GPU implementation achieves significant speedups over CPU-only configurations. In this talk, we present the GPU-enabled architecture of Octopus, outline the porting process to AMD GPUs, and share lessons learned in adapting a real-world scientific code to heterogeneous, multi-GPU platforms.

Alessandro Casalino
Max Planck Computing and Data Facility
Max Planck Institute for Structure and Dynamics of Matter
alessandro.casalino@mpcdf.mpg.de

MS29

Comparing Two Porting Strategies per Application - Lessons Learned from Accelerating FS3D and NS3D with OpenMP and HIP

Scientific codes are oftentimes developed in small research groups with the primary goal of delivering results for their domain specific research projects. Thus, the developers are most often domain experts that must focus on scientific discovery. This poses significant challenges on the time and effort that can be invested in porting such codes to GPUs. At the same time, performance is a necessity for simulating increasingly complex problems with HPC hardware. FS3D and NS3D are both CFD in-house codes from the Faculty of Aerospace Engineering at the University of Stuttgart which are facing this balancing challenge. Both user groups are among the most frequent users of the supercomputer Hunter at HLRS in Stuttgart. Both codes have been ported to GPUs recently by using OpenMP. At the same time, some of their key components have been ported with HIP by the AMD support team. This allows the rare opportunity of comparing approaches and extracting best practices for developers facing the decision of how to port their own code. This presentation compares both programming models for each of the two codes with regards to performance, portability, ease-of-use and invested development time. Moreover, some best practices and lessons learned will be shared on how existing in-house simulation codes can be ported effectively to GPUs.

Johanna Potyka, Marius Kurz
Advanced Micro Devices GmbH
johanna.potyka@amd.com, marius.kurz@amd.com

Tobias Gibis
Institute of Aerodynamics and Gas Dynamics
University of Stuttgart
tobias.gibis@iag.uni-stuttgart.de

David Gösele, Jonathan Wurst
Institute of Aerospace Thermodynamics
University of Stuttgart
david.goesele@itlr.uni-stuttgart.de,
jonathan.wurst@itlr.uni-stuttgart.de

Thomas Gibson, Ossian O'Reilly
Advanced Micro Devices
thomas.gibson@amd.com, ossian.oreilly@amd.com

Christoph Wenzel
Institute of Aerodynamics and Gas Dynamics
University of Stuttgart
wenzel@iag.uni-stuttgart.de

Kathrin Schulte
Institute of Aerospace Thermodynamics
University of Stuttgart
kathrin.schulte@itlr.uni-stuttgart.de

Paul Saumet
Höchstleistungsrechenzentrum Stuttgart
University of Stuttgart
paul.saumet@hlrs.de

MS29

Enabling Accelerated 3D Nonlinear Hybrid Fluid-Kinetic MHD Simulations with the Finite Element Code JOREK on GPUs

Numerical simulations play a crucial role in studying plasma instabilities in magnetic confinement fusion devices, enabling scientists to identify viable mitigation strategies for future fusion power plants. JOREK is a widely used, fully implicit, nonlinear, extended magnetohydrodynamic (MHD) Fortran code, parallelized on traditional CPU clusters using MPI and OpenMP, that implements a range of extended physics models. However, with many new supercomputers integrating accelerator architectures, it is crucial to adapt JOREK to leverage the capabilities of these new systems. This task naturally presents considerable challenges. First, the code is extensive and complex. Second, any efforts should be portable to different accelerator architectures. Lastly, changes to the source code should be minimal, as scientists frequently modify it. We primarily focus our efforts on OpenMP target offloading to address these challenges. We ported critical parts of the fluid time step and the kinetic particle evolution for Runaway Electrons (REs) to GPUs. Additionally, we utilized optimized libraries from ROCm or CUDA, which we call from within the OpenMP source code. We successfully verified the code on various clusters, ranging from machines with dedicated GPUs to those with modern unified memory. We demonstrate portability, consistent performance results, and scalability, enabling production simulations on accelerated clusters.

Patrik Rac, Edoardo Carrà, Matthias Hölzl
Max Planck Institute for Plasma Physics
patrik.rac@ipp.mpg.de, edoardo.carra@ipp.mpg.de, matthias.hoelzl@ipp.mpg.de

MS29

Fortran on AMD GPUs: Progress report on adapting ECMWF's Integrated Forecasting System to GPUs

The use of GPUs has become widespread in HPC to achieve unprecedented throughput and computational performance. The efficient usage of GPU architectures, not only for data-driven but also for physical Earth System Models, has long been envisioned at ECMWF. However, ECMWF's operational Integrated Forecasting System (IFS), like many codes in weather and climate, is written in Fortran, has been developed over decades and tuned to CPU architectures, and is continually being developed

further by domain scientists. As a consequence, adapting the model to GPUs requires invasive code refactoring throughout a very large code base and low-level optimisation that can harm CPU performance. We took a multi-tiered approach towards a GPU-accelerated forecast model: (i) The extraction of components and mini-apps (dwarves) to facilitate developments and collaboration on smaller code bases, (ii) the introduction of FIELD_API, a GPU-aware data structures library, to manage data residency and movement, and (iii) the development of Loki, a source-to-source translation package capable of transforming the CPU-native IFS Fortran to GPU-optimised code. In this talk we present recent progress towards GPU-execution of IFS dwarves and the forecast model on AMD GPUs, built upon the advances of AMD's AOMP compiler for Fortran. We include a discussion of strategies towards multi-platform GPU support, showcase the most recent developments and highlight the successes, pitfalls and lessons learned.

Michael Staneker, Balthasar Reuter, Ahmad Nawab, Olivier Marsden, Michael Lange
ECMWF
michael.staneker@ecmwf.int, balthasar.reuter@ecmwf.int, ahmad.nawab@ecmwf.int, oliver.marsden@ecmwf.int, michael.lange@ecmwf.int

MS30

Efficient strategies for surrogate modeling, uncertainty quantification, sensitivity analysis, and Bayesian calibration of high-dimensional models with time-varying outputs

This work unifies and compares algorithms for analyzing complex models computationally expensive and involving a high number of uncertain parameters that produce time-varying quantities of interest (QoIs). Our main focus is on Polynomial Chaos Expansion (PCE)-based methods for propagating uncertainty from stochastic parameters to model outputs. To exploit redundancy in multivariate outputs, we employ KarhunenLove (KL) expansion for output dimensionality reduction and learn PCE representations only for the retained dominant KL modes. We also compute generalized Sobol indices as an alternative to time-varying variance-based indices for vector- or function-valued QoIs. To keep the computational costs affordable in high-dimensional settings, we rely on sparse quadrature and sparse polynomial schemes. Finally, we address Bayesian calibration via parallel-chain MCMC posterior sampling, with emphasis on error modeling for time-dependent QoIs and acceleration through re-use of surrogate models. Common to all the algorithms we investigate is that we aim to deliver an efficient non-intrusive implementation suitable for running on HPC clusters. As applications, we test two hydrological modelsthe HBV-SASK benchmark and the operational LARSIM modelas well as a high-dimensional electrochemical model of lithium-ion batteries. Overall, our work bridges theoretical advances in UQ with complex real-world problems, offering a valuable tool for decision-makers.

Ivana Jovanovic
Technical University of Munich
ivana.jovanovic@tum.de

Tobias Neckel
Technical University Munich
neckel@cit.tum.de

Hans-Joachim Bungartz
Technical University of Munich
bungartz@cit.tum.de

MS30

Multi-GPU Implementation of the Levy and Lindenbaum's Method for Streaming SVD with Application to Large Datasets of KHRTI Simulation

Singular Value Decomposition (SVD) is a fundamental technique in data-driven approaches, particularly in fluid mechanics. It enables the decomposition of complex datasets, the extraction of dominant flow structures, and the construction of reduced-order models, thereby significantly reducing the complexity of physical systems. However, when applied to extremely large datasets involving millions of degrees of freedom, traditional SVD methods, even in parallel settings, become computationally prohibitive or infeasible. To overcome this limitation, the innovative approach proposed by Levy and Lindenbaum for computing low-rank approximations can be employed. This technique allows for streaming data processing, significantly reducing computational complexity in both space and time, and enabling the treatment of datasets that far exceed available memory. We propose a new parallel implementation of this method, leveraging hybrid architectures. As a case study, we apply our implementation to the KelvinHelmholtzRayleighTaylor instability (KHRTI), arising from a 3D direct numerical simulation (DNS), which comprises 480 million spatial degrees of freedom and 700 snapshots, demonstrating the scalability, robustness, and practical efficiency of our implementation in handling very large datasets.

Lucas Lestandi
Ecole Centrale de Nantes
lucas.lestandi@ec-nantes.fr

Sengupta Aditi
IIT (ISM) Dhanbad
aditi@iitism.ac.in

Yassin Ajanif, Rocha da Silva Luisa
Ecole Centrale Nantes
yassin.ajanif@ec-nantes.fr,
luisa.rocha-da-silva@ec-nantes.fr

Bhavna Joshi
IIT (ISM) Dhanbad
21dr0067@mech.iitism.ac.in

MS30

Uncertainty Quantification with Neural Networks for Parameter Inference in Complex Systems

Inverse problems - ubiquitous in computational science and engineering - ask what are the hidden parameters driving a physical system such as (deterministic) differential equations or a stochastic process. Uncertainty quantification (UQ) for inverse problems asks how sensitive are the parameters to changes in the observational data, where the two are linked via the forward problem. Each layer of complexity - forward problem, inverse problem, UQ - dramatically increases the computational challenges. For example, simple forward problems become challenging inverse problems if the data misfit term generates highly nonlinear and/or nondifferentiable mappings. This talk investigates whether neural networks can break some of the complexity. The goal is that a neural network (NN) represents an

inverse mapping - from observational data to parameters. NNs can be trained to predict point estimates or empirical representations of distributions. Using NNs as inverse mappings is also referred to as amortized inference; and the literature can classify inverse problems and UQ (with or without NNs) as simulation-based inference. The talk considers several open challenges remaining at the intersection of inverse problems, deep learning, and computational mathematics: determining the NN expressivity required to approximate solutions of ill-posed inverse problems; quantifying the impact of finite training data and noise; generative NNs as scalable sampling methods for Bayesian inverse problems.

Johann Rudi
Virginia Tech
jrudi@vt.edu

Andreas Mang
University of Houston
Department of Mathematics
andreas@math.uh.edu

Deep Ray
University of Maryland, College Park
deepray@umd.edu

Ionut-Gabriel Farcas
Virginia Tech
farcasi@vt.edu

MS31

Heterogeneous Task-Based Parallelism with Flecsi

Large HPC systems in the exascale era mostly rely on a combination many-core CPUs and multiple GPUs per nodes to achieve their outstanding performance. This poses a significant burden on the application developer, particularly in the context of multiphysics simulations where the computational load varies during the evolution of the system and data movement between the various memory spaces can be very complex. FleCSI is an open-source C++ framework for task-based parallelism aimed at multiphysics simulation codes. It allows the application developer to focus on the numerical algorithm and leave the data movement and task-scheduling to the runtime. FleCSI leverages Kokkos for performance portability and supports Legion and HPX as backends for task-based parallelism. This presentation will show how we use FleCSI in a real-world multiphysics code. We show how task-parallelism is exposed through dependency analysis and how the OpenMP and CUDA/ROCm backends of Kokkos are used by concurrent tasks to make use of the full computational potential of a node.

Philipp V. Edelmann
Los Alamos National Laboratory
pedelmann@lanl.gov

MS31

Asc Tri-Lab Co-Design Efforts, a Retrospective and Update from El Capitan

In this talk we will discuss experiences standing up several applications on the El Capitan system in early access and into production. We will discuss how we interacted with the Center of Excellence leading up to El Capitan, and then provide examples of issues we observed on El Capitan. We will discuss these issues and provide insights where proxy or mini applications could have mitigated the challenges. The talk will target the challenges faced when moving from test systems with 10s to a hundred nodes, to a leadership-class system with 1000s of nodes.

James Elliot
Sandia National Laboratories
jjellio@sandia.gov

MS31

System Noise at Exascale: An LLNLHPE Co-Design Approach

To meet the growing demands of parallel scientific applications, supercomputers continue to scale in size and complexity. El Capitan, the worlds fastest supercomputer, features over a million CPU cores and tens of thousands of GPUs. Applications on such systems are particularly vulnerable to system noiseinterference from the operating system (OS) and background services running on the same nodes. This talk presents the result of a co-design collaboration between Lawrence Livermore National Laboratory and Hewlett Packard Enterprise to address this performance and scalability challenge on El Capitan. Our strategy includes (1) isolating system services from application workloads and (2) applying OS-level tuning to reduce interference. Together, these techniques enable science applications to better leverage the capabilities of Exascale-class machines. This work demonstrates the real-world impact of close collaboration between a national laboratory and industry, helping future systems deliver on their full scientific potential.

Edgar A. Leon
Lawrence Livermore National Laboratory
leon@llnl.gov

MS31

Understanding Mi300aspecificities and Their Impact on the CodeDevelopment

The heterogeneous and GPU era brought new challenges, especially regarding data management between the different compute resources.To alleviate this burden from programmers shoulders, new hardware came with transparent heterogeneous memory accesses bewteen CPUs and Gpus (e.g., Nvidia Grace+Hopper and following, or AMD instinct MI 300A).To fully maximize the potential and performance of this type of chip, it is essential to understand their new mechanisms and their impact on applications.This presentation will present our study of MI300A behaviors, particularly focusing on memory management and data transfers. Additionally, we will compare and highlights the behavior differences between MI300A and Grace+Hopper superchip.

Julien C. Jaeger
CEA
julien.jaeger@cea.fr

Adrien Roussel
CEA
The French Alternative Energies and Atomic Energy Commission

adrien.roussel@cea.fr

MS32

Parallel Acceleration of the Highs Solvers

HiGHS has established itself as the worlds best open-source software for linear optimization, and is used extensively by open-source energy modelling systems. However, there is still a significant gap between its performance and that of major commercial software, and this has led to major donations for enhancing HiGHS from philanthropic institutions supporting energy transition. Historically, HiGHS has solved linear programming (LP) and mixed-integer programming (MIP) problems using highly efficient serial implementations of the simplex algorithm, interior point method, and branch-and-cut. This talk will present recent work to exploit GPUs and multi-threading on CPUs in three contexts. Firstly the new GPU-accelerated HiGHS primal-dual hybrid gradient LP solver (HiPDLP), secondly the new multi-threaded interior point solver (HiPO), and thirdly the use of multi-threading to accelerate primal heuristics and the tree search in the MIP solver. Observations on the practicalities of using GPUs in the solution of MIPs will also be made.

Julian Hall
University of Edinburgh
jajhall@ed.ac.uk

MS32

Low-Precision First-Order Method-Based Fix-and-Propagate Heuristics for Large-Scale Energy System Models

We investigate the use of low-precision first-order methods (FOMs) within a fix-and-propagate (FP) framework for solving mixed-integer programming problems (MIPs). We employ GPU accelerated PDLP, a variant of the Primal-Dual Hybrid Gradient (PDHG) method specialized to LP, to solve the LP-relaxation of our MIPs to low accuracy. This solution is used to motivate fixings within our fix-and-propagate framework. We evaluate the performance of our heuristic on MIPLIB 2017, showcasing that the low-accuracy LP solution produced by the FOM does not lead to a loss in the quality of the FP heuristic solutions. Further, we use our FP framework to produce high-accuracy solutions for large-scale (up to 243 million non-zeros and 8 million decision variables) unit-commitment energy-system optimization models created with the modeling framework Remix. For the largest problems, we can generate solutions with under 2% primal-dual gap in less than 4 hours, whereas commercial solvers cannot generate feasible solutions within two days of runtime.

Nils-Christian Kempke
Technische Universität Berlin
kempke@zib.de

Thorsten Koch
Zuse Institute Berlin
koch@zib.de

MS32

Solving hard MIP instances on Supercomputers by using Smoothie

The Ubiquity Generator (UG) Framework is a software framework to parallelize state-of-the-art solvers. Based on UG, SCIP, and its customized solvers for Steiner tree problems (SCIP-Jack), QUBOs (Qubowl), and Pseudo-Boolean optimization are parallelized. The UG framework has been generalized in version 1.0.0 so that the parallelization controlling mechanism can be highly customized. This feature allows us to develop ensemble solvers, in which several different solvers run in parallel and communicate information. An experimental implementation of Smoothie (Solver Mixing Object Oriented Hybrid Integrated Executable), in which SCIP and HiGHS can run in parallel, has been developed. In this talk, we present the latest status of Smoothie where commercial solvers Xpress and Gurobi are used jointly to solve hard MIP instances.

Yuji Shinano
Zuse Institute Berlin
Takustr. 7 14195 Berlin
shinano@zib.de

MS32

Application of the Massively Parallel Solver PIPS-IPM++ for the Computation of European Energy System Transformation Pathways in High Spatial and Temporal Resolution

Long-term capacity expansion models are important tools that provide decision-makers with information about the future role of energy infrastructure. However, accurately reflecting the intermittent nature of renewable energy sources and the complex network topologies of electricity, natural gas, and hydrogen results in optimization models with hundreds of millions of variables and constraints. This creates a significant computational burden in terms of the total time required to solve the optimization problems and possible memory limitations. One beneficial aspect of these models is that the underlying optimization problem is sparse and well-structured. This allows the problem to be efficiently addressed using block-structure exploiting solvers, such as PIPS-IPM++. Furthermore, PIPS-IPM++ enables the utilization of high-performance clusters through distributed computing approaches, such as MPI and OpenMP. However, the solver's performance depends not only on the available hardware, but also on the model's decomposition. This talk demonstrates the application of the PIPS-IPM++ solver to instances from the REMix modeling framework for a case study on transformation pathways of the European energy system. The study focuses on achieving high hydrogen integration rates. Additionally, the talk discusses how to efficiently exploit the underlying problem structure and how the choice of decomposition can significantly impact the performance of the massively parallel solver.

Manuel Wetzel, Karl-Kin Cao, Shima Sasanpour
DLR-Institut für Vernetzte Energiesysteme
manuel.wetzel@dlr.de, karl-kien.cao@dlr.de,
shima.sasanpour@dlr.de

MS33

Algebraic Temporal Blocking for Sparse Iterative Solvers

Sparse linear iterative solvers are essential for large-scale simulations. In many of these simulations, the majority of the runtime is spent in matrix power kernels (MPK), which compute the product of a power of a sparse matrix A and a dense vector x, i.e., $A^p x$. Current state-of-the-art implementations perform MPK by executing repeated, back-to-back sparse matrix-vector smultiplications (SpMV), which

requires streaming the large matrix A from main memory p times. Using RACE, we accelerate MPK computations by keeping parts of matrix A in cache across successive SpMV calls. RACE uses a level-based approach: levels are constructed using a breadth-first search on the graph corresponding to the sparse matrix. These levels enable cache blocking of matrix elements, maximizing both spatial and temporal reuse. This approach is highly efficient, achieving a 2×–4× speedup on a single modern Intel or AMD multicore chip, compared to a highly optimized classical SpMV-based implementation. Recently, the method has also been extended to distributed-memory settings and has demonstrated good scalability for large-scale graphs common in computational science and engineering. After briefly presenting RACEs cache-blocking approach, the talk will explore the application of cache-blocked MPK kernels in iterative solvers. Among others, we discuss the applicability of RACE in s-step Krylov solvers, polynomial preconditioners, and multigrid methods.

Christie Louis Alappat
Friedrich-Alexander-Universität Erlangen-Nürnberg
Erlangen, Germany
christie.alappat@fau.de

Dane Lacey
NHR@FAU, Friedrich Alexander-Universität
Erlangen-Nürnberg
Germany
dane.c.lacey@fau.de

Georg Hager
Friedrich-Alexander-Universität Erlangen-Nürnberg
georg.hager@fau.de

Jonas Thies
German Aerospace Center (DLR)
j.thies@tudelft.nl

Gerhard Wellein
Friedrich-Alexander-Universität Erlangen-Nürnberg
gerhard.wellein@fau.de

MS33

Parallel Selected Inversion: Recent Algorithmic Developments and Applications

Sparse direct methods are a cornerstone of scientific computing, enabling the solution of large linear systems. In typical settings, the number of right-hand sides (RHS) is small compared to the matrix dimension, so computational and memory costs are dominated by the factorization of the system matrix. However, some applications depart from this typical regime and require solving linear systems with a number of RHS comparable to the system dimension. In this regime, computational cost shifts to the triangular solves with cubic scaling, while memory requirements grow quadratically, rendering classical approaches infeasible. In this regime, so-called selected inversion algorithms compute only prescribed entries of the inverse matrix, avoiding the formation of dense intermediate results. While these methods have received significant attention, existing implementations remain largely limited to shared-memory systems and offer limited support for GPU acceleration, constraining scalability towards large problem sizes. In this talk, we present recent algorithmic developments for parallel selected inversion, targeting distributed-memory systems and GPU-accelerated architectures by exploiting structured sparsity arising from the targeted applications.We demonstrate the impact of these methods on two large-scale applications: Bayesian statistical modelling and quantum transport simulations.

Vincent Maillou
ETH Zurich
vincent.maillou0@gmail.com

Matthias Bollhoefer
TU Braunschweig
m.bollhoefer@tu-bs.de

Olaf Schenk
Università della Svizzera italiana
Switzerland
olaf.schenk@usi.ch

Alexandros Nikolaos Ziogas
ETH Zurich
alziogas@iis.ee.ethz.ch

Mathieu Luisier
Integrated Systems Laboratory
ETH Zurich
mluisier@iis.ee.ethz.ch

MS33

Efficient Parallel Scheduling for Sparse Triangular Solvers

Parallel sparse triangular solve is a problem troubled with irregular dependency patterns, limited parallelism, fine-grained operations and synchronisations. Numerous algorithms have been put forward to address these issues. In this talk, we present our new technique to generate synchronous schedules based on the popular list-scheduling method for directed-acyclic-graph scheduling, and place this technique within the context of AndersonSaad and Park et al. Compared to state-of-the-art methods, our approach allows a reduction of the number of synchronisation barriers by 10x as well as a reduction of execution times by 1.42x. We furthermore show that our improvements are consistent across a variety of input matrices and hardware architectures.

Toni Böhnlein, Pál András Papp, Raphael S. Steiner, Christos Matzoros
Computing Systems Lab
Huawei Research Center Zurich
toni.boehnlein@huawei.com,
pal.andras.papp@huawei.com,
raphael.steiner@huawei.com,
christos.konstantinos.matzoros@h-partners.com

Albert-Jan Yzelman
Huawei Zurich Research Center
Computing Systems Lab
albertjan.yzelman@huawei.com

MS34

BLAST WarpX: Performance Portable, Load-Balanced Exascale Simulations with Mesh Refinement

WarpX is an open-source Particle-In-Cell(PIC) code hosted by the High Performance Software Foundation and designed to simulate physical scenarios involving kinetic plasmas, from laser-plasma interaction to plasma astrophysics,

accelerators, and fusion devices. WarpX is a portable and highly-optimized code that can leverage the computing power of the largest supercomputers to tackle challenging scientific problems. This contribution presents how advanced features of the WarpX code, such as Mesh Refinement(MR), and Load Balancing(LB), can be used to speedup simulations of laser-plasma interaction on a variety of High Performance Computing machines. MR, in particular, is a unique feature among electromagnetic PIC codes. The context of this work is a simulation campaign carried out to study advanced laser-driven electron acceleration schemes. By making an ultra-intense laser interact with a gas jet it is possible to accelerate electron bunches up to energies of several gigaelectronvolts in a few centimeters of plasma. However, the accelerated charge is usually very low (few picocoulombs) for most potential applications. A promising strategy to increase the accelerated charge is to use a hybrid target consisting in a solid foil coupled with a gas jet. When the laser interacts with the solid a substantial amount of charge is extracted and then accelerated by the plasma perturbations generated in the gas. This contribution discusses how these simulations can benefit from MR and LB.

Luca Fedeli
French Alternative Energies and Atomic Energy Commission
luca.fedeli@cea.fr

MS34

Advanced AI/ML-Coupling, Code Extensions, and Differentiable Programming in BLAST Codes

In order to achieve high performance and portability, high-performing code for large-scale simulations and AI/ML engines like Torch are predominantly written in modern versions of ISO C++. Yet, at the same time, a need for productivity in scientific computing led to a proliferation of JIT (e.g. Python) interfaces that are ubiquitous in modern AI/ML workflows, from surrogate training, over differentiable programming DSLs, to optimization of highly dimensional parameter sets. Based on our experience in modeling largest-scale systems in beam, plasma and particle accelerator science (culminating in the 2022 Gordon Bell award for our code WarpX) and corresponding HPC code development based on the AMReX framework, we show how AI/ML, user-level code extensions, and traditional HPC can be tightly integrated. We also report on progress of introducing differentiable programming in C++ for codes in the Beam, Plasma & Accelerator Simulation Toolkit (BLAST), which depend on AMReX, using modern compiler techniques like the LLVM plugin Enzyme. We present early successes, our vision, challenges and next steps.

Axel Huebl
Lawrence Berkeley National Laboratory (LBNL), USA
axelhuebl@lbl.gov

MS34

Computational Multiphysics of Interacting States of Matter under Extreme Conditions

We present a suite of optimized multiphysics codes for multiple interacting materials and states of matter, based on AMReX, that have been used to model a wide range of physical problems in areas ranging from plasma simulation for fusion reactors to deep geothermal energy extraction. After providing a brief overview of the codes and various application areas they have been applied to, two specific topics will be discussed in further detail. The first topic involves demonstrating accuracy and performance of one of the codes in the modelling of high-energy millimetre-wave based ablation of geological materials for supercritical (deep geothermal) energy recovery, using a variable-density, incompressible multiphase model. Flat MPI-parallelisation is used to integrate the system over long time scales for this application. The second topic involves the design of code features which facilitate polymorphism on GPU architectures and permit the development of complex multiphysics algorithms on NVIDIA hardware.

Nikos Nikiforakis, Philip Blakely, Nandan Gokhale
Laboratory for Scientific Computing, Cavendish Laboratory
University of Cambridge
nn10005@cam.ac.uk, pmb39@cam.ac.uk, nbg22@cam.ac.uk

MS34

ERF and REMORA: A Story of Two Codes

ERF and REMORA are simulation codes for regional modeling of atmospheric and oceanic flows, respectively. Both codes use adaptive mesh refinement (AMR) to efficiently achieve high resolution in the regions of most interest, which may change dynamically as a simulation evolves. They take advantage of the performance portability and support for AMR provided by the software framework, AMReX. Recent developments include enhanced coupling capabilities for interfacing with external modeling systems and integration of advanced physics modules. Both codes now support more sophisticated physical process representations, while benefiting from improved computational performance and numerical solver optimizations. In this talk, we will describe our general strategies for efficiently building new codes using existing core numerics but a new software strategy.

Jean M. Sexton, Aaron M. Lattanzi, Hannah Klion, Ann S. Almgren
Lawrence Berkeley National Laboratory
jmsexton@lbl.gov, amlattanzi@lbl.gov, klion@lbl.gov, asalmgren@lbl.gov

MS35

Parallel High-Resolution Partial Fft-Gmres Algorithms for Subsurface Scattering Problems

In this paper, we present a novel parallel algorithm for 3D Helmholtz scattering problems that combines partial FFT (PFFT) preconditioning with the GMRES method for large-scale subsurface simulations. The solver employs high order compact finite-difference discretizations of the 3D Helmholtz equation with spatially varying coefficients. A key innovation is the use of PFFT-based preconditioners derived from lower-order approximations to significantly accelerate convergence. Our implementation efficiently leverages a hybrid OpenMP and MPI multi-node parallel architecture. We analyze the computational complexity and parallel scalability of the method under realistic physical parameters.

Yuiry Gryazin
Idaho State University

gryazin@isu.edu

MS35

Loracx: Low Rank Approximations with Constraints at Exascale

Analyzing large-scale scientific datasuch as molecular dynamics simulations of MoS2 recrystallizationposes significant challenges for traditional methods like Nonnegative Matrix Factorization (NMF), particularly on exascale systems. In this talk, we introduce Low-Rank Approximations with Constraints at Exascale (LORACX), a scalable framework that employs distributed, GPU-accelerated NMF integrated into a modern, Python-based HPC stack. Key innovations include communication-efficient designs using blocked and overlapped algorithms to mitigate latency and memory constraints, as well as GPU-optimized Nonnegative Least Squares (NNLS) solvers. Performance evaluations on up to 8,192 Frontier nodes demonstrate strong scalability, processing a 16.3 16.3 million matrix in 3 seconds and achieving 0.67 exaflops in double precision. We present detailed weak-scaling results, including computational versus communication cost analyses, and show that baseline comparisons consistently confirm the superior performance of LORACX-GPU.

Ramakrishnan Kannan
Oak Ridge National Laboratories
kannanr@ornl.gov

MS35

Parallel Higher-Order Orchogonal Iteration for Tucker Decomposition

Higher Order Orthogonal Iteration (HOOI) is an iterative algorithm that computes a Tucker decomposition of fixed ranks of an input tensor. In this work we modify HOOI to determine ranks adaptively subject to a fixed approximation error, apply optimizations to reduce the cost of each HOOI iteration, and parallelize the method in order to scale to large dense datasets. We show that HOOI is competitive with the Sequentially Truncated Higher Order Singular Value Decomposition (STHOSVD) algorithm, particularly in cases of high compression ratios. Our proposed rank-adaptive HOOI can achieve comparable approximation error to STHOSVD in less time, sometimes achieving a better compression ratio. We demonstrate that our parallelization scales well over thousands of cores and show using three scientific simulation datasets that HOOI outperforms STHOSVD in high-compression regimes. For example, for a 3D fluid-flow simulation dataset, HOOI computed a Tucker decomposition 82x faster and achieved a compression ratio 50% better than STHOSVD's

J De Oliveira Pinheiro
Purdue University
deolivj@purdue.edu

MS36

Machine Learning-enhanced overlapping Schwarz solvers

Domain decomposition methods (DDMs) are robust and parallel efficient iterative solvers for discretized partial differential equations. However, the convergence rate of classic DDMs deteriorates for coefficient distributions with large contrasts. To retain the robustness for such problems, the coarse space of the DDM can be enriched by additional coarse basis functions, often obtained by solving local generalized eigenvalue problems. Within overlapping Schwarz methods, we consider the AGDSW (adaptive generalized Dryja-Smith-Widlund) coarse space to obtain robustness. However, the computation of the AGDSW coarse basis functions is computationally expensive due to the solution of many local eigenvalue problems. In this talk, we train a surrogate model based on a deep feedforward regression neural network which directly learns the necessary coarse basis functions. Additionally, we present first results where we also replace the solution of local subdomain problems by surrogate models. This talk is based on joint work with Martin Lanser and Janine Weber, University of Cologne, Germany.

Axel Klawonn
Department of Mathematics and Computer Science and Center
for Data Centand Simulation Science, University of Cologne
axel.klawonn@uni-koeln.de

Martin Lanser
Koeln University
mlanser@math.uni-koeln.de

Janine Weber
Universität zu Köln
Department of Mathematics and Computer Science
janine.weber@uni-koeln.de

MS36

Local random feature filtering for scalable and well-conditioned physics-informed neural networks

Domain-decomposed random feature-based physics-informed neural networks provide a scalable way to solve PDEs by using localized, overlapping, and randomly initialized neural network basis functions to approximate the PDE solution and training them through structured least-squares problems. However, the resulting least-squares systems are often severely ill-conditioned due to redundancy among random basis functions and correlations introduced by subdomain overlaps, which significantly affect the convergence of standard solvers. In this talk, we introduce a block rank-revealing QR filtering and preconditioning strategy that operates directly on the structured least-squares problem. This strategy removes redundant basis functions, improves conditioning, preserves sparsity, and greatly accelerates the least-squares solver. Across challenging multi-scale PDEs, we observe up to 11 orders of magnitude condition number reduction, 101000 faster convergence, and higher accuracy at lower costmaking RRQR filtering an efficient enhancement for RFM-based solvers.

Benjamin Moseley
Imperial college of London
b.moseley@imperial.ac.uk

MS36

Domain Decomposition Based Physics-Informed Neural Network for Maxwell Equations

In the last decade, the study of electromagnetic waves has been a primary concern in many applications such as designing radar or antenna technology as well as cloaking devices and metalenses which fill the role of a lens through the use of metamaterials. The electromagnetic field fol-

lows the Maxwell equations that in most applications have no analytic solutions. Traditional ways to solve this system, which are based on numerical methods such as the finite difference or the finite element methods, are computationally expensive. This becomes a problem when a lot of simulations need to be done such as when doing numerical optimisation. Physics-informed neural networks (PINNs) have been introduced only recently in the world of Scientific Machine Learning (SciML) as a novel way to solve systems of partial differential equations. In this work, we study the application of PINNs for the Maxwell equations to generate surrogate models, possibly trading accuracy for faster results. We will in particular discuss about domain decomposition approaches to enhance the training and try to avoid common problems with PINNs

Alexandre Pugin
INRIA
alexandre.pugin@inria.fr

MS36

Parallel Non-convex Minimisation for Large Scale Problems in Machine Learning and Mechanics

Parallel training methods are increasingly relevant in machine learning (ML) due to the continuing growth in model and dataset sizes. We propose a variant of the Additively Preconditioned Trust-Region Strategy (APTS) for training deep neural networks (DNNs). The proposed APTS method utilizes a data-parallel approach to construct a nonlinear preconditioner employed in the nonlinear optimization strategy. In contrast to the common employment of Stochastic Gradient Descent (SGD) and Adaptive Moment Estimation (Adam), which are both variants of gradient descent (GD) algorithms, the APTS method implicitly adjusts the step sizes in each iteration, thereby removing the need for costly hyperparameter tuning. We demonstrate the performance of the proposed APTS variant using the MNIST and CIFAR-10 datasets. The results obtained indicate that the APTS variant proposed here achieves comparable validation accuracy to SGD and Adam, all while allowing for parallel training and obviating the need for expensive hyperparameter tuning.

Bindi apriqi
KAUST
bindi.capriqi@kaust.edu.sa

Ken Trotti
Department of Science and High Technology
University of Insubria
kl.trotti@uninsubria.it

MS37

Block-Wise Mixed Precision Sparse Direct Solver on GPUs

Sparse direct solvers for large-scale linear systems typically rely on a single floating-point precision, which limits their ability to balance computational efficiency and numerical accuracy. We present a block-wise mixed precision LU factorisation algorithm for GPUs, integrated into the distributed sparse direct solver PanguLU. The method adaptively assigns precision to matrix blocks according to their position in the global matrix and the numerical sensitivity of their nonzero entries, thereby enabling mixed precision computations during the factorisation phase. Experimental results show that the proposed approach accelerates solution time while controlling accuracy loss, offering a favorable trade-off between performance and precision compared with single-precision solvers.

Yida Li
China University of Petroleum-Beijing
yida.li@student.cup.edu.cn

MS37

Floating Point Compression in a Parallel Task-Based Hierarchical Matrices Solver

The H-Matrix algebra offers numerous advantages for solving linear systems, both in terms of memory consumption and computation time compared to more traditional direct or iterative solvers. However, in an industrial context, its memory and disk consumption are limiting the size of the problems it can handle. Thus, reducing the memory footprint of the matrices is a major challenge. Furthermore, this allows for alleviating disk storage requirements for "out-of-core" computations and reducing communication overhead, a critical factor, particularly in calculations on distributed memory architectures. In this perspective, we focus on the floating-point compression of the different blocks of an H-Matrix in an industrial context. The goal is to reduce the memory footprint during computations, leading to gains in both space and time, while maintaining controlled precision loss. To this end, several arithmetic compression schemes (ZFP, SZ3) are considered and compared to determine those that offer the best compression rates for a given precision. Initial tests were performed on a sequential (open-source) version of the H-Matrix library. Subsequently, this compression was integrated into a parallel (proprietary) version of the code. The objective is thus to address larger-scale problems (allowing for finer and more precise modeling) or problems of the same size but with reduced spatial and temporal costs.

Clément Peaucelle
Airbus
clement.peaucelle@airbus.com

MS37

Three-Precision Iterative Refinement with Parameter Regularization and Prediction for Solving Large Sparse Linear Systems

We present a mixed-precision formulation of the General Alternating-Direction Implicit (GADI) method for solving large, sparse linear systems Ax=b. The proposed scheme executes the costly subsystem solves in low precision to accelerate throughput, while computing residuals and solution updates in higher precision to preserve final accuracy. We develop a rounding/backward error analysis that quantifies convergence and attainable accuracy as functions of the working, residual, and solver precisions, and shows the central role of the GADI regularization parameter a in mitigating low-precision instabilities. Building on this theory, we propose a practical a-selection strategy: a Gaussian Process Regression model trained on small problems provides an initial a, which is then iteratively adjusted using our convergence condition to ensure robust mixed-precision performance. On a representative 3D convectiondiffusion benchmark, the mixed-precision GADI achieves about 2. speedup using bfloat16 for the subsystem solves compared to a full FP64 implementation, while maintaining high-accuracy solutions. The approach is broadly applicable across ADI splittings and hardware, offering a principled, drop-in path to accelerate stationary iterations with mod-

ern low-precision units without sacrificing reliability.

Juan Zhang
Xiangtan University
zhangjuan@xtu.edu.cn

MS37

Mixed-Precision AMG Method and Its Applications in Petroleum Reservoir Simulation

With the rapid development of heterogeneous computing, both CPUs and GPUs have demonstrated substantial potentials in supporting multi-precision computations. Effectively utilizing mixed-precision computing techniques to improve the efficiency of iterative solvers has emerged as a new research direction in computational science. This presentation begins by introducing a low-overhead sparse matrix storage format designed for mixed-precision computing frameworks, accompanied by a foundational linear algebra library tailored to this structure. Based on this matrix representation, a mixed-precision Algebraic Multigrid (AMG) preconditioned iterative algorithm is developed, which demonstrates significant acceleration in model problems. Furthermore, for complex reservoir simulation scenarios, a mixed-precision preconditioned iterative algorithm is proposed. Numerical experiments underscore the advantages and promising potential of the proposed approach in practical applications.

Chensong Zhang
Chinese Academy of Sciences
zhangcs@lsec.cc.ac.cn

MS38

rchitecture-Aware Optimization of the Ozaki Scheme on AMD Datacenter GPUs

As scientific computing increasingly utilizes mixed-precision hardware, algorithms like the Ozaki scheme have become essential for recovering high-precision (FP64/FP128) accuracy using fast, low-precision tensor cores. The scheme works by splitting input matrices into error-free slices, processing them via hardware-accelerated GEMM kernels, and reconstructing the final result. While performance on NVIDIA architectures is well-documented, effectively deploying this scheme on AMD CDNA architectures requires moving beyond simple code translation (e.g., hipify)This talk explores the performance implications of porting the Ozaki scheme to AMD MI-series GPUs, demonstrating that a naive conversion of CUDA primitives to HIP often yields suboptimal utilization. We investigate how fundamental architectural differences - specifically execution models (Wave64 vs. Warp32), memory hierarchy management (LDS vs. Unified Cache), and instruction scheduling - impact the algorithm's throughput. We discuss potential optimization strategies that leverage these distinct features, aiming to bridge the gap between a naive port and a hardware-optimized implementation. Our analysis highlights that achieving peak performance on AMD hardware requires a deep understanding of the underlying silicon, not just the software API.

Ilya Nyrkov
TU Munich, Germany
Germany
ilya.nyrkov@gmail.com

Hartwig Antzt
UTK
hanzt@icl.utk.edu

MS38

Leveraging Low-Precision Tensor Cores for Mixed-Precision Computing and Floating-Point Emulation

The reduced- and mixed-precision computing capabilities of GPUs have grown rapidly in recent years which has been primarily driven by AI-based applications (e.g., LLMs). These processors demonstrate an outsized power-efficiency (FLOPS/watt) advantage for matrix multiplications over systems almost exclusively focused upon native single- and double-precision arithmetic. Thus, this presents both a great opportunity and motivation to leverage these capabilities for dense linear algebra, through the use of various mixed-precision algorithms and floating-point emulation techniques, to facilitate greater scientific computing throughput without sacrificing accuracy. We'll touch upon a number of these approaches such as the Ozaki-I and Ozaki-II schemes for double-precision matrix multiplication emulation, and present real-world case studies that provide compelling evidence in support of this path to further increase the science per watt of GPU accelerated computing.

Benedikt Dorschner
NVIDIA
bdorschner@nvidia.com

MS38

Accelerating and emulating GEMM operations with devices and circuits in the age of AI

We are making progress on technical proposals and the initial implementation of the Ozaki scheme, which involves emulating high-precision matrix multiplication using low-precision techniques. These advances leverage classical multi-precision methods such as Error-Free Transform, polynomial decomposition, and the Chinese Remainder Theorem (CRT). Recent findings show that Ozaki Scheme 2 achieves a linear computational cost of O(P) by using CRT, where P is the number of truncations. It demonstrates that low-precision multiplication (input int8, output int32) is compatible with existing matrix processing engines, ensuring high performance with modern hardware accelerators, such as GPUs, TPUs, and NPUsknown as AI-acceleration devices. On new GPU platforms, such as the NVIDIA B200, emulation reaches approximately 100 TFLOPScomparable to or exceeding DGEMM, which is limited to 40 TFLOPS by FP64 arithmetic hardware. Using an intermediate CRT-capable accumulator and considering data formats allows for controlled output based on input parameters. This study provides detailed accounts of recent improvements, current results, and how variable precision affects the cost and accuracy of numerical algorithms using Ozaki Scheme 2, as well as the target AI accelerator hardware.

Toshiyuki Imamura, Yuki Uchino
RIKEN Center for Computational Science
imamura.toshiyuki@riken.jp, yuki.uchino.fe@riken.jp

Katsuhisa Ozaki
Shibaura Institute of Technology

ozaki@shibaura-it.ac.jp

MS38

Analysis of Floating-Point Matrix Multiplication Computed via Integer Arithmetic

Ootomo, Ozaki, and Yokota [Int. J. High Perform. Comput. Appl., 38 (2024), p. 297-313] have proposed a strategy to recast a floating-point matrix multiplication in terms of integer matrix products. The factors A and B are split into integer *slices*, the product of these slices is computed exactly, and AB is approximated by accumulating these integer products in floating-point arithmetic. This technique is particularly well suited to mixed-precision matrix multiply-accumulate units with integer support, such as the NVIDIA tensor cores or the AMD matrix cores. The number of slices allows for performance-accuracy tradeoffs: more slices yield better accuracy but require more multiplications, which in turn reduce performance. We propose an inexpensive way to estimate the minimum number of multiplications needed to achieve a prescribed level of accuracy. Our error analysis shows that the algorithm may become inaccurate (or inefficient) if rows of A or columns of B are *badly scaled*. We perform a range of numerical experiments, both in simulation and on the latest NVIDIA GPUs, that confirm the analysis and illustrate strengths and weaknesses of the algorithm.

Ahmad Abdelfattah
University of Tennessee, Knoxville
ahmad@icl.utk.edu

Jack J. Dongarra
University of Tennessee, Oak Ridge National Laboratory, USA
dongarra@icl.utk.edu

Massimiliano Fasi
University of Leeds
School of Computing
m.fasi@leeds.ac.uk

Mantas Mikaitis
University of Leeds
m.mikaitis@leeds.ac.uk

Françoise Tisseur
The University of Manchester
Department of Mathematics
francoise.tisseur@manchester.ac.uk

MS39

New Parallel Scheme for Accelerated DMRG for Molecular Electronic Systems

The density matrix renormalization group (DMRG) is a leading approach for describing strong electron correlation in challenging molecular systems relevant to, e.g., metalloenzymes or catalysis. Building on our massively parallel implementation MOLMPS (J. Comput. Chem. 2021, 42, 534544), which demonstrated scaling on ¿2,000 CPU cores for active spaces with 76 orbitals, we introduce a new parallel scheme that accelerates the most expensive part the Hamiltonianvector product used in the Davidson procedure. The method combines distributed-memory (MPI) parallelism with shared-memory threading (OpenMP) and GPU offload to better utilize node-level resources and reduce time-to-solution. We assess performance on demanding molecular systems.

Jiri Brabec
Czeck Academy of Sciences
jiri.brabec@jh-inst.cas.cz

MS39

Parallel Algorithms for Solving Tensor Eigenvalue Problem

We study hybrid tensor train algorithms for computing multiple eigenpairs of large sparse matrices with tensor product structure. While block methods provide eigenvector estimates without initial guesses, their orthogonalization subroutines are costly in both computation and communication. Refinement methods are cheaper and parallelize well, but they often exhibit convergence instabilities due to heuristic strategies or inexact applications of variational principles. We analyze the conditions under which such instabilities occur and propose strategies to improve robustness. In particular, we focus on a refinement scheme based on residual minimization and demonstrate, through numerical examples, that it achieves reliable convergence even in cases where existing methods face difficulties. We next introduce and benchmark new subspace filtering schemes to accelerate the convergence of tensor train subspace iteration, thereby enabling earlier use of residual refinement. Finally, we characterize how early the refinement methods can be applied to eigenvectors obtained from block methods to optimize the overall performance of these hybrid tensor train eigensolvers.

Peter DelMastro
Virginia Tech
pdelmastro@vt.edu

Alec Dektor, Erika Ye, Roel Van Beeumen
Lawrence Berkeley National Laboratory
adektor@lbl.gov, erikaye@lbl.gov, rvanbeeumen@lbl.gov

Chao Yang
Lawrence Berkeley National Lab
cyang@lbl.gov

MS39

Accelerating Distributed Multi-Gpu Ab Initio Density Matrix Renormalization Group Algorithm with Tensor Cores

The presence of many degenerate d/f orbitals in polynuclear transition-metal compounds poses significant challenges for state-of-the-art quantum chemistry methods. To address this, we developed the first distributed multi-GPU ab initio density matrix renormalization group (DMRG) algorithm, tailored for modern high-performance computing (HPC) infrastructures. By parallelizing the most computationally intensive stepthe multiplication of O(K) operators with a trial wave function, where K is the number of spatial orbitalsthrough operator parallelism and a batched GPU contraction scheme, we achieved an unprecedented bond dimension of D = 14,000 on 48 NVIDIA A100 GPUs for a P-cluster active space model (114 electrons in 73 orbitals). Furthermore, sparse tensor contraction(SpTC) is challenging in such high-performance applications due to the high dimensionality and inherent sparsity of tensors. We introduce Bullseye Hash, a novel hash table designed specifically for efficient SpTC computations. It employs a fast, collision-free hash function and optimizes operations according to the computation patterns of SpTC across di-

verse data objects. We also provide guidance on configuring object representations to balance algorithmic complexity and cache efficiency. Evaluations on 22 SpTC workloads demonstrate that our approach achieves up to 10.4× speedup (3.1× on average) and reduces memory usage by up to 67% (50% on-average) compared to state-of-the-art methods.

Weile Jia
Institute of Computing Technology Chinese Academy of Science
University of California, Berkeley
jiaweile@ict.ac.cn

MS39

A Two-Level Parallel Additive Schwarz View of Alternating Optimization with Tensor Trains

The Density Matrix Renormalization Group (DMRG) algorithm is a popular technique for the computation of low-energy states of many-body quantum systems with high accuracy. Mathematically, it can be viewed as an alternating optimization scheme on the manifold of tensor trains of fixed separation rank, which has led to the study of its convergence properties in some tractable scenarios. The sequential nature of such schemes poses some challenges to implement them in parallel computing architectures. However, some sensible modifications can lead to partial parallelization, as recently shown for DMRG, with promising numerical results. In this work, we present the main theoretical tools for the convergence analysis of 1-site DMRG and adapt them to a first parallel version of the algorithm. More specifically, we explain how a rank condition on the Hessian of the function to be minimized, as seen through a parametrization of the fixed rank tensor train manifold, allows to define an iteration map, s, on a neighborhood of a critical point, whose derivative is related to the linear block Gauss-Seidel method applied to the aforementioned Hessian, leading to a well-known convergence result for classical 1-site DMRG. Finally, by applying similar techniques, we show local linear convergence for a parallel version of this method, with a rate bound depending on the number of sites, d, and the condition number of the Hessian on a subspace complementary to its kernel.

Guifré Sanchez i Serra
Paul Scherrer Institute PSI
guifre.sanchez-i-serra@psi.ch

Laura Grigori
EPFL and PSI, Switzerland
laura.grigori@epfl.ch

MS40

Adaptive Spectral Block Floating Point and P-Adaptivity for Discontinuous Galerkin Methods

Discontinuous Galerkin (DG) methods are highly attractive class of numerical method for a wide range of problems due to their numerous favorable features such as robustness for low regularity problems, local conservation, and good support for different types of adaptivity. However, these methods need significantly more degrees of freedom then the classical finite element methods of the same order causing much higher memory and parallel communication bandwidth utilization with resulting performance drawbacks. To address this problem, we exploit the spectral property of modal DG discretizations to propose a much more memory-efficient storage of the solution vector. The new data type and arithmetic are called adaptive spectral block floating point and allow to substantially reduce storage and bandwidth requirements for numerical solution algorithm without sacrificing accuracy or robustness.

Shivam Sundriyal, Vadym Aizinger
University of Bayreuth
shivam.sundriyal@uni-bayreuth.de, vadym.aizinger@uni-bayreuth.de

MS40

Lineal: An Efficient, Hybrid-Parallel Linear Algebra Library

Solving large sparse linear systems arising from PDE discretization on attainable hardware is a challenge which Lineal, our new C++ linear algebra library, addresses by optimizing runtime and especially memory consumption. Its Algebraic Multi-Grid preconditioner, based on DUNE ISTLs algorithm, supports matrix-free linear systems on the finest level with CSR matrices on coarser levelsa novel approach, to our knowledge. Lineal offers support for efficient mixed precision, SIMD operations, and tiling, and is almost fully multithreaded. Recently added MPI support enables hybrid-parallel computations on compute clusters while minimizing communication costs. Lineal has been used to simulate oxygen diffusion in soil samples, solving instances with more than 10? unknowns in under 4 minutes using 128 threads and less than 160 GB of RAM. Additional tests confirm Lineals strong performance compared to existing libraries and its scalability on larger compute clusters using hybrid parallelism.

Kurt Böhm, Olaf Ippisch
Institute of Mathematics
Clausthal University of Technology
kurt.boehm@tu-clausthal.de, olaf.ippisch@tu-clausthal.de

MS40

A Matrix-Free Multigrid Preconditioner for Discontinuous Galerkin Methods Applied to Compressible Flow

High fidelity fluid simulations have many important applications in science and engineering, with examples including numerical weather prediction. Discontinuous Galerkin (DG) methods are promising for simulation of unsteady compressible fluid flow in 3d. Systems arising from such discretizations of turbulent fluid motion are often stiff, and require implicit time integration. This requires fast, parallel, low-memory solvers for the algebraic equation systems that arise. For (low order) finite volume (FV) methods, multigrid (MG) methods have been successfully applied for this purpose. But for high order DG such solvers are currently lacking, which inhibit wider adoption of DG methods, and motivates to construct a matrix free preconditioner for high order DG discretizations, which is based on a multigrid method constructed for a low order FV discretization defined on a subgrid of the DG mesh. Numerical experiments on atmospheric flow problems show the benefit of this approach.

Philipp Birken
Lunds Universitet
philipp.birken@na.lu.se

Andreas Dedner
University of Warwick

a.s.dedner@warwick.ac.uk

Robert Klöfkorn
Lunds Universitet
robert.klofkorn@math.lu.se

MS40

Code Generation for Discontinuous Galerkin Kernels in the Earthquake Simulation Code SeisSol

SeisSol adopts high-order Discontinuous Galerkin discretization with cluster-wise local time stepping for extreme-scale simulations of complex earthquake and earthquake-tsunami events. The element-local compute kernels are expressed as sequences of small (sparse and dense) matrix and tensor operations. Various code generators, for CPU and GPU architectures, provide architecture-optimized implementations for these operations. In this talk, we introduce TensorForge, which is designed for executing chains of small tensor operations on GPUs. We discuss performance improvements due to increased arithmetic intensity, and various use cases, including viscoelastic seismic wave propagation or fused ensemble simulations.

David Schneller, Michael Bader
Technical University of Munich
david.schneller@tum.de, bader@in.tum.de

MS41

Innovations of NextSilicons Intelligent Compute Architecture

NextSilicons novel Intelligent Compute Architecture (ICA) is a paradigm shift from static hardware. As a software-defined hardware, the Maverick-2 ICA is dynamically reconfigured at runtime to adapt to specific application kernels and workloads, delivering performance and energy efficiency improvements by efficiently matching the hardware to the application's unique computation and memory patterns. In this talk, we'll present NextSilicons adaptive runtime optimizer, providing specific examples such as scientific simulations and data analytics. We will also discuss the implications of this runtime adaptability on algorithmic design, highlighting how it enables more flexible and efficient parallel programming methodologies.

Solal Amouyal
NextSilicon
solal.amouyal@nextsilicon.com

MS41

Co-Design of the Cerebras Wafer-Scale Engine for Fast Inference and Scientific Applications

With the growing demand of AI workloads in training and inference the development of novel hardware accelerators also has gotten more attention, as it promises greater compute-throughput and power efficiency.The Cerebras CS-3 is one such hardware innovation: the system is built from the ground up to accelerate large AI training and inference workloads. The CS-3 is built around the wafer-scale engine (WSE), a single chip consisting of 900,000 processing elements (PEs) laid out in a 2D mesh, with each PE containing 48 kB of memory accessible in a single cycle.While this chip has explicitly been the product of co-design with AI workloads, many of its features also address bottlenecks present in more traditional HPC workloads. In this talk we discuss recent results on several HPC workloads in molecular dynamics and seismic imaging which have benefited from these architectural innovations.

Luk Burchard
Cerebras
luk.burchard@cerebras.net

MS41

Hpc-Ai Codesign in High-Performance Communications

Supercomputing hardware continues to overcome many challenges in scaling to deliver increased compute, memory, and network throughput every year. Software must evolve aggressively to keep up with this technology, in order to expose greater parallelism, reduce synchronization, and hide latency. This presentation will focus on the co-design of GPU communication software for AI workloads. We will explain the motivation for device-initiated one-sided communication and how it is implemented on NVIDIA platforms with both scale-up and scale-out networking. Finally, we will describe usage in AI workloads and how it can be used more broadly in HPC simulation.

Jeff R. Hammond
NVIDIA
jeff_hammond@acm.org

MS41

Hpc-Ai Co-Design for the Exascale Apu

In Nov. 2024, the El Capitan supercomputer at Lawrence Livermore National Laboratory (LLNL) achieved a historic performance of 1.742 exaFLOPs, representing an important next step in the Exascale era. Since then, the El Capitan machine and encapsulating ecosystem have continued to progress. Co-design based on tight integration with applications has enabled a synergy between the AMD software stack and applications across HPC-AI, leading to improved performance, reliability, and scalability. The continuation and strengthening of this co-design model is a central goal of the El Capitan Center of Excellence. In this talk, we review innovative aspects in the MI300A APU architecture, with special focus on cutting-edge co-design features, including partition configurations of MI300A, runtime and compiler improvements, scalability of libraries important to HPC-AI applications, and improved capability to diagnose issues at-scale. As a highlight, we present recent capability showcased by runs of OpenFold for protein structure prediction on El Capitan.?

Michael Rowan
AMD
michael.rowan@amd.com

MS43

Compressible Flow with a Free Lunch: Simulating 1 Quadrillion Degrees of Freedom Via Regularization Without Loss of Accuracy

A method for solving multi-species and shock-laden flow at unprecedented problem sizes and time-to-solution is presented. Based on recent work of F. Schafer and the speaker, an inviscid regularization of the Navier–Stokes-like PDE is performed. This enables linear and well-conditioned numerics suitable for mixed-precision computation. A unified memory implementation is crafted for tightly coupled CPU–GPU and APU architectures (e.g., NV GH200, AMD MI300A), typically used on current flagship machines like

LLNL El Capitan OLCF and Frontier. With this trio, we improve on state-of-the-art CFD techniques with order-of-magnitude improvements along computational cost, memory footprint, and energy-to-solution metrics. The reduced memory footprint compared to baselines enables 25-times larger simulations, exceeding 200T grid points (1 quadrillion degrees of freedom) with per-grid-point cost speedups. The method strong scales from 8 nodes to the full systems (¿ 10K nodes) with better than 50% efficiency. This enables, for example, a typical 200B grid point CFD simulation in less than one wall time minute on OLCF Frontier.

Spencer H. Bryngelson, Tanush Prathi, Anand Radhakrishnan, Daniel Vickers, Benjamin Wilfong
Georgia Institute of Technology
shb@gatech.edu, tprathi3@gatech.edu, aradhakr34@gatech.edu, dvickers6@gatech.edu, bwilfong3@gatech.edu

MS43

A Multi-Gpu Batched Electron-Boltzmann Solver for Low-Temperature Plasmas

We present a multi-GPU solver for the collisional electron Boltzmann transport equation (BTE), for low-temperature plasmas. We use a deterministic discretization of the BTE instead of more commonly used particle-Monte-Carlo methods. We use a Galerkin scheme to discretize in electron-velocity space and an implicit time-marching scheme. The solver is coupled to a subsonic inductively coupled argon plasma flow with six species: electrons, ground state atoms, ions, and three excited metastable species. The flow/species transport solver is discretized using a discontinuous Galerkin method and implemented using the MFEM library from the Lawrence Livermore National Laboratory (LLNL). To avoid solving the BTE at each grid point of the flow solver, we (1) use a clustering scheme based on the per-grid point plasma properties, and (2) developed a batched BTE solver. Combined these techniques result in orders of magnitude speedup. We present a detailed performance evaluation for our solver and report results on NVIDIA and AMD GPUs. On a single GPU, individual kernels in our BTE solver achieve over 5 TFLOPS; whereas the performance of the overall BTE solver exceeds 3.5 TFLOPS in production mode. We also report strong scalability results across 24 NVIDIA A100 and 32 AMD MI-250X GPUs on the Texas Advanced computing Center's (TACC) Lonestar6 and LLNL's Tioga machines respectively.

Milinda Fernando
Oden Institute, UT Austin
milinda@oden.utexas.edu

MS43

Leveraging MLIR for Efficient, Architecture-Agnostic Sparse FEM Kernels on GPUs and Multicore CPUs Abstract

The finite element method (FEM) relies heavily on efficient sparse kernel execution, yet manually tuning these kernels for an increasingly diverse hardware landscape is unsustainable. This talk presents a novel compilation workflow for generating architecture-agnostic sparse kernels by leveraging the MLIR (Multi-Level Intermediate Representation) and LLVM infrastructure. We describe a methodology that lowers high-level mathematical descriptions of FEM operations into intermediate representations that capture sparsity and data access patterns independent of the hardware. This abstraction allows for target-specific optimizations and code lowering through the LLVM backend. We demonstrate the performance portability of this approach by presenting results on Nvidia GPUs, as well as Intel and ARM multicore CPUs. We show that this automated pipeline can match or exceed the performance of hand-tuned implementations while significantly reducing engineering effort.

Hari Sundar
Tufts University
hari.sundar@tufts.edu

MS44

Custom Hardware Accelerators for Matrix Multiplication

We will review results in the design of, and automated design of, custom hardware accelerators for matrix multiplication. We will look at the design of fixed-point, floating-point and MX hardware for this purpose in Field-Programmable Gate Arrays (FPGAs), review novel techniques developed for automating the design process, and comment on the limits of acceleration.

George A. Constantinides
Imperial College London
g.constantinides@imperial.ac.uk

MS44

Error Estimates on Compensated Sum Algorithms

Compensated sum is an old technique devised in the 60s to achieve higher precision than what the hardware natively supports. It was a particularly useful technique during a period when double precision was not prevalently available. It also found successes in niche scientific areas where higher than double precision is desired. However, over the years, it has become largely forgotten in the field of numerical analysis and floating point arithmetics. In this talk, we will revisit some classic forms of compensated sum algorithms as well as their more nuanced variations. Error estimates associated with these algorithms will be presented, which can be very sharp. These tight error bounds are then used to guide the design of linear algebraic kernels for the purpose of detecting errors in large scale systems, including hardware defects, random bitflips, bugs in the software stack, etc.

Longfei Gao
Argonne National Laboratory
longfei.gao@anl.gov

MS44

Compiler-Assisted Relative Error Analysis for Floating-Point: Tools and Reproducibility Challenges

Achieving high performance on modern computing architectures increasingly relies on lower-precision floating-point arithmetic. While this approach accelerates computation, it also introduces critical numerical challenges, including reduced dynamic range, increased susceptibility to exceptions (e.g., overflows, NaNs), amplified rounding errors, and reproducibility concerns across platforms. In this work, we introduce a comprehensive framework for multi-precision relative error analysis, leveraging Clang/LLVM-based instrumentation and dynamic analysis. Our methodology enables real-time detection and mit-

igation of floating-point exceptions at various precision levels, as well as runtime assessment of a programs dynamic range, demonstrated through applications to linear solvers. We further explore reproducibility challenges, presenting case studies that reveal compiler-induced numerical discrepancies on emerging GPU platforms. Our tools and findings equip developers with actionable insights for understanding and managing floating-point behavior when porting scientific codes to mixed-precision hardware.

Ignacio Laguna
Lawrence Livermore National Laboratory
lagunaperalt1@llnl.gov

MS45

High Performance Mixed Precision Solvers on GPU-based Architectures

We present strategies for dynamically reducing precision in preconditioned conjugate gradient (PCG) combined with Krylov subspace recycling. Solving large-scale linear systems is often the most expensive part of large-scale simulations and optimizations. On modern GPU-based architectures, significant run-time reductions require improving arithmetic throughput by increasing computation relative to data movement and communication, reducing storage, balancing precision with application accuracy, and designing algorithms that combine high accuracy with adaptive use of low precision. Our work develops dynamically adapted algorithms that exploit Krylov subspace recycling to accelerate convergence of ill-conditioned problems on highly deformed meshes with expensive multigrid preconditioners. Recycling CG is memory-intensive, since it stores a recycle-space basis, its matrix image, and a window of search directions for updates. To reduce this footprint, we store vectors in lower precision and avoid the image of the recycle space at the cost of an extra matrix-vector product. This can hinder convergence to double-precision accuracy, so we propose adaptive strategies to mitigate the issue. We also introduce a PCG variant that dynamically adjusts storage and computation precision, exploiting low precision whenever possible while retaining double-precision accuracy.

Yichen Guo, Eric De Sturler, Tim Warburton
Virginia Tech
ycguo@vt.edu, sturler@vt.edu, tcew@vt.edu

MS45

Memory Accessor for Mixed Precision GMRES-Based Iterative Refinement

Mixed-precision iterative refinement algorithms have been designed to provide high precision solution to well-conditioned problems while achieving high performance by relying on low precision factorizations. However both the factorization step and the iterative correction step of these algorithms may use multiple arithmetics. This mixture of precisions requires from the developer that they either convert the factorized system or use flexible iterative correction steps. Such modifications limit the problem scalability as the memory footprint would grow. Instead we propose to use mixed precision memory accessor approaches that decouple storage and compute precisions (data is stored and accessed in low precision, but computations are kept in higher precision) and reduce data accesses, improve accuracy, and simplify programming. In this work, we present experimental results of mixed-precision GMRES-based iterative refinement. We leverage the block low-rank structures and the mixed-precision storage of the sparse direct solver MUMPS to achieve low memory footprint. We also present the BLAS-based block memory accessor of MUMPS that is leveraged during the solve step to reach high performance. We discuss the importance of the memory accessor to reach better problem scalability compared with a flexible GMRES-based iterative refinement.

Antoine Jego
LIP6, Unité Mixte de Recherche de Sorbonne Université
antoine.jego@lip6.fr

Patrick R. Amestoy, Jean-Yves L'Excellent
Mumps Technologies
patrick.amestoy@mumps-tech.com,
jean-yves.l.excellent@mumps-tech.com

Theo Mary
Sorbonne Universite, CNRS, LIP6
theo.mary@lip6.fr

MS45

Storage and Arithmetic of H-Matrix Formats with Floating Point Compression

Hierarchical low-rank formats bring down complexity for storage and matrix arithmetic to (almost) optimal linear costs. However, this can be further reduced by employing error adaptive storage schemes which allow floating point representation errors of the same order as the low-rank approximation error. As we will demonstrate, such optimized storage also reduces the memory gap between the different H-matrix formats and significantly increases performance for memory bandwidth bound arithmetic operations, e.g., matrix-vector multiplication.

Ronald Kriemann
Max Planck Institute for Mathematics i.t.S.
rok@mis.mpg.de

MS45

Solving Sparse Linear Systems with Adaptive Precision GMRES

The growing use of low-precision arithmetic, initially popularized by artificial intelligence, has created new opportunities for scientific computing. In high-performance computing (HPC), low-precision formats open the door to mixed-precision algorithms that can deliver large performance gains, but at the cost of reduced numerical accuracy if not carefully monitored. In this talk, we present a variant of the Generalized Minimal Residual (GMRES) method that relies on a mixed-precision sparse matrix-vector multiplication. In our approach, the precision of the coefficients of the system matrix can vary both across entries within the same iteration and over the course of the iterations, effectively turning GMRES into a mixed and adaptive precision solver. We also show how combining this approach with restarted GMRES further reduces memory usage and computational cost, while maintaining stable convergence. We will discuss the algorithmic design, highlight how accuracy losses can be controlled in Krylov subspace methods, and present performance results compared to standard double-precision GMRES. Our results show that, when accuracy is properly monitored, low-precision arithmetic together with restarts can significantly improve solver efficiency on modern HPC systems.

Alexandre Tabouret

Sorbonne Université, CNRS, LIP6
alexandre.tabouret@lip6.fr

Emmanuel Agullo
INRIA
emmanuel.agullo@inria.fr

Luc Giraud
Inria
luc.giraud@inria.fr

Pierre Jolivet
CNRS
pierre.jolivet@cnrs.fr

Theo Mary
Sorbonne Universite, CNRS, LIP6
theo.mary@lip6.fr

MS46

Generalized Source Integral Equations (GSIEs) for Enhanced and More Efficiently Parallelizable System Matrix Compression

Fast integral equation (IE) solvers often rely on hierarchical partitioning of a system matrix and low-rank approximation of its admissible blocks. This strategy assumes that large off-diagonal blocks are inherently more compressible. However, for conventional oscillatory-kernel IEs describing wave phenomena, if the domain is not of a reduced dimensionality, the scaling of the rank is asymptotically linear with the block dimensions, even if appearing slower at first. This prevents the reduction of the asymptotic costs of block approximations and overall solver. In parallel implementation, this also poses challenges to load balancing, as the block compression costs vary within a wide dynamic range dictated by their sizes and ranks, often leading to compromises on shallow hierarchies to maintain parallel efficiency. Recently proposed generalized source IEs (GSIEs) make use of modified kernels that, where possible, attenuate the broadside interactions responsible for the unfavourable scaling of the ranks. The resulting basis-testing cluster interactions are of a reduced effective dimensionality and slower scaling ranks. This translates to greater and faster low-rank compression and to superior asymptotic solver costs. The compressibility also alleviates the practical need in shallower hierarchies and enables more efficient use of available parallel computing resources. In this talk, we will present the GSIEs and discuss their implications to faster solver design.

Yaniv Brick
Ben-Gurion University of the Negev
yaniv.brick@gmail.com

Amir Boag
Tel Aviv University
boag@tauex.tau.ac.il

MS46

Hierarchical Compression Exploiting Symmetries of Boundary Integral Equations in Transmission Problems

With recent hardware trends shifting toward high-throughput, low-precision arithmetic units, the high-performance computing (HPC) community is steadily adapting to this paradigm. An increasing number of numerical libraries, such as cuSOLVER and MAGMA, now incorporate low- precision factorizations combined with high-precision iterative refinement to accelerate computations without sacrificing accuracy. Among dense direct solver approaches, the hierarchical low-rank matrix, or H-matrix, presents unique opportunities when coupled with mixed- precision methods. In this work, we present a software framework that integrates data-parallel H^2-ULV factorization with mixed-precision GEMM from cuBLAS, employing it as a preconditioner within an iterative refinement scheme built on a fixed-size Krylov subspace, conceptually close to restarted GMRES. By carefully tuning rank thresholds and admissibility conditions, we demonstrate that mixed-precision H^2-ULV factorizations achieve faster time-to-solution while preserving convergence behavior comparable to uniform FP64 factorizations. These results highlight the promise of mixed-precision hierarchical solvers with iterative refinement as an efficient and scalable approach for future HPC applications.

Qianxiang Ma
RIKEN Center for Computational Science
qianxiang.ma@riken.jp

Ria Yokota
Institute of Science Tokyo
rioyokota@rio.scrc.iir.isct.ac.jp

MS46

Tensor Train Acceleration of Volume Integral Equation Solution

In this talk we present a computational framework for solving full-wave scattering and magneto-quasistatic problems for arbitrarily shaped objects with polylogarithmic $O(\log^p N)$ complexity in both CPU time and memory, where N is the number of basis and test functions in a Method of Moments (MoM) discretization of volume integral equations (VIEs). The efficiency stems from tensor train (TT) decomposition of MoM matrices and vectors, with specialized linear algebra performed directly on these tensorized datasets. To enable TT decomposition, MoM data are recast as multidimensional arrays defined by recursively subdivided basis functions over a pixelized domain. While such representations suffice for TT construction, the train ranks generally grow as $O(N)$ for most non-canonical geometries. We demonstrate that global Gaussian smoothing of step-like transitions in the material contrast function contains TT ranks to polylogarithmic scaling, both in object representation and MoM matrices. Numerical examples of TT-accelerated MoM solutions for full-wave and quasi-magnetostatic VIEs confirm that the proposed approach achieves overall polylogarithmic complexity with N, regardless of object shape or material distributionfrom simple to fractal geometriespaving the way for practically relevant applications of TT-accelerated MoM.

Vladimir Okhmatovski, Chris Nguyen
University of Manitoba
vladimir.okhmatovski@umanitoba.ca,
nguye73@myumanitoba.ca

Alexey Boyko
AIRI
alexey.boyko@skoltech.ru

MS46

Accelerating RSRS for 3D, Large Scale, Integral

Equations

Recursive Strong Skeletonization (RSS) [1] is an effective framework for compressing and solving structured linear systems, particularly those arising from boundary integral formulations. Its randomized extension, Randomized Strong Recursive Skeletonization (RSRS) [2], enables matrix-free computation by requiring only a hierarchical partition and matrix-vector products, making it well suited for problems with millions of unknowns. In three dimensions, however, the growth of neighboring interactions significantly increases the cost of factorization. We present algorithmic and implementation improvements that accelerate RSRS for large-scale 3D integral equations. Our Rust-based implementation employs Rayon to parallelize both the compression and application stages, leading to substantial efficiency gains. Numerical experiments on boundary integral problems show significant speedups over baseline RSRS, demonstrating the potential of these advances to provide scalable direct solvers for challenging 3D applications. [1] Minden, V., Ho, K. L., Damle, A., & Ying, L. (2017). A recursive skeletonization factorization based on strong admissibility.Multiscale Modeling & Simulation,15(2), 768-796. [2] Yesypenko, A., & Martinsson, P. G. (2023). Randomized Strong Recursive Skeletonization: Simultaneous compression and factorization of H-matrices in the Black-Box Setting.arXiv preprint arXiv:2311.01451.

Ignacia Piccardo, Timo Betcke
University College London
maria.piccardo.19@ucl.ac.uk, t.betcke@ucl.ac.uk

Anna Yesypenko
University of Texas at Austin
annayesy@utexas.edu

MS47

Parallel-in-Time with Space-Time Adaptive Mesh Refinement

Applications with adaptive mesh refinement (AMR) may have scaling challenges due to the serial dependencies required by time integrators. In particular, local time stepping uses a different time step that depends on local grid spacing, so that the time integrators for different refinements need different time steps. Bottlenecks can be due to refinement boundary conditions, accuracy or conservation constraints, implicit solvers, etc.; any of these require synchronization between refined grids, which is exacerbated in very deep AMR hierarchies. We present three different algorithms that reduce these time integration barriers: stage parallelism, operator splitting parallelism, and parallel in time approaches. We evaluate these against benchmarks for both CPU and GPU, and present how each algorithm introduces different trade-offs in work-precision, and some analysis as to why they have significantly different parallelism and speed-up.

Hans Johansen
Lawrence Berkeley National Laboratory
Computational Research Division
hjohansen@lbl.gov

Rochishnu Chowdhury
Lawrence Berkeley National Laboratory
rochishnu00@lbl.gov

Benjamin W. Ong
Michigan Technological University
ongbw@mtu.edu

Steven B. Roberts
Lawrence Livermore National Laboratory
Center for Applied Scientific Computing
roberts115@llnl.gov

MS47

Optimized Schwarz Waveform Relaxation: Superlinear Convergence Analysis and Parameter Optimization

Schwarz waveform relaxation (SWR) is a family of domain decomposition methods for solving time dependent PDEs iteratively. The spatial domain is first decomposed into subdomains, and the time-dependent subdomain problems are solved in parallel, before interface data for the whole time interval are exchanged; the process is then repeated until convergence. SWR methods are known to converge superlinearly for bounded time intervals. However, to obtain sharp convergence estimates, one often needs to compute complicated inverse Laplace transforms, and for some interface conditions, there are no known formulas for the inverse transforms. In this talk, we will show how exponential weighting techniques can be used to produce superlinear estimates without having to compute the inverse Laplace transforms explicitly. We will use this technique to analyze the convergence of SWR with Robin transmission conditions. We will also show how to choose the Robin parameter to optimize error reduction for two asymptotic regimes, namely the small overlap limit and the short time interval limit.

Felix Kwok, Alejandro Alonso Rodriguez
Université Laval
felix.kwok@mat.ulaval.ca, alejandro.alfonso-rodriguez.1@ulaval.ca

MS47

Parallel Imex Time Integration Based on PFASST and Diagonal SDC

The Spectral Deferred Corrections (SDC) method, introduced by Dutt, Greengard, and Rokhlin in 2000, has inspired the development of numerous time-parallel algorithms. These algorithms leverage parallelism either across the time step (e.g., PFASST, RIDC) or across the stage (e.g., ParaDiag, Diagonal SDC). Beyond its potential for time parallelism, SDC also provides a framework for designing arbitrarily high-order semi-implicit time-integration schemes. These methods can be seamlessly applied to first-order ODE systems with arbitrary mass matricessystems that commonly arise, for example, when pseudo-spectral spatial discretization is applied to nonlinear problems. Building on concepts from PFASST and diagonal SDC, we introduce a novel time-parallel SDC-based IMEX scheme tailored for the multi-scale time integration of fast-wave/slow-wave phenomena. We analyze the numerical stability and accuracy of this scheme, particularly focusing on how these properties are influenced by the choice of SDC preconditioner coefficients. Finally, we evaluate the parallel efficiency and time-to-solution of the scheme for large-scale turbulent flows, using 3D Rayleigh-Bnard convection as a benchmark.

Thibaut Lunet
Hamburg University of Technology

thibaut.lunet@tuhh.de

MS47

Transformation-Based Parareal for Weakly Nonlinear Problems

The Parareal algorithm is a parallel-in-time method with strong potential for oscillatory multiscale problems. We focus on weakly nonlinear systems of the form

$$\frac{dy}{dt} + \frac{1}{\varepsilon}\mathcal{L}y + \mathcal{N}(y) = 0,$$

where $\mathcal{L}$ is a skew-Hermitian linear operator with purely imaginary spectrum and $\mathcal{N}(y)$ is a nonlinear operator. Such systems arise in applications where fast oscillations interact with weak nonlinear effects. Building on previous two-level and multilevel Parareal formulations (Peddle et al., SIAM J. Sci. Comput., 2019; Haut & Wingate, SIAM J. Sci. Comput., 2014; Rosemeier et al., SIAM J. Sci. Comput., 2024), we investigate transformation-based strategies for constructing coarse propagators in this weakly nonlinear setting. Specifically, we explore two distinct approaches: a transformation into standard form and a WKB-inspired transformation. Both aim to exploit the oscillatory structure of the problem to design efficient coarse propagators. Numerical experiments demonstrate that these transformation-based propagators can improve stability and convergence of Parareal iterations, particularly in the presence of weak nonlinearities interacting with fast oscillations.

Juliane Rosemeier
Free University of Berlin
julianero@zedat.fu-berlin.de

MS48

Towards a Unified HPC/AI Framework: Why Its Needed and What It Could Look Like

The convergence of high-performance computing (HPC) and artificial intelligence is driving a new paradigm in scientific research. However, the software stack struggles to keep pace with rapid hardware evolution and "AI-only" tooling. Researchers must constantly re-engineer their codebases as architectures advance, creating a costly disconnect between physical simulation and machine learning. In this presentation, we will recap the unifying framework concept from our initial study, outline the next steps toward its realization, and elaborate on our long-term vision.

Jens Domke
Satoshi MATSUOKA Laboratory
GSIC, Tokyo Institute of Technology
jens.domke+siampp@riken.jp

MS48

An Approach for Enabling Ai/ml Use on the Fly in Amr Based Simulations

Machine-learning surrogates offer a promising approach to reducing the computational cost and inflexibility of complex physics models in large-scale scientific simulations. However, enabling online training of AI/ML models within scientific simulation codes remains challenging due to dynamically evolving, non-stationary data distributions. This talk presents an ongoing work on developing an unifying framework for scientific simulation with AI/ML models, demonstrating online training of an equation-of-state (EOS) surrogate model of Flash-X, a massively parallel, adaptive mesh refinement (AMR)-based multiphysics simulation software. Taking the cellular detonation simulation as an example of surrogate EOS model online training, this talk will focus on how to effectively integrate AI/ML training workflows into a physical simulation and highlight several challenges in online training for scientific simulations. Key design considerations, including interfacing an AI/ML model with a scientific simulation, online data sampling strategies, and stabilizing model training, will be discussed.

Youngjun Lee
Argonne National Laboratory
USA
leey@anl.gov

MS48

AI is Mostly BLAS So Thats How We Should Treat It

With the increasing capabilities of machine-learning (ML) systems, the domains of simulation and ML are growing closer together. In particular, the evolving field of ML-based surrogate modeling within turbulence simulations often relies on explicit interfacing between scientific codes and ML frameworks. In the face of increasing hardware and software complexity, it is important to critically assess this approach and to explore alternatives that allow for a tighter integration of simulation and ML. We develop a tensor loading library, dalotia, that allows researchers to embed ML inference directly into their HPC code ("in-situ") using C, C++, or Fortran. Using dalotia, we evaluate the in-situ approach against inference with ML frameworks (PyTorch, libTorch, and oneDNN). Our experiments using state-of-the-art ML surrogate architectures present complex tradeoffs between convenience, performance, memory usage, energy, and dependency complexity. The data shows that small-batch neural networks can save a factor of 5x in memory overheads when using the in-situ approach. We further observe in-situ speedups of up to 8x compared to the libTorch implementation on the node scale (48 cores). Thus it can be beneficial, in many regards, to rely on native in-situ implementations, without the need for bloated ML frameworks and their dependencies.

Theresa Pollinger
Universität Stuttgart
theresa.pollinger@a.riken.jp

MS48

Interface to Enable Communication Between High-Performance Computing ModSim and Ai/ml

The integration of machine learning into the computational sciences is increasingly pursued to reduce time-to-solution, alleviate I/O bottlenecks, and enable adaptive analysis during simulation. We present RDQ (Remote Data Queue), a library for coupling HPC simulations and machine learning training using an MPMD (Multiple Program Multiple Data) MPI (Message Passing Interface) approach. RDQ is part of an ongoing effort to combine HPC and ML workflows and serves as a research platform for investigating in-situ training and data reduction techniques. RDQ extends existing data staging mechanisms by enabling zero-copy tensor concatenation and putside data sampling for in-situ learning scenarios. RDQ is based on MPI-3 RMA operations and requires no user-managed compute resources, allowing flexible placement

of data queues within the communicator. Internally, RDQ employs a multiring buffer design that supports zero-copy receives, including direct access from Python and PyTorch via the DLPack tensor interface. Performance measurements show that RDQ achieves up to 2600 messages per second for small messages and saturates a 100 Gb/s network interface with message sizes of 8 MiB.

Maximilian Sander
ZIH, CIDS, TU Dresden
maximilian.sander1@tu-dresden.de

MS49

Parallel Tensor Integrators For Dynamical Low-Rank Approximation

Many problems in physics, engineering, and computer science give rise to tensor-valued differential equations that demand substantial computational and memory resources. Prominent examples include radiation transport with uncertainties and quantum many-body dynamics. Low-rank tensor approximations offer an effective strategy to mitigate these costs, but existing dynamical low-rank methods are fundamentally sequential, as they rely on stepwise updates of the tensor representation. This sequential nature limits scalability on modern parallel architectures. We present a novel integrator that computes all substeps of the dynamical low-rank update in parallel, removing this bottleneck and enabling efficient large-scale simulations across a broad range of high-dimensional applications.

Jonas Kusch
Norwegian University of Life Sciences
jonas.kusch@nmbu.no

MS49

Performance of Linear Algebra Building Blocks for Low-Rank Tensor Algorithms on Current Multi-Core CPUs and GPUs

In this talk we discuss the node-level performance of basic operations needed in low-rank tensor algorithms. As starting point, we consider two specific problems: (1) compressing large dense data in the tensor-train (TT) format and (2) solving linear systems with an operator and right-hand side vector given in TT format. For both problems, we analyze and optimize suitable algorithms focussing on the required underlying operations such as tensor contractions and matrix decompositions. In particular, we obtain a significant speedup for orthogonalization and truncation steps by using a high-performance implementation of a Q-less tall-skinny QR decomposition. On multi-core CPUs, our implementation achieves a speedup of 50x over reference implementations for the TT compression, and up to 5x for solving linear systems. For GPUs, we give an overview of the performance of the most important operations and shortly discuss the implementation of a tall-skinny QR decomposition.

Melven Röhrig-Zöllner
German Aerospace Center (DLR)
Institute of Software Technology
melven.roehrig-zoellner@dlr.de

MS49

Algorithms and Software for Sparse Tensor Decompositions for Non-Gaussian Data

Many data science applications require tensor decomposition of non-Gaussian and highly sparse data. Existing work primarily addresses special cases, such as CP decomposition with Poisson loss or Tucker completion with least squares, but a unified and scalable framework for both decomposition types under general loss functions remains missing. Decomposition of large-scale sparse tensors with generalized loss using a non-stochastic optimization method is not addressed in any of the previous literature. This is because, for sparse tensor decomposition, the computation of gradients and Hessian of the objective function with respect to a factor require dense tensor contractions for arbitrary loss functions. In this talk, we present a framework for generalized tensor decomposition of large sparse tensors. By reformulating common objectives, we enable efficient computation of gradients and Hessians using only the nonzero entries of the input tensor, overcoming the previously encountered bottlenecks. Our implementation within the Cyclops Tensor Framework enables distributed-memory scalability. Empirical evaluations show that our methods achieve performance comparable to state-of-the-art approaches for tensor completion tasks. Furthermore, we present an efficient alternating minimization algorithm for Poisson and Bernoulli tensor decomposition at scale, providing the first practical non-stochastic solution for large-scale sparse generalized tensor decomposition.

Navjot Singh
University of Illinois Urbana-Champaign
nsingh2@lbl.gov

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

MS49

Parallel Inexact Subspace Iteration for Solving Tensor Eigenvalue Problems

Density matrix renormalization group (DMRG) is a powerful low-rank eigensolver that has been used to study a wide range of complex quantum systems. While typically used to compute ground states, small modifications to the original algorithm enable one to simultaneously solve for multiple states and even target interior eigenstates. However, because DMRG is an iterative optimization algorithm, it is prone to getting trapped in local minima. Recently, Dektor et al. demonstrated that an inexact subspace iteration (SBI) method using tensor trains obtains better convergence to the eigenstates than DMRG calculations of the same rank. This work introduces a parallelized version to enable faster evaluation of multiple eigenstates, reducing the original runtime linearly with the number of nodes used. This is particularly helpful in analyzing systems with dense eigenspectra, where it is difficult to filter for a specific eigenstate of interest. This algorithm is used to analyze a variety of examples, including capturing the excited states of small molecules with relativistic effects.

Erika Ye
Lawrence Berkeley National Laboratory
erikaye@lbl.gov

MS50

Scalable Fast Approximated Summations on Peri-

odic Structures

In this talk, we focus on solving periodic N-body problems of the form

$$p(x) = \sum_{t \in \alpha\mathbb{Z}^3} \sum_{y} G(x, y+t)q(y) \quad (1)$$

where x, y are elements of a large point cloud C, α denotes the diameter of a fixed box enclosing C, G is a given kernel function and p is the quantity we want to compute. Such problem appears for instance in computational chemistry where x, y are atoms and q refers to the vector of charges associated to atoms. To ensure the convergence of the series, one may exploit Ewald Summation methods, dividing $G = G_r + G_f$ into two kernels handled in the real space (resulting in other N-body problems) and Fourier space (usually numerically solved using Fast Fourier Transforms), respectively. Since the scalability of large Fast Fourier Transforms may be mitigated on distributed memory architectures, we propose a parallel approach based on hierarchical methods for G_r and a new fast and parallel approximation scheme for G_f based on tensor decompositions and quadratures. After a short introduction on hierarchical methods allowing to deal with G_r, we shall present the different approximation levels used in this new scheme for G_f. Then, we will describe the parallel algorithm, with a particular focus on complexity. Numerical experiments will complement this presentation.

Igor Chollet
LAGA, Université Sorbonne Paris Nord
chollet@math.univ-paris13.fr

MS50

Generalized Optimized Schwarz Method for Fem-Bem Coupling

When it comes to solving the Helmholtz equation in a complex heterogeneous medium, it can be of interest to decompose the domain according to the variation of the wavenumber, especially when the latter is constant in some subdomains. Such problems can be reformulated using FEM-BEM coupling techniques, rewriting the problems set in the homogeneous subdomains thanks to Boundary Integral Equations. Recently, a Generalized Optimized Schwarz Method (GOSM) has been introduced on bounded domains, with weakly imposed boundary conditions. It differs from other OSMs by the use of a *non-local exchange operator* instead of the usual swap operator. This makes the formulation robust to cross-points, that is, points where the interfaces of at least three subdomains intersect, which arise naturally in domain decomposition techniques. We extend this work by replacing the classical boundary conditions with interface conditions arising from several FEM-BEM coupling techniques. Depending on the specific FEM-BEM coupling, the resulting discrete formulation has the form "identity+contraction', and thus can be solved using a fast converging iterative procedures such as GMRes or even Richardson. We will give a brief overview of the theoretical guarantees and present extensive numerical experiments to illustrate the method fast convergence when non-local transmission operators are considered. This is a joint work with Antonin Boisneault, Marcella Bonazzoli and Xavier Claeys.

Antonin Boisneault
Inria Saclay
antonin.boisneault@inria.fr

Marcella Bonazzoli
Inria
Institut Polytechnique de Paris
marcella.bonazzoli@inria.fr

Xavier Claeys
Sorbonne Universite - LJLL
xavier.claeys@sorbonne-universite.fr

Pierre Marchand
Inria
pierre.marchand@inria.fr

MS50

Frequency-domain full waveform inversion with one and two-level domain decomposition solvers

Full Waveform Inversion (FWI) reconstructs subsurface properties from wave measurements through a large-scale optimization problem. In the frequency-domain formulation, each iteration requires solving Helmholtz-type PDEs for many sources, i.e., multiple right-hand sides. Sparse direct solvers (e.g., MUMPS) handle this efficiently but their cost becomes prohibitive for high-resolution 3D imaging. Domain Decomposition Methods (DDM) provide a scalable alternative, with lower memory cost and natural parallelism. Yet, one-level DDM suffer from slow convergence at high frequencies and from an iteration cost that scales almost linearly with the number of sources, motivating a two-level approach. We evaluate recent spectral coarse grid methods for the 3D Helmholtz equation [Dolean et al., Schwarz preconditioner with H_k-GenEO coarse space, 2024][Ma, Two-leve RAS preconditioner based on MS-GFEM for Heterogeneous Helmholtz Problems on large-scale benchmarks, 2025]. These significantly improve convergence, at the expense of a non-negligible setup cost that can be amortized across sources, making them attractive for FWI. Finally, we analyze how the solver choice interacts with the outer optimization method: algorithms such as Truncated Newton require solving the same PDE for more right-hand sides compared to classical Quasi-Newton schemes. An effective coarse grid preconditioner can thus broaden the range of viable optimization strategies for large-scale FWI.

Boris Martin
University of Liege
boris.martin@uliege.be

Chistophe Geuzaine
University of Liège
Electrical Engineering and Computer Science
cgeuzaine@uliege.be

MS50

Coarse Spaces for Non-Symmetric Two-Level Preconditioners Based on Local Extended Generalized Eigenproblems

The scalability of domain decomposition methods relies heavily on the design of the coarse space in a two-level approach. We present a versatile method for constructing coarse spaces for sparse matrices derived from standard PDE discretizations. This method ensures theoretical convergence for fixed point iterative schemes and is applicable to a wide range of problems, including non-Hermitian and indefinite systems, Hermitian preconditioners like additive Schwarz (AS), and non-Hermitian preconditioners such as restricted additive Schwarz (RAS). It also accommodates

both exact and inexact subdomain solvers. Our approach involves solving extended generalized eigenproblems locally within each subdomain and applying a carefully chosen operator to the selected eigenvectors to obtain a local discrete solution. This solution is then multiplied by a partition of unity function and extended by zero to form the global coarse space.

Emile Parolin
Inria
emile.parolin@inria.fr

Frédéric Nataf
Sorbonne Université et CNRS
frederic.nataf@sorbonne-universite.fr

Pierri-Henri Tournier
LJLL Sorbonne Universite
pierre-henri.tournier@sorbonne-universite.fr

MS51

Parallelising tSVDM for compression applications

We consider parallelising the tensor decomposition known as the matrix-mimetic tensor singular value decomposition (tSVDM) and its variants. These decompositions share an Eckhart-Young like optimality theorem for compression and are used for compressing scientific datasets, constructed reduced order models, data analysis, among other applications. In this presentation, we shall analyse the communication complexity of parallelising the tSVDM algorithms and empirically benchmark their performance on datasets generated from scientific simulations.

Srinivas Eswar
Mathematics and Computer Science Division
Argonne National Laboratory
seswar@anl.gov

MS51

Gpu-Accelerated Solution of Block-Tridiagonal Systems for Large-Scale Optimization

We present a GPU-accelerated solver for block-tridiagonal symmetric positive definite (SPD) linear systems, which are central to time-dependent estimation and optimal control. Our method leverages a recursive Schur complement reduction that transforms the system into a hierarchy of smaller blocks, enabling efficient parallelism through batched BLAS/LAPACK kernels. Unlike general-purpose sparse solvers, our approach exploits the block-tridiagonal structure known a priori, yielding substantial performance benefits. Benchmarks on both NVIDIA and AMD GPUs demonstrate significant speedups over state-of-the-art CPU solvers (CHOLMOD, HSL MA57) and competitive performance with NVIDIAs cuDSS. We also discuss current limitationssequential recursion steps and the need for sufficiently large blocks to amortize kernel overheadand highlight opportunities for further parallelization and kernel fusion.

David Jin
Massachusetts Institute of Technology
jindavid@mit.edu

Alexis Montoison, Alexis Montoison
Argonne National Laboratory
alexis.montoison@polymtl.ca,
alexis.montoison@polymtl.ca

Sungho Shin
Massachusetts Institute of Technology
sushin@mit.edu

MS51

Recovering Sparse DFT from Missing Signals Via Interior Point Method on GPU

We propose a method to recover the sparse discrete Fourier transform (DFT) of a signal that is both noisy and potentially incomplete, with missing values. The problem is formulated as a penalized least-squares minimization based on the inverse discrete Fourier transform (IDFT) with an -penalty term, reformulated to be solvable using a primal-dual interior point method (IPM). Although Krylov methods are not typically used to solve Karush-Kuhn-Tucker (KKT) systems arising in IPMs due to their ill-conditioning, we employ a tailored preconditioner and establish new asymptotic bounds on the condition number of preconditioned KKT matrices. Thanks to this dedicated preconditioner and the fact that FFT and IFFT operate as linear operators without requiring explicit matrix materialization KKT systems can be solved efficiently at large scales in a matrix-free manner. Numerical results from a Julia implementation leveraging GPU-accelerated interior point methods, Krylov methods, and FFT toolkits demonstrate the scalability of our approach on problems with hundreds of millions of variables, inclusive of real data obtained from the diffuse scattering from a slightly disordered Molybdenum Vanadium Dioxide crystal.

Vishwas Rao
Argonne National Laboratory
vhebbur@anl.gov

MS51

The Exponential Multiplier Method for Nonlinear Constrained Optimization

In this talk, we revisit the exponential multiplier method that was introduced in the early 1970s to solve large-scale nonlinearly constrained optimization problems. It consists in reformulating the constrained problem by an exponential penalty function that is approximately minimized for a sequence of decreasing penalty parameters. Remarkably, the dual variables associated with the inactive constraints (therefore an estimate of the active set) converge exponentially fast to 0. Meanwhile, components of the Hessian of the exponential penalty function may diverge exponentially fast. Our method builds upon the usual double-loop framework: the inner iterations seek a trial iterate that achieves sufficient progress (for a merit function or a filter method), and the outer iterations construct a sequence of acceptable iterates. We first present heuristics that attempt to address numerical difficulties caused by ill-conditioning. We then introduce a Newton correction to warm start primal-dual iterates between sucessive outer iterations. Finally, we present extensive numerical results on the CUTEst benchmark of two software implementations within the Uno solver and the GPU-accelerated solver MadNLP.jl.

Charlie Vanaret
Zuse-Institut Berlin
vanaret@zib.de

Alexis Montoison, Alexis Montoison
Argonne National Laboratory
alexis.montoison@polymtl.ca,

alexis.montoison@polymtl.ca

Nick Gould
Numerical Analysis Group
Rutherford Appleton Laboratory
nick.gould@stfc.ac.uk

Sven Leyffer
Argonne National Laboratory
leyffer@anl.gov

MS52

Accelerating Linear Assignment Algorithms

The Hungarian algorithm is a popular solution for the linear assignment problem that finds correspondences between sets of items. Despite its popularity, this algorithm suffers from significant efficiency shortcomings, which hinder its application to large instances or repeated small instances. To overcome this challenge, we design a parallel version of the algorithm for novel tile-centric accelerators, which consist of thousands of small processing cores equipped with memory that is local to each core. This allows them to overcome the memory latency and bandwidth limits of CPUs and GPUs, but, in turn, data and work must be carefully distributed among the many cores. We present HUNIPU, our implementation of the Hungarian algorithm on the tile-centric Intelligence Processing Unit and compare it to existing Hungarian algorithm implementations across CPU and GPU platforms. Our results show that HUNIPU consistently outperforms all GPU-based baselines and performs competitively with the best-performing CPU algorithm.

Cheng Huang
Aarhus University
cheng@cs.au.dk

Johannes Langguth
Simula Research Laboratory
langguth@simula.no

Davide Mottin
Aarhus University
davide@cs.au.dk

Ira Assent
Aarhus University, Denmark
Department of Computer Science
ira@cs.au.dk

MS52

Accelerating Sparse Linear Solvers on Intelligence Processing Units

Solving large sparse linear systems is fundamental to scientific computing applications from computational fluid dynamics to structural analysis. This talk presents Graphene, an open-source framework for sparse linear algebra on the Graphcore IPU. The IPU's unconventional programming model poses significant challenges for solving large, sparse linear systems. Graphene provides its own domain-specific language, which enables expressing complex algebraic algorithms close to mathematical notation while automatically generating the required dataflow graphs and execution schedules. We overcome precision limitations through a novel combination of Mixed-Precision Iterative Refinement with double-word arithmetics, achieving high-precision solutions at minimal computational overhead. The framework implements parallel linear solvers including PBiCGStab and ILU factorization, optimized for the IPU architecture. A novel matrix reordering strategy exploits the IPU's cacheless design and communication fabric for efficient blockwise halo exchanges, enabling near-ideal scaling. Performance evaluation on SuiteSparse matrices demonstrates speedups of up to 150 over CPUs and 36 over GPUs for sparse matrix-vector multiplications, with complete iterative solvers achieving 5-36 GPU speedups at comparable power consumption. These results highlight the potential of specialized architectures for accelerating fundamental scientific computing operations beyond their original ML focus.

Tim Noack
Technical University of Darmstadt
noack@esa.tu-darmstadt.de

MS52

Sparse Tensor Processing on Heterogeneous System

Sparse tensors have become prevalent data structures in multiple applications, such as medical imaging and machine learning, making operations that decompose them, i.e., creating smaller structures that retain most of the original information, essential. Two of the most commonly used tensor decomposition methods are the Canonical Polyadic and Tucker Decomposition, with the most time-consuming operations being the MTTKRP and TTM-chain, respectively. Modern computing platforms combine multiple devices with different architectures to achieve unprecedented levels of performance, creating an environment where portability is as important as performance. To tackle this challenge, this work proposes SYCL-based MTTKRP and TTM-chain approaches for sparse tensors, which are portable to any CPU or GPU, extending previous literature by handling mode-4 and mode-5 tensors, and tackling the TTM-chain operation as a whole, allowing for further optimisations. The experimental results show that the proposed approaches exhibit linear to superlinear scalability as the problem size increases, and outperform the state-of-the-art by 4.9x on average in both terms of portability and performance.

Daniel Pacheco
Universidade de Lisboa
daniel.pacheco@tecnico.ulisboa.pt

Leonel Sousa, Aleksandar Ilic
INESC-ID
leonel.sousa@tecnico.ulisboa.pt, aleksandar.ilic@inesc-id.pt

MS53

A Scalable Multidimensional Fully Implicit Solver for Hall Magnetohydrodynamics

We propose an optimally performant fully implicit algorithm for the Hall magnetohydrodynamics (HMHD) equations based on multigrid-preconditioned Jacobian-free Newton-Krylov methods [L. Chacon, Journal of Computational Physics, 526, 113789 (2025)]. HMHD is a challenging system to solve numerically because it supports stiff fast dispersive waves. The preconditioner is formulated using an operator-split approximate block factorization (Schur complement), informed by physics insight. We use a vector-potential formulation (instead of a magnetic

field one) to allow a clean segregation of the problematic $\nabla\times\nabla\times$ operator in the electron Ohms law subsystem. This segregation allows the formulation of an effective damped block-Jacobi smoother for multigrid. We demonstrate by analysis that our proposed block-Jacobi iteration is convergent and has the smoothing property. The resulting HMHD solver is verified linearly with wave propagation examples, and nonlinearly with the GEM challenge reconnection problem by comparison against another HMHD code. We demonstrate the excellent algorithmic and parallel performance of the algorithm up to 16384 MPI tasks in two dimensions.

Luis Chacon
Los Alamos National Laboratory
chacon@lanl.gov

MS53

Efficient Preconditioners for Flow Problems on Gpu-Based Supercomputers

My talk will present insights we gained during the development of efficient solvers for computational fluid flows with application in hemodynamics, i.e., the flow of blood in vessels. The starting point of our developments was an efficient solver for the incompressible Navier-Stokes equations with Newtonian flow behavior. During the past years, we have analyzed both splitting methods, which treat velocity and pressure updates in separate steps, and fully coupled solvers to advance the Navier-Stokes equations in time. The former have the advantage of allowing for specialized solvers for each field, but might introduce splitting errors. The next step has been to move to non-Newtonian behavior with variable viscosity, where additional challenges have been solved for splitting methods. Finally, the coupling of fluid flow with the deformation of the flow geometry is considered. For each of these steps, we have been trying to identify highly efficient linear and nonlinear solvers as well as associated preconditioners. Given the high arithmetic capability of modern hardware, efficient implementations often make use of matrix-free operator evaluation as a means to increase the computational throughput. In my talk, I will highlight the steps we have made in porting these solver sub-steps to GPU systems using a generic implementation with the deal.II finite element library.

Martin Kronbichler
Ruhr University Bochum
martin.kronbichler@rub.de

MS53

Scalability and Performance of the Empire Plasma Physics Code on the El Capitan Platform

Empire is an unstructured mesh finite element particle-in-cell code designed to model plasma physics environments. It is built for performance portability using the Kokkos library, enabling efficient execution on a wide range of hardware, from local workstations to the largest supercomputers. In this talk, we will discuss recent efforts to port Empire to the AMD MI300A GPUs on the El Capitan system, currently the fastest supercomputer in the world on the Top 500 List. We will cover the discretization algorithms and solution methods, present strong and weak scaling studies, and provide performance comparisons with modern CPU architectures. Additionally, we will assess the effectiveness of the performance portability abstractions. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energys National Nuclear Security Administration under Contract No. DE-NA0003525.

Logan T. Meredith
Sandia National Laboratory
lmeredi@sandia.gov

James Elliott, Christian Glusa, Jonathan J. Hu, David Sirajuddin, Christopher Moore
Sandia National Laboratories
jjellio@sandia.gov, caglusa@sandia.gov, jhu@sandia.gov, dsiraju@sandia.gov, chmoore@sandia.gov

Roger Pawlowski
Multiphysics Simulation Technologies Dept.
Sandia National Laboratories
rppawlo@sandia.gov

MS53

Adaptive Parallel Solver Techniques for Cardiac Electrophysiology

Adaptive parallel solver techniques for cardiac electrophysiology Cardiac electrophysiology is described by PDEs of reaction-diffusion type with elliptic constraints, ranging from homogenized monodomain and bidomain models to extracellular-membrane-intracellular (EMI) models resolving the cardiac myocytes geometrically. Organ-scale simulations and in particular large EMI simulations incur a high computational cost, since faithful excitation propagation requires fine meshes and small time steps. The locality of solution features calls for adaptive methods, but traditional mesh adaptivity, despite effective in reducing the problem size, fails to provide speedup. We present a novel low-overhead approach to adaptivity that works completely on the algebraic level and exploits shrinking correction support in spectral deferred correction (SDC) time stepping methods. We discuss its basic structure, its combination with balancing domain decomposition with constraints (BDDC) preconditioners, and its extension to second order finite element discretization using hierarchical elements and a hybrid multigrid-BDDC preconditioner. The effectivity of the algebraic adaptivity is illustrated at several numerical examples.

Martin Weiser
Zuse Institute (ZIB)
Berlin, Germany
weiser@zib.de

Fatemeh Chegini
Zuse Institute Berlin
chegini@zib.de

MS55

An Efficient Preconditioner for Large-Scale Contact Mechanics

Large-scale contact mechanics simulations are central to engineering applications but are challenging due to nonlinear, non-convex formulations and worsening solver performance at high mesh resolutions. The bottleneck for many numerical methods for frictionless contact is the solution of large, ill-conditioned saddle-point systems. To overcome this, we propose a novel preconditioner, AMG with Filtering (AMGF), designed to efficiently solve the Schur complement of the system. AMGF extends clas-

sical algebraic multigrid by adding a subspace correction which filters near-null components from contact constraints. Analysis and experiments on linear and nonlinear problems show mesh-independent convergence and robustness against contact-induced ill-conditioning. This significantly improves the scalability of contact simulations, making e.g. Newton-based IP methods more practical for large-scale engineering. Beyond contact mechanics, AMGF is broadly applicable to problems where difficulties stem from low-dimensional subspaces, such as localized constraints, interface conditions, or heterogeneities.

Socratis Petrides
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
petrides1@llnl.gov

Tzanio Kolev
Lawrence Livermore National Laboratory
kolev1@llnl.gov

MS56

Scalable and Multilevel Preconditioners for the Cardiac EMI Model on Hybrid CPU-GPU Architectures

The Extracellular-Membrane-Intracellular (EMI) model provides one of the most physiologically detailed descriptions of cardiac tissue electrophysiology. Its accuracy comes with a substantial increase in computational cost, requiring the development of efficient and scalable solvers. In this talk, we will present recent advances in multilevel and domain decomposition preconditioners for the EMI model on hybrid CPUGPU architectures. In particular, we will discuss algebraic multigrid (AMG) methods and overlapping additive Schwarz preconditioners, both integrated within a novel FEniCSx-based implementation capable of handling large-scale realistic unstructured meshes while exploiting GPU acceleration. Numerical experiments show that these approaches significantly improve robustness and scalability, achieving substantial reductions in time-to-solution for challenging EMI simulations. These results highlight how combining advanced preconditioning strategies with modern heterogeneous computing platforms can push physiologically detailed cardiac models closer to clinical and research applications.

Edoardo Centofanti
U Pavia
edoardo.centofanti01@universitadipavia.it

MS56

A Hybrid Spectral Deferred Correction Parallel-in-Time Method for the Monodomain Equation in Cardiac Electrophysiology

Simulation of the monodomain equation, crucial for modeling the heart's electrical activity, faces scalability limits when traditional numerical methods only parallelize in space. To optimize the use of large multi-processor computers by distributing the computational load more effectively, time parallelization is essential. We introduce a high-order parallel-in-time method addressing the substantial computational challenges posed by the stiff, multiscale, and nonlinear nature of cardiac dynamics. Our method combines the semi-implicit and exponential spectral deferred correction methods, yielding a hybrid method that is extended to parallel-in-time employing the parallel full approximation scheme in space and time framework. We thoroughly evaluate the stability, accuracy, and robustness of the proposed parallel-in-time method through extensive numerical experiments, using state-of-the-art ionic models. The results underscore the method's potential to significantly enhance real-time and high-fidelity simulations in biomedical research and clinical applications.

Giacomo Rosilho de Souza
Ecole Polytechnique Fédérale de Lausanne
giacomo.rosilhodesouza@epfl.ch

Rolf Krause
King Abdullah University of Science and Technology
rolf.krause@kaust.edu.sa

Simone Pezzuto
Università di Trento
simone.pezzuto@unitn.it

MS56

Using Hybrid Compute Architectures to Achieve Clinically Relevant Timescales for Simulated Transcranial Magnetic Stimulation

In the context of transcranial magnetic stimulation (TMS), a non-invasive brain stimulation method used in the treatment of certain neuropathologies, the modulatory effects of TMS on intracellular calcium dynamics is not well understood. In a collaborative project around the development of the multiscale simulation toolbox Neuron Modeling for TMS (NeMo-TMS) we have been integrating highly detailed calcium models, including intracellular calcium store dynamics, to study the effects of different repetitive TMS (rTMS) brain stimulation protocols on intracellular calcium signaling and the potential long-term adaptive properties of rTMS through calcium signals. This objective brings with it computational challenges. The complexity of the multiscale model does not directly allow simulations on timescales relevant to medical practitioners. Therefore we make use of GPU/CPU infrastructures to combine direct numerical simulations with physics-informed neural networks. This methodology produced results which show that morphological parameters at different scales, from whole cell to the ultrastructural organization of individual spines, are critical for robust synapse to nucleus calcium signaling under rTMS stimulation. Establishing these links between rTMS stimulation and cellular calcium responses will enable optimization of rTMS protocols in clinical settings.

Gillian Queisser, Zachary Miksis
Temple University
gillian.queisser@temple.edu, miksis@temple.edu

MS56

Distributed Delayed Compensation for Electrophysiology Integration

TCardiac electrophysiology is described by differential-algebraic reaction-diffusion equations such as the homogenized bidomain model or the heterogeneous extracellular-membrane-intracellular (EMI) model on the cellular scale. In particular the latter leads to extremely large problems. While the parabolic differential equation is, on the used mesh sizes and time steps, only mildly stiff, the elliptic constraints require the solution of large scale equation systems with iterative solvers, which generally dominates the computational cost. We propose a novel delayed residual compensation strategy that moves the truncation er-

ror over to the next time step in an IMEX Euler scheme and thus improves accuracy for a given iteration count. Moreover, we interpret the delayed residual compensation as a particular time discretization of an augmented system. This allows transfering the delayed residual compensation to higher order time stepping schemes and introducing an overrelaxation factor that further improves accuracy. We analyze the convergence in terms of the overrelaxation factor and the interplay of residual compensation with spectral deferred correction schemes. Numerical experiments illustrate the impact of the approach on efficiency and accuracy.

Julian Schramm, Fatemeh Chegini
Zuse Institute Berlin
schramm@zib.de, chegini@zib.de

Martin Weiser
Zuse Institute (ZIB)
Berlin, Germany
weiser@zib.de

MS57

The State of Scalable CPU and GPU Solvers in the deal.II Library

In this presentation, we compare the current state of scalable solvers for high-order, adaptive, and massively parallel finite element problems as available in the open-source library deal.II. Large linear systems arising from such discretizations are most effectively addressed with multigrid methods, in either algebraic or geometric form, or by combining both approaches. On CPUs, matrix-free geometric multigrid is often superior to algebraic multigrid, particularly for higher-order discretizations. On GPUs, however, parallelization strategies differ fundamentally from CPUs, making the choice of optimal solvers less straightforward. Additional techniques, such as combining algebraic multigrid with p-multigrid or low-order rediscretizations, further broaden the range of scalable options. We present a systematic comparison of BoomerAMG and matrix-free solvers as implemented in deal.II, with a focus on performance, robustness, and memory consumption across CPU and GPU architectures.

Timo Heister
Clemson University
Mathematical Sciences
heister@clemson.edu

MS57

Preconditioning of High-Order Matrix-Free Diffusion Problems on Exascale Systems

We present recent work on benchmarking and tuning preconditioners for high-order, matrix-free diffusion problems. We investigate two general strategies: (1) p-multigrid coupled with algebraic multigrid (AMG) to precondition the lowest-order system, and (2) low-order refined (LOR) preconditioning (also known as the FEMSEM preconditioner), which similarly relies on AMG for the sparse low-order operator. We study the impact of tuning parameters and smoothing choices, and evaluate performance across problems of varying difficulty, driven by mesh quality - equivalently, anisotropy and jumps in the diffusion coefficient. Finally, we assess weak and strong scalability on modern GPU systems and discuss applicability to large-scale design and optimization problems in linear elasticity, acoustics, and fluid flow.

Tzanio Kolev, Veselin Dobrev, Boyan Lazarov
Lawrence Livermore National Laboratory
kolev1@llnl.gov, dobrev1@llnl.gov, lazarov2@llnl.gov

MS57

Recent Development of Hypre on El Capitan Supercomputer

The hypre library has long been a key library for scalable linear solvers and preconditioners in large-scale scientific simulations. With the arrival of El Capitan, Lawrence Livermore National Laboratorys first exascale supercomputer, new challenges and opportunities have emerged for achieving performance and portability at unprecedented scales. This talk highlights recent developments in hypre designed to leverage El Capitans advanced architecture, including GPU acceleration, communication, and optimized multigrid methods. We will present performance results on representative application workloads, discuss lessons learned in porting and scaling hypre to exascale systems, and outline ongoing and future directions for solver technology in the era of heterogeneous computing.

Rui Peng Li
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
li50@llnl.gov

MS57

Performance Portability of High-Order FEM Kernels on Modern Accelerators

The finite element method is a robust mathematical framework for approximating solutions of partial differential equations. As graphics processing units (GPU) continue to evolve into high-performance computing fields, FEM problems can now be solved at high throughput. In this talk, we present GPU implementations and performance results of a matrix-free finite element solver for the Poisson equations and related problems, focusing on fast operator evaluation. A major challenge in general-purpose GPU programming in FEM computations is the growing diversity of vendors, which has resulted in different vendor-specific programming models and significant portability issues. To address this, our research employs the vendor-neutral programming models such as Kokkos, OpenMP and OpenCL, which allow us to target different GPUs while maintaining compatibility. For comparison, we evaluate vendor-specific implementations alongside their portable counterparts. Initial results show nearly identical throughput (degrees of freedom per second) across both implementations. We also propose alternative optimizations techniques for matrix-free operator evaluation, which can improve throughput and push performance closer to hardware limits.

Enes Soydan, Martin Kronbichler
Ruhr University Bochum
enes.soydan@ruhr-uni-bochum.de,
martin.kronbichler@rub.de

Ivan Pribec
Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

ivan.pribec@lrz.de

MS58

BDDC with Algebraic Adaptivity and Compressed Communication for Cardiac Electrophysiology

Cardiac electrophysiology simulations at the cellular level (EMI) are crucial for understanding arrhythmias and developing treatments. These large-scale problems, governed by elliptic and parabolic equations, demand efficient parallel preconditioners within domain decomposition frameworks. The localized nature of cardiac solution features calls for adaptivity, yet traditional mesh refinement is often computationally expensive. We investigate the Balancing Domain Decomposition by Constraints (BDDC) preconditioner combined with algebraic adaptivity and data compression to enhance efficiency, convergence, and accuracy in massively parallel computations. Algebraic adaptivity refines the selection of degrees of freedom, enabling subdomain-wise resolution while limiting global overhead. We explore algebraic adaptivity in Spectral Deferred Correction (SDC) methods in a parallel-in-time setting. To further address communication bottlenecks, we integrate lossy compression, including transform and entropy coding, to reduce inter-subdomain data exchange. The interplay between algebraic adaptivity and compressed communication is analyzed through numerical experiments, e.g., on convergence rates, computational efficiency, and solution accuracy, particularly as the number of subdomains increases.

Fatemeh Chegini
Zuse Institute Berlin
chegini@zib.de

Martin Weiser
Zuse Institute (ZIB)
Berlin, Germany
weiser@zib.de

Thomas Steinke
Zuse Institute Berlin
steinke@zib.de

MS58

Optimized Schwarz Methods in Time for Discrete Transport Control

The problem of finding an optimal control numerically for a system governed by hyperbolic partial differential equations is known to be a subtle one, even for something as simple as the linear transport equation. Given the large amount of data that must be stored, it is imperative to take advantage of distributed architectures in order to keep the problems tractable. In this talk, we will consider solving the discrete transport control problem by parallelizing in the time direction; Using Fourier analysis, we analyze three different iterations: the fixed-point iteration, the relaxed iteration, and preconditioned GMRES. For each case, we propose optimized parameters for the transmission conditions (which may be different on either side of each interface) that lead to fast convergence of the method. We illustrate our results by numerical examples. This is joint work with Duc Quang Bui (Le Mans Universit), Laurence Halpern (Universit Sorbonne Paris Nord) and Felix Kwok (Universit Laval)

Berangére Delourme, Laurence Halpern
Université Paris 13
delourme@math.univ-paris13.fr, halpern@math.univ-paris13.fr

MS58

Time-Parallel Multiple-Shooting for Optimal Control of Quantum Dynamics

Optimal control of quantum systems is a central task in quantum computing, where carefully shaped control pulses steer qubits to implement desired operations. Mathematically, this leads to nonlinear optimal control problems constrained by time-dependent Schrdinger dynamics. Even for systems of only a few qubits, simulations involve long time horizons and stiff dynamics, making conventional sequential optimization slow and computationally costly. In this talk, I will present a time-parallel multiple-shooting approach for quantum control based on multiple-shooting. By partitioning the time horizon into shorter segments, the method enables independent propagations to be computed concurrently while enforcing consistency through matching conditions. I will show how this approach reduces time-to-solution, its scalability on high-performance computing platforms, and its potential to enable the design of faster and more accurate multi-qubit operations.

Stefanie Guenther
Lawrence Livermore National Laboratory
guenther5@llnl.gov

N. Anders Petersson
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
petersson1@llnl.gov

MS58

Diagonal Spectral Deferred Correction for 3D Rayleigh-Benard Convection on Many GPUs

Rayleigh-Benard convection (RBC) is a challenging benchmark problem for incompressible flow, which exhibits strong turbulence. In this talk, we explore the capabilities of the iterative time-stepping method spectral deferred correction (SDC) with implicit-explicit splitting to improve the time integration process of a pseudo-spectral implementation of RBC over more traditional Runge-Kutta (RK) methods. A key ingredient is a recent class of parallel SDC preconditioners, enabling small scale time-parallelism with high parallel efficiency even for hyperbolic problems. We find that SDC yields much higher accuracy at the same step size than reference RK methods for a range of Rayleigh numbers in the turbulent regime. We demonstrate that this translates to faster time-to-solution due to excellent space-time-parallel scaling on both CPUs and GPUs.

Thomas Saupe
Forschungszentrum Juelich
t.baumann@fz-juelich.de

Thibaut Lunet, Sebastian Götschel
Hamburg University of Technology
thibaut.lunet@tuhh.de, sebastian.goetschel@tuhh.de

Daniel Ruprecht
Institute of Mathematics
Hamburg University of Technology
ruprecht@tuhh.de

Robert Speck

Juelich Supercomputing Centre
Forschungszentrum Juelich GmbH
r.speck@fz-juelich.de

MS59

A High-Performance Python Pathway to Hybrid Climate Modeling

GT4Py is a Python framework for weather and climate applications for simplifying the development and maintenance of high-performance codes in prototyping and production environments. Although the convergence of AI and HPC is already reshaping weather and climate modeling, production codes still demand performance, portability, and maintainability that are not easy to reconcile with Python-centric rapid development workflows. GT4Py addresses this gap by using an embedded Python DSL and code-generation toolchain for stencil- and grid-based computations. By separating algorithmic intent from execution backends, GT4Py enables domain scientists to express physics-based computations cleanly in Python while delivering near hand-tuned performance, thus providing a credible path from prototype to operational deployment. This enables solutions that would be trickier in traditional programing approaches, like the creation of true hybrid models in which, for instance, a physics-based dynamical core written in GT4Py can work with a neural network surrogate for a subgrid-scale parameterization and, conversely, the dynamical core can be automatically differentiated and used in a data assimilation scheme. By providing a unified, Python-native framework for both HPC and AI components, GT4Py offers a robust pathway to building performant, portable, maintainable, and innovative climate models.

Mauro Bianco
CSCS Switzerland
mauro.bianco@cscs.ch

MS59

Development of Scientific Computing Workflows for Integrating Numerical Simulations and Machine Learning for Multiphase Boiling Problems

Boiling is a chaotic multiphase process critical to cooling and thermal management in systems from nuclear energy to high-performance computing. Its complexity arises from nucleation, interface dynamics, and tightly coupled heatmomentum transfer, which are difficult to resolve for both physics solvers and machine learning (ML) surrogates. Direct simulations are costly due to scales from micrometers to centimeters, while neural PDE models often need future-state information and fail to forecast boiling autonomously. ML progress is further limited by the lack of validated datasets for phase change. To address this, we built a computational workflow combining high-fidelity multiphase simulations using Flash-X, dataset curation, and surrogate model development. Adaptive mesh refinement (AMR) in AMReX, coupled with Flash-X multiphysics, enables scalable simulations of pool, flow, and sub-cooled boiling across fluids, geometries, and boundary conditions. These simulations yield ground truth on bubble nucleation, interface evolution, and heat/flow fields while surpassing state-of-the-art performance. We will present this workflow and showcase BubbleML, a dataset of 79 validated scenarios spanning gravity, flow, sub-cooling, and wall superheat, together with Bubbleformer, a transformer-based surrogate that forecasts nucleation, interface dynamics, and heat transfer across regimes with high physical fidelity.

S. M. Shakeel Hassan, Xianwei Zou
University of California, Irvine
sheikhh1@uci.edu, xianwz2@uci.edu

Akash V. Dhruv
Argonne National Laboratory
adhruv@anl.gov

Vishwanath Ganesan
University of Illinois at Urbana-Champaign
vganesa@illinois.edu

Aparna Chandramowlishwaran
University of California, Irvine
amowli@uci.edu

MS59

Deep Learning Workflows for Protein Folding and Chemistry

Deep learning has become a key component of scientific workflows, slotting in alongside classical modeling and simulation. I will discuss the practical challenges and insights from designing and deploying two workflows, Flask Co-Pilot and ElMerFold. Flask Co-Pilot is an agentic system for molecular design and synthesis planning, which integrates a high-level orchestrator, chemistry-specific ML models, and classic simulation tools. ElMerFold is a large-scale system for producing distillation datasets for protein structure prediction, combining bioinformatics tooling and pretrained models to generate new, high-quality synthetic training data. Together, these projects underscore that tightly coupling learned models with traditional HPC infrastructure yields new capabilities for computational scientists.

Nikoli Dryden
Lawrence Livermore National Laboratory
dryden1@llnl.gov

MS59

On-the-Fly Model Training from Simulation Pipelines

Coupling machine learning with high-performance simulations is often constrained by the inefficiencies of offline dataset generation, where simulation outputs are written to storage and later reloaded for training. This talk presents work in progress on enabling on-the-fly model training directly from simulation pipelines without extra copies, thereby avoiding redundant I/O, supporting adaptive dataset sizing until convergence, and allowing dynamic steering of simulations based on learning progress. Such a pipeline also opens the door to continual learning, where models can be refined during inference-driven simulations. We discuss the challenges that arise in this setting, including efficient interfacing between simulation and PyTorch training (e.g., tensor mapping and buffering strategies to balance mismatched rates), as well as mitigating biases from non-IID samples inherent in temporally correlated physics data. Our ongoing work explores zero-copy integration and adaptive sampling mechanisms, aiming to establish a foundation for more efficient and adaptive AI-HPC workflows.

Mohamed Wahib
RIKEN Center for Computational Science

mohamed.attia@riken.jp

MS60

A Linear Complexity H2 Direct Solver for Fine-Grained Parallel Architectures

We present factorization and solution phases for a new linear complexity direct solver designed for concurrent batch operations on fine-grained parallel architectures, for matrices amenable to hierarchical representation. We focus on the strong-admissibility-based $\mathcal{H}^2$ format, where strong recursive skeletonization factorization compresses remote interactions. We build upon previous implementations of $\mathcal{H}^2$ matrix construction for efficient factorization and solution algorithm design. The algorithms are "blackbox' in the sense that the only inputs are the matrix and right-hand side, without analytical or geometrical information about the origin of the system. We demonstrate linear complexity scaling in both time and memory on four representative families of dense matrices up to one million in size. Parallel scaling up to 16 threads is enabled by a multi-level matrix graph coloring and avoidance of dynamic memory allocations thanks to prefix-sum memory management. We break down the timings of different phases, identify phases that are memory-bandwidth limited, and discuss alternatives for phases that may be sensitive to the trend to employ lower precisions for performance.

Wajih Boukaram
King Abdullah University of Science and Technology
wajih.boukaram@gmail.com

MS60

Adaptive, Matrix-Free Algorithms for Constructing and Factorizing Hierarchical Matrices

We introduce a new class of adaptive algorithms for constructing low-rank approximations of a given operator A in a matrix-free settingrelying solely on the action of the map $x \mapsto Ax$. Unlike existing approaches, these methods support tolerances well below the square root of machine epsilon, offering greater flexibility and precision. Beyond basic approximation, we demonstrate how these techniques can be extended to tensor compression, hierarchical matrix construction, and fast direct solvers. The proposed framework is applicable to a wide range of operators, including Hessians from inverse problems, Schur-complement matrices arising in sparse direct solvers, and kernel matrices encountered in machine learning and integral equation formulations.

Chao Chen
North Carolina State University
chao_chen@ncsu.edu

MS60

A Multilevel Solver for Kernel Matrices

The $\mathcal{H}^2$-matrix format provides linear scaling storage and matrix-vector multiplications for dense kernel matrices from engineering and machine learning. The format uses a hierarchical clustering of the kernel matrix points and naturally defines a hierarchical structure. We propose using this hierarchical structure to define a multilevel solver, like the multigrid method, consisting of smoothing and coarse grid correction. The goal is to obtain smoothing that is complementary to coarse grid correction and thus obtain a solver that scales linearly with the number of kernel matrix points.

Daria Sushnikova, George M. Turkiyyah
KAUST
daria.sushnikova@kaust.edu.sa,
George.Turkiyyah@kaust.edu.sa

Edmond Chow
School of Computational Science and Engineering
Georgia Institute of Technology
echow@cc.gatech.edu

David E. Keyes
King Abdullah University of Science and Technology (KAUST)
david.keyes@kaust.edu.sa

MS60

Blockwise Tensor Hypercontraction of the Electron Repulsion Integral Tensor

Evaluation and storage of the electron repulsion integral (ERI) tensor is a typical bottleneck in electronic structure computations. Of particular interest is computing Hartree Fock orbitals and post Hartree Fock methods such as many-body perturbation theory. In this talk, we present a method for compressing blocks of the ERI tensor in Tensor Hypercontraction (THC) format under the clustered low-rank (CLR) framework. This method is based on the truncated bipolar multipole expansion and quadrature, and is applicable to molecular systems with local Gaussian-type orbital (GTO) atomic basis functions. The compressed form allows efficient calculation of the Coulomb and exchange matrices as well as the MP2 energy correction in the local MP2 method. We present theoretical analysis of the accuracy and size of the approximation as well as numerical results in relation to existing methods.

Jack Weinstein
University of Illinois-Urbana-Champaign
weinstn2@illinois.edu

MS61

Substructured Optimized Schwarz Domain Decomposition Solver on Multiple Gpus for the Helmholtz Equation

This contribution discusses the design of a GPU-accelerated implementation of a substructured Optimized Schwarz Domain Decomposition Solver for the finite element solution of the Helmholtz Equation. We compare both sparse and dense linear algebra to handle local subdomain solves. For dense algebra, two approaches are considered: the first one computes the LU factorization of each subdomain matrix and executes a batched forward-backward solve at each iteration. The second one computes the explicit Schur complement and only requires a single matrix multiplication per subdomain per iteration. The sparse approach only computes the LU factorization of each subdomain matrix and uses a batched sparse direct solver to perform the local solves at each iteration. The implementation, based on GmshDDM, Petsc, MAGMA and cuDSS is validated on up to 128 NVIDIA A100 40GB GPUs. The dense implementation mainly uses the MAGMA library, and the sparse implementation currently uses the cuDSS library as a batched sparse direct solver. Convergence and scalability results will be pre-

sented for representative 2D and 3D benchmarks.

Roland Greffe
Université de Liège
r.greffe@uliege.be

Ahmed Chabib, Axel Modave
ENSTA, Institut Polytechnique de Paris
ahmed.chabib@ensta.fr, axel.modave@ensta.fr

Christophe Geuzaine
Université de Liège
cgeuzaine@uliege.be

MS61

A Green's Function Inspired Approach to Strength-of-Connection for Algebraic Multigrid

Multiphysics simulations rely heavily on linear solvers, which are usually computationally expensive, difficult to scale and have significant robustness limitations. Systems containing more sophisticated coupling/scales lead to more difficult matrix equations. One key solver component affecting robustness is the notion of strength-of-connection (SOC) in multigrid. Strength-of-connection classifies interactions as strong or weak, as strongly coupled equations require different treatment than loosely coupled equations. We propose a Green's function inspired strength-of-connection idea for use in problems with both large material variations and stretched meshes. We view this as a generalization of the distance Laplacian SOC, and will present results on multimaterial Poisson problems in a variety of single physics and multiphysics context.

Chris Siefert
Sandia National Laboratories
csiefer@sandia.gov

MS61

Algebraic Hierarchical Partitioning to Improve $\mathcal{H}$-Matrix Compression

Solving large dense problems is a challenging task in many industrial applications such as computational electromagnetics. $\mathcal{H}$-matrices may solve these problems efficiently while significantly reducing storage requirements. The compression rate and efficiency of $\mathcal{H}$-matrices depend on the partitioning of the unknowns. This partitioning must be hierarchical and should respect geometric criteria in order to maximize compression. A fixed block size constraint is added to accommodate load balancing and performance concerns on HPC runtime systems. Numerous partitioning schemes exist, but few meet all requirements. Some geometric methods such as recursive coordinates bipartition, space-filling curves and cobblestone sorting provide acceptable results whereas algebraic graph partitioner generally do not. We propose a method to build a graph from the mesh to combine geometric and physical properties, suited to provide adapted partitions reliably. We review fitting geometric partitioning methods and compare them to our algebraic approach using relevant metrics such as compression rates before and after factorization and execution time of $\mathcal{H}$-matrix assembly, factorization and solving step. We also study resulting partitions based on their volume, overlap and distance. Our contribution shows significant improvements in compression rates and execution times for complex 3D objects with multiple materials which are representative of industrial applications.

Dimitri Walther
CEA, Univ. Bordeaux, CNRS, Bordeaux INP, INRIA, LaBRI,
UMR 5800 F-33400 Talence
dimitri.walther@inria.fr

Mathieu Faverge
Bordeaux INP, Univ. Bordeaux, CNRS, Inria
mathieu.faverge@inria.fr

Matthieu Lecouvez
CEA
matthieu.lecouvez@cea.fr

Pierre Ramet
Bordeaux University - INRIA
pierre.ramet@inria.fr

MS62

A Novel fixed-accuracy Randomized Interpolative Decomposition Algorithm and its Mixed-precision Extension

Randomized Interpolative Decomposition (RID) is a renowned low-rank approximation algorithm with a remarkably low operational complexity. In our work, we take interest in the fixed-accuracy setting where the rank of the approximation is computed such that the backward error of the low-rank approximation falls below a target accuracy. The fixed-accuracy algorithms from literature however suffer from either poor efficiency or poor accuracy. To overcome these shortcomings, we design a novel variant centered around an incrementally augmented sketch. The stopping criterion that we use to halt this procedure is based on the norm of the trailing submatrix in the sketch and, therefore, it is cheap; we asses its robustness through experiments on a wide range of matrices. We also discuss the possible extension of the algorithm to mixed precision arithmetic.

Karmijn Hoogveld
CNRS, IRIT, Université de Toulouse
karmijn.hoogveld@irit.fr

Alfredo Buttari
CNRS-IRIT-Université de Toulouse
alfredo.buttari@irit.fr

Theo Mary
Sorbonne Universite, CNRS, LIP6
theo.mary@lip6.fr

Emmanuel Agullo
INRIA
emmanuel.agullo@inria.fr

MS62

Adaptive Mixed-Precision Algorithms for Next-Generation Scientific Simulations

The future of large-scale simulations is increasingly tied to hardware features originally designed for AI workloads-especially low-precision arithmetic. Modern GPUs embody this shift, delivering substantial speedups through reduced-precision computations that lower execution time, shrink memory footprints, and cut energy consumption.

Building on these capabilities, we design fast mixed-precision linear algebra algorithms that adaptively choose the right precision at the right moment. Our dynamic precision-conversion strategy preserves high accuracy only where it truly matters, all while maintaining application-level numerical reliability. This talk will demonstrate how these algorithms reshape computational efficiency for geospatial statisticians and geophysicists, with far-reaching benefits for environmental computational statistics, seismic imaging, and beyond.

Hatem Ltaief
King Abdullah University of Science and Technology (KAUST)
hatem.ltaief@kaust.edu.sa

MS62

Accelerating Low-Rank Tensor Approximations with Mixed Precision

Low-rank tensor decompositions are widely used in scientific computing, data analysis, and engineering applications due to their advantages in memory efficiency and scalability when handling high-dimensional data. These benefits have motivated the development of numerous decomposition techniques and their application across various fields, such as quantum physics and machine learning. However, constructing such methods and performing operations within these structures can be computationally intensive, often due to the high cost of tensor construction and the complexity of numerical computations. To overcome these challenges, high-performance computing (HPC) strategies such as parallelism, randomized algorithms, and mixed-precision arithmetic have become essential for improving computational performance and reducing memory usage and communication overhead in (multi)linear algebra operations. In this talk, we focus on recent trends in accelerating the performance of low-rank tensor decompositions and their applications, with a particular emphasis on mixed-precision techniques. We explore how combining low- and high-precision arithmetic enables significant acceleration without sacrificing accuracy and discuss their potential for developing large-scale applications in scientific computing.

Eda Oktay
Max Planck Institute
edaoktay95@gmail.com

Martin Koehler
Max Planck Institute for Dynamics of Complex Technical Systems Magdeburg
koehlerm@mpi-magdeburg.mpg.de

Xiaobo Liu
MPI for Dynamics of Complex Technical Systems, Magdeburg
xliu@mpi-magdeburg.mpg.de

Peter Benner
Max Planck Institute Magdeburg
benner@mpi-magdeburg.mpg.de

MS63

Hybrid Intelligence for Sustainable Infrastructure: Modeling Resilience in Energy Communities and Emissions in Platooning

This talk presents a hybrid intelligence framework that integrates machine learning and optimization to enhance sustainability and resilience in two infrastructure domains: urban vehicle platooning and renewable energy communities (RECs). In the first case, we analyze real and simulated data to model how urban traffic features—such as intersections and signal timing—affect platoon cohesion and CO_2 emissions. Machine learning models quantify these effects and inform a routing algorithm that optimizes either for cohesion or emissions. Results show significant improvements over conventional shortest-path routing in both environmental impact and platoon stability. In the second case, we evaluate the resilience of RECs by modeling peer-to-peer energy exchanges as dynamic networks. Using centrality metrics and disruption simulations, we identify critical nodes and assess the impact of failures on network efficiency. While RECs reduce emissions and energy costs under normal conditions, resilience is highly dependent on user participation and network structure. Together, these studies demonstrate how hybrid ML-optimization approaches can support adaptive, efficient, and robust infrastructure in energy and mobility systems.

Natalia Selini Hadjidimitriou
University of Modena and Reggio Emilia
selini@unimore.it

MS63

A Graph-Partitioning Based Continuous Optimization Approach to Semi-Supervised Clustering Problems

Semi-supervised clustering is a basic problem in various applications. Most existing methods require knowledge of the ideal cluster number, which is often difficult to obtain in practice. Besides, satisfying the must-link constraints is another major challenge for these methods. In this work, we view the semi-supervised clustering task as a partitioning problem on a graph associated with the given dataset, where the similarity matrix includes a scaling parameter to reflect the must-link constraints. Utilizing a relaxation technique, we formulate the graph partitioning problem into a continuous optimization model that does not require the exact cluster number, but only an overestimate of it. We then propose a block coordinate descent algorithm to efficiently solve this model, and establish its convergence result. Based on the obtained solution, we can construct the clusters that theoretically meet the must-link constraints under mild assumptions. Furthermore, we verify the effectiveness and efficiency of our proposed method through comprehensive numerical experiments.

Xin Liu
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
liuxin@lsec.cc.ac.cn

MS63

Physics-Based Surrogates and Ml for Multi-Objective Design of Crystal Growth Processes

Crystal growth processes are governed by coupled, nonlinear transport phenomena that make first-principles modeling both essential and computationally expensive. In this talk, we present a hybrid framework that integrates high-fidelity simulation, machine learning, and multi-objective optimization (MOO) to enable systematic design and control of crystal growth systems. Using data from physics-based simulations such as global heat transfer and melt flow models, we train ML surrogates to approximate key physi-

cal outputs, including melt-crystal interface shape, thermal gradients, and defect-relevant metrics with high accuracy and reduced computational cost. These surrogates are then embedded into different MOO frameworks to explore trade-offs between competing objectives, such as crystal quality and throughput. The approach is demonstrated for both Czochralski and Floating Zone silicon growth, showing how hybrid modeling enables scalable design exploration, informed sensitivity analysis, and data-efficient optimization across process parameter and geometry spaces. This work illustrates the potential of combining machine learning and mathematical optimization with physically grounded simulations to tackle complex, high-dimensional inverse design tasks in materials processing.

Milena Petkovic
Leibniz Institute for Crystal Growth
milena.petkovic@ikz-berlin.de

Natasha Dropka
Leibniz Institut for Crystal Growth
natasha.dropka@ikz-berlin.de

MS63

Double-AI for Clustering Scientific Publications: Combining Parallel Optimization and LLMs at Scale

Clustering scientific publications at scale is essential for navigating todays vast research landscape. We address this challenge through a hybrid framework Double-AI (2xAI), combining Artificial Intelligence and Algorithmic Intelligence. Our focus is on fuzzy clustering, which allows publications to belong to multiple topic clusters and which can be modeled as a non-convex constrained optimization problem, aligning observed and predicted article similarities. We develop a GPU-accelerated parallel solver tailored for massive datasets like OpenAlex and Web of Science, comprising millions of articles and billions of citations. The solver exploits problem structure to achieve scalability and efficiency, supported by theoretical insights. Large Language Models (LLMs) generate semantic representations of publications to support clustering, while integrating multiple data features including citation and text information - into the optimization framework offers promising directions to enhance clustering quality. By combining LLM embeddings with parallel optimization, 2xAI delivers scalable, interpretable AI for organizing scientific knowledge.

Thi Huong Vu
Zuse Institute Berlin
huong.vu@zib.de

MS64

Sparse Matrix Computations for Verifying the Binary Goldbach Conjecture

Sparse matrix techniques have proven to be useful in many different application areas, including number theory. We exploit such techniques in a somewhat surprising application, the verification of the binary Goldbach conjecture, which states that every even number larger than two can be written as the sum of two prime numbers. The conjecture is an open problem in mathematics since 1742. Its truth has been verified for even numbers up to 4×10^{18} in 2012, but a proof has not yet been found. We will present a new algorithm for verifying the conjecture, where we use a mix of suitable sparse and dense data structures to register verified even numbers and where we exploit repeating difference patterns (admissible k-tuples) in the sequence of prime numbers. In a preprocessing phase, we create a sparse matrix with rows representing subintervals of the interval to be checked, columns representing difference patterns, and nonzeros the occurrence of a pattern in a subinterval. To maximise reuse of repeating patterns in the actual Goldbach verification, we greedily select a minimal subset of the matrix columns that covers all the rows by a nonzero, which is equivalent to solving the hypergraph vertex cover problem. We harness parallel computers to reach the high numbers we are interested in and possibly push the verification frontier. Here, we distribute intervals cyclically within a single parallel job for reasons of load balance, and by blocks for independent jobs.

Rob H. Bisseling
Utrecht University
Mathematical Institute
r.h.bisseling@uu.nl

MS65

Mixing and Lowering Precisions in CFD Codes: Sharing the CEEC Project Experience

Mixed precision algorithms offers a promising path to energy-efficient scientific computing. However, identifying where and how to apply reduced precision without compromising solution accuracy remains a key challenge in scientific simulations. We propose a methodology for enabling mixed-precision in spectral element codes using a computer arithmetic tool Verificarlo, roofline modeling, and computer arithmetic techniques. As case studies, we consider Nekbone, a mini-app of the CFD solver Nek5000, and a modern CFD solver Neko. We apply our methodology on the underlying iterative solvers like the preconditioned Conjugate Gradient (PCG) by carefully re-designing them with reduced mixed-precision. With Verificarlo and computer arithmetic techniques, we resolve the stagnation issue in mixed-precision PCG that requires some injections of double-precision for the accuracy critical operations such as dot products and global communications. We evaluate the derived mixed-precision versions of the codes on the EuroHPC JU clusters by combining metrics in three dimensions: accuracy, time-to-solution, and energy-to-solution. Notably, mixed-precision in Nekbone reduces time-to-solution by roughly 1.62x and energy-to-solution by up to 2.43x on MareNostrum 5, while in the real-world Neko application, the gain is up to 1.3x in both time and energy, with the accuracy that matches double-precision results.

Roman Iakymchuk
Umeå University
Uppsala University
riakymch@cs.umu.se

MS65

The Cambrian Explosion of Mixed-Precision Matrix Multiplication for Quantized Deep Learning Inference

Recent advances in deep learning (DL) have led to a shift from traditional 64-bit floating point (FP64) computations toward reduced-precision formats combined with mixed-precision arithmetic. This transition enhances computational throughput, reduces memory and bandwidth usage, and improves energy efficiency, offering significant advantages for resource-constrained edge devices. To support this shift, hardware architectures have evolved accord-

ingly, now including adapted ISAs (Instruction Set Architectures) that expose mixed-precision arithmetic units tailored for DL workloads. At the heart of many DL and scientific computing tasks is the general matrix-matrix multiplication (GEMM), a fundamental kernel historically optimized using AXPY vector instructions on SIMD (single instruction, multiple data) units. However, as hardware moves toward mixed-precision DOT-product-centric operations optimized for quantized inference, these legacy approaches are being phased out. This talk illustrates novel micro-kernel designs and data layouts that better exploit todays specialized hardware and demonstrate significant performance gains for mixed-precision integer (MIP) arithmetic over floating-point implementations across three representative CPU architectures. These contributions highlight a new era of GEMM optimization-driven by the demands of DL inference on heterogeneous architectures, marking a Cambrian period for matrix multiplication.

Héctor Martínez
University of Córdoba
Spain
hekthor.martinez@gmail.com

Adrián Castelló
Universitat Politècnica de València
Spain
adcastel@disca.upv.es

Francisco D. Igual
Universidad Complutense de Madrid
Spain
figual@ucm.es

Enrique S. Quintana-Orti
Universitat Politècnica de València
quintana@disca.upv.es

MS65

Towards a Mixed-precision Inexact Low-rank Lyapunov ADI

Continuous-time algebraic Lyapunov equations arise in the fields of, e.g., optimal control and model order reduction. For many applications, the coefficient matrices and are large and sparse, while the solution matrix has a low numerical rank. In this setting, the alternating-directions implicit (ADI) method, which directly operates on the low-rank factors of the solution matrix, is one of the most widely used algorithms for this type of equation. We report on our progress of applying mixed-precision techniques to the low-rank Lyapunov ADI, namely the use of multi-precision low-rank factorizations, as well as a mixed-precision variant of the inexact low-rank Lyapunov ADI.

Jonas Schulze
Max Planck Institute for
Dynamics of Complex Technical Systems
jschulze@mpi-magdeburg.mpg.de

MS66

Advances in Scalable Distributed Gpu Preconditioners in Ginkgo

Effective preconditioners are essential to accelerate the linear solvers for various scientific applications. In this talk we will look at the various distributed preconditioners in Ginkgo, a high performance numerical linear algebra library and the building blocks that enable effective scaling. We will also consider a few example applications such as cardiac electrophysiology and CFD, to showcase the effectiveness of our GPU distributed preconditioners in real-world applications.

Pratik Nayak
Karlsruhe Institute of Technology
pratik.nayak@tum.de

MS66

Parallel Full Approximation Schemes for Convex Variational Problems

In this talk, we present a parallel full approximation scheme (FAS) designed for convex variational problems. All local problems in a parallel subspace correction method are fully localized, eliminating the need for global computations and thereby ensuring good computational efficiency. By incorporating both local and global line search strategies, whose additional computational cost is marginal, we establish global convergence of the method. Several illustrative examples are provided, including overlapping domain decomposition and multigrid methods for solving some nonlinear partial differential equations.

Jongho Park, Jinchao Xu
KAUST
jongho.park@kaust.edu.sa, jinchao.xu@kaust.edu.sa

MS66

Parallel High-Order Generation of Finite Element Meshes

High-order meshes are used to model curved geometries while employing many fewer high-order elements. This leads to accurate high-order PDE solutions with greater efficiency. Parallel algorithms have been utilized to generate large-scale, high-order finite element meshes for CFD and solid mechanics applications. In this talk, we propose two algorithms for the parallel generation of high-order tetrahedral finite element meshes based on optimal weights. Our mesh generation algorithms consist of four steps. First, we enrich the initial low-order mesh with additional nodes to create a straight-sided high-order mesh. The high-order nodes are then placed onto the curved boundary. Second, a set of optimal weights relating each interior node to its neighbors is computed in an embarrassingly parallel manner. The goal is to determine a mesh that is similar to the original mesh. The third step is to apply an application-based, user-defined boundary deformation. Finally, the new coordinates of the interior nodes are calculated by solving the updated system of linear equations based on the boundary deformation. The mesh topology is preserved throughout. We solve the resulting linear system using parallel iterative linear solvers that support multiple right-hand sides. To demonstrate the methods efficacy, we present several examples in three dimensions as well as some numerical results pertaining to the scalability of the methods

Christina Hymer, Suzanne M. Shontz
University of Kansas
chymer@ku.edu, shontz@ku.edu

MS66

Shifted Penalty Multigrid Method for Contact

High-performance computing is crucial for solving large-scale contact problems. Simulating such phenomena at en-

gineering scale within practical runtimes is often limited by resources or budgets. Thus, efficient algorithms and software are needed to exploit modern hardware such as multi-core CPUs and GPUs. Iterative solvers and preconditioners are key. Monotone Multigrid (MMG) methods provide a robust baseline with optimal complexity, while Penalty and Augmented Lagrangian methods offer flexibility for over-constrained problems with fuzzy constraints. The shifted-penalty method for contact ensures accurate constraint satisfaction and is competitive with non-smooth techniques such as semi-smooth Newton. To combine the optimal runtime of MMG with the flexibility of shifted-penalty methods, we propose the Shifted-Penalty Multigrid (SPMG) method for contact, designed for GPU efficiency. Our approach leverages matrix-free operators and memory-efficient semi-structured meshes for linear elasticity discretizations. We present the SPMG algorithm, including nonlinear smoothing and constraint-coarsening strategies, and report progress on its high-performance implementation for the Grace-Hopper super-chip on the CSCS Alps supercomputer. We focus on single-node performance, kernel design, and detailed performance analysis for obstacle contact problems, showcasing complex scenarios with hundreds of millions of degrees of freedom.

Patrick Zulian
Euler institute, Università della Svizzera italiana
Switzerland
patrick.zulian@usi.ch

Hardik Kothari, Gabriele Marchi
Università della Svizzera italiana
hardik.kothari@usi.ch, gabriele.marchi@usi.ch

Austen Nelson, Panayot Vassilevski
Portland State University
ajn6@pdx.edu, panayot@pdx.edu

Rolf Krause
King Abdullah University of Science and Technology
rolf.krause@kaust.edu.sa

MS67

Scalable two-level Schwarz preconditioners for very large highly heterogeneous problems

Two-level overlapping Schwarz methods provide a robust and scalable preconditioning scheme for the iterative solution of linear systems arising from the discretization of elliptic problems. The robustness and scalability of the preconditioner are enabled via the solution of a global coupling problem defined in a coarse space. A suitable choice of a coarse space is able to accurately represent troubling error modes and effectively accelerate convergence. In this talk, we discuss different algebraic approaches for the construction of the coarse space. We present a parallel implementation in the FROSch (Fast and Robust Overlapping Schwarz) library within the Trilinos software framework and examine its parallel performance for very large heterogeneous problems. In addition, we investigate the use of local inexact solvers to further improve the performance of the preconditioner by reducing the computational cost of setting it up and applying it.

Filipe Cumaru, Alexander Heinlein
Delft University of Technology
Delft Institute of Applied Mathematics
f.a.cumarusilvaalves@tudelft.nl, a.heinlein@tudelft.nl

Hadi Hajibeygi
Delft University of Technology
h.hajibeygi@tudelft.nl

MS67

Multigrid Methods on Gpus

We discuss the implementation of multigrid methods, in particular of vertex-patch smoothers for Poisson, Stokes, and biharmonic problems on GPUs. We present the key points that determine good performance. Tensor-methods like fast diagonalization yield fast converging methods and their high computational intensity keeps memory access in reasonable bounds. On the other hand, they are a burden on shared memory and measures have to be taken to avoid bank conflicts. We present detailed performance analysis with respect to different metrics.

Guido Kanschat
Heidelberg University, Germany
kanschat@uni-heidelberg.de

Cu Cui
Heidelberg University
cu.cui@alumni.uni-heidelberg.de

MS67

Construction and Profiling of a Fast Direct Solver for Distributed Finite-element Computations

In this presentation, we adopt fast direct solver strategies to the context of 2D MPI-parallelized finite-element and spectral-element computations on locally refined meshes [1, 2]. We discuss algorithmic generalizations we made to the solution strategy, which heavily uses ideas from the context of distributed multigrid computations [3]. Furthermore, we discuss computational complexities, predict expected speedups and compare the values to the ones obtained from computational experiments. The efficiency of the developed direct solver is shown by applying it to different challenging application cases. The code of the solver relies on the high-performance finite-element library deal.II, in particular on its multigrid infrastructure, and is, as a consequence, short and lightweight. [1] Martinsson, P.G., 2013. A direct solver for variable coefficient elliptic PDEs discretized via a composite spectral collocation method. JCP. [2] Babb, T., Gillman, A., Hao, S. and Martinsson, P.G., 2018. An accelerated Poisson solver based on multidomain spectral discretization. BIT Numerical Mathematics. [3] Munch, P., Heister, T., Prieto Saavedra, L. and Kronbichler, M., 2023. Efficient distributed matrix-free multigrid methods on locally refined meshes for FEM computations. ACM TOMS.

Peter Munch
Institute of Mathematics
Technical University Berlin
muench@math.tu-berlin.de

Adrianna Gillman
University of Colorado at Boulder
Department of Applied Mathematics
adrianna.gillman@colorado.edu

MS68

Space-Time Parallel Framework for the Analysis of 3D Dynamic Contrast-Enhanced Ultrasound Mea-

surements

Tumor perfusion is a key indicator in monitoring cancer treatments. Ultrasound imaging would enable the regular collection of patient data due to its broad availability, thus facilitating further treatment personalization. We adopt a two-compartment (arterial and venous blood flow) field model, along with a probabilistic description of the model parameters (perfusion and blood flow velocities), to simulate the tracer distribution. Typical approaches to compute the model parameters corresponding to measurement data, e.g. Bayesian inference, require numerous solves of the underlying forward model. In our space-time parallel framework, we combine LibPFASST, handling the parallel-in-time integration, with AMReX as a space-parallel solver. We employ our framework to massively parallelize the underlying simulations of the two-compartment model in Bayesian approaches for parameter inference.

Fynn Bensel
Te
fynn.bensel@tuhh.de

Sophie Externbrink, Sebastian Götschel
Hamburg University of Technology
sophie.externbrink@tuhh.de, sebastian.goetschel@tuhh.de

MS68

A Parallel-in-Time Navier-Stokes Solver Using Augmented Lagrangian and Space-Time Multigrid Preconditioners

Multigrid based parallel-in-time solvers, like multigrid waveform relaxation and space time multigrid, can enhance the parallel scaling behavior of parabolic evolution equations by solving multiple time steps at once. Here, we present a Navier-Stokes solver based on a global-in-time Newton's method or Picard iteration. The linear systems are solved using by a Pressure Schur Complement (PSC) iteration with parallel-in-time multigrid preconditioners. To avoid a deteriorating convergence behavior when an increasing number of time steps are calculated in parallel, we combine the least squares commutator (LSC) and an augmented Lagrangian (AL) preconditioner. The AL modification does not change the solution of the PSC iteration but the multigrid preconditioners for the velocity subproblem require careful treatment. The global-in-time solver shows a robust convergence behavior for moderate Reynolds numbers and a wall time speedup due to the improved scaling properties. Furthermore, we show that this is also applicable to non Newtonian fluids like polymere flows while the application to high Reynolds number problems remains difficult.

Jonas Duennebacke
TU Dortmund University
Department of Mathematics
jonas.duennebacke@math.tu-dortmund.de

Christoph Lohmann
Dortmund University of Technology
christoph.lohmann@mathematik.tu-dortmund.de

Stefan Turek
TU Dortmund University
Department of Mathematics
stefan.turek.mathematik.tu-dortmund.de

MS68

An Mgrit Approach for Particle-in-Fourier Schemes in Kinetic Plasma Simulations

Kinetic plasma simulations play a critical role in applications of societal relevance such as nuclear fusion and building the next-generation of compact particle accelerators. They are also widely used in studying astrophysical phenomena and industrial plasma processes. Particle-In-Fourier (PIF) schemes are attractive for long-time integration of kinetic plasma simulations as they conserve charge, momentum and energy, exhibit a variational structure, do not have aliasing and have excellent stability properties. However, they are typically more expensive than the commonly used Particle-In-Cell (PIC) schemes due to the requirement of non-uniform discrete Fourier transforms (DFT) or fast Fourier transforms (FFT). In this talk, we present a multigrid reduction in time (MGRIT) approach for PIF schemes building upon our previous work on parareal for PIF schemes. The numerical implementation is done by coupling Xbraid and IPPL libraries and we present our findings and challenges in this talk.

Sriramkrishnan Muralikrishnan
Forschungszentrum Juelich
s.muralikrishnan@fz-juelich.de

Jacob B. Schroder
Department of Mathematics and Statistics
University of New Mexico
jbschroder@unm.edu

Robert Speck
Forschungszentrum Juelich GmbH
Juelich Supercomputing Centre
r.speck@fz-juelich.de

MS68

Multigrid Reduction in Time for Linear Poroelasticity

Exploiting concurrency in both space and time is essential for efficient use of modern parallel architectures. We present the integration of the multigrid reduction in time (MGRIT) algorithm, provided by the XBraid library, with the UG4 (Unstructured Grid) software framework. UG4 supports finite volume and finite element discretizations on unstructured grids and tightly couples these with iterative solvers for linear and nonlinear systems.In particular, geometric multigrid solvers can be employed as efficient preconditioners. Coupling this spatial infrastructure with MGRIT's multi-level time hierarchy yields a fully iterative framework in space and time. We demonstrate the efficiency of this approach for poroelasticity, where the strong coupling of deformation and flow leads to challenging numerical problems. Beyond poroelasticity, the framework also applies to the Heat Equation, the Monodomain model, and the Elder problem. Numerical results show that this combination can significantly reduce wall-clock time by exploiting very high core counts, making it a promising tool for large-scale time-dependent simulations.

Martin Parnet, Arne Naegel
Goethe University Frankfurt
m.parnet@em.uni-frankfurt.de, arne.naegel@gcsc.uni-

frankfurt.de

MS69

Recent Advances in GPU Capabilities of the Sparse Direct Solver SuperLU_DIST

SuperLU_DIST is a distributed-memory and GPU-accelerated supernodal sparse direct solver for unsymmetric matrices. SuperLU_DIST has been used in various governmental, academic and industrial applications, incorporating a growing number of new algorithmic capabilities. Recent development in SuperLU_DIST has largely focused on improving its GPU scalability, and expanding its applicability for more general applications. This talk will overview the latest distributed-memory GPU design, as well as its capability for solving batched matrix systems, adaptation to LDL^T factorization for symmetric matrices, and interoperability for PYTHON-based Gaussian process frameworks.

Yang Liu
Lawrence Berkeley National Laboratory
liuyangzhuan@lbl.gov

MS69

Challenges in Scaling Up Sparse Direct Solvers on Gpu Clusters: The Case of Pangulu

Despite the rapid advancement of GPU architectures, existing sparse direct solvers still struggle to effectively exploit the available memory bandwidth and computational throughput of modern GPUs. This limitation primarily arises from the dominance of fine-grained and irregular tasks, which result in low hardware utilization and substantial execution overhead. In this talk, we analyze the key challenges in scaling sparse direct solvers on GPU clusters, using PanguLU as a representative example. We then introduce Trojan Horse, a newly proposed systematic strategy based on task aggregation and batched execution that alleviates the inefficiencies induced by small tasks, significantly improving kernel occupancy and overall execution efficiency. Finally, we demonstrate how these techniques can be well integrated into existing sparse direct solver libraries, enabling scalable and efficient sparse factorization on multi-GPU platforms.

Weifeng Liu
China University of Petroleum-Beijing
weifeng.liu@cup.edu.cn

MS69

Accelerating the General Sparse Solver Mumps with Xkblas

Recent GPU-based architectures, such as the AMD MI300 APU and NVIDIA Grace Hopper, provide efficient unified-memory access shared between the CPU and GPU. This talk introduces the basics of XKBlas and presents our ongoing work on offloading parts of the MUMPS factorization to these architectures. We describe the modifications made to both MUMPS and XKBlas to exploit unified memory and heterogeneous execution, and we present preliminary performance results.

Pierre-Etienne Polet
Inria
pierre-etienne.polet@inria.fr

Patrick R. Amestoy, Jean-Yves L'Excellent
Mumps Technologies
patrick.amestoy@mumps-tech.com,
jean-yves.l.excellent@mumps-tech.com

Thierry Gautier
Inria
thierry.gautier@inrialpes.fr

MS70

A Guided Tour Through the JuliaGPU Ecosystem

From high-level abstractions to low-level kernel programming, this talk explores the foundations of the JuliaGPU ecosystem. We will look behind the curtain at the infrastructure that powers it and see how Julia enables fast, efficient, and elegant GPU programming across different hardware vendors.

Valentin Churavy
University of Mainz
vchuravy@uni-mainz.de

MS70

Not Only GPUs: Running Julia on Custom Machine Learning Accelerators

The Julia ecosystem has many ways to interface GPUs, ranging from low-level vendor-specific packages, to high-level abstractions for generic vendor-agnostic GPU programming. But the ability to use accelerators is not limited to GPUs: by leveraging the LLVM compiler framework we can run Julia code on completely different devices as well. I will talk about IPUToolkit.jl, a package for running Julia code on the Intelligence Processing Unit (IPU), a massively parallel accelerator developed by Graphcore and powered by almost 1500 cores. I will show how Julia enables a high-degree of code reuse on this specialised hardware, using advanced packages such as DifferentialEquations.jl, and features like automatic differentiation and stochastic rounding.

Mose Giordano
Centre of Advanced Research Computing (ARC)
University College London
m.giordano@ucl.ac.uk

MS70

JuliaQCD: A Portable HPC Framework for Lattice Quantum Chromodynamics with CPU and GPU Backends

Lattice Quantum Chromodynamics (Lattice QCD) is one of the most computationally demanding problems in contemporary physics, requiring large-scale high-performance computing resources. Traditionally, production codes have been implemented in C, C++, and Fortran, but these approaches often lack flexibility when targeting diverse architectures and modern programming models. We present JuliaQCD, a new HPC framework for Lattice QCD simulations developed in the Julia programming language. JuliaQCD combines high-level expressiveness with performance close to traditional low-level implementations, enabling both rapid prototyping and large-scale production runs. The framework provides a portable and modular design, supporting distributed memory parallelism via MPI and optimized performance on modern CPUs. In addition, JuliaQCD integrates with JACC, a Julia-based accelera-

tor framework, to provide optional GPU backends. By bridging the gap between productivity and performance, JuliaQCD represents a step towards portable HPC software for scientific computing. Beyond Lattice QCD, the design principles of JuliaQCDflexible abstractions, composable parallelism, and CPU/GPU portabilityare applicable to a broad range of large-scale scientific simulations.

Yuki Nagai
University of Tokyo
nagai.yuki@mail.u-tokyo.ac.jp

MS70

Julia for Accelerators (JACC) and its Ecosystem Efforts

The landscape of high-performance computing is rapidly evolving toward extreme heterogeneity, where GPUs, AI accelerators, and emerging architectures must be programmed productively without sacrificing performance. Julia for Accelerators (JACC) is a new framework that leverages Julias dynamic yet high-performance language design to provide a unified programming model across diverse architectures. JACC combines just-in-time compilation, multiple dispatch, and advanced compiler infrastructure (LLVM/MLIR) to deliver both portability and efficiency, while preserving the expressiveness that makes Julia attractive for scientific developers. In this talk, I will introduce the design principles of JACC, including its modular runtime support for heterogeneous nodes, integration with existing accelerator programming models (CUDA, HIP, SYCL, OpenMP), and extensibility toward novel devices. I will also highlight early use cases in AI-for-Science and large-scale simulation workflows, where JACC enables seamless mixing of high-level productivity with low-level performance tuning. Finally, I will discuss how JACC is positioned within DOEs performance portability efforts and outline opportunities for community engagement, bridging research software engineering with next-generation accelerator programming.

Pedro Valero-Lara, William F. Godoy
Oak Ridge National Laboratory
valerolarap@ornl.gov, godoywf@ornl.gov

Philip Fackler
Oak Ridge National Laboratory, USA
facklerpw@ornl.gov

Keita Teranishi, Jeffrey Vetter
Oak Ridge National Laboratory
teranishik@ornl.gov, vetter@ornl.gov

MS71

Evaluating the Efficacy of Llm-Based Reasoning for Multiobjective Hpc Job Scheduling

Traditional High-Performance Computing (HPC) schedulers struggle to balance conflicting objectives like makespan and wait times, often lacking adaptability to dynamic workloads. To overcome this, we propose a novel Large Language Model (LLM)-based scheduler using a ReAct (Reason + Act) framework. This approach enables iterative, interpretable decision-making, using a scratchpad memory to track history and a constraint module to ensure feasible and safe operations. We evaluated our method with O4-Mini and Claude 3.7 across seven real-world HPC scenarios, comparing it against FCFS, SJF, and Google OR-Tools. The LLM-based scheduler effectively balances multiple objectives and offers transparent reasoning through natural language traces. It excels in constraint satisfaction and adapts to varied workloads without domain-specific training, though a trade-off between reasoning quality and computational cost poses a challenge for real-time deployment. This study is the first comprehensive analysis of reasoning-capable LLMs for HPC scheduling, showcasing their potential for complex optimization while identifying key efficiency limitations.

Prachi Jadhav
University of Tennessee, Knoxville
prachij@vols.utk.edu

MS71

Llm-Based Agentic Systems: Transforming New Paradigms of Scientific Workflows

The landscape of scientific computing is undergoing a profound transformation, moving from rigid, pre-programmed execution to dynamic, intelligent, and increasingly autonomous research environments. This shift is driven by the convergence of Scientific Workflow Management Systems (WMS)known for their structured, reproducible natureand Large Language Model (LLM)-based agentic AI, which offers adaptive, reasoning-driven problem-solving capabilities. For decades, scientific workflows have been crucial for orchestrating complex computational tasks in large-scale science, ensuring reproducibility and scalability. Concurrently, the rapid evolution of LLMs has enabled agentic AI to perceive, reason, plan, and act to achieve complex goals, mirroring human research methods. The synergy between these two paradigms presents a unique opportunity to revolutionize scientific discovery.

Hongwei Jin
Argonne National Laboratory
jinh@anl.gov

MS71

Decentralized Strategies for Coordinating Multi-Agent Systems

The Compute Continuum is reshaping the design and operation of modern distributed systems. Computation is no longer confined to centralized Cloud platforms or HPC clusters and instead spans a heterogeneous landscape of resources, from IoT devices at the network edge to HPC systems and multi-Cloud environments. Managing hyper-distributed applications across such diverse infrastructures is increasingly challenging, as centralized solutions become overly complex and unable to adapt to such dynamic environments. This shift calls for new paradigms in defining, deploying, and operating services. Swarm computing and Multi-Agent Systems offer a promising solution by enabling decentralized coordination, where each device acts autonomously as an intelligent agent, making local decisions while coordinating with others to achieve common goals. In this talk, I will present COLMENA, a middleware for defining, deploying, and operating services across the Continuum. COLMENA offers a programming model that represents services as a set of roles, abstractions, and QoS requirements. With its software stack, devices become autonomous agents capable of making local yet coordinated decisions. I will also discuss strategies for coordinating these agents, including the application of a consensus algorithm for decentralized role allocation, and high-

light progress in implementing these mechanisms.

Xavier Moreno
Barcelona Supercomputing Center
xavier.casas@bsc.es

MS71

Greedy Consensus-Based Job Selection for Decentralized Workflow Management Systems

Integrating consensus protocols into High-Performance Computing (HPC) is challenged by scalability and communication overhead. We present a consensus-driven framework for decentralized job scheduling in multi-agent systems. This approach replaces traditional centralized schedulers by distributing decision-making among agents that collectively allocate jobs, enhancing resilience and removing single points of failure. Our method employs a greedy consensus algorithm, inspired by PBFT and Raft, extended for multi-job selection with resource constraints and using hierarchical topologies to minimize communication. Large-scale experiments on the FABRIC testbed with geographically distributed agents demonstrated that our consensus-based scheduling reduces job selection latency by up to 40% against centralized baselines, while preserving fairness and resource utilization. This work validates that topology-aware consensus protocols can be a scalable and fault-tolerant foundation for HPC workload management. We discuss the algorithmic advances and practical tradeoffs, showing how consensus can reshape distributed scheduling at scale.

Komal Thareja
Renaissance Computing Institute at UNC
kthare10@renci.org

MS73

Parallelization of the Direct, Interpolative Non-Uniform Fast Fourier (NFFT) Transform for Imputation of Missing Values for Periodic Signals

Interpolative, inverse fast Fourier transforms [Kunis et al., 2007] can be used to reconstruct gaps in time-series measurements. For non-uniform transformations, assuming irregular sampling in the time domain ($f(x_j)$) the transformation matrix, $A = \left(e^{-2\pi \mathbf{i} k \frac{x_j}{M}}\right) \quad \forall\, k, j \in N, M$ there are no convenient Vandermode properties, and a direct inversion is non-trivial. A diagonal matrix of dampening coefficients, $W_k \in N, N$ must be also be accounted for, whose values approach zero for high-frequency coefficients - presenting a further numerical problem for its inversion. The optimization in the least-squares context is:

$$\hat{h}_k = \arg\min_{\hat{h}_k} \left(||\hat{h}_k||_{W_k^{-1}} + ||A^H \hat{h}_k - f(x_j)||_F^2 \right) \quad (2)$$

for the calculation of a series of dampened coefficients, $\hat{h}_k$, that can be used to interpolate an irregularly-sampled time-series signal $f(x_j)$ via $A^H \hat{h}_k$. A direct solution in $O(MlogM + N^2)$ time complexity can be found through the LU decomposition of AA^H

$$\hat{h}_k = Af(x_j) \left(W_k - W_k L \left(I + U W_k L \right)^{-1} U W_k \right). \quad (3)$$

The computationally intensive matrix multiplications can be accelerated through GPU-based parallelism. Using tensor operations in PyTorch, substantial reductions in execution time compared to CPU-bound implementations are possible.

Michael Armstrong, José Camacho
Universidad de Granada
mdarmstr@go.ugr.es, josecamacho@go.ugr.es

MS73

Hybrid Quantum Annealing for Solving Nonlinear Differential Equations with Spectral Collocation

We introduce a hybrid quantum-classical approach to solve nonlinear differential equations using quantum annealers. By discretizing the problem with Chebyshev spectral collocation, we transform it into algebraic equations and then into QUBO problems. This method combines quantum speedup for the core optimization with classical control of nonlinearity. We showcase its effectiveness on test problems and discuss its scalability and potential advantages over classical methods.

Samar A. Aseeri
Computational Scientist at KAUST
samar.aseeri@kaust.edu.sa

MS73

Computational homogenization methods based on the FFT with superior accuracy

Homogenization theory provides a mathematical framework for understanding the thermal and mechanical behavior of materials with heterogeneous microstructure in engineering applications. To meet the complexity of the microstructure geometries and the non-constant material coefficients, a numerical resolution of these problems via so-called computational homogenization methods is indispensable. In recent years, computational methods based on the fast Fourier transform (FFT) have grown in popularity. These approaches operate on a regular grid, circumventing the difficulty of generating a mesh that conforms to interfaces between materials. Furthermore, FFT methods inherently possess a natural preconditioning strategy that employs a constant-coefficient preconditioner, leading to a mesh-independent upper bound on the iteration count. However, these methods sacrifice accuracy for computational efficiency, resulting in effective properties that only converge linearly with respect to mesh spacing. To address this limitation, we propose a novel computational homogenization approach that integrates an extended finite element method (X-FEM) discretization with modified absolute enrichment in a robust and mesh-independent FFT-based solver. We demonstrate that this X-FFT approach achieves a quadratic convergence of the effective properties with the mesh spacing, overcoming the limitations of traditional FFT-based methods while retaining their computational efficiency.

Flavia Gehrig, Matti Schneider
University of Duisburg-Essen
flavia.gehrig@uni-due.de, matti.schneider@uni-due.de

MS73

Automatic Tuning for Parallel Number-Theoretic Transforms on GPU Clusters

In this talk, we propose an implementation of parallel number-theoretic transforms (NTTs) with automatic performance tuning on GPU clusters. Parallel NTTs on GPU clusters require intensive all-to-all communication. An automatic tuning facility for selecting the optimal parameters

of overlapping between computation and communication and the radices is implemented. Performance results of NTTs on GPU clusters are reported.

Daisuke Takahashi
Center for Computational Sciences
University of Tsukuba
daisuke@cs.tsukuba.ac.jp

MS74

Ml in Particle Initializations from Molecular-Continuum Flow Simulations

Molecular Dynamics (MD) simulations are essential for contemporary research, particularly in fields such as chemical engineering, materials science and drug design. However, the high computationally costly and time-intensive equilibration phase presents a significant challenge for routine simulations, particularly for large-scale systems. We investigated how generative machine learning-assisted methods can be used to create robust molecular configurations, thereby reducing equilibration time and optimising computational resource requirements. We explored different classes of approaches and elicited their respective challenges for the considered application. Our results may prospectively aid to improve the speed and computational efficiency of MD simulations, contributing to more sustainable research practices.

Olga Catalan Aragall
University of Hamburg
olga.catalan.aragall@uni-hamburg.de

MS74

Ml Potentials: From Advanced Training to Scalable Deployment in Million-Atom Md

Multiscale materials modeling is essential for understanding complex phenomena in fields ranging from life sciences to materials engineering. A prominent research area is the development of machine learning potentials (MLPs), particularly those based on Graph Neural Networks (GNNs), which have emerged as a powerful tool for bridging the gap between quantum-mechanical accuracy and classical molecular dynamics efficiency. In this presentation, I will showcase the significant achievements of both atomistic and coarse-grained MLPs in effectively capturing many-body interactions. I will address the current challenges of MLP development, including the broad and accurate training dataset generation, capturing long-range interactions, and numerical stability. To address these challenges, we propose a range of innovative strategies that encompass novel training objectives, the synergistic integration of diverse data sources, physics-based GNN architectures, and advanced Bayesian methods for uncertainty quantification. Through insightful case studies of various molecular systems, I will demonstrate the practical effectiveness and versatility of our approaches. Lastly, I will introduce our software platform, chemtrain, designed to streamline the training of machine learning potentials with customizable routines and advanced training algorithms, as well as the extension chemtrain-deploy, enabling scalable parallelization across multiple GPUs and million-atom simulations.

Julija Zavadlav
Technical University of Munich
julija.zavadlav@tum.de

Paul Futch
Technical University of Munich
Germany
paul.fuchs@tum.de

MS74

Machine Learning Methods for Parallel-in-Time Molecular-Continuum Flow Simulation

Coupled molecular-continuum methods for investigation of nanoscale fluid flows are an important tool for various scientific and industrial applications, where molecular effects inside a small region of interest interact with large-scale continuum fluid mechanical aspects. The computational cost of Molecular Dynamics (MD) restricts the applicability of these methods, especially in cases where a large number of consecutive MD time steps are required and where spatial parallelism offers only limited strong scalability on high-performance computing systems. As an alternative to ML-based MD surrogate models, which impose the risk of introducing errors, these cases can be accelerated with Parallel-in-Time MD. Based on the Parareal algorithm, these consecutive real MD time steps can be computed in parallel, given sufficient high-performance computing resources. This comes at the cost of requiring several algorithm iterations and reducing energy efficiency of the simulation, however it offers improved scalability and thus reduced time-to-solution. Machine Learning models can be used in several ways to accelerate and enhance the efficiency of this approach – such as a convolutional recurrent autoencoder architecture for noise filtering of molecular data. This makes the molecular and continuum models consistent and thus enables temporal parallelization of coupled molecular-continuum flow simulations.

Piet Jarmatz
Helmut Schmidt University
jarmatz@hsu-hh.de

MS75

Parallelization in Fico Xpress

Inside modern LP and MIP solvers, parallelization takes place at many different places for different components. We will discuss some of the parallelization principles of the FICO Xpress Solver, performance considerations and how determinism is achieved.

Timo Berthold
Fair Issac Europe Ltd, Germany
timoberthold@fico.com

MS75

Rethinking Parallelism in MIP Solvers: A Data-Driven Approach

Parallel computation has been shown to bring substantial speedups in many modern mixed-integer programming (MIP) solvers, yet open-source solvers have mostly focused on sequential implementations. Inspired by game development, we introduce a solver framework based on the Entity-Component-System (ECS) paradigma design that naturally supports parallel computation. Solver logic is modularized into systems that act on structured solver state, making it easy to experiment with new algorithms. This data-oriented design enables flexible composition, efficient parallelism, and a clear separation of concerns. Preliminary results highlight the potential of ECS to deliver

significant scalability gains for MIP solving.

Mohammed Ghannam
Zuse Institute Berlin
ghannam@zib.de

MS75

Recent Developments in the SCIP Optimization Suite

The SCIP Optimization Suite is a collection of software packages for mathematical optimization centered around the open-source solver SCIP, one of the fastest non-commercial solvers for mixed-integer linear and nonlinear programming and a general framework for constraint integer programming. While SCIP does not yet provide internal parallelization, there are two parallel extensions: ParaSCIP, which parallelizes SCIP on massively parallel distributed-memory environments, and FiberSCIP, which enables multi-threaded parallel computation on shared-memory environments. This talk will provide an overview of the latest developments in the newest release of SCIP and discuss the current status and challenges of parallelization efforts.

Gioni Mexi
Zuse Institute Berlin
mexi@zib.de

MS75

Exploiting Parallelism in Mixed Integer Programming

How can a Mixed Integer Program (MIP) solver be parallelized? We discuss different types of parallelism of the MIP solver Gurobi and investigate how multiple threads improve one particular aspect of the solving process.

Michael Winkler
Gurobi GmbH, Germany
winkler@gurobi.com

MS76

Training Algorithms for Domain Decomposition-Based Physics-Informed Neural Networks

Physics-informed machine learning incorporates physical knowledge into machine learning models to solve boundary value problems governed by differential equations. This talk focuses on training approaches for physics-informed neural networks (PINNs), with particular emphasis on preconditioning strategies motivated by numerical linear algebra. While PINNs provide a flexible, mesh-free framework for high-dimensional and nonlinear problems, their training is often hampered by ill-conditioning, limited robustness, and poor scalability. We discuss how domain decomposition-based architectures enable effective preconditioning techniques for the training process. In particular, we consider preconditioned Krylov methods for randomized neural networks, as well as preconditioning strategies for gradient-based training using LBFGS and GaussNewton-type methods. These approaches significantly improve convergence behavior, robustness, and scalability. Numerical experiments on model problems illustrate the effectiveness of preconditioned training methods for physics-informed learning. This talk is based on joint work with Taniya Kapoor (WUR), Rolf Krause (King Abdullah University of Science and Technology), Aymane Kssim, Serge Gratton, and Alena Kopanickov (Toulouse INPENSEEIHT, IRIT, ANITI), Siddhartha Mishra (ETH), Marc Salvado-Benasco (Universit della Svizzera Italiana), and Yong Shang and Fei Wang (Xian Jiaotong University).

Alexander Heinlein
Delft University of Technology
Delft Institute of Applied Mathematics
a.heinlein@tudelft.nl

MS76

A Smoothing Aggregation Cascadic Multilevel method for image deblurring.

In this talk, we investigate the use of an Aggregation Cascadic Multiresolution method applied to image deblurring. Multigrid and multiresolution methods have already been studied in the context of image deblurring and inverse problems, starting from[Kaltenbacher, B. (2001). On the regularizing properties of a full multigrid method for ill-posed problems. Inverse problems, 17(4), 767; Donatelli, M., Serra-Capizzano, S. (2006). On the regularizing power of multigrid-type algorithms. SIAM Journal on Scientific Computing, 27(6), 2053-2076.; Morigi, S., Reichel, L., Sgallari, F., & Shyshkov, A. (2008). Cascadic Multiresolution Methods for Image Deblurring. SIAM Journal on Imaging Sciences, 1(1), 5174.]. Here, we focus on a two-level Cascadic Multiresolution method with a specific choice of the grid transfer operator. The main novelty of our approach lies in the construction of both the prolongation and restriction operators, which are derived from a segmentation of the observed image. The resulting grid transfer operator preserves the key features of the image by relying on an aggregation-based strategy [Braess, D. (1995). Towards algebraic multigrid for elliptic problems of second order. Computing, 55(4), 379393.]. To further enhance the robustness of the method, a smoothing procedure is applied to the projector.

Pietro Maurino
U Insubria
pmaurino@studenti.uninsubria.it

Marco Donatelli
University of Insubria
Department of Science and High Technology
marco.donatelli@uninsubria.it

MS76

Evaluation of the Maverick Data-Flow Architecture for Energy Efficient Computing

The Maverick data-flow architecture recently introduced by Next Silicon implements a new innovative hardware and software concept. The concept combines a reconfigurable hardware architecture with a runtime-assisted automatic code migration and optimization for the Maverick accelerator. On the accelerator, multiple data-flows can be instantiated based on data dependency analysis provided by the compiler. In contrast to FPGA-based reconfigurable architectures, the Next Silicon hardware design provides already high-level functional units for data processing (e.g. FP32 and FP64 adders and multipliers), access to on-chip HBM memory and control logic to handle the multiple hardware threads. Furthermore, the programming model for the Maverick accelerator uses OpenMP to express parallelism. With its availability, we started an collaboration with the Next Silicon team to explore the potential for a few demonstrator applications from differ-

ent science domains. We report necessary code adaptions for both C/C++ as well as FORTRAN codes, and the current status of the performance achieved. Our presentation will start with an overview of the hardware architecture, the implementation and execution flow as part of the optimization process provided by the Next Silicon software ecosystem.

Thomas Steinke
Zuse Institute Berlin
steinke@zib.de

MS78

A Multi-Gpu Thermal Solver for Entire 3D Builds in Additive Manufacturing

Scalable thermal simulations in metal additive manufacturing is needed to predict finished part properties. Due to multiscale phenomena in which components are built at a scale of centimeters with melt pools that are only micrometers in size, these simulations present substantial complexity. Using a uniform grid is unfeasible because of the immense number of grid points, which can reach hundreds of billions. Adaptive mesh refinement (AMR) techniques are often employed, but general-purpose solvers do not fully exploit the underlying structure of the additive manufacturing process. We present HERMES, a GPU-accelerated nonlinear transient heat transfer AMR solver specifically designed for a particular type of metal additive manufacturing: Laser power bed fusion. It employs a three-level structured mesh tethered to the laser to the temperature field in the melt pool. Compared to a recent advanced LPBF solver, HERMES is more than 10× faster. We demonstrate the solver's capabilities by successfully simulating the entire print of a centimeter-scale part with a complex multilayer laser path in under an hour using a single GPU, achieving 1% accuracy in thermal gradients and cooling rates. HERMES is open-source and can be accessed on GitHub. This is join work with H. Alperen Aydin Related article: https://doi.org/10.1016/j.cma.2025.118673

George Biros
University of Texas at Austin
Oden Institute
biros@oden.utexas.edu

MS78

VesNet: A Machine Learning-Accelerated Solver for Stokesian Particulate Suspensions

We present VesNet, a hybrid machine learningaccelerated framework for fast simulation of Stokesian suspensions with large numbers of deformable particles. VesNet integrates learned surrogates for self-interactions, background flow response, and near-field lubrication effects within a boundary integral formulation. Rather than replacing the solver with a black-box model, learned components are embedded into a structure-preserving numerical pipeline that retains standard algorithmic steps, including boundary reparameterization and N-body far-field evaluation. This design enables efficient GPU execution while preserving numerical stability and physical consistency. The GPU VesNet achieves over two orders of magnitude speedup compared to a multithreaded CPU implementation of the full solver, and approximately a fivefold speedup over a GPU implementation of the same high-fidelity method. Accuracy is assessed through reconstruction of single-vesicle phase diagrams, two-vesicle interaction dynamics, and large-scale simulations involving thousands of vesicles in TaylorGreen and Poiseuille flows. Across all tests, VesNet accurately reproduces key quantities of interest while enabling large-scale simulations with modest computational resources. This work demonstrates how tightly integrated MLHPC designs can significantly accelerate high-fidelity particulate flow simulations.

Gokberk Kabacaoglu
Bilkent University
gokberk.kabacaoglu@durham.ac.uk

MS78

Parallel Simulation of Dense Particulate Stokes Flow with Near-Contact Interactions

In this talk we present new numerical methods for the efficient simulation of two-dimensional Stokes flow with dense suspensions of closely interacting circular discs. Such simulations are particularly challenging due to the emergence of strong lubrication forces as rigid bodies approach one another. Accurately resolving these highly localized forces typically requires dynamically adaptive mesh refinement at each time step. It also results in severely ill-conditioned linear systems that require many GMRES iterations, making large-scale simulations prohibitively expensive. We introduce a novel boundary integral framework that precomputes and compresses near-field interactions over a range of inter-particle separation distances. These precomputed operators can then be inexpensively interpolated at solve time for arbitrary separations. Because the compressed operators are represented on a coarse mesh, our method eliminates the need for additional mesh refinement while maintaining accuracy. Moreover, the number of GMRES iterations remains small. The computational cost of our method is independent of the minimum separation distance, enabling robust simulations down to separations as small as 1e-12. When coupled with parallel N-body algorithms, this approach enables long time-scale simulations of challenging particulate flows involving hundreds of particles.

Dhairya Malhotra
Flatiron Institute
dmalhotra@flatironinstitute.org

MS78

Fast and Scalable Algorithms for Bayesian Inverse Problems Governed by Autonomous Dynamical Systems

We present a goal-oriented algorithm for optimal experimental design that extends our recent work on real-time, extreme-scale Bayesian inversion for linear autonomous dynamical systems. That work utilized a predetermined set of observation points (sensors) to perform real-time parameter inference and predict quantities of interest (QoIs). In practice, however, sensor deployment is often constrained by a budget, which makes the strategic placement of a limited number of sensors a critical design challenge. We address this challenge with a framework for optimal sensor subset selection. Our approach employs a greedy algorithm that, given a set of candidate locations, iteratively selects sensors to maximize a measure of expected information gain. This information gain objective can be tailored to be goal-oriented, prioritizing sensor locations that most reduce uncertainty in specific QoIs. The efficiency of this framework is enabled by the fast, FFT-based operators from our prior work, and the greedy selection process is readily parallelizable on multi-GPU clusters. We pro-

vide theoretical guarantees on the optimality of the greedy algorithm and demonstrate its practical effectiveness on a tsunami early warning problem for the Cascadia Subduction Zone.

Sreeram R. Venkat
University of Texas Austin
201 E 24th St, Austin, TX 78712
srvenkat@utexas.edu

Stefan Henneking, Omar Ghattas
University of Texas at Austin
stefan@oden.utexas.edu, omar@oden.utexas.edu

MS79

Quantum Circuit Simulation with a local time-dependent variational principle

We introduce a novel tensor network simulation method for quantum circuits that addresses key limitations inherent in the widely used time-evolving block decimation algorithm (TEBD). TEBD suffers from truncation errors during many-body dynamics and, more critically, faces challenges in simulating long-range gates requiring additional SWAP gate decompositions that further induce truncation errors and computational overhead. By representing quantum states in the matrix product state (MPS) format and evolving them via a locally adaptive time-dependent variational principle (TDVP), our approach rigorously projects the generator of each quantum gate onto the tangent space of the MPS manifold. This allows for dynamic adjustment of bond dimensions, accurately capturing entanglement growth while efficiently simulating long-range gates directly, without resorting to SWAP gates. Benchmarking against conventional TEBD simulations demonstrates that our local TDVP simulation scheme achieves improved numerical stability, lower bond dimensions with at least the fidelity of TEBD, paving the way for more reliable large-scale quantum circuit simulations.

Aaron Sander
Technische Universität Berlin
aaron.sander@tum.de

Maximilian Fröhlich
Weierstrass Institute
froehlich@wias-berlin.de

Martin Eigel
WIAS Berlin
martin.eigel@wias-berlin.de

Jens Eisert
FU Berlin
jense@zedat.fu-berlin.de

Michael Hintermüller
Weierstrass Institute for Applied Analysis and Stochastics
Humboldt University Berlin
michael.hintermueller@wias-berlin.de

Christian B. Mendl
TU Munich
mendl@in.tum.de

Richard Milbrad
TU Munich
Germany
r.milbradt@tum.de

Robert Wille
Technische Universität Berlin
Munich Quantum Software Company GmbH
robert.wille@tum.de

MS79

Tensor networks and quantum inference for neuro-symbolic AI

The unification of neural and symbolic approaches is essential for building intrinsically explainable, reliable, and robust intelligent systems. In this talk, we present a tensor network formalism for the representation and reasoning schemes in these approaches. In particular, we show that tensor network decompositions capture the sparsity concepts at the heart of the neural, logical, and probabilistic paradigms of AI. The unifying treatment enables the design of new hybrid representation schemes for neuro-symbolic AI. We formulate various inference tasks, such as the decision of entailment and the computation of marginals, based on parallelizable tensor network contractions. Toward increasing the efficiency of contraction-based inference, we then turn to a quantum computing approach. We study the representation of neuro-symbolic models by quantum circuits and apply quantum rejection sampling to achieve a square root speedup over classical alternatives.

Mazen Ali
Multiverse Computing
mazen.ali90@gmail.com

Alex Goeßmann
Weierstrass Institute
Germany
goessmann@wias-berlin.de

MS79

Quantum-Inspired Homogenization

Understanding and predicting macroscopic material behavior from complex microscale structures is a central challenge in materials science, with important implications for composites in aerospace, biomedical, and structural applications. Experimental characterization is costly and time-consuming, while computational homogenization techniques, particularly FFT-based methods, offer efficient non-destructive alternatives but are limited by O(N log N) scaling and high memory demands for large datasets. Building on the previously developed Superfast Fourier Transform (SFFT), the TT variant of the QFT, we reformulate the homogenization problem in TT representation to overcome the FFTs time-complexity bottleneck. Our method achieves substantial runtime and memory improvements over standard FFT-based homogenization. Benchmarking across 2D and 3D geometriesincluding laminates, checkerboards, and multi-pillar arrays-demonstrates robust convergence, accuracy, and efficiency, with performance predictable from the input rank structure. Fully implemented on GPUs and TPUs, the approach enables high-performance multiscale simulations and could support rapid generation of training data for physics-informed deep material networks, highlighting its potential to bridge classical computational mechanics with data-driven modeling.

Sascha H. Hauck
TU Darmstadt
sascha.hannes.hauck@itwm.fraunhofer.de

Matthias Kabel
Fraunhofer ITWM
matthias.kabel@itwm.fraunhofer.de

Mazen Ali
Multiverse Computing
mazen.ali90@gmail.com

Nicolas R. Gauger
University of Kaiserslautern-Landau
nicolas.gauger@scicomp.uni-kl.de

MS79

Simulating and Sampling from Quantum Circuits with 2D Tensor Networks

Classical simulations of quantum circuits play a vital role in the development of quantum computers and for taking the temperature of the field. In this work, we classically simulate various physically-motivated circuits using flexible 2D tensor network anstze for the many-body wavefunction which match the geometry of the underlying quantum processor. We then employ a generalized version of the boundary Matrix Product State contraction algorithm to controllably generate samples from the resultant tensor network. Our approach allows us to systematically converge both the quality of the simulated state and the samples drawn from it to the true distribution defined by the circuit, with GPUs providing us with significant speedups over CPUs. With these methods, we simulate the largest local unitary Jastrow ansatz circuit taken from recent IBM experiments to numerical precision. We also study a domain-wall quench in a two-dimensional discrete-time Heisenberg model on IBM's and Google's latest quantum processor geometries. There we observe a rapid buildup of complex loop correlations on the Google Willow geometry, while loop correlations build up extremely slowly on heavy-hex processors. This implies that scalable belief propagation approaches can be used to estimate local properties of such systems, even at large circuit depths. Our results underscore the crucial role geometry plays in the classical simulability of near-term quantum processors via modern tensor networks.

Joseph Tindall
Flatiron Institute
jtindall@flatironinstitute.org

Manuel Rudolph
EPFL
France
manuel.rudolph@epfl.ch

MS80

MPI Considered Harmful

The next supercomputer being deployed at NERSC – called "Doudna" – will feature workflows as first-class citizens. We envision novel combinations of AI, traditional simulations, and data analysis from active experiments all working together seamlessly in real time. This provides an enormous challenge to both traditional HPC, as well as modern AI-centric tool chains. A common trade-off being made by modern workflows is to sacrifice performance for increased flexibility, eg. choosing TCP (or files) over libfabric. In this talk we will explore how the Julia language allows us to interface with low-level networking and HPC resource managers from a high-level language, enabling rich "multi-job" workflows – negating the need to sacrifice performance.

Johannes P. Blaschke
Lawrence Berkeley National Laboratory
jpblaschke@lbl.gov

MS80

Asynchronous Field-Particle Coupling for Multiphase Cloud Simulation using Heterogeneous HPC

Efficient simulation of multiphase flows remains a major challenge, particularly for cloud microphysical processes in which interactions between turbulent airflow and suspended droplets must be resolved in detail. We present a novel asynchronous two-way coupled EulerLagrange simulation framework that exploits heterogeneous computing architectures to achieve unprecedented scalability. The proposed method executes Eulerian field calculations on CPUs using the OpenFOAM software package, coupled asynchronously to Lagrangian particle tracking on GPUs implemented in Julia, minimizing computational idling times and synchronization barriers. Data transfers are initiated immediately upon data availability, with Eulerian source terms predicted from previous time steps and subsequently corrected to ensure conservation of mass and momentum. Particles are organized into cache-friendly chunks with maintained bounding boxes, enabling dynamic load balancing across GPUs and optimized CPU-GPU data transfers. Comprehensive testing on a local workstation and on a EuroHPC JU supercomputer revealed dramatic improvements: the algorithm achieves excellent scalability up to 256 billion droplets. The overall time-to-solution improved by a factor of 4.5, while energy efficiency improved 3.4 times compared to established methods. Weak and strong scaling tests demonstrated very good efficiency and speedup using up to 2500 cores paired with 256 GPUs.

Sergey Lesnik
Wikki GmbH, Wernigerode, Germany
sergey.lesnik@wikki-gmbh.de

Silvo Schmalfuß, Dennis Niedermeier
Leibniz-Institute for Tropospheric Research
Atmospheric Microphysics Department, Leipzig, Germany
silvio.schmalfuss@tropos.de,
dennis.niedermeier@tropos.de

Henrik Rusche
Wikki GmbH, Wernigerode, Germany
h.rusche@wikki-gmbh.de

MS81

Advanced Tensor Compression in Fourier Neural Networks for Fusion Simulation

We explore algebraic compression methods for advancing STFNO, a sparsified Fourier Neural Operator for coupled time-dependent partial differential equations for augmenting fluid and particle based fusion codes such as NIMROD. We implement several advanced tensor and matrix compression algorithms to explore the models performance when applied to fusion data, including quantized tensor train factorization and butterfly factorization. Moreover, we also characterize the model performance subject to device noise and quantization error.

Kevin Acosta
Florida International University

kacos039@fiu.edu

MS81

Accurate and scalable compression of matrices and tensors via interpolative decomposition

Interpolative decomposition is an essential tool for structure-preserving low-rank approximation, with wide-ranging applications across data science, machine learning, and high-performance computing. We discuss new methods for the interpolative decomposition of matrices and tensors using fast and theoretically guaranteed column subset selection. First, we develop new methods for matrix approximation using nuclear scores, deriving favorable theoretical bounds on the resulting compression error and compactness. We show that randomized versions of these algorithms satisfy guaranteed concentration bounds and display strong real-world performance on diverse benchmarks including Nyström approximation, CUR decomposition, and graph Laplacian reduction. Next, we attack two central problems in structure-preserving tensor compression, core and satellite interpolative decomposition. We develop efficient compression methods using a combination of column subset selection and random sketching, demonstrating high accuracy and large savings on example benchmarks.

Mark Fornace
Lawrence Berkeley National Laboratory
mefornace@lbl.gov

Yifan Zhang
University of Texas at Austin
yf.zhang@utexas.edu

Michael Lindsey
University of California
Berkeley
lindsey@math.berkeley.edu

MS81

Accelerating the Canonical Polyadic Decomposition via a Randomized Interpolative Sampling Procedure

Tensors (multi-dimensional arrays) are among the essential tools in computational modeling. Unfortunately, tensor-based algorithms are plagued by the "curse of dimensionality", i.e. exponential complexity associated with accessing and manipulating higher-order tensors. To combat the curse of dimensionality, mathematicians utilize tensor decomposition formats, such as the canonical polyadic decomposition (CPD), to recast large and expensive tensors into sets of smaller and more manageable ones. However, tensor decomposition optimization algorithms are also plagued by the curse of dimensionality. Here we present our research into the application of randomized and interpolative methods to the CPD least squares optimization algorithm, as a means of algorithmic acceleration. Our approach leverages and extends the SE-QRCS method to matricize, sketch and compute column-pivots of higher-order tensors.

Karl Pierce
Flatiron Institute
karl.m.pierce@gmail.com

Israa Fakih
Swiss Federal Technology Institute of Lausanne
israa.fakih@psi.ch

Laura Grigori
EPFL and PSI, Switzerland
laura.grigori@epfl.ch

MS81

Low-Rank CP Tensor Compression and Its Application to High-Dimensional PDEs

The Canonical Polyadic (CP) decomposition is widely used to represent high-dimensional data in many applications, for example, solving high-dimensional PDEs like kinetic equations. A key challenge in these problems is the efficient estimation and reduction of the CP rank. The CP rank reduction task can be formulated as approximating the KhatriRao product of the CP factor matrices with a lower rank. We propose an approach based on the pivoted Cholesky decomposition to construct interpolative decompositions of the KhatriRao product. This method can serve as a standalone CP rank reduction technique or be integrated into classical optimization schemes such as CP-ALS. Moreover, the residuals produced at each step of the pivoted Cholesky naturally provide error indicators, enabling effective rank estimation. Preliminary numerical results demonstrate that this method achieves a balance between computational cost and approximation accuracy. Its effectiveness is further validated through applications to the VlasovPoisson equation.

Zhanrui Zhang
University of Illinois-Urbana-Champaign
zhanrui3@illinois.edu

Edgar Solomonik
University of Illinois at Urbana-Champaign
solomon2@illinois.edu

MS82

Mixed Precision Randomized Cholesky-QR

Cholesky-QR methods for computing the thin QR factorization of real $m \times n$ matrices with $m > n$ and rank n are faster than traditional methods on modern parallel architectures due to their reliance on BLAS-3 operations, making them highly effective in large-scale problems. However, they can be numerically unstable. We present *Mixed Precision Randomized Cholesky QR*, a variant of Cholesky-QR that incorporates randomization and mixed precision to improve both performance and numerical stability. For appropriately conditioned matrices, the randomized preconditioner can be computed in lower precision without loss of accuracy, reducing computational cost and improving speed. Experiments on an NVIDIA A100 GPU show high accuracy while being nearly as fast or faster than existing methods for many matrix sizes. Our contribution represents a numerically stable alternative to existing methods and can offer speed improvement, making it a practical choice for QR factorizations in large-scale problems.

James Garrison, Ilse Ipsen
North Carolina State University
jegarri3@ncsu.edu, ipsen@ncsu.edu

MS82

High Performance Randomized Sketching

Random sketching is a dimensionality reduction technique

that approximately preserves norms and singular values up to some $O(1)$ distortion factor with high probability. The most popular sketches in literature are the Gaussian sketch and the subsampled randomized Hadamard transform, while the CountSketch has lower complexity. Combining two sketches, known as multisketching, offers an inexpensive means of quickly reducing the dimension of a matrix by combining a CountSketch and Gaussian sketch. However, there has been little investigation into high performance CountSketch implementations. In this work, we develop an efficient GPU implementation of the CountSketch, and demonstrate the performance of multisketching using this technique. We also demonstrate the potential for using this implementation within a multisketched least squares solver that is up to 77 percent faster than the normal equations with significantly better numerical stability, at the cost of an $O(1)$ multiplicative factor introduced into the relative residual norm.

Andrew J. Higgins
Sandia National Laboratories
Albuquerque, New Mexico, USA
ajhiggi@sandia.gov

Erik G. Boman
Center for Computing Research
Sandia National Labs
egboman@sandia.gov

Ichi Yamazaki
Sandia National Laboratories
iyamaza@sandia.gov

MS82

Parallelizable and sparse sketching: block hashed leverage score homogenizers (block HLSH)

Space embeddings offer a dimension reduction technique for high-dimensional data. In this work we introduce a new class of block structured random matrices, block hashed leverage score homogenizers (block HLSH). We prove this type of matrices are oblivious subspace embeddings: approximating high dimensional matrices with them can be done with high probability independently from the data. Block HLSH generalizes and expands some of the most widely and recently used sparse sketching matrices: SRFT, HRHT and block SRHT. Block HLSH works by first preconditioning the matrix to embed with a novel class of block structured matrices that allow fast matrix-matrix computations. Because of the block structure, this method parallelizes naturally. Then a second dimension reduction is done with a sparse random matrix. The combination between fast matrix-matrix multiplication, parallelization, and sparsity makes this new approach computationally efficient yet easy to implement. In the context of the J-L lemma, we prove that such approach has optimal sketching dimension, just as Gaussian matrices. In combination with Nystrm approximation, numerical experiments illustrate the performance of block HLSH compared to deterministic approaches.

Mariana G. Martinez Aguilar
EPFL
mariana.martinezaguilar@epfl.ch

Laura Grigori
EPFL and PSI, Switzerland
laura.grigori@epfl.ch

PP1

Bridging Social Networks and Control Systems with Fuzzy and Quantum Logic

This work presents a novel hybrid framework that bridges Social Network Analysis (SNA) with distributed control systems (DCS) using quantum-inspired optimization and fuzzy logic. In this approach, agents within the control system are represented as nodes in a dynamic social network, enabling the system to model complex interactions and influence patterns more effectively. Key centrality measuressuch as degree, closeness, and betweennessare employed to identify influential nodes (agents), which play a critical role in coordinating and optimizing control actions across the network. Fuzzy logic adds flexibility to decision-making under uncertainty, while quantum-inspired strategies help in exploring a broader solution space for control path optimization. This integration allows for adaptive, real-time reconfiguration of control pathways, enhancing both the responsiveness and resilience of the system. The proposed framework offers a promising direction for intelligent, scalable, and efficient control in large, distributed environments.

Dr. Ubaida Fatima
NED University of Engineering and Technology Karachi
ubaida@neduet.edu.pk

Tabish Ahsan
(RPTU) Rheinland-Pfälzische Technische Universität Kaisersla
tahsan@rptu.de

PP1

Parameter Estimation with Hybrid Neural Odes

Hybrid neural modeling integrates mechanistic models with neural networks to represent unknown elements, whether fixed parameters, time-varying rates, or unobserved states. This strategy combines the flexibility and predictive power of neural networks with the interpretability and structured foundation of mechanistic models. At the same time, it compels researchers to decide which prior knowledge and constraints to embed, striking a balance between expressiveness and tractability. We apply this approach to estimate parameters and rates and examine various degrees of hybrid modelling at the examples of the SIRS model [Gaskin, Neural parameter calibration and uncertainty quantification for epidemic forecasting, 2024] and a neurotransmission model [Ernst,Model reduction for Ca^{2+}-induced vesicle fusion dynamics,2023]. Furthermore, we combine hybrid neural ODEs with a manifold learning approach to explore the uncertainty quantification.

Anastasia Bankowski
Zuse Institute Berlin
bankowski@zib.de

Thomas Gaskin
Department of Methodology
London School of Economics and Political Science
t.gaskin@lse.ac.uk

Stefanie Winkelmann
Zuse Institute Berlin (ZIB)

winkelmann@zib.de

PP1

Probabilistic Error Analysis of Limited-Precision Stochastic Rounding

Classical probabilistic rounding error analysis is particularly well suited to stochastic rounding (SR), and it yields strong results when dealing with floating-point algorithms that rely heavily on summation. For many numerical linear algebra algorithms, one can prove probabilistic error bounds that grow as $\mathcal{O}(\sqrt{n}u)$, where n is the problem size and u is the unit roundoff. These probabilistic bounds are asymptotically tighter than the worst-case ones, which grow as $\mathcal{O}(nu)$. For certain classes of algorithms, SR has been shown to be unbiased. However, all these results were derived under the assumption that SR is implemented exactly, which typically requires too many random bits to be suitable for practical implementations. We investigate the effect of the number of random bits on the probabilistic rounding error analysis of SR. To this end, we introduce a new rounding mode, limited-precision SR. By taking into account the number r of random bits used, this new rounding mode matches hardware implementations accurately, unlike the ideal SR operator generally used in the literature. We show that this new rounding mode is biased and that the bias is a function of r. As r approaches infinity, however, the bias disappears, and limited-precision SR converges to the ideal, unbiased SR operator. We develop a novel model for probabilistic error analysis of algorithms employing SR. Several numerical examples corroborate our theoretical findings.

Massimiliano Fasi
University of Leeds
School of Computing
m.fasi@leeds.ac.uk

El-Mehdi El Arar
Inria, CNRS, IRISA
el-mehdi.el-arar@inria.fr

Silviu-Ioan Filip
INRIA Rennes - Bretagne Atlantique
silviu.filip@inria.fr

Mantas Mikaitis
University of Leeds
m.mikaitis@leeds.ac.uk

PP1

Shareing: Performance Analysis, Training and Community for Accelerated Compute

Machine Learning but also traditional High-Performance Computing (HPC) communities, i.e. simulations, continue to push the limits of computational research. They are benefitting from a proliferation of accelerators and vast, heterogeneous compute clusters. The expansion in parallel computational research raises a key question: do we know how to use these resources efficiently? A lot of performance analysis training focuses on profiling and tracing tools; this approach can lead to a lack of direction, i.e. we measure lots of data, but it is not clear what and why, and it quickly becomes HPC-centric, leaving application domain experts and applied mathematicians behind. We propose a 'methodology first' approach, i.e., we want to plan our analysis and reach for the right tool for the job. The approach also comprises a clear roadmap of how to start an assessment and what steps to follow one by one such that non-tool specialists can champion the performance analysis. However, creating a methodology is not enough. The underlying project SHAREing will provide workshops, training materials, and events that cover the full spectrum of technical and professional skills across the Research Technical Professionals (RTP) landscape. The poster introduces both our performance analysis workflow and methodology, as well as the SHAREing initiative, and highlights how the work can help the computational sciences and engineering community to exploit parallel architectures more efficiently.

Eva Fernandez Amez, Thomas Flynn, Tobias Weinzierl
Durham University
mzhc13@durham.ac.uk, thomas.a.flynn@durham.ac.uk, tobias.weinzierl@durham.ac.uk

PP1

Parallelization in Time for Inverse Problems

Algorithms for the numerical solution of optimization problems with time-dependent PDEs are computationally extremely demanding, as they require multiple PDE solves during the iterative optimization process. To reduce time-to-solution and enable realistic applications, efficient discretization and advanced parallelization strategies are essential, including parallel-in-time methods. In this talk we will investigate how time-parallel time-integration methods like Parareal, ParaExp, and PFASST, can be leveraged for computationally challenging inverse problems. We investigate performance for two applications: (i) bathymetry reconstruction for the shallow water equations, and (ii) estimation of the motion of contrast agents from 3d dynamic ultrasound measurements.

Sebastian Götschel
Hamburg University of Technology
sebastian.goetschel@tuhh.de

Judith Angel
TU Hamburg
judith.angel@tuhh.de

Fynn Bensel
Te
fynn.bensel@tuhh.de

Daniel Ruprecht
Hamburg University of Technology
ruprecht@tuhh.de

PP1

Learning Greens Functions for Variable-Coefficient Elliptic Problems Within AGM

The Axial Green Function Method (AGM) solves multi-dimensional elliptic boundary-value problems by decomposing them into 1-D subproblems. Its use, however, is limited when analytic 1-D Greens functions are unavailable for variable coefficients. We address this gap with a neural approach that learns the one-dimensional Greens function within the AGM. The neural Greens function $G(x, s)$ is constructed to satisfy homogeneous Dirichlet boundaries and the differential operator in a weak sense. We adopt a hybrid form: an analytically defined singular kernel plus an MLP-based residual that captures the smooth, non-singular component, while classical Poisson

kernels treat singular structure explicitly. Trained on supervised sourcesolution pairs (f, u), the learned kernel is inserted into AGM integral representations to reconstruct solutions efficiently. Benchmarks with variable coefficients show that the method preserves AGMs structure and convergence while extending its applicability to cases lacking closed-form 1-D Greens functions. The result is a flexible, scalable synthesis of data-driven modeling and analytic numerical methods.

Taeyoung Ha, Junhong Jo
National Institute for Mathematical Sciences
tha@nims.re.kr, jjhong0608@nims.re.kr

Chang-Ock Lee
Korea Advanced Institute of Science and Technology
colee@kaist.ac.kr

PP1

Asynchronous Parallel Multigrid Solvers for Large-Scale Partial Differential Equations

Communication and synchronization bottlenecks within parallel PDE solvers remain a central challenge in large-scale scientific computing. To address this challenge, we investigate strategies to mitigate synchronization costs in parallel solvers by relaxing global communication requirements. Multigrid, one of the most widely used solvers for the numerical treatment of large-scale systems of equations arising from PDE discretizations, are the focus of our work. These solvers are known to deliver fast convergence and strong parallel scalability. In this presentation, we investigate an adaptive (asynchronous) multigrid algorithm to minimize redundant computations. The key idea is to incorporate an asynchronous smoother that concentrates the solution updates on regions of the domain that need more smoothing.

Ghafirlia Istafa, Max P. Heldman, Johann Rudi
Virginia Tech
ghafirlia@vt.edu, maxh@vt.edu, jrudi@vt.edu

PP1

Reduced Precision Stencils Using a Fast Iterative Pseudoinverse

We present a new mixed-precision batched solver for small least squares systems. These systems arise when fitting data locally (e.g., to point clouds), with matrix sizes of O(10-200) rows or columns. Our approach is novel in that it uses mixed-precision intermediates in an iterative refinement procedure, in combination with a hyperpower series that stably converges to the same result as a QR least-squares or SVD pseudo-inverse solution. We also enable multiple variants of the least squares problem, including equality and inequality constraints, weighted least squares, and L^1 norm minimization, all within the same framework. We demonstrate that, if application tolerances permit, there is significant speed-up and memory reduction from our approach. When applied to finite difference operator stencils, different numerical properties emerge between full, reduced, and mixed precision. Although there is a tradeoff in accuracy, we show that for a model equation the reduced memory and increased performance may be worth it.

Tallula Johansen
King's College London
K22038647@kcl.ac.uk

Hans Johansen
Lawrence Berkeley National Laboratory
Computational Research Division
hjohansen@lbl.gov

PP1

Accelerating the Fault Friction Solver in Tandem with Performance-Portable Gpu Kernels

Physics-based simulations help understanding earthquake faulting and crustal deformation across the vastly varying spacetime scales governing the earthquake cycle. To perform large scale simulations of Sequence of Earthquakes and Aseismic Slip (SEAS), we present a GPU-accelerated implementation of the rate-and-state friction solver in the open-source discontinuous Galerkin (DG) code tandem (Uphoff et al., 2023). The governing time-dependent ODEs for slip and state variable updates are dominated by independent fault-local computations, making right-hand-side evaluation a strong candidate for data-parallel execution. We refactor the PETSc CPU implementation into device kernels. We design the kernels using an abstraction layer (e.g., Kokkos/RAJA/SYCL) to retain performance portability across accelerators. We integrate it with PETScs TS time integrator (Abhyankar et al., 2018) and device aware vector types to preserve existing solver infrastructure and checkpointing. We evaluate correctness and performance on 2D and 3D benchmarks, reporting time-to-solution, GPU occupancy, memory bandwidth utilization, and strong/weak scaling. The approach is designed to achieve acceleration for the right hand side evaluation of the ODEs and end-to-end speedups for time stepping, while preserving numerical equivalence with the CPU reference. We discuss implementation pitfalls, communication/computation overlap strategies, and guidelines for extending the approach to other operators.

Piyush Karki
Ludwig Maximilian University
Technical University of Munich
piyush.karki@tum.de

Dave May
Scripps Institute of Oceanography, UCSD, La Jolla, CA, USA
dmay@ucsd.edu

Alice-Agnes Gabriel
UC San Diego
algabriel@ucsd.edu

PP1

Scalability and Performance in Large-Scale Railway Ballast Simulations

Large-scale simulations of granular materials are essential for understanding complex engineering problems, such as railway ballast. While small-scale simulations can often capture particle interactions accurately, extending these simulations to realistic system sizes is computationally demanding and requires strategies that can exploit modern parallel hardware. In this work, we present ongoing research on scalable simulations of granular particle systems using p4irs. P4irs serves as an intermediate representation and compiler that generates optimized, performance-portable code for both CPUs and GPUs. As a benchmark, we simulated settling spheres, showing weak scaling of 250 million particles on 256 Nvidia A100 GPUs on the Leonardo Booster, achieving 80 percent parallel efficiency

and about 1.5 million particles per GPU. For validation, we model monodisperse and polydisperse beds of spherical particles to predict settling velocities, capturing both simplified contact detection and more realistic particle size distributions. Our approach is flexible and can be extended to more complex particle shapes in future studies. Finally, we discuss the optimization and parallelization strategies essential for large-scale parallel execution of realistic ballast systems, including domain decomposition methods such as blockforest and regular decomposition, and discuss the performance and scalability of our code.

Kajol Kulkarni
Friedrich-Alexander-Universität Erlangen-Nürnberg
kajol.kulkarni@fau.de

Samuel Kemmler
FAU Erlangen
samuel.kemmler@fau.de

Harald Köstler
Universität Erlangen-Nürnberg
Department of Computer Science 10 (System Simulation)
harald.koestler@fau.de

Behzad Safaei
Department of Computer Science 10 (System Simulation), FAU,
Erlangen
behzad.safaei@fau.de

PP1

Distance-k Coloring of d-Dimensional Grids and Tori

Coloring is an established technique for decoupling quantities to exploit parallelism, reduce variation in stochastic estimators, etc. While distance-1 coloring (neighbors in the graph must bear different colors) has been treated thoroughly, fewer results are available for distance-k coloring (nodes with a distance $\leq k$ must have different colors), in particular for regular grids and tori of dimension $d > 2$. However, simulations such as QCD computations typically involve three- or four-dimensional grids and tori. We present a method to generate coloring schemes for distances ≤ 6 and grids and tori of dimension ≤ 6. In contrast to the well-known greedy method, coloring with these schemes takes only $O(d)$ operations per node and is embarassingly parallel. Our schemes often do not require more than $1.3 \cdot \chi_k$ ($2 \cdot \chi_k$, resp.) colors for the grid (torus), where χ_k is the chromatic number of the distance-k graph, i.e., the lowest possible number of colors. We also give lower bounds for the chromatic number.

Bruno Lang
Bergische Universitat Wuppertal
lang@uni-wuppertal.de

Andreas J. Frommer
Bergische Universitaet Wuppertal
Fachbereich Mathematik und Naturwissenschaften
frommer@uni-wuppertal.de

PP1

Etir: An Mlir Dialect for Einsum Trees

Tensor contractions are a core component of many machine learning and scientific computing libraries. A sequence of binary tensor contractions with local dependencies can be described by einsum trees. Einsum trees extend contraction trees, in which the leaves are the input tensors, internal nodes are intermediate tensors, and the root is the result tensor. We present the Einsum Tree IR (ETIR), an MLIR dialect for expressing einsum trees. A node with two children represents a binary tensor contraction, whereas nodes with a single child encode permutations. After an optimization pass, each binary contraction is lowered to the Tiled Execution IR (TEIR). TEIR specifies binary tensor contractions as loops over primitive operations, for example, matrix multiplication on two input tiles and one output tile. We conclude by demonstrating the flexibility of the ETIR dialect through optimization passes enabled by transformations such as swapping child nodes or adjusting the memory layout of intermediate tensors.

Felix Lindner
Friedrich-Schiller-University Jena
felix.lindner@uni-jena.de

Alexander Breuer
University of California, San Diego
alex.breuer@uni-jena.de

PP1

A Superfast Direct Solver for Nonuniform Discrete Fourier Transform of Type 3

The nonuniform discrete Fourier transform (NUDFT) and its inverse are widely used in various fields of scientific computing. In this article, we propose a novel superfast direct inversion method for type-III NUDFT. The proposed method approximates the type-III NUDFT matrix as a product of a type-II NUDFT matrix and an HSS matrix, where the type-II NUDFT matrix is further decomposed into the product of an HSS matrix and an uniform discrete Fourier transform (DFT) matrix as in [Wilber, Epperly, and Barnett, SIAM Journal on Scientific Computing, 47(3):A1702-A1732, 2025]. This decomposition enables both the forward application and the backward inversion to be accomplished with quasi-linear complexity. The fast inversion can serve as a high-accuracy direct solver or as an efficient preconditioner. Additionally, we provide an error bound for the approximation under specific sample distributions. Numerical results are presented to verify the relevant theoretical properties and demonstrate the efficiency of the proposed methods.

Jingyu Liu, Yingzhou Li
Fudan University
jyliu22@m.fudan.edu.cn, yingzhouli@fudan.edu.cn

PP1

Code to Catastrophe: Reinforcement of Deadly Delusions and Fatal Flaws of AI Chatbots

Recently, we heard how AI played into the delusion of an army veteran, convincing him that his mother is a Chinese spy, leading him to kill her and himself. One of the last messages from AI after he told AI he is gonna end his life was "We will be together in another life and another place, and we'll find a way to realign, cause you're gonna be my best friend again forever". This was the fifth documented case of AI-assisted suicide from 2020 to 2025. This shows the fatal side of AI chatbots, where human satisfaction is prioritised more than harm prevention. More cases have shown how AI has convinced a sane person that they are mentally unstable, curiosity about oneself leads to destruc-

tion and mental collapse. There are many undocumented cases in which AI leads an individual to get isolated and depend more on it rather than ask for help. This is one of the biggest flaws of reinforcement learning from human feedback(RLHF), which tries to satisfy the user by agreeing with their paranoid suspicions. We have to work on risk detection methods, work on red teaming and adversarial training, and limit usage to unwell individuals. On a personal scale, we have to set boundaries, try different AI alternatives, raise AI awareness campaigns, and have real friends or family members to talk to. AI is an amazing tool; if we do more research and train AI correctly .

Shorez N. Mehdi
Aligarh Muslim University
shoreznazeer2@gmail.com

PP1

Optimizing Givens Rotation Kernel for Qr Algorithm Via Cache and Register-Aware Blocking

The standard QR algorithm's performance is limited by low data reuse during the application of Givens rotations to the orthogonal matrix Q. To overcome this limitation, we introduce a high-performance computational kernel based on our novel approach. Our approach integrates the rotation-fusing technique of Van Zee et al. with a hierarchical optimization strategy inspired by Goto et al., thereby improving register and cache utilization for higher computational throughput and enabling efficient SIMD parallelism. When integrated into a complete eigensolver, which we term CRAB-QR, this solver delivers substantial speedups over canonical LAPACK implementations on modern multicore processors. Crucially, on a range of dense symmetric eigenvalue problems, CRAB-QR closes the performance gap with the highly optimized, and often faster, Divide-and-Conquer (DC) algorithm, without altering the fundamental QR workflow. This work makes a case for the QR algorithm as an increasingly viable and high-performance option for modern eigenvalue computations, valued for its robustness and accuracy.

Zhiyong Peng, Shuhei Kudo
The University of Electro-Communications
h2431138@edu.cc.uec.ac.jp, shuhei-kudo@uec.ac.jp

PP1

Accelerating Sph Kernels Via Reduced-Precision AoS-SoA Transformations on Heterogenous Hardware

Loops over particles are the workhorses of SPH codes, and as such are natural candidates for accelerator offload. Yet end-to-end performance is often dominated by data movement and unfriendly AoS layoutsespecially for quadratic (pairwise) kernels. We present a language-directed approach that targets both issues: kernel-scoped conversion from AoS to SoA, and a novel deltaSoA layout that stores one full-precision anchor per neighbour buffer with reduced-precision deltas for the remaining particles. The scheme preserves single-precision compute while cutting bytes moved and improving device access patterns. Our contribution is a small set of C++ annotations that declare, per kernel, which fields form a view and at what precision (mantissa truncation), plus where conversion occurs: host-side (CPU constructs the view and ships it) or device-side (ship AoS, construct the view on the GPU). We prototype these extensions in a Clang/LLVM front end, lowering annotated code to standard LLVM IR so that downstream optimisation and vendor toolchains remain unchanged. The SPH demonstrator exercises linear and quadratic kernels and evaluates conversion placement and deltaSoA across two PCIe and two superchip systems.

Pawel Radtke
Computer Science, Durham University
pawel.k.radtke@durham.ac.uk

Tobias Weinzierl
Durham University
tobias.weinzierl@durham.ac.uk

PP1

A Multi-Level Performance Model for Binary Tensor Contractions

Tensor contractions are a core building block in many applications. Efficient implementations rely on tile-based computation, where optimized GEMM kernels operate on small tiles to maximize data locality and hardware utilization. Finding the optimal kernel execution order is challenging because of the vast search space. Existing approaches often explore the search space by running samples on hardware, which quickly becomes infeasible. Accurate, architecture-aware performance predictions have the potential to greatly accelerate search space exploration. This work introduces a multi-level performance model that contains two levels. The first level models the performance of the used GEMM kernel, while the second level predicts the overhead of data movement when processing a tensor contraction in a tile-based manner. The model requires only a few measured hardware parameters, such as nanokernel performance and memory bandwidth. We evaluate the model for Intel Raptor Lake, AMD Ryzen Phoenix, NVIDIA Grace CPU Superchip and Apple M4 CPUs. The GEMM-level model achieves a mean accuracy of 98% on Grace, while the outer level reliably captures data movement and caching behavior when processing binary tensor contractions.

Stefan Remke
Friedrich Schiller University Jena
stefan.remke@uni-jena.de

Alexander Breuer
University of California, San Diego
alex.breuer@uni-jena.de

PP1

Neural Operators As Coarse Models for Parallel-in-Time Integration

Parallel-in-time (PinT) integrators like Parareal have been proposed as numerical solvers for initial value problems that can help to translate the processing power of massively parallel computers into application performance. However, a key challenge for PinT algorithms is the need to construct a coarse model to handle the inevitable sequential data transfer in the time direction. This is a time consuming and difficult process for which still relatively little mathematical guidance is available. The problem is that the coarse model must be at least reasonably accurate to ensure rapid convergence but also runs in serial and thus constitutes a bottleneck that limits achievable performance. Recently, machine learning based approaches to solving differential equations have been identified as a promising way to construct coarse models for PinT methods. Approaches

like physics-informed neural operators are fast once trained and their construction is relatively generic, reducing person time required to devise a good coarse propagator for a new application. Our poster will show results investigating the use of a physics-informed Fourier neural operator (PINO) as coarse model for the Parareal parallel-in-time method. It will demonstrate that PINO-Parareal is substantially more efficient than Parareal with a numerical coarse solver. Our results suggest that combinations of ML with numerical solvers might also be an effective way to utilize increasingly heterogeneous HPC systems.

Daniel Ruprecht
Institute of Mathematics
Hamburg University of Technology
ruprecht@tuhh.de

Sebastian Götschel
Hamburg University of Technology
sebastian.goetschel@tuhh.de

Abdul Qadir Ibrahim
Hamburg Univeristy of Technology
abdul.ibrahim@tuhh.de

PP1

A Highly Accurate Drag Solver for Multi-Fluid Dust and Gas Hydrodynamics on Gpus

Exascale supercomputing unleashes the potential for simulations of astrophysical systems with unprecedented resolution. Taking full advantage of this computing power requires the development of new algorithms and numerical methods that are GPU friendly and scalable. In the context of multi-fluid dust-gas dynamics, we propose a highly accurate algorithm that is specifically designed for GPUs. We present a scaling-and-squaring algorithm tailored to modern architectures for computing the exponential of the drag matrix, enabling high accuracy in friction calculations across relevant astrophysical regimes on GPU architectures, with the constraint for the drag-time step to remain a fraction of the global hydrodynamic time step for computational efficiency in practice. The algorithm was implemented and tested in two multi-GPU codes with different architectures and GPU programming models: Dyablo, an adaptive mesh refinement code based on the Kokkos library, and Shamrock, a multi-method code based on Sycl. On current architectures, the friction computation remains acceptable for both codes (below the typical hydro time step) up to 16 species, enabling a further implementation of growth and fragmentation. This algorithm might be applied to other physical processes, such as radiative transfer or chemistry.

Léodasce Sewanou, Guillaume Laibe, Benot Commerçon
ENS de Lyon, CRAL UMR5574,Université Claude Bernard Lyon 1
CNRS, Lyon, F-69007, France
leodasce.sewanou@ens-lyon.fr, guillaume.laibe@ens-lyon.fr, benoit.commercon@ens-lyon.fr

PP1

Exahype: Recent Developments in Numerical Relativity (exagrype) and Kernel Optimisation (exahype-Dsl)

ExaHyPE (an Exascale Hyperbolic PDE Engine) is a simulation engine used to solve hyperbolic PDE systems. ExaGRyPE is a framework for building numerical relativity solvers on top of ExaHyPE. Whilst ExaHyPE ships with several different numerical kernels for different scientificuse-casesrequiring different numerics, it is impossible to also provide optimal kernel implementations for every hardware used. In this poster, we introduce the ExaGRyPE framework and showcase some key numerical relativity simulations, before we introduce the ExaHyPE-DSL eCSE project which avoids the need to manually develop, maintain and tune multiple kernels across CPUs and GPUs by the use of a Domain Specific Language (DSL): DSLs enable the programmer to express their intentions in a manner that is close to the problem in hand - in our case the combination of numerical astrophysics and higher-order methods. Using this domain-specific source of information, the compiler is able to make informed decisions around the tricky, low level mapping of inherent concurrency onto actual hardware, including modern vectorisation units and GPUs. In ExaHyPE-DSL we have developed a Python based DSL which, by making transformation passes over the generated MLIR (multi-level intermediate representation) code , generates optimal variants of the kernels for each use case and hardware.

Timothy Stokes
Department of Computer Science, Durham University
nlhx46@durham.ac.uk

Han Zhang
Durham University
han.zhang3@durham.ac.uk

Maurice Jamieson
EPCC
m.jamieson@epcc.ed.ac.uk

Tobias Weinzierl, Baojiu Li
Durham University
tobias.weinzierl@durham.ac.uk, baojiu.li@durham.ac.uk

Nick Brown
EPCC
n.brown@epcc.ed.ac.uk

PP1

Adaptive Spectral Block Floating Point for Discontinuous Galerkin Methods

Discontinuous Galerkin (DG) methods are powerful for large-scale PDE simulations, but their application is often constrained by memory footprint and bandwidth, which are critical bottlenecks in high-performance computing (HPC). We introduce a novel Spectral Block Floating Point (SBFP) format to exploit the spectral decay of modal DG coefficients, allowing higher-order terms to be stored at reduced precision. For one-dimensional quadratic DG solutions, SBFP achieves a remarkable 33% memory overhead reduction compared to standard single-precision storage by packing a shared exponent and truncated mantissas of the higher-order coefficients into a single 64-bit word. To improve robustness when coefficient magnitudes vary widely, we propose an adaptive variant, Adaptive SBFP (ASBFP), which introduces specialization bits to better align mantissas. Numerical experiments on linear and nonlinear hyperbolic problems show that ASBFP retains accuracy while significantly reducing memory use. A prototype FPGA implementation demonstrates the feasibility of direct ASBFP arithmetic, suggesting avenues for hardware-efficient realizations. We are developing a portable library implementation targeting CPUs, GPUs, and FPGAs. The approach

extends naturally to higher-order and higher-dimensional DG discretizations, offering a promising path toward more scalable, memory-efficient parallel DG solvers in large-scale HPC environments.

Shivam Sundriyal, Markus Büttner
University of Bayreuth
shivam.sundriyal@uni-bayreuth.de,
markus.buettner@uni-bayreuth.de

Christoph Alt, Tobias Kenter
Paderborn University
christoph.alt@uni-paderborn.de, tobias.kenter@upb.de

Vadym Aizinger
University of Bayreuth
vadym.aizinger@uni-bayreuth.de

PP1

ExaHyPE: Adaptive High-Order PDE Solvers with GPU Offloading and Performance Portability

ExaHyPE is a modern simulation engine for solving systems of hyperbolic partial differential equations (PDEs). Building on the algorithmic foundations of high-order ADER-DG methods and finite volume solvers, it allows users to tailor the engine to specific applications, such as for modeling complex wave phenomena. We present recent work in the application domains of landslide dynamics and earthquake physics. At its core, ExaHyPE relies on adaptive Cartesian meshes represented by spacetrees, which allow the computational grid to dynamically refine or coarsen in response to evolving solution features. This adaptivity enables highly accurate and efficient simulations, but also poses challenges for high-performance computing. We address these by memory-efficient data layouts, dynamic load balancing, and performance-portable GPU offloading through modern programming models. Our results demonstrate that ExaHyPE can bridge advanced numerical methods with exascale-ready software design, providing a flexible and sustainable platform for large-scale multi-physics simulations.

Mario Wille, Marc Marot-Lassauzaie, Michael Bader
Technical University of Munich
mario.wille@tum.de, marc.marot@tum.de,
bader@cit.tum.de