#	Last name	Last name	Presentation title	Authors	Abstract
M01	Motamarri	Phani Sudheer	Scalable finite-element based methods for quantum modelling of materials using density functional theory in the exascale era	Phani Sudheer Motamarri	The predictive capability offered by quantum modelling of materials, especially using density functional theory (DFT), has opened up a gateway for gaining crucial insights into materials' behaviour, leading to an accurate prediction of mechanical, transport, chemical, electronic, magnetic and optical properties of materials. However, the stringent accuracy requirements required to compute meaningful material properties and the asymptotic cubic-scaling computational complexity of the underlying DFT problem demand enormous computational resources. Thus, these calculations are routinely limited to periodic simulation domains with material systems containing a few hundred atoms. Additionally, these DFT calculations occupy a sizeable fraction of the world's computational resources today but mostly remain in the high throughput calculation mode as the widely used DFT implementations struggle to keep up with evolving heterogeneous architectures in today's exascale era. To this end, the talk introduces the recent advancements in finite-element (FE) based methods for DFT calculations -via- the DFT-FE code, the workhorse behind the ACM Gordon Bell Prize 2023. These methods provide a systematically convergent, computationally efficient and scalable hybrid CPU-GPU framework for large-scale norm-conserving pseudopotential DFT calculations that overcomes these limitations with no restrictions on the boundary conditions that can be applied. Subsequently, we will discuss our group's very recent efforts in developing a fast and scalable approach combining the efficiency of projector-augmented wave (PAW) formalism involving smooth electronic fields with the ability of systematically improvable higher-order FE basis facilitating substantial reduction in degrees of freedom to achieve significant computational gains (~8x-10x) compared to the current DFT-FE calculations for medium to large-scale material systems. These recent developments have wide-ranging implications for tackling critical scientific and technological problems.
M02	Nataf	Frederic	Adaptative spectral coarse spaces for domain decomposition methods	Frederic Nataf	Convergence of domain decomposition methods rely heavily on the efficiency of the coarse space used in the second level. The GenEO coarse [1,2] space has been shown to lead to a robust two-level Schwarz preconditioner which scales well over thousands of cores. The robustness is due to its good approximation properties for problems with highly heterogeneous material parameters. It is available in the finite element package FreeFem++ [3] and as a standalone library in HPDDM [4] as well as a PETSc preconditioner. Numerical results as well as, if time permits, extensions to saddle point problems or to the time dependent Maxwell system will be discussed. ; [1] Spillane N., Dolean V., Hauret P., Nataf F., Pechstein C. and Scheichl R., Abstract robust coarse spaces for systems of PDEs via generalized eigenproblems in the overlaps, Numer. Math., 2014; [2] Dolean, V. and Jolivet, P. and Nataf, F., An Introduction to Domain Decomposition Methods: algorithms, theory and parallel implementation, SIAM, 2015; [3] Hecht F., New development in FreeFem++, J. Numer. Math., 2012; [4] Jolivet, P. and Nataf, F., HPDDM: High-Performance Unified framework for Domain Decomposition methods, MPI-C++ library, https://github.com/hpddm/hpddm, 2014
M03	Babbar	Arpit	Admissibility preserving Lax-Wendroff Flux Reconstruction schemes for compressible flows	Arpit Babbar, Praveen Chandrashekar, Sudarshan K	Lax-Wendroff Flux Reconstruction (LWFR) is a single-stage, high order, quadrature free method for solving hyperbolic conservation laws. The single step nature of the method enables evolution to the next time level with only one interface communication, making the method arithmetically intense and MPI efficient. Since solutions to hyperbolic conservation laws often contain shocks, in this work, we develop a subcell based limiter by blending LWFR with a lower order scheme, either first order finite volume or MUSCL-Hancock scheme. While the blending with a lower order scheme helps to control oscillations, it may not guarantee admissibility of discrete solution, e.g., positivity property of quantities like density and pressure. By exploiting the subcell structure and admissibility of lower order schemes, we devise a strategy to ensure that the blended scheme is admissibility preserving for the mean values and then use a scaling limiter to obtain admissibility of the polynomial solution. For MUSCL-Hancock scheme on non-cell-centered subcells, we develop a slope limiter, time step restrictions and suitable blending of higher order fluxes, that ensures admissibility of lower order updates and hence that of the cell averages. By using the MUSCL-Hancock scheme on subcells and Gauss-Legendre points in flux reconstruction, we improve small-scale resolution compared to the subcell-based RKDG blending scheme with first order finite volume method and Gauss-Legendre-Lobatto points. We demonstrate the performance of our scheme on compressible Euler's equations, showcasing its ability to handle shocks and preserve small-scale structures.
M04	Goswami	Shubham Kum	A scalable asynchronous discontinuous Galerkin method for massively parallel flow simulations	Shubham Kumar Goswami, Vidyesh Dapse, Konduri	Accurate simulations of turbulent flows are crucial for understanding complex phenomena in engineered systems and natural processes. These simulations often require the use of supercomputers due to their high computational cost. However, scalability at extreme scales can be significantly affected by the communication overhead. To address this challenge, an asynchronous computing approach for time-dependent partial differential equations (PDEs) that relaxes communication and synchronization at a mathematical level has been developed with finite difference schemes that are ideal for structured meshes. This work proposes an asynchronous discontinuous Galerkin (ADG) method, which combines the benefits of the DG method with asynchronous computing and has the potential to provide high-order accurate solutions for various flow problems on both structured and unstructured meshes. The numerical properties of the proposed method, including local conservation, stability, and accuracy, are investigated, where the method is shown to be at most first-order accurate. To recover accuracy, new asynchrony-tolerant (AT) fluxes that utilize data from multiple time levels are developed. To validate these theoretical findings, several numerical experiments are conducted based on both linear and nonlinear problems. Finally, a parallel PDE solver based on the ADG method is developed within an open-source finite element library deal.II, using a communication-avoiding algorithm. The accuracy of the solver is validated, and scalability benchmarks demonstrate a speedup of up to 80% with the ADG method at an extreme scale with 9216 cores.
M05	Aizinger	Vadym	p-adaptive discontinuous Galerkin method for the shallow water equations on hybrid CPU-GPU architectures	Vadym Aizinger, Sara Faghih-Naini, Richard Angersb	Exploiting integrated CPU-GPU architectures allows to achieve computational performance benefits by creating specialized kernels optimized to different hardware. We propose to combine hybrid execution mode with a p-adaptive quadrature-free discontinuous Galerkin (DG) method and employ this scheme to solve the shallow water equations. Our approach separates the computations of the non-adaptive (lower-order) and adaptive (higher-order) parts of the discretization from each other and executes them on separate hardware. We use automatic code generation to create optimized compute kernels and distribute them between the CPU and GPU. Several setups, including a prototype of a tsunami simulation in a tide-driven flow scenario, are investigated, and the results show that significant performance improvements can be achieved in suitable setups.
M06	Heinlein	Alexander	Domain decomposition for neural networks	Alexander Heinlein	Scientific machine learning (SciML) is a rapidly evolving field of research that combines techniques from scientific computing and machine learning. In this context, this talks focuses on the enhancement of machine learning using classical numerical methods, in particular, on improving neural networks using domain decomposition-inspired architectures. In the first part of this talk, the domain decomposition paradigm is applied to the approximation of the solutions of partial differential equations (PDEs) using physics-informed neural networks (PINNs). It is observed that network architectures inspired by multi-level Schwarz domain decomposition methods can improve the performance for certain challenging problems, such as multiscale problems. Moreover, a classical machine learning task is considered, that is, image segmentation using convolutional neural networks (CNNs). Domain decomposition techniques offer a way of scaling up common CNN architectures, such as the U-Net. In particular, local subdomain networks learn local features and are coupled via a coarse network which incorporates global features.
M07	Pleiter	Dirk	HPC-based Federated Digital Infrastructures	Dirk Pleiter	Different areas of computational science have seen tremendous progress in recent years, in parts due to the remarkable continuation of the growth of computational performance. Various relevant use cases show, however, that there is a need for considering the full workflows that extend far beyond a single HPC system. This can be addressed by integrating HPC systems in federated digital infrastructures. In this talk, I discuss the relevance of this in parts conceptional approach for different use cases, including those that involve the realisation of HPC-based digital twins. Based on an analysis of the current state and ongoing challenges, I will discuss how to improve design methodologies for such kind of infrastructures.
M08	Anzt	Hartwig	It is all in the GPUs - How the Hardware Architecture impacted Scientific Software in the US Exascale Computing Project	Hartwig Anzt	The US Exascale Computing Project (ECP) just ended and succeeded with the goal to deliver a capable exascale computing ecosystem to provide breakthrough modelling and simulation to address the most critical challenges in scientific discovery, energy assurance, economic competitiveness, and national security. This venture came with the challenge to prepare a scientific computing software stack to a changing hardware landscape that increasingly treats GPUs as the workhorse of supercomputers. In this talk, we describe the path towards a GPU-centric scientific computing software stack, discuss the programming and numerical challenges coming with it, and outline which concepts enabled sustainability beyond the project completion.
M09	Pericas	Miquel	Challenges and Opportunities of Long Vector Architectures in HPC	Miquel Pericas	The European Union's drive to establish a domestic HPC industry has led to considerable R&D around RISC-V and its vector extension, embodied in projects such as EPI and EUPilot. By leveraging long vector registers, RISC-V offers the potential for significant performance and energy efficiency gains. However, achieving scalable execution with long vectors is challenging. In this talk, I will present our research on co-designing several CNN algorithms (direct, Winograd, im2col) with long vector architectures, using ARM-SVE and RISC-V as case studies. I will delve into the intricacies of exploiting long vector registers effectively and highlight the challenges in achieving scalable execution when using approaches such as autovectorization or compiler intrinsics. In the second part of this talk I will discuss our work on SYCL as an alternative for generating multithreaded and vectorized code. SYCL, an explicitly parallel programming model, can facilitate the generation of long vectors, potentially improving performance and programmability. We are currently developing a microbenchmark to evaluate the parallelization and vectorization capabilities of various SYCL implementations, specifically when targeting multicore processors with vector units. I will present the status of this microbenchmark and discuss how the major two SYCL implementations (DPC++ and AdaptiveCPP) perform over a range of ARM and x86 systems.
M10	Alpay	Aksel	AdaptiveCpp: Portable Heterogeneous Computing in C++	Aksel Alpay, Vincent Heuveline	Increasing diversity in the HPC hardware landscape requires development tools that address the associated programmability challenge. This is particularly relevant when accelerators such as GPUs are involved, where every hardware vendor typically supports different programming models. AdaptiveCpp is a modern and highly competitive compiler and runtime stack for heterogeneous computing in C++. AdaptiveCpp can target any LLVM-supported CPU, as well as GPUs from Intel, NVIDIA and AMD through either the SYCL or C++ standard parallelism offloading programming models. It is used in production by large scale HPC applications such as the Gromacs molecular dynamics software. Unlike most other heterogeneous C++ compilers, AdaptiveCpp is entirely community-driven rather than vendor-driven, with the Engineering Mathematics and Computing Lab (EMCL) at Heidelberg University leading its development. We argue that community-driven compilers are a key component to make scientific software immune to vendor lock-in. In this talk, we give an overview of AdaptiveCpp and the programming models it supports, and discuss how AdaptiveCpp can help scientists develop high-performance, portable applications in C++. We also provide insight into some of the unique features of AdaptiveCpp, such as its unified JIT compilation infrastructure, the ability to conveniently generate a portable binary that runs on GPUs from all vendors, and the ability to adapt kernel code at runtime. This allows AdaptiveCpp to automatically take into account runtime information during code optimization, often resulting in superior performance compared to vendor compilers.
M11	Beutel	Moritz	Temporal blocking in practice	Moritz Beutel, Robert Strzodka	Temporal blocking allows to exploit the spatio-temporal locality of stencil codes in order to surpass the system bandwidth which bounds the performance of naïve cache-blocking schemes. While the technique has been subject to extensive theoretical study, the complexity it entails has so far impeded its widespread application in iterative computations. In this talk, we will give an introduction to the concept of temporal blocking and discuss strategies for applying the technique in existing numerical codes. We will investigate what it takes to adapt a simple hydrodynamic solver to a temporal blocking scheme, concluding the talk with an evaluation of the speed-up attained.
M12	Chhabra	Adil	Buffered Streaming Edge Partitioning	Christian Schulz, Marcelo Fonseca Faraj, Daniel See	Addressing the challenges of processing massive graphs, which are prevalent in diverse fields such as social, biological, and technical networks, we introduce HeiStreamE and FreightE, two innovative (buffered) streaming algorithms designed for efficient edge partitioning of large-scale graphs. HeiStreamE utilizes an adapted Split-and-Connect graph model and a Fennel-based multilevel partitioning scheme, while FreightE partitions a hypergraph representation of the input graph. Besides ensuring superior solution quality, these approaches also overcome the limitations of existing algorithms by maintaining linear dependency on the graph size in both time and memory complexity with no dependence on the number of blocks of partition. Our comprehensive experimental analysis demonstrates that HeiStreamE outperforms current streaming algorithms and the re-streaming algorithm 2PS in partitioning quality (replication factor), and is more memory-efficient for real-world networks where the number of edges is far greater than the number of vertices. Further, FreightE is shown to produce fast and efficient partitions, particularly for higher numbers of partition blocks.

T01	Behera	Ratikanta	Tensor factorizations with applications	Ratikanta Behera	Abstract: In the era of BIG data, artificial intelligence, and machine learning, we need to process multiway (tensor-shaped) data. These data are mainly in the three or higher-order dimensions, whose orders of magnitude can reach billions. Huge volumes of multidimensional data are a big challenge for processing and analyzing; the matrix representation of data analysis is not enough to represent all the information content of the multiway data in different fields. In this talk, we will discuss tensor factorization as a product of tensors. To address the factorizations, we define a closed multiplication operation between tensors with the concept of transpose, inverse, and identity of a tensor. We will conclude with a few colour image applications in a tensor-structure domain.
T02	Carson	Erin	Balancing Inexactness in Matrix Computations	Erin Carson	On supercomputers that exist today, achieving even close to the peak performance is incredibly difficult if not impossible for many applications. Techniques designed to improve the performance of matrix computations - making computations less expensive by reorganizing an algorithm, making intentional approximations, and using lower precision - all introduce what we can generally call "inexactness". The questions to ask are then: 1. With all these various sources of inexactness involved, does a given algorithm still get close enough to the right answer? 2. Given a user constraint on required accuracy, how can we best exploit and balance different types of inexactness to improve performance? Studying the combination of different sources of inexactness can thus reveal not only limitations, but also new opportunities for developing algorithms for matrix computations that are both fast and provably accurate. We present a few recent examples of this approach, in which mixed precision computation is combined with other sources of inexactness.
T03	Hasani	Pouria	Software Controlled Hardware Approximation for General-Purpose Computing	Pouria Hasani, Nima Amirafshar, Ekrem Altuntop, Ste	High performance and low power consumption are two of the main goals of computer architecture design, both of which encounter several challenges. To tackle these demands, emerging computing paradigms have been introduced, one of which is approximate computing. It has been shown that many applications are error-tolerant with little to no loss in their levels of functional or required accuracy. Therefore, using approximate computing methods deliberately can reduce hardware complexities and lead to considerable improvements in speed, power consumption, and area. However, the majority of approximate hardware in the literature is not only limited to basic computational units of processing systems but also have been application-specific. Moreover, using them required knowledge and expertise in hardware, which is out of the reach for many users. To tackle these challenges, we propose the Approximate and Exact Multi-Processor (AxE) platform for general-purpose applications which would want to benefit from hardware-level approximation, controlled by software. The AxE platform is a multiprocessor system on chip where each processor can possess either an exact or an approximate hardware core. An approximate core refers to a processor with an approximate accelerator (so far multipliers), whereas an exact core refers to a processor with an exact multiplier. The processors are connected through a Network on Chip (NoC), and the number of processors is easily adjustable. The programs to be executed are loaded onto a common memory, where each program is independent of the others and has its own exclusive memory space. The task of assigning programs to each core is managed by the controller processor, allowing each core to execute a program in parallel and independently of the other cores on the network. The AxE platform is capable of executing general-purpose applications, both legacy and new approximate ones. Error-tolerant applications can benefit from the approximate hardware (multipliers) embedded in the approximate cores to save power and reduce execution time, while non-error-tolerant applications can be executed on the exact cores. The user can make the decision by appropriate modification of their legacy code. This architecture provides a robust platform for experimental research in approximate computation for general-purpose applications. It enables users to quickly evaluate whether and which portion of their software could be offloaded to an approximate hardware and what the gain and potential quality costs of such an approach may be. In this presentation, we will showcase AxE platform.
T04	Pathak	Ritik Kumar	Optimal control of PDEs	Ritik Kumar Pathak	Optimal control of partial differential equations (PDEs) is a sophisticated and dynamic field that lies at the intersection of mathematics, engineering, and applied sciences. This discipline focuses on determining control functions that optimize a certain performance criterion governed by PDEs. These equations are fundamental in describing various physical phenomena such as heat conduction, fluid dynamics, electromagnetism, and financial modeling. The primary objective in optimal control problems is to find a control strategy that minimizes or maximizes a cost functional while satisfying the constraints imposed by PDEs. This involves the interplay between the theory of PDEs, calculus of variations, and numerical optimization techniques. Key aspects include the existence and uniqueness of solutions, the derivation of optimality conditions through methods like the Pontryagin Maximum Principle and the Lagrange multipliers, and the development of efficient computational algorithms for solving large-scale problems. Recent advancements have expanded the applicability of optimal control of PDEs to complex and high-dimensional systems, incorporating state-of-the-art methods such as machine learning for model reduction, and employing parallel computing to handle the computational intensity. This presentation will delve into the fundamental concepts, mathematical formulations, and practical applications of optimal control in PDEs, showcasing examples from engineering, economics, and environmental science. Additionally, we will explore contemporary challenges and future directions in this vibrant research area, emphasizing the integration of new computational techniques and interdisciplinary approaches. By understanding the theoretical underpinnings and computational strategies in optimal control of PDEs, researchers and practitioners can better tackle real-world problems, leading to innovations and improvements in various technological and scientific domains.
T05	Seelinger	Linus	Enabling Bayesian Inference for Large-Scale Simulation: Methods and Software	Linus Seelinger, Max Kruse (Karlsruhe Institute of Te	Bayesian inference of complex simulation models enables important scientific insight, but comes at great computational cost and technical complexity.; We present UM-Bridge, a universal software framework that breaks down this complexity. It enables straightforward linking of any high-level algorithm with any simulation model, makes models portable, and enables scalability on supercomputers and cloud clusters. As a result, we can now rapidly develop and scale applications that combine state-of-the-art methods from Bayesian inference and numerical simulation.We further present recent developments around Multilevel Delayed Acceptance, a state-of-the-art inference method capable of handling compute intensive simulation models. Finally, we demonstrate its effectiveness in UM-Bridge powered applications in astrophysics and earthquake research.
T06	Engwer	Christian	Parallel-in-time Block-Krylov methods to improve node-level performance	University of Münster	Combining paralle-in-time and Block-Krylov methods we present a promissing approach to increase the node-level performance of PDE solvers for a wide range of instationary problems on modern hardware. In many applications low order methods are still the most common approach to solve PDEs. While they are easy to implement, they are inherently memory bound due to a low arithmetic intensity and thus don't benefit from the high level of concurrency of modern hardware architectures. To increase the arithmetic intensity, it is necessary to increase the work per matrix entry. For applications which require solving different linear systems with the same operator, Block-Krylov methods offer a mathematical tool to increase the arithmetic intensity and in previous work we added corresponding support to our DUNE linear algebra library. Conceptually, a system for multiple time steps of a time stepping method or multiple Runge-Kutta stages can be reformulated to solve a single matrix equation instead of many linear systems which may be solved using Block-Krylov methods. We introduce the underpinning ideas, present first performance results and discuss implementational aspects.
T07	Sundaresan	Vaanathi	Scientific machine learning techniques and their applications for precision medicine	Vaanathi Sundaresan	The talk will provide an overview of application of deep learning (DL) techniques for medical imaging applications, especially for identifying medical imaging biomarkers for precision medicine. This includes development of DL models for accurate automated segmentation of pathological findings in medical imaging modalities and analysis of their clinical impact at the population-level. Another key discussion point of the talk will be to improve the robustness of the deep learning tools by tackling various practical challenges in the DL-based tool development such as limited availability of manually labelled data for training, domain shift in the data acquired from different centres and lack of variability in the low data regimes. Future avenues of the research include for the detection of anomalies (abnormalities) using multiple diverse imaging modalities, and their classification and quantification.
T08	Westermann	Josephine	Neural Networks vs Sparse Polynomials in Spectral Operator Surrogates	Josefine Westermann, Jakob Zech, Thomas O'Leary	Problems in uncertainty quantification and stochastic simulation typically require large numbers of evaluations of the operator associated with the system in question. In cases where these evaluations are computationally expensive, such as in systems governed by Partial Differential Equations, it is advantageous to construct a surrogate that is cheap to evaluate. Given that these problems are usually high- or infinite-dimensional, encoder and decoder functions are employed to map inputs and outputs into lower-dimensional coefficient representations. The operator surrogate is then constructed from an approximation to the corresponding mapping between these coefficient spaces. While neural networks have been applied successfully to such tasks, sparse multivariate polynomial expansions offer a promising alternative since they allow for deterministic convergence rates in many scenarios. In this work, we empirically evaluate the performance of both approaches across several test problems. In particular, we examine the cost-accuracy trade-off for both surrogate types and investigate how it is influenced by factors such as parametric smoothness and variance. Our results show that although neither method consistently outperforms the other, specific conditions favor one approach over the other. We discuss the underlying reasons for these observations and their practical implications.
T09	Branca	Lorenzo	Emulating the Interstellar Medium Chemistry with Neural Operators	Lorenzo Branca	Understanding the interstellar medium (ISM) chemistry is pivotal for the study of galaxy formation and evolution. Traditional computational models rely on costly ordinary differential equation (ODE) solvers to simulate complex photo-chemical processes. This study introduces a novel approach using DeepONet neural operators to emulate a non-equilibrium chemical network, significantly reducing computational costs while maintaining precision. Unlike conventional methods, our approach approximates the differential operator directly, enabling the model to generalize beyond the specific conditions encountered during training. This capability ensures the robustness and flexibility of the emulation across a broader parameter space, including varied densities, temperatures, species abundances, and radiation fields. Remarkably, our method maintains an accuracy within 1% while achieving speed-ups of up to 128× compared to traditional methods. This makes it a powerful tool for large-scale astrophysical simulations and advancing our understanding of ISM dynamics.
T10	Baumgarten	Niklas	A High-Performance Multi-level Stochastic Gradient Descent Method with Applications in Optimal Control under Uncertainty	Niklas Baumgarten	We present a high-performance multi-level stochastic gradient descent method to optimally control the state of systems guided by partial differential equations under uncertain input data. The gradient descent method used to find the optimal control leverages a parallel budgeted multi-level Monte Carlo method as stochastic sub-gradient estimator. As a result, we get tight control over the sub-gradient's bias, introduced by numerical discretizations, and the sub-gradient's variance with respect to the invested computational resources. We provide empirical evidence that the method outperforms the standard stochastic gradient descent method in terms of convergence speed and accuracy. The method is particularly well-suited for high-dimensional control problems by exploiting the parallelism and the distributed data structure of the budgeted multi-level Monte Carlo method. Furthermore, we establish a connection to the batched gradient descent and the ADAM optimizer methods. Lastly, we study the method's performance at hand of a three-dimensional elliptic subsurface diffusion problem with log-normal coefficients and Matérn covariance functions.
T11	Turek	Stefan	The Future of CFD Simulations (from a numerical & computational perspective) - Faster and more reliable predictions are needed to compete with AI	Stefan Turek	The main aim of this talk is to discuss how modern High Performance Computing (HPC) techniques regarding massively parallel hardware with millions of cores together with very fast, but lower precision accelerator hardware can be applied to numerical simulations of PDEs so that a much higher computational, numerical and hence energy efficiency can be obtained. Here, as prototypical extreme-scale PDE-based applications, we concentrate on nonstationary flow simulations with hundreds of millions or even billions of spatial unknowns in long-time computations with many thousands up to millions of time steps. For the expected huge computational ressources in the coming exascale era, such type of spatially discretized problems which typically are treated sequentially in time, that means one time after the other, are often too small to exploit adequately the huge number of compute nodes, resp., cores so that further parallelism, for instance w.r.t. time, might get necessary. In this context, we discuss how "parallel-in-space & global-in-time" Newton-Krylov Multigrid approaches can be designed which allow a much higher degree of parallelism. Moreover, to exploit current accelerator hardware in lower precision (for instance, GPUs from NVIDIA built for AI applications), we discuss the concept of "prehandling" (in contrast to "preconditioning") of the corresponding ill-conditioned systems of equations, for instance arising from Poisson-like problems. Here, we assume a transformation into an equivalent linear system with similar sparsity but with much lower condition numbers so that the use of lower precision hardware gets feasible. In our talk, we provide for both aspects numerical results as "proof-of-concept" and discuss the challenges, particularly for incompressible flow problems, also in view of comparisons with AI predictions.

T12	Chamakuri	Nagaiah	A Computational Framework for Optimal Control of Cardiac Defibrillation	Nagaiah Chamakuri	This talk will showcase a computational framework designed for the optimal control of cardiac defibrillation, specifically using a detailed three-dimensional anatomical model of the rabbit ventricle with bilateral control constraints. Our approach addresses the numerical challenges of multi-scale, multi-domain simulations of the bidomain equations, utilizing the primal-dual active set method to tackle large-scale PDE-constrained optimization problems. The bidomain model comprises a system of elliptic partial differential equations coupled with a nonlinear parabolic reaction-diffusion equation, alongside ordinary differential equations that capture ionic transport dynamics. Additionally, we address an extra Poisson problem to simulate scenarios where the heart is immersed in a conductive medium, such as a tissue bath or the surrounding torso. Given that the ODEs describe ionic currents in the tissue, the PDE component typically dominates the computational effort. This raises questions about whether traditional splitting methods can achieve the accuracy required for the bidomain and bidomain-bath models compared to a fully coupled approach. In the first part of the presentation, we will compare results from our coupled solver with those from commonly used splitting schemes for more complex physiological models. In the second part, we will present our optimal control framework for effective cardiac defibrillation, focusing on minimizing a specifically designed cost functional that incorporates various types of cost metrics.
W01	Varbanescu	Ana Lucia	Towards zero-waste computing by	Ana Lucia Varbanescu	"Computation" has become a massive part of our daily lives: in science, a lot of experiments and analysis rely on massive computation, in AI we use vast resources to train and use massive models, and in engineering we use complex simulations and digital twins to increase efficiency and productivity. Under the assumption that computation is cheap, and time-to-result is the only relevant metric, we often use significant computational resources at low efficiency. In this talk, I argue this approach is an unacceptable waste of computing resources, and demonstrate we can do better! By means of a couple of case-studies, I will show how performance engineering can be used towards zero-waste computing. I will further propose a co-design methodology that leverages such performance engineering methods to enable the selection of algorithms _and_ their effective deployment on suitable infrastructure. The approach relies on design-space exploration, driven by efficient search methods and compositional performance models. I will conclude by reflecting on the next steps and open questions that need answers to make this co-design approach feasible and applicable for more applications and systems.
W02	Konduri	Aditya	Dimensionality reduction based on Cokurtosis-PCA: application to chemical kinetics	Aditya Konduri	Direct numerical simulations of turbulent reacting flows resolve the detailed chemical kinetics that provide insights into the turbulence-chemistry interactions. While the kinetics models should accurately represent the chemistry, their sizes need to be small enough for the computations to be tractable. Dimensionality reduction aims to reduce the feature space of high-dimensional data while retaining the information and dynamics of the original system effectively. Widely used principal component analysis (PCA) achieves this for combustion data by transforming the original thermo-chemical state space into a low-dimensional manifold with eigenvectors of the covariance matrix of the input data. However, this may not effectively capture stiff chemical dynamics when the reaction zones are localised in space and time. Alternatively, a co-kurtosis PCA (CoK-PCA), wherein the principal components are obtained from the singular value decomposition (SVD) of the matricized co-kurtosis tensor, demonstrate greater accuracy in capturing stiff dynamics. In this study, we demonstrate the efficacy of a CoK-PCA-based reduced manifold using a posteriori analysis. Simulations of spontaneous ignition in a homogeneous reactor that pose a challenge in accurately capturing the ignition delay time as well as the profiles of the scalar within the reaction zone are considered. The principal components are evolved from the initial conditions using two ODE solvers. First, a standard solver that uses an aritifical neural network to estimate the source terms. Second, a neural ODE solver that incorporates the time integration of PCs in training ANNs, which predict their source terms. The time-evolved profiles of the PCs and ANN-reconstructed thermo-chemical scalars demonstrate the robustness of the CoK-PCA-based low-dimensional manifold in accurately capturing the ignition process. Furthermore, we observed that the neural ODE solver provides more accurate results than the standard ODE solver. The results from this study demonstrate the potential of CoK-PCA-based manifolds to be implemented in massively parallel reacting flow solvers.
W03	Ippisch	Olaf	Lineal: An new efficient, hybrid-parallel Linear Algebra Library	Kurt Böhm, Olaf Ippisch, Institute of Mathematics, Cla	Efficiently solving large sparse systems of linear equations arising from the discretization of PDEs is still a challenging problem. To solve large problems on attainable hardware, the new linear algebra library Lineal has been developed. Lineal uses the preconditioned CG method as its main solver, with an Algebraic Multigrid solver (an optimized version of the AMG from DUNE ISTL) as its main preconditioner. However, Lineal uses a number of techniques to achieve very low memory requirements while providing low runtimes as well as generic and extensible interfaces. For stencil-based problems, Lineal can compute the matrix elements on the finest grid on the fly and only needs to store the coarse grid hierarchy explicitly. In this case, only a single value (a single byte for some problems) per cell is needed on the finest grid, which drastically reduces memory consumption compared to explicit matrices. Additionally, matrix-vector products are computed using tiling to improve cache utilization. Alternatively, Compressed Row Storage (CRS) matrices can be used, which support indices consisting of an arbitrary number of bytes to reduce memory consumption. Furthermore, floating point types can be mixed (almost) arbitrarily to save memory. Elementary operations are represented as classes that perform element-wise computations, using inlining to combine operations efficiently. Almost all components are fully multithreaded and use explicit SIMD operations to improve performance. Additionally, recent work has added support for distributed memory parallelism using MPI, allowing for hybrid-parallel computations that utilize a compute cluster while minimizing communication costs. Lineal has been successfully used to simulate oxygen diffusion in X-ray scans of soil samples, solving instances with more than 10^9 unknowns in 10 to 120 minutes on a single AMD EPYC system with 32 cores and 256~GB of RAM. Further tests using this problem show that Lineal performs well compared to existing libraries in terms of runtime and memory consumption. Tests demonstrating its scalability on larger compute clusters using hybrid parallelism will be shown as well.
W04	Cui	Cu	Efficient and High-Performance Finite Element Methods on GPUs	Cu Cui, Guido Kanschat	We present a GPU-accelerated implementation of vertex-patch smoothers for higher-order finite element methods in 2D and 3D. Optimizing multigrid operations with on-chip memory reduces global data transfers and achieves conflict-free access. Tests on Nvidia A100 GPUs show our optimized kernel is twice as fast as the baseline for the Poisson problem, reaching up to 36% peak performance in single and double precision. Additionally, we introduce a matrix-free multigrid method for high-order discontinuous Galerkin finite element methods, achieving up to 39% peak performance with shared memory and mixed-precision approaches. MPI parallelization further enhances efficiency and robustness in 2D and 3D Poisson problems. We also accelerate tensor product operations using Nvidia A100 GPU Tensor Cores. Inline PTX instructions with conflict-free shared memory access yield a 2.3-fold increase in double precision and a fourfold enhancement in half-precision for solving the Poisson equation with FGMRES. These results highlight Tensor Cores' benefits in balancing computational speed and precision for finite element operators.
W05	Büttner	Markus	Performance portability across CPUs, GPUs and FPGAs for a SYCL implementation of a discontinuous Galerkin shallow water solver	Christoph Alt, Friedrich Alexander University Erlangen-Nuremberg and Paderborn Center for Parallel Computing Tobias Kenter, Paderborn Center for Parallel Computing Vadym Aizinger, University of Bayreuth	We will present our work on a performance-portable implementation of the 2D shallow water equations. The discretization is based on the modal discontinuous Galerkin method on unstructured triangular grids. Unstructured meshes are particularly well suited for representing complex geometries like coastal regions. We use the open SYCL 2020 standard to achieve performance portability across CPUs, GPUs and FPGAs. With the two major SYCL compilers currently available, oneAPI and AdaptiveCpp, we can demonstrate that our code runs on a wide variety of hardware. Tested systems include notebooks, workstations and high-end servers with x86 and ARM CPUs, consumer grade GPUs as well as high-end data center GPUs and even Intel FPGAs. Furthermore, we can compare the tested hardware also in terms of energy efficiency by measuring and comparing consumed energy and sustained power draw for a realistic domain.
W06	Schmalfuß	Jonathan	Block-structured grids: avoiding indirect memory access while retaining solution accuracy – a technology demonstration	Sara Faghih-Naini, Daniel Zint, Julian Stahl, Roberto	Numerical simulations on domains with complex boundaries, such as coastal ocean areas, often rely on unstructured triangular grids. Computation on such grids typically involves some overhead due to irregular memory access patterns. On the other hand, a regular grid structure has performance optimization potential not only for CPUs but also GPUs and FPGAS. Block-structured grids (BSG) allow to exploit this potential using a topologically unstructured collection of blocks, each containing a structured mesh. We present performance evaluations, methods for generation and validation for a range of different BSG techniques: (Standard BSG) a method for automatic generation for ocean domains with a variable number of blocks; (Masked BSG) an enhancement by permitting masking of elements and (Hybrid BSG) a generalization, combining both unstructured and structured blocks in one block-structured grid. The utilization of BSGs in our shallow water equations solver reveals varying degrees of complexity. First, addressing the problem block-wise as an unstructured grid, similar to cache blocking, avoids cache misses. Second, employing explicit offsets to bypass indirect memory accesses. Third applying algorithmic changes exposes optimization potential for the compiler, improving single-core performance. Concurrently, BSGs are designed for parallelism, due to the flexible block number and number of elements per block. The potential gains are accurate and faster ocean simulations, resulting in enhanced understanding and prediction of climate events.
W07	Subramani	Deepak	Physics-Informed Neural Models: Recent Results in Boundary Generalization, Turbulence Modeling, and Ocean State Forecasting	Deepak Subramani	We report three advances from our lab in developing physics-informed neural models: boundary condition generalization, turbulence modelling, and ocean state forecasting. First, we introduce a transformer-based neural operator to learn generalized solutions for various initial and boundary conditions of a PDE. Details of the new architecture and benchmarks are presented. Second, we developed a PINN model using the 2-equation $k-\epsilon$ model and the actuator disc method to simulate wind turbine wakes, reducing training time by not relying on high-fidelity data. Simulations of HOLEC WPS 30/3 and Nibe 630 kW turbine wakes were compared with field data. Third, we developed a discretization-invariant Fourier neural operator (FNO) model to predict the surface fields of the Bay of Bengal in an autoregressive generative framework using the initial ocean state and boundary forcing data.
W08	Ganesan	Sashikumaar	FastVPINNs: Tensor-Driven Acceleration of VPINNs for Complex Geometries	Sashikumaar Ganesan	Variational Physics-Informed Neural Networks (VPINNs) utilize a variational loss function to solve partial differential equations, mirroring Finite Element Methods. Existing implementation of hp-VPINNs, while generally more effective than PINNs, are computationally intensive and scale poorly with increasing element counts. Moreover, their applications thus far have been limited to simple geometries that can be decomposed by regular quadrilateral elements, which limits their applications on complex geometries with skewed quadrilateral elements. This work introduces FastVPINNs, a tensor-based framework that significantly reduces training time and handles complex geometries. Using optimized tensor operations, FastVPINNs can achieve up to 100-fold reduction in the median training time per epoch compared to traditional hp-VPINNs. With the proper choice of hyperparameters, FastVPINNs can surpass conventional PINNs in both speed and accuracy, especially in problems with high-frequency solutions. Further, we have developed FastVPINNs for vector-valued problems, such as the 2D stationary Burgers' equation. We have also extended the implementation to the 2D incompressible Navier-Stokes equation and demonstrated its performance on low Reynolds number flow regimes in problems like lid-driven cavity and Kovasznay flow. Demonstrated effectiveness in solving inverse problems on complex domains underscores FastVPINNs' potential for widespread application in scientific and engineering challenges, opening new avenues for practical implementations in scientific machine learning.