PPAM 2024

15th International Conference on **Parallel Processing & Applied Mathematics**

Book of abstracts

Czech Republic, Ostrava
September 8-11, 2024

Contents

KEYNOTE TALKS

Oak Ridge National Laboratory's AI Initiative: Advancing Secure, Trustworthy, and Energy-Efficient AI at Scale for Scientific Discovery

Prasanna Balaprakash, Massimiliano Lupo Pasini
Oak Ridge National Laboratory, US
pbalapra@ornl.gov, lupopasinim@ornl.gov

We will present an overview of the Oak Ridge National Laboratory's Artificial Intelligence Initiative, which aims to advance the domains of science, energy, and national security. At the core of this initiative are two fundamental thrusts: transformative science applications and cross-cutting assurance. The application thrust focuses on developing AI methods to accelerate scientific discoveries, while the cross-cutting assurance thrust ensures that AI systems are secure, trustworthy, and energy-efficient. Secure approaches include alignment, privacy preservation, and robustness testing for AI models. Trustworthiness is achieved through validation and verification processes, coupled with advanced techniques in uncertainty quantification and causal reasoning. Meanwhile, energy efficiency is prioritized by developing scalable solutions, integrating edge computing technologies, and adopting a holistic co-design approach that optimizes the synergy between software and hardware resources. Through this initiative, we will demonstrate how ORNL is advancing the development and implementation of assured AI for scientific discovery that is impactful, secure, and sustainable.

Keywords: AI, Machine Learning, Trustworthy AI

Hits, Flops, and More

Jack Dongarra
University of Tennessee and ORNL, USA

In this talk we will delve into a retrospective examination of several significant moments and achievements that I've been privileged to contribute to and be a part of. We'll explore these highlights in detail, offering insights into the circumstances, challenges, and outcomes that have shaped these experiences. Through this reflective journey, we aim to glean valuable lessons, celebrate accomplishments, and perhaps even uncover new perspectives that may inspire future endeavors.

Keywords: high performance computing, cloud computing, supercomputer

Building Blocks of Quantum-Accelerated Supercomputing

Ivona Brandic
Vienna University of Technology, Austria

As data volumes are growing faster than the computing power, the computer science community is forced to look for alternatives beyond von Neumann architecture. Among different architectures that are currently being developed, Quantum Computing is one of the most promising ones. In this talk we discuss the concept of hybrid Classic-Quantum architecture and challenges when executing an application on a hybrid computational continuum where parts of the application are executed on the classic machine and parts of the application are executed on the quantum machine. We discuss the problems and challenges caused by the complexities of noise, hyperparameter optimization and data encoding.

Keywords: quantum computers, supercomputing, hybrid Classic-Quantum architecture

Neuromorphic Computing: Towards Brain-like Energy Efficiency

Suma George Cardwell
Sandia National Laboratories, USA

Neuromorphic computing is an area of active research to build next-generation computing systems by mimicking key computational principles from the brain with a brain-like energy footprint. Neuromorphic processing has demonstrated advantages in artificial intelligence/machine learning (AI/ML), scientific computing, remote sensing, graph problems, data analytics and is an ideal bridge between neuroscience and artificial intelligence. Neuromorphic computing will have an impact at both large and small computing scales for efficiency and computational density. Neuromorphic systems currently include digital, analog, and emerging beyond-CMOS devices that enable highly parallel computation, collocated memory and compute elements, event driven computation, and stochasticity. Neuromorphic Computing gives a path forward for computational efficiency scaling and meeting future demands for HPC (high energy physics simulations, climate simulations, scientific computing) and edge computing (remote systems, robotics, unmanned autonomous agents, satellite & space systems applications).

Keywords: artificial intelligence, neuromorphic computing, brain-like energy efficiency

Design, Code-generation, and Synthesis for Next Generation Science Codes

Anshu Dubey
Argonne National Laboratory, USA

Using high-performance computing (HPC) resources effectively has become more challenging than ever due to increasing heterogeneity in both hardware and software. A positive feedback loop of more scientific insight leading to more complex solvers which in turn need more computational resources has been a continuous driver for development of more powerful platforms. The field of computer architecture is poised for more radical changes in how future platforms are likely to be designed, especially because scientific workflows themselves are growing more complex and diverse. Additionally, machine imbalance in HPC has put the focus squarely on the software architecture of application codes. Abstraction with C++ templates that have succeeded until now can only go so far. Generative AI fails miserably at writing any non-trivial scientific code. For the foreseeable future it will take thoughtful software design with judicious task decomposition and a battery of tools for abstraction, code generation, and verification with a human in the loop orchestrating and directing the development to continue to make scientific advances. I will present our efforts along these lines with Flash-X, a multi-physics multi-domain scientific application software.

Keywords: high performance computing, science codes, software engineering

Extreme Scale Earthquake Simulation with Crossing Multi-faults and Topography

Lin Gan
Tsinghua University & National Supercomputing Center in Wuxi, China

A high-scalable and fully optimized earthquake model is presented. The proposed novel optimizing techniques, help fully exploit the hardware potential of all aspects and enable us to perform large earthquake simulations with flexibility. Several real-world earthquakes are successfully simulated with a maximum resolution of 12-m. Precise hazard evaluations for the hazardous reduction of earthquake-stricken areas are also conducted.

Keywords: earthquake simulation, high performance computing, scientific computing

Pushing RISC-V into HPC

Jesus Labarta
Barcelona Supercomputing Center, Spain

The talk will present the philosophy and results of the activity within the European Processor Initiative (EPI) to design a RISC-V vector accelerator. I will briefly present the overall project structure but then focus on the vision of how long vector architectures address fundamental issues in HPC computing such as expressing concurrency and dealing with latency. I will also discuss how the Open Standard RISC-V ISA provides a foundation on which that vision can be deployed while at the same time leveraging contributions of a growing community.

I will describe the architecture of the RISC-V processor designed in the project and it software environment. I will present performance analysis results obtained on an FPGA emulator implementing the same RTL of the taped out test chip.

The FPGA emulator constitutes a Software Development Vehicle (SDV) where a standard Linux environment is available, as well as an LLVM compiler supporting both intrinsics and automatic vectorization. A powerful performance analysis framework is available to understand the behavior of real applications. This environment seamlessly covers a very wire range of levels of detail, from full application coarse grain to microscopic micro-architectural behavior. The emulation SDV allows us to us perform wide parametric sweeps and gain detailed insight on the impact on performance of micro-architectural features.

The bring up processes of the manufactured chip has also been successfully done. The chip now boots Linux and it is possible to run applications and measure performances on it.

I will try to convey some of the experiences and learning of the cooperative work by the different partners in the RISC-V vector activity within the EPI project.

Keywords: HPC, RISC-V, vector computing

Beyond Fugaku/Exascale - Evolving Computing Towards the Next Generation at Riken R-CCS: AI for Science TRIP-AGIS, JHPC-Quantum, and FugakuNEXT

Satoshi Matsuoka
RIKEN Center for Computational Science, Japan

Keywords: exascale, HPC, Fugaku

From the Edge to HPC – Harnessing the Computing Continuum for Science

Manish Parashar
University of Utah, USA

Emerging data-driven, AI-enabled scientific workflows integrate distributed data sources with pervasively availably computing resources, spanning HPC to the edge, to understand end-to-end phenomenon, drive experimentation, and facilitate important decision making. Despite the exponential growth of available digital data sources at the edge, and the ubiquity of non-trivial computational power for processing this data, realizing such science workflows remains challenging. This talk will explore a computing continuum spanning resources at the edges, in HPC centers and clouds, and in-between, and providing abstractions that can be harnessed to support science. The talk will also introduce recent research in programming abstractions that can express what data should be processed and when and where it should be processed, and autonomic middleware services that automate the discovery of resources and the orchestration of computations across these resources.

Keywords: HPC, edge computing, computing continuum

Collaborative continuous benchmarking for HPC

Olga Pearce
Lawrence Livermore National Laboratory, USA

Benchmarking is integral to procurement of HPC systems, communicating HPC center workloads to HPC vendors, and verifying performance of the delivered HPC systems. Currently, HPC benchmarking is manual and challenging at every step, posing a high barrier to entry, and hampering reproducibility of the benchmarks across different HPC systems. In this talk, we describe collaborative continuous benchmarking which enables functional reproducibility, automation, and community collaboration in HPC benchmarking. We develop a common language to streamline the interactions between HPC centers, vendors, and researchers, further enabling the previously unimaginable large-scale improvements to the HPC ecosystem. We introduce an open source continuous benchmarking repository, Benchpark, for community collaboration. We believe collaborative continuous benchmarking will help overcome the human bottleneck in HPC benchmarking, enabling better evaluation of our systems and enabling a more productive collaboration within the HPC community.

Keywords: HPC, benchmarking, collaborative continuous benchmarking

Strengthening AI to Enable Scientific Discovery

Amarda Shehu
George Mason University, USA

What do AlphaFold2, ESM1-2, ChatGPT, GPT-4, DALL-E, Hawk, Claude Opus, Birdie, Llamas, and Alpacas have in common? The word disruption comes to mind. Those of us who started our love affair with AI because we wanted to advance scientific enquiry and the human condition are familiar with disruptions. Molecular biology has the honor of experiencing many disruptions due to groundbreaking findings by Darwin, Miescher, Rosalind, Watson and Crick, Anfinsen, Scheraga, Perutz and Kendrew, Karplus, McCammon, Levitt, Warshel, Scheraga, and many others due to rapid computational advances. Through representative examples, I will showcase some of my laboratory's work on generative AI before and after deep learning. I will tie this work to a central question in AI research, learning the right representation, and will instantiate it on diverse wicked problems in molecular biology and human health that necessitate or benefit from strengthening AI by integrating domain knowledge. I will showcase representative examples of our AI research on grounding, instructibility, and alignment.

Keywords: AI, transformers, scientific discovery

Analytics4NN: Accelerating Neural Architecture Search through Modeling and High-Performance Computing Techniques

Michela Taufer
University of Tennessee, USA

This talk addresses challenges and innovations in Neural Architecture Search (NAS) within high-performance computing. Focusing on the substantial computational demands of designing neural network (NN) architectures, we present Analytics4NN, a unified solution that combines advanced modeling and high-performance computing techniques to enhance NAS efficiency. Analytics4NN introduces a novel fitness prediction engine and a composable workflow. It leverages parametric modeling for early fitness prediction of NNs, seamlessly integrating with existing NAS methods to create more flexible and efficient workflows. This strategy enables the early termination of less promising NNs, optimizes the use of computational resources, and increases the evaluation scope of NN models. Demonstrated on the Summit supercomputer, Analytics4NN shows a remarkable increase in throughput, up to 7.1 times, and a reduction in training time by as much as 5.3 times across diverse benchmark datasets and three state-of-the-art NAS implementations. Analytics4NN's approach to distributed training and rigorous documentation significantly aids in the efficient design of NNs. Applied to a dataset generated by an X-ray Free Electron Laser (XFEL) experiment simulation, it reduced training time by up to 37%. It decreased the required training epochs by up to 38%. Analytics4NN represents a significant leap in the scalability and efficiency of NN design for scientific computing, effectively accelerating NAS by combining cutting-edge modeling with robust, high-performance computing techniques.

Keywords: neural networks, HPC, analytics

New Algorithmic Results for Scheduling via Integer Linear Programming

Klaus Jansen
University of Kiel, Germany

In this talk we present an overview about new results for scheduling problems on parallel machines. During the last years we have worked on the design of efficient exact and approximation algorithms for packing and scheduling problems. In order to obtain faster (implementable) algorithms we studied integer linear programming (ILP) formulations for these problems, developed new parameterized algorithms based on the Steinitz lemma and discrepancy bounds, and proved structural results for optimum solutions of the corresponding ILPs and lower bounds on the running time. This is joint work with Sebastian Berndt, Lin Chen, Max Deppert, Kim-Manuel Klein, Lars Rohwedder, José Verschae, and Gouchuan Zhang.

Keywords: scheduling, linear programming, parallel machines

Designing Converged Middleware for HPC, AI, Big Data, and Data Science

Dhabaleswar K. Panda
Ohio State University, USA

This talk will focus on challenges and opportunities in designing converged middleware for HPC, AI (Deep/Machine Learning), Big Data, and Data Science. We will start with the challenges in designing runtime environments for MPI+X programming models by considering support for multi-core systems, high-performance networks (InfiniBand, RoCE, Slingshot), GPUs (NVIDIA, AMD, and Intel), and emerging BlueField-3 DPUs. Features and sample performance numbers of using the MVAPICH libraries over a range of benchmarks will be presented. For the Deep/Machine Learning domain, we will focus on MPI-driven solutions (MPI4DL) and Mix-and-match Communication Runtime (MCR-DL) to extract performance and scalability for popular Deep Learning frameworks (TensorFlow and PyTorch), large out-of-core models, Bluefield-3 DPUs, and parallel inferencing. Finally, we will focus on MPI-driven solutions to accelerate Big Data applications (MPI4Spark) and data science applications (MPI4Dask) with appropriate benchmark results will be presented.

Keywords: HPC, AI, Big Data, Data Science, Converged middleware

Sparse matrix ordering for fine grain parallel triangular solve using SIMD

Aboul-Karim Mohamed El Maarouf[1], Luc Giraud[2], Abdou Guermouche[2,3], Thomas Guignon[1]
[1]IFP Energies nouvelles, France
[2]Inria Centre of Bordeaux University, France
[3]LaBRI, Université de Bordeaux, France
aboul-karim.mohamed-el-maarouf@ifpen.fr, Luc.Giraud@inria.fr
abdou.guermouche@labri.fr, thomas.guignon@ifpen.fr

The evolution of processor hardware increasingly supports fine grain parallelism through SIMD vector instruction sets and hardware threading. For instance, the new ARM SVE instruction set allows for hardware implementation of up to 32 double precision SIMD vector sizes per hardware thread. In this work, we focus on vectorization of the triangular solves required in BiCGStab preconditioned with ILU(0) that is particularly numerically effective for IFPEN applications. In our context, expressing some parallelism can be achieved by changing the sparse structure of the matrices through unknown reordering; that can be recast in terms of graph reordering and coloring. We use a graph coloring method called ColorRCM to exhibit fine grain parallelism to feed the SIMD computing units while improving the convergence of the Krylov solver compared to classical greedy graph coloring method. We first evaluate the performance of SIMD-SpTRSV using the permutation provided by ColorRCM and achieve an acceleration between 1.7 and 6 in AVX2 compared to Intel MKL 21.4. Then we examine the impact of ColorRCM ordering on ILU(0)-BiCGStab performance on 201 matrices, including those from the Suite Sparse matrix collection and from the IFPEN porous media flow simulations. The solver configuration uses the ColorRCM ordering and vectorized with AVX2 instructions showed the best convergence times in 90

Keywords: Graph Multi-Coloring and re-ordering, Reverse Cuthill-McKee, Krylov Solver, Sparse Triangular Solve, SIMD

Stabilizing the Block BiCG with Extended Precision: A Case Study

Alexandre Hoffmann, Yves Durand, Jérome Fereyre
Univ. Grenoble Alpes, France
CEA, Grenoble, France
{alexandre.hoffmann, yves.durand, jerome.fereyre}@cea.fr

Problems that require solving the same equation for multiple Right Hand Sides (RHSs) are ubiquitous in physics and chemistry. They typically result in large scale matrix equations, which can be either solved once for each RHS or all-at-once. Krylov subspace projection solvers have low memory requirement, which make them the preferred choice for large scale applications, but tend to less predictible than direct methods. Block Krylov methods offer an interesting alternative to their classical counterparts. By simultaneously solving the problem for all RHSs, it is possible to achieve faster convergence, thus making them an attractive approach for solving matrix equations.

Block Krylov methods, however, are even more sensitive to round-off errors than their classical counterparts. This may result in slow con- vergence and may even lead to divergence. While multiple regulariza- tion techniques have been proposed to improve the convergence of block Krylov methods, their efficacy remains case-dependent.

Previous work has shown that in some synthetic cases, increasing the working precision improves the convergence rate of classical Krylov meth- ods. In the current work, we evaluate the impact of extending the work- ing precision on block-Krylov methods. We first compare different im- plementations of the BLock-BiConjugate Gradient (BL-BiCG) method with and without extended working precision on various problems from the SuiteSparse matrix collection. We then study the impact of working precision on a synthetic problem relevant to medical imaging.

We show that increasing the working precision enables the convergence of the BL-BiCG in all of our considered cases. Moreover, we show that a straightforward implementation of the BL-BiCG in extended preci- sion, converges almost as fast, or even faster than, more sophisticated implementations in double precisions. Additionally, we illustrate, on a case study relevant to medical imaging, how the BL-BiCG behaves and how extended precision alleviates the usual pitfalls of BL-BiCG such as breakdowns and loss of orthogonality. Finally, using this case study, we analyze, the convergence of the BL-BiCG for several RHSs

and several precision.

Keywords: Extended precision, Krylov methods, Linear algebra

The need for accuracy and smoothness in numerical simulations

Carl Christian Kjelgaard Mikkelsen[1], Lorién López-Villellas[2]
[1]Department of Computing Science, Umeå University, 90187 Umeå, Sweden
[2]Departamento de Informática e Ingeniería de Sistemas / Aragón Institute for Engineering Research (I3A), Universidad de Zaragoza, Zaragoza, Spain
spock@cs.umu.se, lorien.lopez@unizar.es

We consider the problem of estimating the error when solving a system of differential algebraic equations. Richardson extrapolation is a classical technique that can be used to judge when computational errors are irrelevant and estimate the discretization error. We have simulated molecular dynamics with constraints using the GROMACS library and found that the output is not always amenable to Richardson extrapolation. We derive and illustrate Richardson extrapolation using a variety of numerical experiments. We identify two necessary conditions that are not always satisfied by the GROMACS library.

Keywords: error estimation, Richardson extrapolation, numerical integration, external ballistics, multi-body dynamics, GROMACS

Trading-off Energy Consumption and Accuracy of a Spectral Method

Thomas Rauber[1], Gudula Rünger[2]
[1]University Bayreuth, Germany
[2]Chemnitz University of Technology, Germany
rauber@uni-bayreuth.de
ruenger@informatik.tu-chemnitz.de

Spectral methods are effective methods for the solution of time-dependent partial differential equations (PDEs). In this article, a Fourier-Galerkin approach is used leading to an approximation method with two steps, consisting of the truncation of the Fourier-Galerkin series and the solution of the resulting ordinary differential equation with a Runge-Kutta solver. The execution of both steps influences the numerical accuracy of the final solution as well as the performance and energy behavior of the solution process. In this article, the effect of the influencing parameters for both steps is investigated. In particular, the question which combination of the influencing parameters leads to the smallest energy consumption for a required numerical accuracy is addressed. A guideline to select the best combination for a required numerical accuracy is derived.

Keywords: spectral method, Fourier-Galerkin approach, solver for ordinary differential equations, numerical accuracy, energy consumption

Enabling mixed-precision with the help of tools: A Nekbone case study

Yanxiang Chen[1], Pablo de Oliveira Castro[2], Paolo Bientinesi[1], Roman Iakymchuk[1,3]
[1]Umeå University, Sweden
[2]Université Paris-Saclay, UVSQ, LI-PaRAD, France
[3]Uppsala University, Sweden
firstname.lastname@umu.se pablo.oliveira@uvsq.fr

Mixed-precision computing has the potential to significantly reduce the cost of exascale computations, but determining when and how to implement it in programs can be challenging. In this article, we consider Nekbone, a mini-application for the CFD solver Nek5000, as a case study, and propose a methodology for enabling mixed-precision with the help of computer arithmetic tools and roofline model. We evaluate the derived mixed-precision program by combining metrics in three dimensions: accuracy, time-to-solution, and energy-to-solution. Notably, the introduction of mixed-precision in Nekbone, reducing time-to-solution by 40.7

Keywords: Mixed-precision, Computer arithmetic tool, Verificarlo, Roofline model, Conjugate Gradient, Nekbone, Energy-to-solution

Measuring and Interpreting Dependent Task-based Applications Performances

Romain Pereira[1], Thierry Gautier[1], Adrien Roussel[2], Patrick Carribault[2]
[1]Avalon, LIP, ENS, Inria, Lyon, France
[2]LIHPC, CEA, DAM, DIF, F-91297, Arpajon, France

Breaking down the parallel time into work, idleness, and overheads is crucial for assessing the performance of HPC applications, but difficult to measure in asynchronous dependent tasking runtime systems. No existing tools allow its measurement portably and accurately.

This paper introduces POT: a tool-suite for dependent task-based applications performance measurement. We focus on its low-disturbance methodology consisting of task modeling, discrete-event tracing, and post-mortem simulation-based analysis. It supports the OMPT standard OpenMP specifications. We evaluate the accuracy of POT's parallel time breakdown analysis on LLVM and MPC implementations and shows that measurement bias may be neglected above 16 us workload per task, portably across two architectures and OpenMP runtime systems

Keywords: Performances, Tasks, Time Breakdown, OpenMP

Using Parallel Performance Data to Classify Parallel Algorithms

Michael McKinsey[1], Stephanie Brink[2], Olga Pearce[1,2]
[1]Texas A&M University, College Station, TX, USA
[1]Lawrence Livermore National Laboratory, Livermore, CA, USA
olga@llnl.gov

Can we tell which parallel algorithm is executing by looking at the performance of the algorithm? In this work, we design and demonstrate a study of parallel algorithm classification for parallel sorting algorithms. We leverage Caliper to collect the performance data, and Thicket for our exploratory data analysis (EDA). We develop a workflow with PyTorch and Scikit-learn to evaluate the effectiveness of decision trees, neural networks, and support vector machines (SVMs) on parallel performance data. We demonstrate classification accuracy of 92.6

Keywords: Parallel Algorithms, Parallel Performance Data, Algorithm Clasification

Tracing of GPU-aware MPI Applications: the First Benchmarks for the Angara Interconnect

Timur Ismagilov[1], Anatoly Mukosey[1], Vladislav Galigerov[1,2], Yuri Grishichkin[1], Felix Smirnov[2], Vladimir Stegailov[1,2], Alexey Timofeev[1,2]
[1]Joint Institute for High Temperatures of RAS, Moscow, Russia
[2]HSE University, Moscow, Russia
tismagilov@mail.ru, mukosey@nicevt.ru, vladgl3@yandex.ru, {gyg, timofeev}@jiht.ru, fealsmirnov@edu.hse.ru, stegailov@gmail.com

The efficiency of data transfer is one of the most important issues of supercomputer development in the post-Moore era. The rise of heterogeneous computing systems introduces such complicated patterns of data transfers as, for instance, the GPU-aware MPI technology. The practical deployment of this technology in applications requires the analysis tools for tracing the runtime behavior of the corresponding algorithms. In this work we analyze the execution patterns of the rocHPL benchmark using the Score-P infractructure. This analysis allows us to make a comparison of the prototype GPU-aware MPI implementation for the Angara Interconnect with its InfiniBand implementation.

Keywords: Low-latency communication, RDMA, HPL, Score-P, HIP

Exploring the Design Space for Message-Driven Systems for Dynamic Graph Processing using CCA

Bibrak Qamar Chandio, Maciej Brodowicz, Thomas Sterling
Indiana University Bloomington, USA
bibrakc@gmail.com, mbrodowi@iu.edu, tron@iu.edu

Computer systems that have been successfully deployed for dense regular workloads fall short of achieving scalability and efficiency when applied to irregular and dynamic graph applications. Conventional computing systems rely heavily on static, regular, numeric intensive computations while High Performance Computing systems executing parallel graph applications exhibit little locality, spatial or temporal, and are fine-grained and memory intensive. With the strong interest in AI which depend on these very different use cases combined with the end of Moore's Law at nanoscale, dramatic alternatives in architecture and underlying execution models are required. This paper identifies an innovative non-von Neumann architecture, Continuum Computer Architecture (CCA), that redefines the nature of computing structures to yield powerful innovations in computational methods to deliver a new generation of highly parallel hardware architecture. CCA reflects a genus of highly parallel architectures that while varying in specific quantities (e.g., memory blocks), share a multiple of attributes not found in typical von Neumann machines. Among these are memory-centric components, message-driven asynchronous flow control, and lightweight out-of-order execution across a global name space. Together these innovative non-von Neumann architectural properties guided by a new original execution model will deliver the new future path for extending beyond the von Neumann model. This paper documents a series of interrelated experiments that together establish future directions for next generation non-von Neumann architectures, especially for graph processing.

Keywords: Processing In Memory, Post Moore Computing, Non von-Neumann Architectures, Asynchronous Dynamic Graph Processing

Introducing the Arm-membench Throughput Benchmark

Cyrill Burth, Markus Velten, Robert Schoene
ZIH, CIDS, TU Dresden
Dresden, Germany
cyrill.burth@mailbox.tu-dresden.de
{markus.velten, robert.schoene}@tu-dresden.de

Application performance of modern day processors is often limited by the memory subsystem rather than actual compute capabilities. Therefore, data throughput specifications play a key role in modeling application performance and determining possible bottlenecks. However, while peak instruction throughputs and bandwidths for local caches are often documented, the achievable throughput can also depend on the relation between memory access and compute instructions. In this paper, we present an Arm version of the well established x86-membench throughput benchmark, which we have adapted to support all current SIMD extensions of the Armv8 instruction set architecture. We describe aspects of the Armv8 ISA that need to be considered in the portable design of this benchmark. We use the benchmark to analyze the memory subsystem at a fine spatial granularity and to unveil microarchitectural details of three processors: Fujitsu A64FX, Ampere Altra and Cavium ThunderX2. Based on the resulting performance information, we show that instruction fetch and decoder widths become a potential bottleneck for cache-bandwidth-sensitive workloads due to the load-store concept of the Arm ISA.

Keywords: Arm, benchmark, throughput, bandwidth, cache, A64FX, ThunderX2, Ampere Altra, performance analysis, microarchitecture, computer architecture

Segmentation of Aortic Valve Calcium Lesions using FPGA Accelerators

Valentina Sisini[1,2], Andrea Miola[1,2], Giada Minghini[1], Enrico Calore[2],
Armando Ugo Cavallo[3], Sebastiano Fabio Schifano[1,2], Cristian Zambelli[1]
[1]Universitá degli Studi di Ferrara, Ferrara, Italy
[2]INFN Sezione di Ferrara, Ferrara, Italy
[3]Istituto Dermopatico dell'Immacolata (IDI) IRCCS, Rome, Italy
{valentina.sisini, andrea.miola, giada.minghini}@unife.it

Semantic segmentation is the task of assigning a class to every pixel of an image, widely used to automatically locate objects in the context of computer vision applications, such as autonomous vehicles, robotics, agriculture, gaming, and medical imaging. Deep Neural Network models like the Convolutional Neural Networks (CNN) are suitable to this extent. Among the plethora of models, the UNet model is widely adopted in bio-medical imaging. The segmentation using CNNs is efficiently performed using GPU accelerators. FPGA devices are also emerging as novel technologies, especially for performing inferences, promising higher energy efficiency and lower latency solutions. In this contribution, we assess the use of FPGA-based accelerators for the inference task using the UNet model. The calcium segmentation in the cardiac aortic valve computer tomography scans is devised as a benchmark application. In particular, we show how to port and deploy a CNN model on such devices and compare the accuracy, throughput, and energy efficiency benchmarking with recent CPUs and GPUs.

Keywords: Semantic Segmentation, AI, UNet, Accelerated Computing, FPGA, GPU, HPC

Improved GPU memory management in evolutionary decision tree induction for large-scale data

Krzysztof Jurczuk, Daniel Reska, Marcin Czajkowski, Marek Kretowski
Bialystok University of Technology
Bialystok, Polnd
{k.jurczuk,d.reska,m.czajkowski,m.kretowski}@pb.edu.pl

Decision trees (DTs) are explainable machine learning techniques applicable to classification and regression problems. Traditionally, DTs are built through a top-down greedy search, which is usually fast but may lead to sub-optimal solutions. An alternative approach involves the use of evolutionary algorithms (EAs), which allow for more global exploration that can yield simpler and accurate DTs. However, the EA-based DT induction is computationally demanding, especially for large-scale data. To alleviate these high computing requirements, various parallel and distributed accelerations are continuously explored.

In this paper, we focus on a GPU-supported solution and extend it by improving memory management. To reduce synchronization points between GPU threads (both within and across blocks), the new solution avoids atomic functions. However, such an approach requires allocating individual result buffers for each thread, which increases memory requirements. To compensate for this, a compact in-memory representation of DTs is additionally applied. Moreover, an additional level of reduction is necessary. Experimental validation on various datasets, both artificial and real-life, shows that the enhanced solution further accelerates the EA-based DT induction. The results also reveal that the time savings increase as the dataset size grows and are influenced by DT size. Therefore, the hybrid solution provides the best time results, by applying synchronization avoidance for smaller DTs and the opposite strategy for larger ones.

Keywords: Evolutionary algorithms, Explainable machine learning, Decision tree, GPGPU, Memory management

Multi-GPU Accelerated Rendering of Massive Scenes with Out-of-Core Support for CPU Memory

Milan Jaros, Lubomir Riha, Petr Strakos, Tomas Kozubek
IT4Innovations, VSB – Technical University of Ostrava
Ostrava, Czech Republic
{milan.jaros, lubomir.riha, petr.strakos, tomas.kozubek}@vsb.cz

We present a new out-of-core method for multi-GPU path tracing of large scenes based on memory access analysis. Our approach allows us to render massive scenes efficiently on a computer system with multiple GPUs and less total memory of all graphics cards than the total size of the scene. The scene is partitioned among GPU memories and the main CPU memory based on a method that operates at the memory management level. Specific parts of the scene are either replicated or distributed in the memory of GPUs, and the rest is located in the main CPU memory.

Keywords: Multi-GPU Path Tracing, NVLink, CUDA Unified Memory, Data Distributed Path Tracing, Distributed Shared Memory Path Tracing, Out-of-Core

A GPU Implementation of McMurchie-Davidson Algorithm for Two-Electron Repulsion Integral Computation

Haruto Fujii[1], Yasuaki Ito[1], Nobuya Yokogawa[1], Kanta Suzuki[1], Satoki Tsuji[1,2], Koji Nakano[1], Akihiko Kasagi[2]
[1]Hiroshima University, Higashi-Hiroshima, Japan
[2]Fujitsu Limited, Kawasaki, Japan

Computational quantum chemistry employs quantum mechanics to investigate the electronic structure of molecules and atoms. While the Schrödinger equation forms the foundation of this approach, its application to molecules with multiple nuclei and electrons poses significant computational challenges. To address this complexity, approximate methods like the Hartree-Fock method are commonly used. An integral part of Hartree-Fock calculations is the two-electron repulsion integral (ERI), computed for each combination of basis functions representing electron orbitals. Although several algorithms exist for ERI calculations, including the widely used McMurchie-Davidson (MD) algorithm, their implementation on GPUs is constrained by recursive formulas. This paper proposes an eficient parallel MD algorithm for GPUs and presents its implementation. Specifically, we introduce batches as parallel processing units for recursive computation, enabling efficient MD algorithm implementation on GPUs. By utilizing batches, parallel computation becomes feasible even with the small amount of shared memory of GPUs. Moreover, our approach allows computation with a single function without the need for separate functions for each combination of azimuthal quantum numbers representing electron orbitals. Experimental results show that the proposed GPU implementation can perform the ERI computation up to 101 and 24 times faster than the CPU implementation on an AMD EPYC 7702 CPU for monatomic and polyatomic molecules, respectively.

Keywords: Two-electron repulsion integrals, McMurchie-Davidson algorithm, Quantum chemistry, Parallel algorithm, GPU

Flexible algorithms for persistent MPI allreduce communication

Andreas Jocksch[1], C. Nicole Avans[2], Riley Shipley[2], Anthony Skjellum[2]
[1]ETH Zurich / CSCS, Swiss National Supercomputing Centre
Lugano, Switzerland
[2]Tennessee Technological University, Cookeville, TN, USA
andreas.jocksch@cscs.ch
{cnavans42, rpshipley, askjellum}@tntech.edu

Modern supercomputers feature an ever-increasing degree of parallelism, especially in the number of cores per node. These high core counts are considered in our flexible implementation of persistent allreduce (an MPI-4 feature), which was implemented specifically with shared-memory communication in mind. At a high level, our algorithm consists of a reduce scatter stage followed by an all-gather stage, and allows for different factors (multi-radix) to be applied at each. Where barriers are required, they are integrated into the algorithm, using counters to track progress. In order to accommodate the complexity of this approach, our implementation is split into a setup phase and an execution phase. The setup phase only occurs once for a given set of parameters, and is responsible for determining the algorithm that will be run each time the allreduce is called in the execution phase. Using these methods, we achieve speedups of half an order of magnitude compared to the blocking and persistent allreduce implementations of MPICH and OpenMPI, on a dual socket node with AMD EPYC processors and almost an order of magnitude on a four socket node with NVIDIA Grace processors. Our implementation also achieves good performance on multiple nodes.

Keywords: MPI, collective, communication, allreduce

Towards the Democratization and Standardization of Dynamic Resources with MPI Spawning

Sergio Iserte[1], Iker Martín Álvarez[2], Krzysztof Rojek[3], José I. Aliaga[2], Maribel Castillo[2], Antonio J. Peña[1]
[1]Barcelona Supercomputing Center, Spain
[2]Universitat Jaume I, Spain
[3]Czestochowa University of Technology, Poland
{siserte, antonio.pena}@bsc.es
{martini, aliaga, castillo}@uji.es
krojek@icis.pcz.pl

This paper presents an efficient tool for managing dynamic resources in production high-performance computing (HPC) settings, focusing on flexibility, adaptability, and user-friendliness. We introduce a unified dynamic resource management application programming interface (API) that supports a wide range of HPC applications, allowing seamless integration without direct interaction with the dynamic resource manager (DMR). The DMR framework, evolved from the DMRlib structure, now supports various dynamic resource managers and includes the Proteo reconfiguration engine to enhance malleability strategies. This integration addresses previous limitations by allowing diverse reconfiguration methods without respawning all processes or lacking RMS support.

The paper also showcases the solution's performance and coding productivity with the MPDATA (Multidimensional Positive Definite Advection Transport Algorithm) application. Key contributions include an enhanced modular DMR framework supporting different reconfiguration managers, upgraded DMRlib with the Proteo reconfiguration engine, offering extensive reconfiguration strategies, and a malleable version of the MPDATA solver.

A comprehensive evaluation of malleable workloads using various strategies and policies. These advancements underscore the potential for improved dynamic resource management in HPC, fostering better scientific outcomes and resource efficiency.

Keywords: DMR, Proteo, MPDATA, Dynamic Resource Management, Malleability

Compiler support for semi-manual AoS-to-SoA conversions with data views

Pawel Radtke, Tobias Weinzierl
Department of Computer Science, Durham University
Durham, United Kingdom
{pawel.k.radtke,tobias.weinzierl}@durham.ac.uk

The C programming language and its cousins such as C++ stipulate the static storage of sets of structured data: Developers have to commit to one, invariant data model—typically a structure-of-arrays (SoA) or an array-of-structs (AoS)—unless they manually rearrange, i.e. convert it throughout the computation. Whether AoS or SoA is favourable depends on the execution context and algorithm step. We propose a language extension based upon C++ attributes through which developers can guide the compiler what memory arrangements are to be used. The compiler can then automatically convert (parts of) the data into the format of choice prior to a calculation and convert results back afterwards. As all conversions are merely annotations, it is straightforward for the developer to experiment with different storage formats and to pick subsets of data that are subject to memory rearrangements. Our work implements the annotations within Clang and demonstrates their potential impact through a smoothed particle hydrodynamics (SPH) code.

Keywords: Array-of-structs, struct-of-arrays, memory layout transformations, data views, compiler, vectorisation

Cultural heritage 3D object management with integrated Automation Workflows

Michał Orzechowski[1,2], Łukasz Opioła[1], Ignacio Lamata Martínez[3], Marinos Ioannides[4], Panayiotis N. Panayiotou[4], Renata G. Słota[2], Łukasz Dutka[1], Jacek Kitowski[1,2]
[1]Academic Computer Centre Cyfronet AGH, Krakow, Poland
[2]AGH University of Krakow, Faculty of Computer Science, Poland
[3]EGI Foundation, Amsterdam, The Netherlands
[4]Digital Heritage Lab, Cyprus University of Technology, Cyprus
{m.orzechowski, l.opiola, lukasz.dutka}@cyfronet.pl
{rena, kito}@agh.edu.pl, ignacio.lamata@egi.eu
{p.panayiotou, marinos.ioannides}@cut.ac.cy

The complexity of high-quality 3D digitalised cultural heritage objects creates challenges for existing data management systems as they need to develop metadata management and processing capabilities to provide semantic insight into the interconnectivity of data that constitute cultural heritage objects. We propose a data and metadata management system, together with the federated authentication and authorisation mechanism, and an integrated system for designing and executing automated workflows that facilitate the processing of both data and metadata. The solution is evaluated with a 3D digitalised cultural object of Lambousa Fishing Boat and presents the complete process from data upload to the publication of the cultural object.

Keywords: data management, workflow processing, automation workflows, metadata management, paradata, data repository, cultural heritage, cultural heritage objects, 3D digitalisation

Collaborative Learning as a Service – a blueprint for a cloud based rural IoTs deployment facility

Henryk Krawczyk, Bogdan Wiszniewski
Gdansk University of Technology
Gdansk, Poland
{hkrawk,bogwiszn}@pg.edu.pl

Vast spaces with inadequate telecommunications infrastructure pose a challenge to deploy IoT systems. A tech stack is proposed and implemented on TASKcloud operated by Gdansk Tech, based on widely available open-source technology components, making it possible to deploy machine learning models developed on the cloud to constrained end devices, to make them capable of intelligently cleaning measurement data and optimizing their volume to save available bandwidth. Thus processed data are transferred from end devices via a nomadic edge UAV based gateway to the cloud instance for further processing.

Keywords: Constrained device, Intelligent sensor, Nomadic computing, Data cleaning

Parallel maximal common subgraphs with labels for molecular biology

Wilfried Agbeto[1], Camille Coti[2], Vladimir Reinharz[1]
[1]Université du Québec à Montréal, Canada
[2]École de Technologie Supérieure, Canada
agbeto.kossi_wilfried@courrier.uqam.ca,
reinharz.vladimir@uqam.ca, camille.coti@etsmtl.ca

Advances in graph algorithmics have allowed in-depth study of many natural objects from molecular biology or chemistry to social networks. Particularly in molecular biology and cheminformatics, understanding complex structures by identifying conserved sub-structures is a key milestone towards the artificial design of novel components with specific functions. Given a dataset of structures, we are interested in identifying all maximum common connected partial subgraphs between each pair of graphs, a task notoriously NP-Hard.

In this work, we present 3 parallel algorithms over shared and distributed memory to enumerate all maximal connected common sub-graphs between pairs of arbitrary multi-directed graphs with labels on their edges. We offer an implementation of these methods and evaluate their performance on the non-redundant dataset of all known RNA 3D structures. We show that we can compute the exact results in a reasonable time for each pairwise comparison while taking into account a much more diverse set of interactions—resulting in much denser graphs—resulting in an order of magnitude more conserved modules. All code is available at https://gitlab.info.uqam.ca/cbe/pasigraph and results in the branch results.

Keywords: Common subgraphs, Parallel algorithms, Distributed memory, Molecular structure, Multi-directed graphs

PPQSort: Pattern Parallel Quicksort

Gabriel Hévr, Ivan Šimeček
Department of Computer Systems, Faculty of Information Technology
Czech Technical University in Prague, Czech Republic
{hevrgabr,xsimecek}@fit.cvut.cz

This paper presents PPQSort (parallel pattern quicksort), a new parallel quicksort algorithm that provides high performance and ease of use. PPQSort uses C++ threads for parallelization, achieving efficient sorting without external libraries and allowing seamless integration across different computing environments. This paper describes novel quicksort optimizations, including branchless partitioning and their efficient parallel implementation. PPQSort is compared with existing parallel quicksort algorithms on different machines and with different input data. Experimental evaluation results demonstrate that PPQSort is fast and robust, consistently outperforming the fastest available parallel quicksort implementations for almost all inputs.

Keywords: parallel sorting, parallel algorithm, quicksort, C++, in-place sorting, multithreading, shared memory

ACE: Algorithm-independent Acceleration and Parallelization of Clustering Implementations

Muyeed Ahmed, Iulian Neamtiu
New Jersey Institute of technology, USA
{ma234, ineamtiu}@njit.edu,

Clustering is a key technique in a wide range of data analysis tasks. However, algorithms that ensure stable, deterministic, accurate clustering are computationally expensive, having superlinear complexity for both memory and time. Therefore, even when using HPC hardware, there are hard limits to dataset sizes that can be clustered, as clustering implementations can run out of memory or take unacceptably long. We introduce an approach called ACE that applies algorithm-independent, black-box parallelization to superlinear sequential clustering algorithms, thereby making the clustering of substantial datasets feasible, even on commodity desktop/laptop systems. ACE starts by partitioning data to fit onto a given machine, and via divide-and-conquer, reduce the complexity of clustering steps. Next, ACE uses parallel, automated hyper-parameter search to find optimal parameters for the current dataset. Finally, ACE aggregates intermediate results effectively and efficiently so that the final clustering output does not sacrifice clustering quality compared to the original algorithm. An evaluation on four popular clustering algorithms – Affinity Propagation, DBSCAN, Hierarchical Agglomerative Clustering, and Spectral Clustering – shows that ACE substantially reduces memory requirements and achieves linear processing time. ACE was able to process an entire suite of 164 datasets, including substantial datasets with 1.4M points or 1,000 dimensions, whereas the default implementations failed to process between 15 and 149 datasets from the suite. Moreover, for those datasets that could be processed by the default implementations, ACE achieved a 1.13x–102x time reduction.

Keywords: Machine Learning, Clustering, Automatic Parallelization

Exploration of performance-energy correlation for CFD codes on cluster with AMD CPUs

Marcin Lawenda[1], Łukasz Szustak[2], László Környei[3]
[1]Poznan Supercomputing and Networking Center, Poznań, Poland
[2]Czestochowa University of Technology, Częstochowa, Poland
[3]Széchenyi István Egyetem-University of Győr, Győr Egyetem, Hungary
lawenda@man.poznan.pl, lszustak@icis.pcz.pl
laszlo.kornyei@math.sze.hu

This work explores the importance of performance-energy correlation for CFD codes, highlighting the need for sustainable and efficient use of clusters. The prime goal includes the optimization of selecting and predicting the optimal number of computational nodes to reduce energy consumption and/or improve calculation time. In this work, the utilization cost of the cluster, measured in core-hours, is used as a crucial factor in energy consumption and selecting the optimal number of computational nodes. The work is conducted on the cluster with AMD EPYC Milan-based CPUs and OpenFOAM application using the Urban Air Pollution model. In order to investigate performance-energy correlation on the cluster, the CVOPTS (Core VOlume Points per Time Step) metric is introduced, which allows a direct comparison of the parallel efficiency for applications in modern HPC architectures. This metric becomes essential for evaluating and balancing performance with energy consumption to achieve cost-effective hardware configuration. The results were confirmed by numerous tests on a 40-node cluster, considering representative grid sizes. Based on the empirical results, a prediction model was derived that takes into account both the computational and communication costs of the simulation. The research reveals the impact of the AMD EPYC architecture on superspeedup, where performance increases superlinearly with the addition of more computational resources. This phenomenon enables a priori the prediction of performance-energy trade-offs (computing-faster or energy-save setups) for a specific application scenario, through the utilization of varying quantities of computing nodes.

Keywords: HPC, Cluster, AMD EPYC, Performance optimization, Prediction model, Experimental findings

High-Performance Implementation of the Optimized Event Generator for Strong-Field QED Plasma Simulations

Elena Panova[1], Valentin Volokitin[1], Aleksei Bashinov[2], Alexander Muraviev[2], Evgeny Efimenko[1], Iosif Meyerov[1]
[1]Lobachevsky State University of Nizhni Novgorod,
Nizhni Novgorod, Russian Federation
[2]Institute of Applied Physics of RAS,
Nizhni Novgorod, Russian Federation
meerov@vmk.unn.ru

Numerical simulation of strong-field quantum electrodynamics (SFQED) processes is an essential step towards current and future high-intensity laser experiments. The complexity of SFQED phenomena and their stochastic nature make them extremely computationally challenging, requiring the use of supercomputers for realistic simulations. Recently, we have presented a novel approach to numerical simulation of SFQED processes based on an accurate approximation of precomputed rates, which minimizes the number of rate calculations per QED event. The current paper is focused on the high-performance implementation of this method, including vectorization of resource-intensive kernels and improvement of parallel computing efficiency. Using two codes, PICADOR and hi-χ (the latter being free and publicly available), we demonstrate significant reduction in computation time due to these improvements. We hope that the proposed approach can be applied in other codes for the numerical simulation of SFQED processes.

Keywords: HPC, Strong-Field QED, Plasma Simulation, Performance Optimization, Vectorization, SIMD, Parallel Computing, PICADOR, hi-Chi

GPU-Based Interval Optimization in the Context of Optical MIMO Systems

Ekaterina Auer, Andreas Ahrens, Lorenz Gillner
University of Applied Sciences Wismar
Germany
{ekaterina.auer,andreas.ahrens,lorenz.gillner}@hs-wismar.de

In this paper, we examine possibilities for global optimization on the GPU using interval-based libraries. We compare a purely brute-force based approach, a monotonicity test based approach and a SIVIA based approach on the GPU with each other and with popular CPU tools in C-XSC and OCTAVE. For that, we employ a problem in the context of optical miltiple-input multiple-output (MIMO) systems. We demonstrate that, although the relatively simple minimization problem we consider is difficult to solve by means of general global optimization, GPU-based methods are able to provide a solution, which is even verified if GPU implementations of the underlying libraries are verified.

Keywords: Interval analysis, GPU, MIMO

Preconditioning of the One-Sided Block-Jacobi SVD Algorithm by Polar Decomposition

Martin Bečka, Gabriel Okša
Institute of Mathematics, Slovak Academy of Sciences
Bratislava, Slovak Republic
{Martin.Becka, Gabriel.Oksa}@savba.sk

We propose the use of the polar decomposition for the preconditioning of the one-sided block-Jacobi algorithm for the singular value decomposition of a given matrix A. The preconditioner comes from the eigenvalue decomposition of the Hermitian factor H, which is computed by using (partial) Halley's iterations. This approach eliminates the computation of the Gram matrix A^TA, which is not numerically reliable for very ill-conditioned matrices A. The iterated matrix in Halley's iterations has a special structure, and three variants for its QR decomposition are proposed and compared. Numerical experiments show, that this new approach is efficient for very ill-conditioned matrices, whereas the Gram matrix can be safely used in other cases.

Keywords: singular value decomposition, one-sided block-Jacobi algorithm, polar decomposition, Halley's iterations

Comparison of Two QUBO Formulations of Approximate Block Diagonalization and Their Performance on the D-Wave Advantage Quantum Annealing Machine

Koushi Teramoto, Shuhei Kudo, Yusaku Yamamoto
The University of Electro-Communications
Tokyo, Japan
t2331105@gl.cc.uec.ac.jp

We consider the problem of transforming a given symmetric matrix as close to block diagonal as possible by symmetric permutations of its rows and columns. Such a problem arises, for example, as a preprocessing for the block Jacobi method for the symmetric eigenvalue problem. To solve this problem on a quantum annealing machine, Teramoto et al. proposed its QUBO (Quadratic Unconstrained Binary Optimization) formulation and strategies for embedding the resulting QUBO into the Pegasus network of the D-Wave Advantage quantum annealer. In this paper, we propose two more embedding strategies based on the minimum-cost flow and integer multi-commodity flow. We also propose an alternative QUBO formulation for which the connectivity graph has a more local structure. Numerical experiments show that our new QUBO formulation requires less physical qubits when embedding the problem into D-Wave Advantage's Pegasus network.

Keywords: block diagonalization, combinatorial optimization, QUBO, minor embedding, Pegasus network, D-Wave Advantage, block Jacobi method, symmetric eigenvalue problem

HEAPS: a novel energy-based configurable HPC scheduler

Esteban Stafford, Luis Cruz, Jose Luis Bosque
Deparment of Computer Engeeniring and Electronics
Universidad de Cantabria, Spain
{stafforde, luis.cruz, bosquejl}@unican.es

High Performance Computing (HPC) is carried out in large computer clusters to support complex scientific applications and simulations. These infrastructures are notoriously power-hungry, and there is a pressing need to optimise their usage. In addition, modern clusters are often considered heterogeneous as they consist of groups of nodes with different characteristics. This article presents a novel scheduling algorithm, Heterogeneous Energy-Aware Pairing Scheduler (HEAPS), that attempts to reduce energy consumption by performing an a-priori estimation of energy consumption of jobs in the available nodes to obtain the best job-node pairings to reduce energy consumption. The evaluation compares its behaviour to that of classic homogeneous schedulers and state-of-the-art heterogeneous schedulers, showing that some configurations of HEAPS are capable of reducing makespan in 18% and energy consumption in 7%, in different cluster types and sizes.

Keywords: Workload manager, Heterogeneous clusters, Energy consumption, Energy efficiency

Fair-Sharing Simulator for Batch Computing Systems

Dalibor Klusacek
CESNET, Prague, Czech Republic
klusacek@cesnet.cz

Scientific computing centers or private (in-house) cloud data centers do not rely on the standard pay-as-you-go business model which is common in commercial clouds to allocate resources. Instead, the system is typically shared by a set of selected users, and the administrator's job is to ensure that resources are shared fairly given the existing policies of that organization. One common approach, especially in batch systems, is to deploy a fairshare-based prioritization in the scheduler, where a prioritization mechanism balances resource consumption so that individual users get the right shares of resources over time. In this paper, we present a tool developed to simulate the fairshare setting in a batch system. Using a set of experiments, we demonstrate the utility of this tool in tuning fairshare settings in a standard HPC/HTC scheduler and present the impact of often-overlooked additional options for modifying the basic fairshare settings. All the findings in this paper are based on our real-world experience of running and optimizing a distributed national computing infrastructure in the Czech Republic.

Keywords: Fair-sharing, Scheduling, Visualization, HPC, HTC

Scalability and Reliability of Port Simulation Workflow on Slurm

Maciej Drozdowski, Jakub Wawrzyniak, Jakub Marszalkowski
Institute of Computing Science,
Poznań University of Technology, Poland
{Maciej.Drozdowski, Jakub.Wawrzyniak,
Jakub.Marszalkowski}@cs.put.poznan.pl

Parallelizing of container terminal seaside layout optimizer using a slurm cluster is considered in this paper. Maritime container terminals have quays divided into discrete berth segments. The number of berths and their lengths have to be adjusted to the arriving vessel traffic for high quality of service. A particular partition of the terminal quay into berths is called terminal layout. The vessel traffic is represented by the stochastic ship traffic model (STM). An instantiation of the STM is a particular ship arrival scenario. Evaluation of a quay layout for an arrival scenario requires scheduling vessels on the berths which is a classic Berth Allocation Problem (BAP) of maritime logistics. The problem of partitioning terminal quay into berths, subject to vessel traffic, for high quality of service will be called a Stochastic Quay Partitioning Problem (SQPP). In this paper, we present a hill-climber metaheuristic for SQPP. The process of the partition evaluation is conducted on a slurm cluster by a simple workflow. We report on the scalability and reliability of this workflow.

Keywords: port simulation, berth allocation problem, stochastic optimization, slurm, reliability, scalability

AI-Driven Acceleration of Computational Fluid Dynamic Simulations

Alejandro Gonzalez Barbera, Jaume Luis Gómez, Raul Martínez Cuenca, Sergio Chiva Vicent
Universidad Jaume I, Castellón, Spain
gonzalal@uji.es

In this study, our central aim is to enhance Computational Fluid Dynamics (CFD) simulations by integrating Artificial Intelligence (AI), with a specific focus on approximating predicted fields to converged steady-state solutions. We propose a workflow leveraging a Neural Network (NN) as a predictive model, utilizing an incremental training paradigm to address the substantial temporal expenses associated with dataset creation.

Our investigation centers on 2D CFD cases, particularly within a water tank configuration featuring a momentum source to induce fluid mixing. Through iterative application of our methodology, significant acceleration of 2D simulations is achieved, resulting in efficiency gains of approximately two-fold. Furthermore, the iterative process enables the accumulation of an expanded dataset, fostering potential for further acceleration and scalability of the AI-driven workflow. Our findings reveal a cumulative speed-up of 2x, corresponding to an estimated reduction in computation time of 73% in the build of the entire dataset.

Keywords: AI, CFD, HPC

Large scale energy resources data forecasting and synthetic generation using high-performance computing

Tiago Pinto, Hugo Paredes
INESC TEC and University of Trás-os-Montes e Alto Douro
Vila Real, Portugal
{tiagopinto, hparedes}@utad.pt

This study addresses the critical role of energy resource forecasting in modern power systems, emphasising the challenges posed by the integration of renewable energy sources. These sources introduce significant uncertainty due to their dependence on natural conditions, requiring advanced forecasting techniques to maintain system balance. The research surveys various forecasting methods tailored to different scenarios, highlighting that no single method universally excels. It also notes the difficulty in obtaining high-quality data, proposing the use of generative models to create synthetic datasets that reflect real-world trends. However, the generation and usage of synthetic data for forecasting are computationally intensive. This paper explores the use of advanced computing to manage these demands effectively, enabling efficient forecasting and data generation within reasonable time frames.

Keywords: large scale energy systems, data forecasting, high performance computing, synthetic generation, generative models

Exploration of Performance and Accuracy of AI-Accelerated CFD Simulation on Intel and NVIDIA GPU/CPU Platforms

Kamil Halbiniak, Roman Wyrzykowski, Krzysztof Rojek
Department of Computer Science
Czestochowa University of Technology, Czestochowa, Poland
{khalbiniak, roman, krojek}@icis.pcz.p

Computational Fluid Dynamics (CFD) has emerged as a cornerstone in understanding and optimizing fluid flow phenomena across various engineering disciplines. In recent years, the intersection of CFD with artificial intelligence (AI) has sparked new possibilities, promising accelerated simulations with required accuracy [1].

In our previous work [2], we proposed a methodology aimed at enhancing CFD simulations by integrating two main components: the AI supervisor and the AI accelerator. During the inference stage, the AI accelerator uses the previously trained AI model to predict the simulation results based on a relatively small number of iterations generated by the CFD solver. The AI supervisor dynamically switches between traditional CFD simulation and AI predictions, while the AI accelerator module expedites the process by extrapolating simulation results. Initially, the traditional CFD solver runs for a predetermined number of iterations to establish initial data points, subsequently utilized by a machine learning model to forecast fluid flow dynamics. Upon generating output, the AI supervisor directs the CFD solver to resume simulation based on the predicted data, iteratively alternating between CFD and AI components until convergence is reached. The convergence threshold depends on factors such as the complexity of simulated flow dynamics and the quality of training data.

This paper explores the performance/scalability and accuracy of training the deep learning model used in the AI-accelerated CFD Motorbike simulation [3] executed on Intel and NVIDIA CPU/GPU platforms. The model is based on the variational autoencoder architecture.

In the first part of the study, we investigate mixed-precision techniques that allow optimizing the performance of AI-accelerated simulations using a combination of lower and higher-precision data formats. While in the previous paper [3], only FP32, FP16, and BF16 data formats were studied, in this work, we focus on TF32 floating-point format for GPUs. It is one of the newest formats for representing numbers in AI computation, first introduced in tensor cores of NVIDIA Ampere GPU architecture. While FP32 requires 32 bits for representing numbers,

the TF32 format uses only 19 bits with three additional bits for mantissa compared to BF16. This small overhead aims to achieve practically the same accuracy for AI computation as the FP32 format while providing radical performance gain (about 15 times for NVIDIA H100 SXM GPUs).

The second part of our study is devoted to exploring the influence of key parameters of the execution environment on the performance and accuracy of training for both CPU and GPU architectures. The first parameter is the batch size, representing the number of samples used in one forward and backward pass through the DNN network. This parameter has a direct impact on the accuracy and performance of the training process. The second parameter relates to the oneDNN option, which is used by default for TensorFlow on Intel and AMD CPUs. This option forces performance optimizations and allows reducing the training time noticeable. However, the side effects of using TensorFlow with oneDNN optimizations are changes in the execution order of operations and greater sensitivity to floating-point round-off errors. Increasing the batch size and setting oneDNN option, as applied in our work [3], improves the performance at the cost of the reduced accuracy, which leads to an increased number of iterations on CPU compared to GPU during the inference stage. This work thoroughly investigates the relationship between the performance and accuracy of training the AI model for the Motorbike simulations. It is shown that switching off the oneDNN option improves training accuracy considerably with relatively small performance degradation. Moreover, we demonstrate that the architecture differences between CPUs and GPUs straightforwardly impact the selection of the optimal batch size values for both types of computing platforms regarding the performance and accuracy of training and inference stages.

In the last part of our work, we focus on investigating the performance and scalability of high-end Intel and NVIDIA server-class GPUs, namely Intel Data Center 1550 MAX and NVIDIA H100 (PCI version). We perform an extensive benchmarking and performance analysis of distributed deep learning training on multiple Intel and NVIDIA accelerators. The resulting performance is experimentally assessed on computing platforms containing (i) two servers with four Intel Data Center Max 1550 (Ponte Vecchio architecture) devices each and (ii) two servers with eight NVIDIA H100 PCI GPUs each. To run distributed training on multiple Intel and NVIDIA accelerators, we use the Horovod framework. While for the NVIDIA-based platform, a single Horovod process is pinned to a single H100 GPU, in the case of the Intel-based platform, two Horovod processes are executed within a single accelerator since Intel Max 1550 consists of two compute tiles. The achieved performance results show the good scalability of distributed DNN training on both Intel and NVIDIA platforms. In particular, 16 NVIDIA H100 GPUs and 8 Intel Data Center Max 1550 GPUs (in total, 16 compute tiles) allow achieving 82.4

References:
1. Ricardo Vinuesa and Steven L Brunton. The Potential of Machine Learning to Enhance Computational Fluid Dynamics. https://arxiv.org/pdf/2110.02085, 2021.
2. Krzysztof Rojek, Roman Wyrzykowski, and Pawel Gepner. AI-Accelerated CFD Simulation Based on OpenFOAM and CPU/GPU Computing. In Maciej Paszynski, Dieter Kranzlm¨uller, Valeria V. Krzhizhanovskaya, Jack J. Dongarra, and Peter M. A. Sloot, editors, Computational Science – ICCS 2021, pages 373–385, 2021.
3. Kamil Halbiniak, Krzysztof Rojek, Iserte Sergio, and Roman Wyrzykowski. Unleashing the Potential of Mixed Precision in AI-Accelerated CFD Simulation on Intel CPU/GPU Architectures. In Computational Science – ICCS 2024 (in press), pages 373–385, 2024.

Keywords: HPC, CFD, AI/M, DNN, mixed precision, CPU/GPU, Intel and NVIDIA architectures

A Framework and Methodology for Performance Prediction of HPC Workloads

Júlia Orteu[1], Marc Clascà[1], Jesús Labarta[1,2], Elise Jennings[3], Stefan Andersson[3], Marta Garcia-Gasulla[1]
[1]Barcelona Supercomputing Center, Barcelona, Spain
[2]Universitat Politècnica de Catalunya, Barcelona, Spain
[3]ParTec AG, München, Germany
{julia.orteu, marc.clasca, jesus.labarta, marta.garcia}@bsc.es
{elise.jennings, andersson}@par-tec.com

We outline an approach for predicting the performance of High-Performance Computing (HPC) workloads based on fine-grained architectural metrics. The methodology and workflow presented consists of gathering data from runtime hardware counters across a range of HPC applications and benchmarks and developing an artificial intelligence model based on ensemble tree algorithms. This model is capable of forecasting the performance of other unseen HPC applications. The workflow presented is based on automatic instrumentation without manual code changes. It uses the traces used for performance analysis to train the model, is fast to run and integrates the prediction results with the performance tools, providing an easy way to iterate and fine-tune the model. Through this approach, we prove that a prediction of the instructions per cycle (IPC) metric of unseen applications is possible based on hardware counters that can be obtained with standard performance tools.

Keywords: Performance Prediction, Performance Engineering, Performance Modeling, HPC workflows, Parallel applications, hardware counters, machine learning, performance tools, Regression trees

EM algorithm for cluster of multicore nodes using Level 3 BLAS operations to learn Gaussian mixture models

Wojciech Kwedlo
Faculty of Computer Science
Bialystok University of Technology
Bialystok, Poland
w.kwedlo@pb.edu.pl

In this paper, we explore the problem of learning the parameters of Gaussian mixture models using the expectation-maximization (EM) algorithm. We propose a new parallel formulation of the EM algorithm that utilizes a static decomposition and distributes the learning set and the matrix storing posterior probabilities among MPI processes. During both the E-step and M-step of an EM iteration, each MPI process spawns a team of OpenMP threads that process data in blocks sized to fit the last-level cache. The calculations necessary for obtaining weighted Gaussian densities in the E-step and mixture parameters in the M-step are conducted using optimized level 3 BLAS operations. Hierarchical all-reduce operations are employed to compute the sums required in the M-step, initially among the OpenMP threads within each MPI process and subsequently among all the MPI processes in the parallel application. In our computational experiments, we compared this proposed approach with the traditional method that employs level 2 BLAS operations and assessed its strong scaling on 64 nodes of a compute cluster. The results demonstrate that the proposed method is 1.6 to 4.7 times faster than the conventional approach utilizing level 2 BLAS. The parallel efficiency of our approach on 64 nodes ranges from 60% to 73%.

Keywords: Gaussian mixture models, EM algorithm, MPI, OpenMP, BLAS

MLP-based Adaptive Sampling and Optimization of Laser-Ion Acceleration with Ultra-Short Laser Pulses

Thomas Miethlinger[1,2], Michael Bussmann[1], Thomas Kluge[1]
[1]Helmholtz-Zentrum Dresden-Rossendorf, 01328 Dresden, Germany [2]Technische Universit¨at Dresden, 01069 Dresden, Germany t.miethlinger@hzdr.de

Modern science, including the field of laser-ion acceleration, often faces challenges with high-dimensional, computationally intensive problems under budget and computational constraints. Reliably achieving high ion energies with current laser technologies through ultra-high intensity pulses in near-critical to overdense plasmas remains difficult, necessitating detailed, costly simulations to explore various acceleration mechanisms and optimize outcomes. The complexity of these simulations, driven by nonlinear partial differential equations that allow for multiple different physical regimes, requires precise, efficient and scalable adaptive sampling methods in high-performance computing environments that must balance exploration of new mechanisms with exploitation of known parameter space dependencies. In this work, we propose and investigate a general approach to scalable adaptive sampling that also allows parallel processing. This method we investigate on data obtained from particle-in-cell simulations using multilayer perceptrons (MLPs), due to their inherent flexibility and capacity to support complex dependencies. Results are benchmarked and compared to Bayesian optimization, and we showcase restrictions of MLPs for this method once the intrinsic uncertainty of the acceleration process is being taken into account. Submitted

Keywords: Adaptive Sampling, Machine Learning, Parallel Processing, Particle-in-Cell, Laser-Plasma Physics

On the Correct Use of Application Efficiency to Calculate Performance Portability

Ami Marowka
Parallel Research Labs, Israel
amimar2@yahoo.com

The emergence of heterogeneity in high-performance computing, which harnesses under one integrated system several platforms of different architectures, also led to the development of innovative cross-platform programming models. Along with the expectation that these Smodels will yield computationally intensive performance, there is demand for them to provide a reasonable degree of performance portability. Therefore, new tools and metrics are being developed to measure and calculate the level of performance portability of applications and programming models.

The ultimate measure of performance portability is performance efficiency. Performance efficiency refers to the achieved performance as a fraction of some peak theoretical or practical baseline performance. Application efficiency approaches are the most popular and attractive performance efficiency measures among researchers because they are simple to measure and calculate. Unfortunately, the way they are used yields results that do not make sense, while violating one of the basic criteria that defines and characterizes the performance portability metrics.

In this paper, we demonstrate how researchers currently use application efficiency to calculate the performance portability of applications and explain why this method deviates from its original definition. Then, we show why the obtained results do not make sense and propose practical solutions that satisfy the definition and criteria of performance portability metrics.

Keywords: Performance Portability, Application Efficiency, Metrics

Assessing the Performance of Portable Programming Models Across GPU Vendors for the N-Body Problem

Rodrigo Bartolomeu, Rene Halver, Jan Meinke, Godehard Sutmann
Forschungszentrum Jülich
Jülich Supercomputing Centre
Germany
{r.bartolomeu, r.halver, j.meinke, g.sutmann}@fz-juelich.de

With the inclusion of Aurora in the TOP500 list in November 2023 three different GPU (Graphics Processing Unit) vendors are by now represented in the top 10 each with its own preferred model of programming GPUs. For this paper we implemented the N-body problem using portable programming frameworks and the vendors' preferred APIs. We show how the performance of the portable solutions compares to the performance of the native solution on each hardware both in absolute numbers and as fraction of the achievable peak performance

Keywords: GPU, programming models, performance portability, N-body problem

Performance Portability of SpMV for CSR and BSR Storage Formats Implemented Using OpenACC and SYCL

Kinga Stec, Przemyslaw Stpiczynski
Maria Curie-Skłodowska University
Lublin, Poland
kingastec439@gmail.com
przemyslaw.stpiczynski@mail.umcs.pl

The aim of this paper is to study the performance portability of OpenACC and SYCL implementations of sparse matrix-vector product for CSR and BSR storage formats on Intel CPU and NVIDIA GPU platforms. Using the reformulated performance portability metric PP we show how it changes for various sparse matrices and which implementation and format achieves better performance portability. Numerical experiments show that on CPU for CSR and BSR and on GPU for BSR, OpenACC is better for smaller matrices. On GPU the SYCL implementation for CSR allows to achieve better performance portability almost in all cases.

Keywords: sparse matrices, performance portability, CSR and BSR storage formats, SpMV, OpenACC, SYCL, PP metric

The Impact of SYCL Data Management on Performance Portability

Ami Marowka
Parallel Research Labs, Israel
amimar2@yahoo.com

SYCL programming model does not guarantee performance portability across different architectures. However, the HPC community severely needs platform-independent performance portable applications more than ever. Therefore, the main challenge of SYCL implementers and application developers is to look for direct or indirect solutions in order to improve the portability, performance, and performance portability of SYCL applications.

In this paper we study and analyze the impact of the two main SYCL abstractions for data management, i.e., the unified shared memory and buffer-accessor approaches, on the three pillars of performance portability: portability, productivity, and performance. Experiments were carried out on state-of-the-art CPU and GPU platforms in order to shed light on the effect of different SYCL features on performance. The conclusions that emerged from this study show that by avoiding the use of SYCL performance portability inhibitors, it is possible to develop applications with a realistic level of performance portability.

Keywords: SYCL, Performance portability, Productivity, USM

LLM-driven Cross-Platform Code Generation for Polyhedral Optimized NPDP Codes

Marek Palkowski
West Pomeranian University of Technology in Szczecin
Faculty of Computer Science and Information Systems
Szczecin, Poland
mpalkowski@zut.edu.p

This paper explores using large language models for automatic and source-to-source programming across various cross-platform languages and libraries, including generating CUDA code for HPC. Our proposed solution aims to demonstrate the capabilities of ChatGPT in generating valid CUDA code based on existing OpenMP code obtained from a polyhedral compilers Traco and Dapt. Specifically, we intend to showcase the model's proficiency in producing equivalent and valid GPU code. Furthermore, we extend the application of our solution by prompting the AI model to generate similar code for another input benchmark, emphasizing its adaptability and versatility across different scenarios. This approach seeks to highlight the model's capacity to provide automated and efficient solutions for diverse programming challenges. We chose non-trivial kernels from the NPDP benchmark with non-uniform loops. The focus is on ensuring code validity and understanding limitations in obtaining valid CUDA code. Additionally, we assess the efficiency and scalability of NVIDIA A100 GPU codes generated by AI models, comparing them with optimized Intel Xeon Gold CPU codes from polyhedral optimizers.

Keywords: NPDP Codes, Polyhedral Compilers, LLM, Chat GPT, Code Optimization, CUDA

Juliana: Automated Julia CUDA.jl Code Translation Across Multiple GPU Platforms

Enrique de la Calle, Carlos García
Universidad Complutense de Madrid
Madrid, Spain
encalle@ucm.es, garsanca@ucm.es

Julia is a high-level language that supports executing parallel code through various packages. CUDA.jl is prominently used for developing GPU Julia code across a significant amount of libraries and programs. In this paper Juliana is presented, a new tool that translates Julia code utilizing the CUDA.jl package to an abstract multi-backend representation powered by the KernelAbstractions package. The performance impact of this translation is evaluated, using a custom adaptation of the well established Rodinia benchmark suite to Julia CUDA.jl. For ensuring the viability of the tool from a performance perspective, an accurate overhead statistical analysis using the BenchmarkTools Julia's package is performed, comparing the same benchmark code over the same CUDA device before and after the translation. Additionally, the portability of this approach is demonstrated by running the translated code across multiple backends of KernelAbstractions, allowing the execution of the Rodina benchmark suite to different GPUs vendors such as NVIDIA, Intel, AMD or Apple.

Keywords: Julia, Juliana, Portability, CUDA, GPU

Boosting GPGPU virtualization and multiplexing with RDMA communication

Mariano Aponte[1], Gennaro Mellone[1], Ciro Giuseppe De Vita[1], Diana Di Luccio[1], Giuseppe Salvi[1], Sokol Kosta[2], Raffaele Montella[1]
[1]Department of Science and Technologies
University of Naples "Parthenope ", Italy
[2]Department of Electronic Systems, Aalborg University
Copenhagen, Denmark
mariano.aponte001@studenti.uniparthenope.it
gennaro.mellone@uniparthenope.it

Hardware virtualization is essential in High-Performance Computing (HPC) and Cloud Computing. It addresses resource allocation challenges, scalability, and performance isolation in shared environments. With the rise of GPGPU-accelerated HPC clusters, research has increasingly focused on GPGPU virtualization, leading to various solutions. This work explores the Generalized Virtualization Service (GVirtuS), a transparent virtualization solution with a plug-in framework for easy development and choice of communicators and stub libraries. We will discuss GVirtuS' implementation and design choices and compare them with similar solutions. Additionally, we will show how GVirtuS' performance was enhanced with a novel RDMA-based communicator, evaluated through CUDA implementations of SAXPY and Matrix Multiplication. Our RDMA Communicator achieved a 35% performance boost over the TCP Communicator on Infiniband and a 55

Keywords: GPU, virtualization, multiplexing, RDMA

Efficient Load Scheduling of IMRT Planning in Heterogeneous multicore clusters

S. Puertas-Martín[1,2], Juan José Moreno Riado[1], Juani Lopez Redondo[1], Pilar M. Ortigosa[1], Ester Martin Garzón[1]
[1]Informatics Departament, ceiA3, University of Almería, Almería, Spain
[2]Information School, University of Sheffield, Sheffield, United Kingdom
{savinspm,juanjomoreno,jlredondo,ortigosa,gmartin}@ual.es

IMRT uses radiation beams with different angles and intensities to target cancerous tissues while protecting healthy organs. Planning methods based on the generalized Equivalent Uniform Dose metric produce plans with excellent tumor coverage, but necessitate the adjustment of many parameters. To address this challenge, a novel approach, PersEUD, has been proposed for the automated tuning of these parameters. This is achieved by combining solutions from a Gradient Descent algorithm with an evolutionary optimization method to explore the parameter space efficiently. Previous research has demonstrated the effectiveness of this approach in meeting clinical constraints. However, its high computational demands hinder its integration into clinical practice. The goal of this study is to accelerate the optimization processes by distributing the evaluations in the nodes of modern multicore clusters. At the node level, these evaluations can be efficiently computed with the combination of parallelization and batching strategies. As a consequence, the efficiency of the evaluations depends on the node's load, and the distribution of evaluations among the nodes must account for this dependence. In this study, we propose an approach to integrate an efficient scheduling of evaluations on heterogenous multi-core nodes in PersEUD. The proposal has been extensively tested on eight clusters, with nodes of three different microarchitectures. The test data set consisted of three head and neck patients treated with IMRT using nine beams. The results indicate that exploiting the cluster appropriately leads to a substantial acceleration of the computation involved in the planning based on PersEUD. This result facilitates the practical implementation of PersEUD in clinical settings.

Keywords: Radiotherapy, Load scheduling, Heterogeneous clusters

Deploying AI-Based Environmental Monitoring Applications at the Edge: Two Case Studies

Gianluca De Lucia[1], Giuliano Laccetti[2], Marco Lapegna[2], Raffaele Montella[3], Diego Romano[4]
[1]Centro Alti Studi per la Difesa (CASD), Rome, Italy
[2]Dept. of Mathematics and Applications, Univ. of Naples Federico II, Naples, Italy
[3]Dept. of Science and Technolgy, Univ. of Naples Parthenope, Naples, Italy
[4]Inst. for High Performance Computing and Networking (ICAR), National Research Council (CNR), Naples, Italy
gianluca.delucia.94@gmail.com
{giuliano.laccetti, marco.lapegna}@unina.it
raffaele.montella@uniparthenope.it, diego.romano@cnr.it

The Edge Computing environments enable the development of pervasive applications distributed across extensive geographical areas, overcoming specific issues associated with centralized information processing, such as network bandwidth saturation and the need for large computing infrastructures. This work presents two case studies of deploying applications for environmental monitoring on low-energy and high-performance edge computing devices, employing accelerated artificial intelligence techniques based on GPUs. The first problem entails the classification of various materials within hyperspectral images, while the second problem focuses on identifying floating plastic debris. The applications were validated on a Nvidia Jetson Nano sensor board, demonstrating good accuracy, effectiveness, and energy consumption results.

Keywords: Edge Computing, Environmental Monitoring Applications, AI-based Classification Algorithms, Energy Consumption

Parallelism in GNN: possibilities and limits of several current approaches

Valeria Mele[1], Luisa Carracciuolo[2], Diego Romano[2]
[1]University of Naples "Federico II", Naples, Italy
[2]Italian National Research Council - CNR, Rome, Italy
valeria.mele@unina.it

Graph Neural Networks (GNNs) have emerged as powerful tools for learning on graph-structured data, demonstrating state-of-the-art performance in various applications such as social network analysis, biological network modeling, and recommendation systems. However, the computational complexity of GNNs poses significant challenges for scalability, particularly with large-scale graphs. Parallelism in GNNs addresses this issue by distributing computation across multiple processors, leveraging techniques such as data parallelism and model parallelism. Data parallelism involves partitioning the graph data across different processors, while model parallelism splits the neural network's layers or operations. These parallelization strategies, along with optimizations like asynchronous updates and efficient communication protocols, enable GNNs to handle larger graphs and improve training efficiency. This abstract explores the key approaches and benefits of parallelism in enhancing the scalability and performance of GNNs, highlighting recent advancements and their implications for future research.

Keywords: parallel computing, deep learning, Graph Neural Networks, performance, neural networks

Solving Soil Microbiota Growth Problem by PINNs

Salvatore Cuomo, Francesco Piccialli, Vincenzo Vocca, Donato Cerciello
Department of Mathematics and Applications ″Renato Caccioppoli″,
University of Naples Federico II, Italy
{salvatore.cuomo, francesco.piccialli}@unina.it
vocca.vincenzo.98@gmail.com

An understanding of the growth of the microbiota in soil is of great importance for the health of farms. This research introduces a new method, PINNs, which captures the intricate relationships between soil inhabitants and newly introduced species. PINNs are more accurate than traditional methods and, when combined with parallel computation, achieve excellent results. This approach promises to improve microbial growth predictions, leading to a deeper understanding of soil health and function.

Keywords: Deep Learning, Physics-Informed Neural Network, Numerical methods, PDE, Soil Microbiota Growth, PINN, Biology, Arbuscular Mycorrhizal Fungi, AM Fungi, Parallel and distributed computing

Two-Phase Distributed Algorithm for Solving the Bi-Objective Minimum Spanning Tree Problem: A Preliminary Study

Lavinia Amorosi, Mariagrazia Cairo, Lorenzo Di Rocco, Paolo Dell'Olmo, Umberto Ferraro Petrillo
Department of Statistical Sciences
Università di Roma "La Sapienza", P.le Aldo
Rome, Italy
{lavinia.amorosi,mariagrazia.cairo,paolo.dellolmo}@uniroma1.it
{lorenzo.dirocco,umberto.ferraro}@uniroma1.it

In this paper we present a novel distributed algorithm for solving the Bi-Objective Minimum Spanning Tree (BMST) problem using a two-phase method. The proposed approach leverages the MapReduce computing paradigm to define a distributed version of the most computational intensive part of this method, the second phase. In this phase, a recursive algorithm explores non-supported non-dominated points by evaluating spanning trees within specified regions. The distributed nature of our solution allows parallel processing of multiple triangles in the objective space, significantly reducing execution time and improving scalability. Our preliminary experimental results, conducted using an HPC system, confirm the efficiency and effectiveness of our algorithm with respect to its sequential counterpart, highlighting its potential for practical applications in multi-objective optimization.

Keywords: Bi-Objective Minimum Spanning Tree, Distributed Algorithm, Multi-Objective Optimization, MapReduce

Average Schwarz methods are simply effective

Talal Rahman[1], Leszek Marcinkowski[2]
[1]Western Norway University of Applied Sciences
Bergen, Norway
[2]Faculty of Mathematics, University of Warsaw
Warszawa, Poland
Talal.Rahman@hib.no, lmarcin@mimuw.edu.pl

The first Average Schwarz method was an additive Schwarz method, which was proposed by Bjørstad, Dryja, and Vainniko in the early nineties as a preconditioner for the second order elliptic PDE with jump coefficients. The method is one of the simplest of all additive Schwarz preconditioners because it is easy to construct and quite straightforward to analyze. Unlike most additive Schwarz preconditioners, its local subspaces are defined on non-overlapping subdomains, and it requires no explicit coarse grid as its coarse space is simply defined as the range of an averaging operator. In the recent years, the method has been further developed and extended to solve more complex problems, including multiscale problems, fourth order proembls, and problems with nonconfirming finite elements. The goal of this talk is to present a review of the class of Average Schwarz methods, their theoretical results and implementation issues, illustrating their effectiveness, and discuss its potential to solve nonlinear problems and problems with AI/ML.

Keywords: Average Schwarz methods, Domain Decomposition Methods, Parallel precondiioner

Minimization of Nonlinear Energies in Python Using FEM and Automatic Differentiation Tools

Michal Béreš[1,2], Jan Valdman[2,3]
[1]Institute of Geonics of the Czech Academy of Sciences,
Ostrava, Czech Republic
[2]Department of Applied Mathematics,
VSB - Technical University of Ostrava, Ostrava, Czech Republic
[3]Institute of Information Theory and Automation of the Czech Academy of Science
Prague, Czech Republic
[4]Department of Computer Science, Faculty of Science,
University of South Bohemia,Českée Budějovice, Czech Republic
michal.beres@ugn.cas.cz, jan.valdman@utia.cas.cz

This contribution examines the capabilities of the Python ecosystem to solve nonlinear energy minimization problems, with a particular focus on transitioning from traditional MATLAB methods to Python's advanced computational tools, such as automatic differentiation. We demonstrate Python's streamlined approach to minimizing nonlinear energies by analyzing three problem benchmarks - the p-Laplace, the Ginzburg-Landau, and the Neo-Hookean hyperelasticity. This approach merely requires the provision of the energy functional itself, making it a simple and efficient way to solve this category of problems. The results show that the implementation is about ten times faster than the MATLAB implementation for large-scale problems. Our findings highlight Python's efficiency and ease of use in scientific computing, establishing it as a preferable choice for implementing sophisticated mathematical models and accelerating the development of numerical simulations.

Keywords: nonlinear energy minimization, finite element method, autograd, p-Laplacian, Ginzburg-Landau model, hyperelasticity

Adaptive Parallel Average Schwarz Preconditioner for reduced Hsieh-Clouh-Tocher Macro Element

Leszek Marcinkowski[1], Talal Rahman[2]
[1]Faculty of Mathematics, Informatics, and Mechanics,
University of Warsaw, Warszawa, Poland
[2]Faculty of Engineering and Science,
Western Norway University of Applied Sciences, Bergen, Norway
Leszek.Marcinkowski@mimuw.edu.pl, Talal.Rahman@hvl.no

In this paper, we describe and analyze an Average Schwarz Method with spectrally enriched coarse space for a reduced Hsieh-Clough-Tocher (RHCT) finite element discretization of a 4th-order elliptic multiscale problem. The derived symmetric preconditioner is applied and the PCG iterative method is used to solve the preconditioned problem. If the enrichments of the coarse space contain sufficiently many specially constructed eigenfunctions, then the convergence rate of the PCG method is weakly dependent on the ratio of the coarse to fine mesh h/H.

Keywords: Average Schwarz method, Domain Decomposition, Finite Element Method, Reduced Hsieh-Clough-Tocher macro element

Combining domain decomposition techniques with an operator learning network

Lars Fredrik Lund
Western Norway University of Applied Sciences
Bergen, Norway
lars.fredrik.lund@hvl.no

Using machine learning as a tool to solve and approximate solutions to partial differential equation (PDE) problems has gained and retained popularity in recent years. A common and popular approach is using a physics informed neural network (PINN). This has the advantage of being relatively simple to implement, and one has flexibility in terms of the domain if one uses, say, the finite element method as a framework. Several recent papers have proposed ways to incorporate domain decomposition techniques into the machine learning process for PINNs. For instance, one can divide the computational domain into subdomains and solve each part individually in parallel. This allows for a parallel training of the network on several computational clusters, reducing the overall training time by a factor proportional to the number of subdomains. Some boundary flux conditions or other global information conditions are enforced to tie them together.

A major drawback of using PINNs is that they are only trained on specific problems. This means, that a change in domain, forcing terms or boundary conditions requires the network to be retrained. Since the training process for the neural network can be quite lengthy, this defeats some of the purpose of using a neural network for quick evaluations.

A powerful alternative to PINNs are so-called operator-learning networks, for instance DeepONet or Physics Informed Neural Operator. Instead of solving a specific instance of a problem, these methods learn the behavior/inverses of the differential operators themselves, finding many solutions at once, and thus making these methods generalizable without constant retraining.

In this paper, we investigate further the operator network Unsupervised Legendre-Galerkin network, ULGNet. This network is based on a spectral element framework, with the Legendre polynomials as basis functions. Previous work has shown that this network has achieved highly accurate approximations for various complex problems on simple domains. For instance, non-linear and boundary layer problems. Since ULGNet uses the Legendre-Galerkin framework, one gives up flexibility for accuracy. To extend this method to more complex domains, we utilize techniques from domain decomposition. Namely, splitting the computational

domain into several smaller subdomains, where one can easily apply the spectral neural network to each of them. Then, we impose certain conditions to deal with the boundary problem at their interface. What we end up with is a machine learning algorithm that can produce generalizable solutions with high accuracy for complex problems on complex domains. Further, the training time of the network can be reduced by the factor of subdomains if one utilizes several computational units.

Keywords: Domain decomposition, Machine learning, Operator learning

Comparison of multigrid and machine learning-based Poisson solvers

Hadrien Godé[1], Carola Kruse[1], Richard Angersbach[2], Harald Kostler[2], Michaël Bauerheim[3], Ulrich Rude[1,2]
[1]CERFACS, Toulouse, France
[2]Friedrich-Alexander Universtät Erlangen-Nürnberg, Germany
[3]ISAE Supaero, Toulouse, France

We present a comprehensive comparison of the multigrid method and the UNet architecture for solving Poisson's equation. Those two methods show a lot of similarity and also have the very interesting characteristic that their solving time should scale with the number of mesh nodes. Nevertheless, for Poisson's equation, a strict analysis of the number of floating-point operations demonstrates that the multigrid V-Cycle should be faster than the UNet. We have realized a practical comparison of the two methods solving time for different number of mesh nodes on the same computation nodes (with GPU).

Keywords: Multigrid, UNet, Poisson solver

A system of PDEs fo crowd evacuation - numerical experiments

Maria Gokieli
Cardinal Stefan Wyszynski University
Warsaw, Poland
mgokieli@icm.edu.pl

A hyperbolic continuity equation has been used as a model of pedestrian dynamics for the last twenty years. We regularize it by a diffusive term, intended to be responsible for people's collision avoidance, and provide a number of velocity fields, which should direct the pedestrians towards the exit by the approximately shortest path. We present the results of numerical experiments so as to discuss: the validity of these velocity fields as for the modelling, their numerical efficiency, and the stability of the model. The modelling part focuses on considering diverse geometries of the evacuated space, especially containing obstacles, and the so-called Braess paradox, saying that the obstacles should shorten the evacuation time.

Keywords: Finite Elements Method, crowd dynamics, advection, diffusion, p-Laplacian, semi-implicit scheme

A RISC-V vector CPU for High-Performance Computing: architecture, platforms and tools to make it happen

Filippo Mantovani
Barcelona Supercomputing Center, Spain

The European Processor Initiative (EPI) is a project dedicated to developing a general-purpose processor and an accelerator, alongside the necessary software layers for their integration into the High Performance Computing (HPC) ecosystem. The Barcelona Supercomputing Center is contributing to the development of a RISC-V-based accelerator targeted at HPC applications, leveraging the RISC-V vector extension. This talk aims to provide a comprehensive overview of the EPI project, an introduction to RISC-V, and insights into vector supercomputing. Special emphasis will be placed on the RISC-V vector extensions (RVV), with a particular focus on implementations utilizing large vectors. Participants will gain an understanding of how RVV compares with other vector architectures and explore a design approach that utilizes vectors up to 16-kb-wide. Ultimately, the talk aims to present the methodologies, tools, and libraries available for vectorization, while addressing the accompanying challenges and limitations.

Keywords: RISC-V, vector CPU, HPC

Optimizing Neural Network Classification On Resource Constrained Processors through Custom Compute

Tadej Murovic[1], Keith Graham[2], Peter Robertson[3], Thomas Hepworth[4], Bharathwaj Muthuswamy[5]
[1]Codasip, VP of University Program, Slovenia
[2]Codasip, European University Program Lead, United States
[3]Codasip, Lead AI/ML Engineer, United Kingdom
[4]Codasip, Software Engineer, United Kingdom
[5]Codasip, Senior FPGA Instructor, United States
{tadej.murovic, keith.graham, peter.robertson}@codasip.com
{thomas.hepworth, bharathwaj.muthuswamy}@codasip.com

The realization of Artificial Intelligence (AI) in battery or cost sensitive embedded platforms will require implementations in low energy systems which correlates to resource constrained processors. These processors can be constrained by the lack of arithmetic processing elements such as a vector engine or accelerators, limited to a single general purpose register file, local on-chip memory, low frequency, or technology node. For example, integrating an AI application onto a sensor may require manufacturing on a 40nm semiconductor process instead of 12 or 22nm to be compatible with the analog sensor front end. This larger technology node will limit the effective processor's core frequency and area to satisfy the product cost and energy targets. The non-AI embedded processors have become ubiquitous and can be found in both industry and consumer Internet of Things (IoT), home automation, automotive, and the interface between the digital and real-world domains. To effectively enable AI into these embedded systems, it will require optimizing this resource constrained processor's efficiency, developing targeted Digital Signal Processing (DSP) support, and integrating AI custom instructions that effectively process within the given hardware constraints. To succeed, an optimal software-hardware interface must be defined that enables effective AI applications with low-cost hardware realization. The RISC-V ISA and Codasip IPs and EDA tools enable such a venture, as is presented in the paper.

Keywords: RISC-V, custom compute, design automation, signal processing, AI/ML, DSP

All-in-One RISC-V AI compute engine

Josep Sans
Semidynamics, Spain
josep.sans@semidynamics.com

In this talk we will describe Semidynamic's solution for future-proof AI compute, based on the combination in a single element of Semidynamics RISC-V core, vector and tensor unit. We will cover the new tensor instructions implemented by Semidynamics, how these can be used in AI convolutions and matrix multiplication. We will also cover the need for the vector unit in modern AI models, such as LLMs, to properly run activations

Bio: Josep Sans got his Master's degree in High-Performance Computing from the UPC. He has been working in Semidynamics' verification team for three years, focusing on the Memory Pipeline of the RISC-V cores. Previously he worked for almost two years at Barcelona Supercomputing Center, where he worked in the verification of a RISC-V vector accelerator.

Keywords: RISC-V, AI, vector unit, tensor unit

RAVE: RISC-V Analyzer of Vector Executions, a QEMU tracing plugin

Pablo Vizcaino Serrano[1], Filippo Mantovani[1], Jesus Labarta[1,2], Roger Ferrer[1]
[1]Barcelona Supercomputing Center, Barcelona
[2]Universitat Politècnica de Catalunya
{pablo.vizcaino, filippo.mantovani, jesus.labarta}@bsc.es
roger.ferrer@bsc.es

Simulators are crucial during the development of a chip, like the RISC-V accelerator designed in the European Processor Initiative project. In this paper, we showcase the limitations of the current simulation solutions in the project and propose using QEMU with RAVE, a plugin we implement and describe in this document. This methodology can rapidly simulate and analyze applications running on the v1.0 and v0.7.1 RISC-V V-extension. Our plugin reports the vector and scalar instructions alongside useful information such as the vector-length being used, the single-element-width, and the register usage, among other vectorization metrics. We provide an API used from the simulated Application to control the RAVE plugin and the capability to generate vectorization traces that can be analyzed using Paraver. Finally, we demonstrate the efficiency of our solution between different evaluated machines and against other simulation methods used in the European Processor Accelerator (EPAC) project.

Keywords: QEMU, RISC-V, Vector Extension, Instruction tracing, Simulation

Batched DGEMMs for scientific codes running on long vector architectures

Fabio Banchelli, Marta Garcia-Gasulla, Filippo Mantovani
Barcelona Supercomputing Center, Barcelona, Spain
name.surname@bsc.es

In this work, we evaluate the performance of SeisSol, a simulator of seismic wave phenomena and earthquake dynamics, on a RISC-V-based system utilizing a vector processing unit. We focus on GEMM libraries and address their limited ability to leverage long vector architectures by developing a batched DGEMM library in plain C. This library achieves speedups ranging from approximately 3.5x to 32.6x compared to the reference implementation. We then integrate the batched approach into the SeisSol application, ensuring portability across different CPU architectures. Lastly, we demonstrate that our implementation is portable to an Intel CPU, resulting in improved execution times in most cases.

Keywords: Batched DGEMM, RISC-V, Long Vector, Optimization

Vectorization of Gradient Boosting of Decision Trees Prediction in the CatBoost Library for RISC-V Processors

Evgeniy Kozinov, Evgenii Vasiliev, Andrey Gorshkov, Valentina Kustikova,
Artem Maklaev, Valentin Volokitin, Iosif Meyerov
Lobachevsky State University of Nizhni Novgorod
Nizhni Novgorod, Russia
meerov@vmk.unn.ru

The emergence and rapid development of the open RISC-V instruction set architecture opens up new horizons on the way to efficient devices, ranging from existing low-power IoT boards to future high-performance servers. The effective use of RISC-V CPUs requires software optimization for the target platform. In this paper, we focus on the RISC-V-specific optimization of the CatBoost library, one of the widely used implementations of gradient boosting for decision trees. The CatBoost library is deeply optimized for commodity CPUs and GPUs. However, vectorization is required to effectively utilize the resources of RISC-V CPUs with the RVV 0.7.1 vector extension, which cannot be done automatically with a C++ compiler yet. The paper reports on our experience in benchmarking CatBoost on the Lichee Pi 4a, RISC-V-based board, and shows how manual vectorization of computationally intensive loops with intrinsics can speed up the use of decision trees several times, depending on the specific workload. The developed codes are publicly available on GitHub.

Keywords: RISC-V, Gradient Boosting Trees, Decision Trees, Machine Learning, Performance Analysis, Performance Optimization, Vectorization

QR Factorization on a Long-Vector Processor

Andres Tomas[1], Pablo Vizcaino[2], Enrique S. Quintana-Orti[1], Filippo Mantovani[2]
[1]Universitat Politècnica de València, Spain
[2]Barcelona Supercomputing Center, Spain
antodo@upv.es, quintana@disca.upv.es
{pablo.vizcaino, filippo.mantovani}@bsc.es

We build two high-performance implementations of QR the factorization, based on either Householder reflectors or a variant of the Gram-Schmidt (GS) method, on top of highly parallel linear algebra kernels that we optimize to take advantage of the long vector units in the accelerator developed as part of the European Processor Initiative.

The experimental evaluation of our customized implementations early test-chip of this accelerator shows significant acceleration factors over the LAPACK code running on a scalar RISC-V core. It also shows the superiority of the Householder-based solution for square matrices and the advantages of GS for tall-and-skinny matrices when both the triangular and orthogonal factors are desired.

Keywords: QR factorization, Householder reflectors, classical Gram-Schmidt method, vector units

C software and peripherals support for RISC-V

Ondrej Golasowsk, Jan Medek, Michal Stepanovsky
Faculty of Information Technology,
CTU in Prague, Czech Republic
{golasond, medekja5, stepami9}@fit.cvut.cz

This paper discusses adding C software and peripheral support for a custom single-cycle RISC-V RV32I microarchitecture implemented as a soft-core processor on an FPGA. This allows programmers to program this processor and communicate with peripherals simply in C using the provided library. Currently supported peripherals include GPIO, UART and system uptime timer. The paper presents key ideas on how to extend the RV32I microarchitecture to support these peripherals and how to add support for C language in the absence of operating system on the target device. Moreover, several examples are provided to illustrate the functionality of developed library.

Keywords: RISC-V · C language · Peripherals

Porting Memory-Bound CFD Application to RISC-V Architecture

Tomasz Olas[1], Lukasz Szustak[1], Paweł Gepner[2], Roman Wyrzykowski[1]
[1]Czestochowa University of Technology, Czestochowa, Poland
[2]Warsaw University of Technology, Warsaw, Poland

In this paper, we tackle the challenge of adapting HPC applications to RISC-V architectures for real-world problems with memory-bound codes, for which memory performance is the main factor affecting computation time. The application we study as a use case is the Multidimensional Positive Definite Advection Transport Algorithm (MPDATA).

This work explores whether the methodology developed in our previous works for Intel and AMD x86 architecture can address performance tradeoffs and bottlenecks of a resource-constrained multicore RISC-V computing platform while executing the memory-bound MPDATA code. The explored platforms include: (i) HiFive Unmatched Rev B development board from SiFive; (ii) mini-cluster Lichee Cluster 4A, including seven boards with RISC-V TH1520 CPUs, and (iii) Banana Pi BPI-F3 development board with SpacemiT K1 RISC-V chip. Besides performance, energy consumption is investigated as well.

Keywords: RISC-V, SiFive Unmatched board, CFD, MPDATA, memory-bound application, porting

Feedback-Based Quantum Algorithm for Constrained Optimization Problems

Salahuddin Abdul Rahman[1], Özkan Karabacak[2], Rafal Wisniewski[1]
[1]Automation and Control section, Department of electronic systems
Aalborg University, Aalborg, Denmark
[2]Department of Mechatronics Engineering, Kadir Has University,
Istanbul, Turkey
{saabra, raf}@es.aau.dk, ozkan.karabacak@khas.edu.tr

The feedback-based algorithm for quantum optimization (FALQON) has recently been proposed to solve quadratic unconstrained binary optimization problems. This paper efficiently generalizes FALQON to tackle quadratic constrained binary optimization (QCBO) problems. For this purpose, we introduce a new operator that encodes the problem's solution as its ground state. Using Lyapunov control theory, we design a quantum control system such that the state converges to the ground state of this operator. When applied to the QCBO problem, we show that our proposed algorithm saves computational resources by reducing the depth of the quantum circuit and can perform better than FALQON. The effectiveness of our proposed algorithm is further illustrated through numerical simulations.

Keywords: Noisy Intermediate-Scale Quantum (NISQ) Devices, Feedback-Based Algorithm for Quantum Optimization (FALQON), Variational Quantum Algorithms (VQA), Quadratic Constrained Binary Optimization (QCBO), Discrete-Time Quantum Lyapunov Control (DQLC)

Halving the number of qubits of quantum comparators

Laura Donaire[1], Gloria Ortega[1], Ester Garzón[1],
Francisco José Orts Gómez[2], Remigijus Paulavičius[2], Ernestas Filatovas[2]
[1]Informatics Department, Agrifood Campus of International Excellence (ceiA3),
University of Almería, Spain
[2]Institute of Data Science and Digital Technologies,
Vilnius University, Lithuania
francisco.gomez@mif.vu.lt

Quantum comparators are of significant importance within the realm of various quantum algorithms. In this work we improve the number of qubits needed to perform a comparison of two N-bit strings from $2N+1$ qubits to $N+2$. To achieve this, we resort to an encoding of the bits in which one qubit is not wasted for each bit entered. This encoding is based on the one proposed in the work of Pérez-Salinas et al. (2020), but in our case, we adapt it for use in arithmetic operations. We also use the implementation of the AND operation proposed by Gidney (2018) to reduce the number of T gates (significantly more expensive than the rest of the gates) necessary for comparison. The result is a circuit that equals the T-count of the best comparator for quantum computing currently available while halving the number of qubits required.

Keywords: Quantum computing, quantum comparator, T-count

Private Computation of Boolean Functions Using Single Qubits

Zeinab Rahmani[1,2,3], Armando N. Pinto[1,2], Luis S. Barbosa[3,4,5]
[1]Department of Electronics, Telecommunications and Informatics
University of Aveiro, Aveiro, Portugal
[2]Instituto de Telecomunicações, Aveiro, Portugal
[3]International Iberian Nanotechnology Laboratory, Braga, Portugal
[4]Department of Computer Science, University of Minho, Braga, Portugal
[5]Institute of Systems and Computer Engineering,
Technology and Science, Porto, Portugal
zeinab.rahmani,anp@ua.pt, lsb@di.uminho.pt

Secure Multiparty Computation (SMC) facilitates secure collaboration among multiple parties while safeguarding the privacy of their confidential data. This paper introduces a two-party quantum SMC protocol designed for evaluating binary Boolean functions using single qubits. Complexity analyses demonstrate a reduction of 66.7% in required quantum resources, achieved by utilizing single qubits instead of multi-particle entangled states. However, the quantum communication cost was increased by 40% due to the amplified exchange of qubits among participants. Furthermore, we bolster security by performing additional quantum operations along the y-axis of the Bloch sphere, effectively hiding the output from potential adversaries. We design the corresponding quantum circuit and implement the proposed protocol on the IBM QisKit platform, yielding reliable outcomes.

Keywords: Secure multiparty computation, Boolean functions, QisKit

The Fredholm determinants and quantum entanglement

Roman Gielerak[1], Joanna Wiśniewska[2], Marek Sawerwain[1]
[1]Institute of Control & Computation Engineering
University of Zielona Góra, Zielona Góra, Poland
[2]Institute of Information Systems, Faculty of Cybernetics
Military University of Technology, Warsaw, Poland
{R.Gielerak, M.Sawerwain}@issi.uz.zgora.pl
JWisniewska@wat.edu.pl

The paper discusses several mathematical issues related to the numerical calculation of entanglement amount present in states describing bipartite quantum systems. The novel aspect of the presented material is the use of the Fredholm determinants theory methods for computing the amount of entanglement through the use of several, entropy-based entanglement measures. In particular, the continuous variable systems case is covered properly by the obtained here results, including presentation of a suitable renormalisation procedure of extracting the finite part of entanglement (often arises a situation in which the value of the corresponding entropy is infinite). Particular attention has been paid to the application of so-called unified entropy variants known as Rényi, resp. Tsallis entropies in the region of parameter α values in the interval $(0,1)$, where, in the case of infinitely dimensional systems, the infinite value of the corresponding unified entropy is very typical, even for frequently encountered quantum states. A library of functions supporting entropies and the corresponding entanglement calculations are also presented as an extension of the constructed by us EntDetector package for the Python language platform.

Keywords: infinite and finite quantum states, quantum entanglement, fredholm determinants, entropy, numerical calculations

Power Consumption and Energy Efficiency of Quantum Computing Platforms in High Performance Computing Integration

Xiaolong Deng[1], Martin Schulz[2,1], Laura Schulz[1]
[1]Leibniz Supercomputing Centre, Garching, Germany
[2]Computer Science Department, TU Munich, Garching, Germany
{xiaolong.deng, laura.schulz}@lrz.de, schulzm@in.tum.de

In this paper we study the power consumption of quantum computing platforms when integrated into high-performance computing (HPC) centers. We analyze the key components of leading quantum computers (superconducting circuit, trapped ion, neutral atom and photonics), and compare their power consumption. We also introduce a quantum energy efficiency metric that combines multiple qubit performance factors and power consumption. Using this metric, we evaluate and compare the energy efficiency of various quantum computers and assess their performance.

Keywords: Power consumption, Energy efficiency, Quantum computer, Superconducting circuit, Trapped ion, Neutral atom, Photonics

Feasibility Study of a Hybrid Quantum-Classical Setup for Multiple GPUs and Two Photonic Quantum Computers

Mateusz Slysz[1,2], Piotr Rydlichowski[1], Krzysztof Kurowski[1]
[1]Poznań Supercomputing and Networking Center
IBCH PAS Poznań, Poland
[2]Poznań University of Technology
Institute of Computing Science, Poznań, Poland
{mslysz,prydlich,krzysztof.kurowski}@man.poznan.pl

Quantum computing continues to advance steadily, showcasing its potential as a transformative technology. Various quantum computing modalities have emerged, each leveraging different physical implementations and approaches, such as superconducting qubits, trapped ions, neutral atoms, and Boson samplers, to realize the concept of a quantum computer. Researchers have explored over the last few years diverse quantum algorithms, highlighting both the strengths and limitations of different quantum computing architectures. This paper focuses on investigating the feasibility of a novel quantum-classical setup involving two photonic quantum computers recently installed at the Poznan Supercomputing and Networking Center (PSNC). We present a series of application performance tests and comparative analyses, with a specific emphasis on two selected optimization problems: Max-Cut and the resolution of potential conflicts between aircraft paths. Through these investigations, we aim to assess the practicality and efficacy of utilizing photonic quantum computers for addressing real-world optimization problems.

Keywords: Hybrid Quantum-Classical Computing, Photonic Quantum Computer, Boson Sampling, Max-Cut, Quantum Air Traffic Management

QCG-QuantumLauncher: a modular tool for quantum scenarios

Tomasz Pecyna, Dawid Siera, Bartosz Bosak
Poznań Supercomputing and Networking Center, IBCH PAS
Poznań, Poland
{tpecyna, dsiera, bbosak}@man.poznan.pl

The evolving landscape of quantum computing in its early stages presents challenges in predicting which among the many different quantum architectures will become dominant. This uncertainty and diversity has led to the proliferation of various software solutions, resulting in complexity for developers and researchers experimenting with different quantum paradigms. Existing tools attempt to address these challenges, yet they often fall short of meeting the needs of current researchers seeking simplicity in software while retaining access to specific quantum hardware details. In this paper, we introduce QCG-QuantumLauncher, a software library designed to solve specific problems by launching selected quantum algorithms on chosen quantum devices through a simple, uniform interface. Additionally, we delineate the future development trajectories of QCG-QuantumLauncher, aiming to position it as the premier tool choice for any quantum researcher.

Keywords: Quantum Computers, Quantum Software, Optimization

Semi-self-testing Quantum Random Number Generator with CMOS Sensors

Hamid Tebyanian
Department of Mathematics, University of York,
Heslington, York, United Kingdom
hamid.tebyanian@york.ac.uk

Our paper studies a faster, cost-effective, semi-device-independent quantum random number generator using a CMOS sensor, ensuring secure, validated randomness with minimal hardware reliance, ideal for high-speed, secure applications.

Keywords: Quantum Random Number Generator, Semi-self-testing, CMOS Sensors

Sustainable HPC for Global Challenges

Michael Resch
High Performance Computing Center Stuttgart, Germany
resch@hlrs.de

High Performance Computing is a key factor in the development and understanding of a sustainable future. In this talk we present a view on sustainable High Performance Computing. We focus on the role of different factors that allow to optimize sustainability of systems and centers. We then have a look at a number of scientific fields that help to tackle global challenges and specifically the challenge of a sustainable future.

Keywords: HPC, Global Challenges, sustainability of centres, sustainability of systems, supercomputing centres

The EuroHPC JU – collective effort for the development of European HPC Infrastructure and Applications

Rafał Duczmal
National Centre for Research and Development, Poland
rafal.duczmal@ncbr.gov.pl

The EuroHPC JU goal is to coordinate efforts and pool European resources to make Europe a world leader in supercomputing. The presentation will focus on 3 pillars of activity embedded in the design of EuroHPC JU: Infrastructure, Applications and AI Factories. The aim is to shed more light on the problem from the funders' perspective, starting from the current situation and elaborating on possible future developments and actions in the given areas.

Keywords: EuroHPC JU, global challenges, development strategies

Simulation of wildfires using EuroHPC resources: challenges and opportunities

David Caballero, Leydi Laura Salazar, Ángela Rivera, Luis Torres
MeteoGrid, Modesto Lafuente 45, Madrid, Spain
david@meteogrid.com

Recent wildfire events in various parts of the world (Canada, the USA, Europe, Australia, etc.) highlight that climate evolution leads to atmospheric patterns that contribute to the development of more destructive fires, revealing complex fire-atmosphere interactions. The Hidalgo2 project addresses several environmental challenges from the perspective of exascale numerical simulation and explores the possibilities and constraints of using European HPC resources. This paper presents some of the strategies followed in the wildfire pilot study for modeling fire spread and smoke emission and dispersion. The WRF-SFIRE model allows for the prediction of atmospheric behavior and its influence on wildfire spread by coupling the factors governing both phenomena. The design of the number of domains and their resolution, as well as the parameterization for downscaling and fire spread simulation, influence performance on EuroHPC installations. Additionally, each HPC installation has particularities for its use. These aspects condition and limit their use for real-time simulations to support operational decision-making. The pre-calculation of simulation ensembles can help overcome this obstacle and also allow for landscape sensitivity analysis. Finally, some strategies for the photorealistic visualization of the results are presented.

Keywords: Wildfires, Smoke, HPC, EuroHPC, WRF, SFIRE, Virtual Reality, Visual Simulation, Hidalgo2

Ktirio Urban Building: A Computational Framework for City Energy Simulations Enhanced by CI/CD Innovations on EuroHPC Systems

Christophe Prud'Homme, Vincent Chabannes, Luca Berti, Maryam Maslek, Javier Cladellas, Philippe Pincon
Cemosis, IRMA UMR 7501, University of Strasbourg, CNRS
Strasbourg, France
{vincent.chabannes,christophe.prudhomme}@cemosis.fr

The building sector in the European Union significantly impacts energy consumption and greenhouse gas emissions. The EU's Horizon 2050 initiative sets ambitious goals to reduce these impacts through enhanced building renovation rates. The CoE HiDALGO2 supports this initiative by developing high-performance computing solutions, specifically through the Urban Building pilot application, which utilizes advanced CI/CD methodologies to streamline simulation and deployment across various computational platforms, such as the EuroHPC JU supercomputers. The present work provides an overview of the Ktirio Urban Building framework (KUB), starting with an overview of the workflow and a description of some of the main ingredients of the software stack and discusses some current results performed on EuroHPC JU supercomputers using an innovative CI/CD pipeline.

Keywords: HPC, HPCOps, Urban building, City Energy Simulation

Fostering uncertainty quantification in Global Challenges with mUQSA toolkit

Michal Kulczewski, Bartosz Bosak, Piotr Kopta, Wojciech Szeliga, Tomasz Piontek
Poznan Supercomputing and Networking Center
Poznan, Poland
{michal.kulczewski, bbosak, pkopta, wojteks}@man.poznan.pl
piontek@man.poznan.pl

In this paper focus is given on uncertainty quantification (UQ) and sensitivity analysis (SA) and their applicability to multiscale Global Challenges. Based on Renewable Energy Sources Global Challenge study, different approaches of applying UQ and SA are showcased. Naturally, UQ can be used to limit the uncertainties of the model and input data, thus providing results of better quality. SA can be used to estimate the influence of given input or model parameters on obtained results. This study can be also used to limit required parameters, thereby reducing the demand for HPC computational resources and time-to-solution required to run the model. However, UQ and SA can be also used to support solving Global Challenges problems which is showcased in this paper. This paper discusses also how multiscale Uncertainty Quantification and Sensitivity Analysis platform (mUQSA) and its underlying tools can effectively support UQ and SA of Global Challenges.

Keywords: Renewable Energy Sources, Uncertainty Quantification, mUQSA, multiscale, HPC

HPC-CFD BASED OPTIMIZATION OF INDOOR ENVIRONMENT TO MINIMIZE AIRBORNE CONTAMINANTS

Makoto Tsubokura[1,2], Rahul Bale[1,2], Alicia Murga[1], Kazuhide Ito[3],
Mario Ruttgers[3], Andreas Lintermann[4]
[1]Kobe University, Kobe, Japan
[2] RIKEN Center for Computational Science, Kobe, Japan
[3] Kyushu University, Kasuga, Japan
[4] Juelich Supercomputing Centre, Germany
{tsubo@tiger, rbale@harbor, alicia.murga@harbor}.kobe-u.ac.jp
{rahul.bale, mtsubo}@riken.jp, ito@kyudai.jp
{m.ruettgers, a.lintermann}@fz-juelich.de

Inhalation exposure to indoor pollutants can be minimized through the design of adequate ventilation systems. The present study focuses on the evaluation and optimization of widely used ventilation techniques in terms of mitigation of airborne transmission by applying two interacting speaking-breathing virtual manikins to directly predict source-to-dose paths and particle deposition on human tissue. Based on these results, the indoor environment has been further optimized through six generations to minimize airborne viral density and reduce building energy consumption.

Keywords: CFD, Building Cube Method, HPC, Multi-puropose optimization, Genetic Algorithm, airborne infection risk, indoor-environment

Performance portability of various programming models on Particle-In-Cell

Kévin Peyen, Mathieu Lobet, Juan José Silva Cuevas, Edouard Audit
Université Paris-Saclay, UVSQ, CNRS, CEA, Maison de la Simulation,
Gif-sur-Yvette, France
{kevin.peyen2, mathieu.lober, juan-jose.silvacuevas}@cea.fr
edouard.audit@cea.fr

With the emergence of multiple architectures, performance portability has become a key consideration when selecting a programming model for modernizing or writing code. This study examines these new programming models in the context of developing scientific applications for present and hopefully futur ex- ascale architectures. In this work, we are particularly interested in benchmarking SYCL against other option such as Kokkos, Thrust or OpenACC.

For that purpose, we have developed a mini-app implementation of a Particle-In-Cell (PIC) code using several programming models. This allow us to compare their performance portability on various type of kernels. The initial results suggest that Kokkos often achieves better performance across different architectures more easily.

Keywords: HPC, Performance portability, SYCL, Kokkos, PIC

Portability of Multiphysics Applications on Heterogeneous Modular Supercomputers

Daniel Caviedes Voullieme[1], Seong-Ryong Koh[1], Stefan Poll[1], Estela Suarez[1], Takashi Arakawa[2], Kengo Nakajima[2], Shinji Sumimoto[2]
[1]Juelich Supercomputing Centre, Germany
[2]ClimTech/The University of Tokyo, Tokyo Japan
{d.caviedes.voullieme, s.koh, s.poll, e.suarez}@fz-juelich.de
arakawa@climtech.jp
{nakajima, sumimoto}@cc.u-tokyo.ac.jp

The Modular Supercomputer Architecture (MSA) integrates various processing units, with compute modules tailored for specific algorithms, forming a unified heterogeneous system. Modules operate as parallel clustered systems, interconnected via a common or federated network. Optimized resource management allows flexible node selection, aiming for better performance. Hardware heterogeneity poses challenges for developers, who must port and optimize codes for various configurations. This study evaluated the performance of two applications—TSMP (regional Earth-system simulation) and mAIA (fluid dynamics)—across two MSA platforms (DEEP, Wisteria/BDEC-01)), considering both hardware and software configurations.

Keywords: MSA, Heterogeneous Systems, Coupling Library

Efficient allocation of LLM and machine learning tasks on multi-GPU systems

Marcin Lawenda[1], Krzesimir Samborski[1], Kyrylo Khloponin[1], Łukasz Szustak[1,2]
[1]Poznan Supercomputing and Networking Center, Poznań, Poland
[2]Czestochowa University of Technology, Częstochowa, Poland
{lawenda,ksamborski,kkhloponin}@man.poznan.pl
lszustak@icis.pcz.pl

The rise of large language models (LLMs) and machine learning (ML) has brought about a substantial revolution in various domains such as natural language processing, computer vision, and data analysis. These models, renowned for their extensive parameters and intricate structures, demand substantial computational resources for both training and inference. Multi-card GPU systems, which harness the parallel processing capabilities of multiple graphics processing units (GPUs), have become indispensable infrastructure to fulfill these computational demands. This approach results in the development of code specifically designed for parallelization, enabling quicker training of significantly larger models by utilizing the increased computational power and memory of additional accelerators. Nevertheless, effectively allocating tasks in multi-card GPU systems remains a notable challenge due to the inherent complexity in model architecture, data dependencies, and hardware heterogeneity.

This work deals with the parallelization performance evaluation of learning and fine-tuning processes for image classification and large language models. For machine learning model in image recognition, various parallelization methods have been developed based on different hardware and software scenarios: simple data parallelism, distributed data parallelism, and distributed processing. A detailed description of each of the presented strategies is presented, highlighting the challenges and benefits of their use. PyTorch provides several built-in mechanisms to facilitate these tasks, which are studied and tested.

Furthermore, the impact of various dataset styles on the fine-tuning process of large language models is investigated. Experiments demonstrate the extent to which task type affects iteration time in a multi-GPU environment, offering valuable insights into optimal strategies for utilizing data to improve model performance. Additionally, PyTorch/torchtune's built-in parallelization mechanisms are leveraged in this study.

On the top of that, performance profiling was enabled for precise measurement of the impact of memory and communication operations during the training/tuning

procedure. Test scenarios were developed and benchmarked by numerous tests on two NVIDIA card architectures: A100 and H100.

Keywords: GPU, scheduling, machine learning, Large Language Models, data distribution, optimization

Parallel reinforcement learning and Gaussian process regression for improved physics-based nasal surgery planning

Mario Rüttgers[1], Fabian Hübenthal[1], Makoto Tsubokura[3,4], Andreas Lintermann[1]
[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, Jülich, Germany
[2]Institute of Aerodynamics and Chair of Fluid Mechanics (AIA),
RWTH Aachen University, Aachen, Germany
[3]Department of Computational Science, Graduate School of System Informatics,
Kobe University, Kobe, Japan
[4]Complex Phenomena Unified Simulation Research Team,
RIKEN Center for Computational Science, Kobe, Japan
m.ruettgers@fz-juelich.de

Septoplasty and turbinectomy are among the most frequent but also most debated interventions in the field of rhinology. A previously developed tool enhances surgery planning by physical aspects of respiration, i.e., for the first time a reinforcement learning (RL) algorithm is combined with large-scale computational fluid dynamics (CFD) simulations to plan anti-obstructive surgery. In the current study, an improvement of the tool's predictive capabilities is investigated for the aforementioned types of surgeries by considering two approaches: (i) training of parallel environments is executed on multiple ranks and the agents of each environment share their experience in a pre-defined interval and (ii) for some of the state-reward combinations the CFD solver is replaced by a Gaussian process regression (GPR) model for an improved computational efficiency. It is found that employing a parallel RL algorithm improves the reliability of the surgery planning tool in finding the global optimum. However, parallel training leads to a larger number of state-reward combinations that need to be computed by the CFD solver. This overhead is compensated by replacing some of the computations with the GPR algorithm, i.e., around 6% of the computations can be saved without significantly degrading the predictions' accuracy. Nevertheless, increasing the number of state-reward combinations predicted by the GPR algorithm only works to a certain extent, since this also leads to larger errors.

Keywords: Septoplasty, Turbinectomy, Computational fluid dynamics, Reinforcement learning, Surrogate-based optimization

High-Resolution Agent-Based Modeling of Campus Population Behaviors for Pandemic Response Planning

Hiroki Sayama[1,2], Shun Cao[3]
[1]Binghamton University, State University of New York, Binghamton, NY, USA
[2]Waseda University, Tokyo, Japan
[3]University of Houston, Sugar Land, USA
sayama@binghamton.edu, scao7@central.uh.edu

This paper reports a case study of an application of high-resolution agent-based modeling and simulation to pandemic response planning on a university campus. In the summer of 2020, we were tasked with a COVID-19 pandemic response project to create a detailed behavioral simulation model of the entire campus population at Binghamton University. We conceptualized this problem as an agent migration process on a multilayer transportation network, in which each layer represented a different transportation mode. As no direct data were available about people's behaviors on campus, we collected as much indirect information as possible to inform the agents' behavioral rules. Each agent was assumed to move along a shortest path between two locations within each transportation layer and switch layers at a parking lot or a bus stop, along with several other behavioral assumptions. Using this model, we conducted simulations of the whole campus population behaviors on a typical weekday, involving more than 25,000 agents. We measured the frequency of close social contacts at each spatial location and identified several busy locations and corridors on campus that needed substantial behavioral intervention. Moreover, systematic simulations with varying population density revealed that the effect of population density reduction was nonlinear, and that reducing the population density to 40-45% would be optimal and sufficient to suppress disease spreading on campus. These results were reported to the university administration and utilized in the pandemic response planning, which led to successful outcomes.

Keywords: High-resolution agent-based modeling and simulation, Multilayer transportation network, Pandemic response planning, Mechanistic modeling, Case study

A comparison of selected agent-based modelling frameworks

Paulina Gacek, Dominika Bocheńczyk, Maciej Krajewski, Bartosz Kruczek,
Paweł Magnuszewski, Krzysztof Łazarz, Antoni Zięciak, Filip Kamiński,
Jarosław Wąs
AGH University of Krakow,
Faculty of Electrical Engineering, Automatics, IT and Biomedical Engineering
Department of Applied Computer Science
Krakow, Poland
{paulinagacek, bochenczyk, pmagnus, kjlazarz}@student.agh.edu.pl
zieciak@student.agh.edu.pl, {kaminski, jaroslaw.was}@agh.edu.pl
maciejkrajewskiml@gmail.com, bartoszkruczek@gmail.com

In recent years, a variety of new frameworks streamlining the process of agent-based modeling has emerged. These frameworks serve different purposes and each offers a unique set of features. In this practical comparative study we evaluate the performance of various ABM frameworks through a series of benchmark simulations. By comparing the distinct functionalities offered by these tools, we aim to assist in the selection of an appropriate ABM toolkit for developing system models. This review presents a concise overview of seven popular agent-based modeling tools aiming to inspire further exploration and investigation into this subject.

Keywords: Agent-based modelling, Agent-based modelling frameworks, Software agent, Multi-agent computing

Modelling of opinion formation process on dierent social networks

Norbert Borowiec, Tomasz Gwizdałła, Paweł Maślanka
Faculty of Physics and Applied Informatics,
University of Łódź, Łódź, Poland
norbert.borowiec@edu.nui.lodz.pl, tomasz.gwizdalla,
pawel.maslanka@uni.lodz.p

Different processes related to the phenomena of interaction across communities are of particular interest, especially since the beginning of the current century. It concerns such processes as disease spreading, opinion distribution, and formation or recommendation systems. The last two are of special interest in the rapid development of social networks. In our paper we will present the proposition of the update scheme intended for the opinion in continuous space. We show it for the 2D opinion space following Nolan's idea of multidimensional opinion presentation. We implement our mechanism for different types of networks and different parameters. The paper shows that similarities exist between the networks with longer tails of node degree distribution, and we can observe the emergence of two groups around two different points in the space.

Keywords: opinion formation, scale-free networks, social networks

Machine learning approach for detecting potential anomalous cosmic rays particles tracks in Earth-scale Cosmic Ray Extremely Distributed Observatory

Jan Tyc, Marcin Zub, Tomasz Hachaj
Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering
AGH University of Krakow, Krakow, Poland
{jantyc, zub}@student.agh.edu.pl, thachaj@agh.edu.pl

In this paper, we present a novel contribution to the problem of the detection of potential anomalous cosmic rays particle tracks acquired by Earth-scale complex distributed observatory infrastructure. Detection of such signals provides an opportunity to analyze the performance of the entire research infrastructure of a complex system (for example, potential failures) as well as an opportunity to record unknown physical phenomena that can be recorded simultaneously with space radiation observations. The observatory is consisted of a worldwide network of CMOS detectors that cooperate in the citizen science paradigm. We propose, evaluate and compare several data processing, feature extractions, and potential anomaly discovery pipelines based on machine learning approaches. To our best knowledge, our paper is the first to publish the results of a study on such a large dataset (nearly 600 000 images) of this modality on so many novel anomalous discovery approaches. We have published both the source codes and dataset of our research, so our results can be reproduced.

Keywords: Anomaly detection, Cosmic rays, Earth-scale observatory, CMOS detector, Citizen science, Complex system, Cooperative problems solving, Machine learning

A multi-cell cellular automata model for roundabout traffic flow considering the heterogeneity of human delay and acceleration

Krzysztof Małecki, Jakub Gębicz
West Pomeranian University of Technology
Szczecin, Poland
kmalecki@zut.edu.pl, gj49250@zut.edu.pl

Various macroscopic and microscopic road traffic models al- low traffic flow analysis. However, it should be emphasised that standard traffic flow models do not include drivers' behaviour. Thus, we propose a multi-agent microscopic model for analysing the roundabout traffic flow considering the various types of agents. Agents have parameters charac- terising their style of acceleration and braking, as well as the distance to the vehicle in front. The simulation studies show that these parameters are crucial in roundabout road traffic analysis. To accurately reflect the acquired dimensions of the cars, a small-cell cellular automaton (CA) was used, where one car is represented as a set of CA cells.

Keywords: Agent-Based Modeling (ABM), Cellular Automata (CA), Traffic Flow, Data-Driven Model

Kernel estimates of pedestrian density applied in simulation of recreational pedestrian movement

Tomáš Novotný, Jana Vacková, Pavel Hrabak
Czech Technical University in Prague
Faculty of Information Technology
Prague, Czech Republic
{novott37, jana.vackova, pavel.hrabak}@fit.cvut.cz

A model for simulating pedestrian recreational movement along a tourist trail is introduced. The model incorporates inter-agent interaction and density-induced velocity reduction without the necessity to perform costly microscopic simulations. The underlying model reduced the two-dimensional nature of bidirectional flow to one dimension by applying the kernel estimates of pedestrian density to describe the pedestrian mass in the perceived surroundings of an agent. The model applicability is investigated by means of three speed distribution scenarios and three arrival intensity scenarios. The results are in correspondence with expectations and recommendations for D level of pedestrian level of service. The model mimics the expected trail capacity behaviour represented by arrival intensity, leading to speed reduction and stoppages. Heterogeneity of the agents similarly as non-homogeneity of arrival intensity leads to clogging at smaller average arrival intensities.

Keywords: Recreational Model, Fundamental Diagram, Kernel Estimation, Pedestrian Movement Simulations

Personal Space of People in Movement under Different Conditions

Tat'Yana Vitova, Ekaterina Kirik
Institute of Computational Modelling of the Siberian Branch
of the Russian Academy of Sciences
Krasnoyarsk, Russia
{vitova, kirik}@icm.krasn.ru

The paper investigates a space occupied by a person (personal space) in movement under different conditions. The different movement scenarios are studied: normal and competitive movement, movement in a straight corridor without bottleneck and with bottleneck. For the personal space to estimate the Voronoi diagram is used. It is assumed that the area of the Voronoi cell corresponds to the area of the personal space. Time series and histograms of the Voronoi cell area were constructed. Personal space is considered both in transition and steadystate regimes, as well as in a static position of people while waiting. There are fluctuations in the Voronoi cells area. The minimum area value does not persist for a long time. This is the result of a temporary compaction of human flow. People keep a space around them while waiting. During competitive movement, people can move in relatively limited conditions. Transition regimes require more space for movement compared to steadystate ones.

Keywords: Personal space, pedestrian dynamics, Voronoi diagram, experimental study

Monitoring and Analysis of Energy Consumption in HPC systems

Ilsche Thomas
TU Dresden, Germany
thomas.ilsche@tu-dresden.de

A robust understanding of energy consumption is essential for efficiently operating High-Performance Computing (HPC) systems as well as data center capacity planning. This talk discusses various instrumentation points as well as exemplary power measurement solutions and their accuracy and time resolution. Additionally, the talk will introduce a unified infrastructure for collection, storage, analysis and visualization.

Keywords: HPC, Monitoring, Energy Efficiency

Smart energy efficiency and management with EAR

Oriol Vidal, Julita Corbalan
Barcelona Supercomputing Centre, Spain
{oriol.vidal, julita.corbalan}@bsc.es

EAR is a European open-source system software for energy efficiency and management in Data Centres. This talk will present the EAR architecture, emphasizing in the core components in charge of the energy efficiency and management in computational nodes. These components are the EAR Job Manager, the EAR Node Manager and the EAR Scheduler plugin. These three components share events, node telemetry, application significant performance and power metrics to implement smart energy optimization policies for HPC and AI workloads and node power cap guided by application activity phases.

Keywords: HPC, energy efficiency management, smart energy optimisation

Data-driven and AI-driven models for sustainable computing

Andrea Bartolini
University of Bologna, Italy
a.bartolini@unibo.it

The efficiency and sustainability of high-performance computing systems have never been so important for societal development. In this talk, I will cover the recent research results as well as lessons learnt in applying AI techniques for modelling and predicting key operational parameters essential for increasing the sustainability of large-scale high-performance computing installations.

Keywords: HPC, AI for job prediction, operational parameters, modelling, energy efficiency

Improving HPC system energy efficiency using MERIC runtime system

Ondrej Vysocky
IT4Innovations, VSB – Technical University of Ostrava
Ostrava, Czech Republic
ondrej.vysocky@vsb.cz

An HPC system can be optimized for energy efficiency at several levels, while the highest level of dynamicity comes from the power management of computing components controlled at the job level. Complex parallel applications show different hardware requirements during their execution. This dynamic behavior can be exploited for energy savings by changing the hardware power knobs to fit the configuration to the application's needs.

Energy-efficient runtime systems provide administrator- and user-friendly ways to perform such hardware power management without requiring a deep understanding of the topic. EuroHPC Center of Excellence Performance Optimisation and Productivity (POP) flagship code MERIC is a light-weight tool designed to provide a detailed analysis of application behavior, identify the optimal hardware settings concerning energy consumption and runtime, and provide dynamic tuning during the application runtime. Thanks to complex execution time coverage by regions of interest, high tuning granularity starting at the level of ten milliseconds, and a large set of controlled power knobs, it pushes the achievable energy savings to the limit.

Keywords: HPC, Energy efficiency, Resource management, MERIC

Towards Energy-efficient System-level Scheduling for Modular Supercomputers

Simon Pickartz
ParTec AG, Germany
pickartz@par-tec.com

Today's HPC systems offer a variety of mechanisms to measure and control the energy and power consumption of the hardware. However, it is up to the system software to take advantage of these features to optimize system utilization in terms of throughput and energy efficiency. This is even more true for heterogeneous systems comprising modules with different capabilities to be matched with the diverse requirements of today's HPC workloads. Therefore, all levels of the system software stack, from the system level to the workflow and job level to the node level, have to be considered when designing the stack.

The system level has a holistic view on the resource availability and requirements of all workloads. With the goal of globally optimizing system utilization, the Resource and Job Management System (RJMS) shall be able to dynamically adapt the schedule as jobs come and go. This talk provides an overview of state-of-the-art RJMSs and an analysis of how to evolve them into a production-quality solution for dynamic scheduling and resource management in HPC.

Keywords: HPC, Job scheduling, Resource management, Modular supercomputing, Energy efficiency

Application of Hybrid Parallelism in Finite Physical Systems Modelling

Łukasz Kucharski[1], Michał Antkowiak[2]
[1]Adam Mickiewicz University, Poznań, Poland
[2]Faculty of Physics, A. Mickiewicz University, Poznań, Poland
luk32@o2.pl, antekm@amu.edu.pl

The accurate modeling of finite physical systems, such as molecular nanomagnets and single-molecule magnets (SMMs), involves complex optimization problems that demand significant computational resources. Determining parameters like exchange coupling and single-ion anisotropy is crucial for understanding these systems' magnetic properties and harnessing their potential applications in quantum computing and magnetic storage. This abstract outlines the integration of advanced computational techniques—genetic algorithms combined with exact diagonalization—within a high-performance computing (HPC) framework to solve these optimization problems effectively.

Finite physical systems, particularly those involving molecular nanomagnets, present a formidable challenge due to the need for precise modeling of magnetic interactions. The complexity of these systems necessitates extensive computational power to explore large parameter spaces and solve quantum mechanical equations accurately.

To address these computational challenges, we implemented a hybrid approach that integrates genetic algorithms with exact diagonalization, optimized for HPC environments. Genetic algorithms are well-suited for exploring vast parameter spaces efficiently, while exact diagonalization provides precise solutions necessary for accurate modeling.

Our methodology leverages a two-level parallelism strategy to maximize the use of HPC resources. This involves parallelizing both the genetic algorithm's population evaluation process and the exact diagonalization calculations, ensuring optimal workload distribution across computing nodes for effective resource utilization. The Cr8Ni molecule, a part of the chromium rings family, exemplifies a complex system with antiferromagnetic interactions and magnetic frustration. Utilizing our HPC-optimized approach, we were able to achieve more accurate non-uniform exchange coupling parameters than previously reported. The genetic algorithm efficiently navigated the parameter space, and the exact diagonalization provided the necessary precision.

Our results demonstrated that the exchange coupling parameters for the Cr8Ni

molecule were systematically overestimated in earlier studies. The refined parameters not only fit the experimental magnetic susceptibility data more accurately but also provided new insights into the molecule's magnetic behavior. Applying our methodology to the tetranuclear oxalato-bridged Re3(IV)Ni(II) complex, known for its SMM behavior, further validated our approach. This complex is characterized by high anisotropy and significant spin-orbit coupling, making it an ideal candidate for advanced computational modeling.

Using a combination of exact diagonalization, genetic algorithms, and EPR resonance data, we identified unique anisotropy parameters that accurately describe the magnetic properties of the complex. These parameters (DRe/kB = -8.8 K, DNi/kB = 9.9 K, E = 0) provide a robust framework for understanding the magnetic behavior of SMMs and enhance our ability to predict and manipulate their properties for technological applications.

The integration of genetic algorithms with exact diagonalization within an HPC framework showcases the significant computational resources required to solve these optimization problems. Our approach demonstrates the effective use of HPC resources to achieve precise and reliable parameter estimations. By employing parallel computing strategies, we maximized computational efficiency and reduced the time required for simulations, making it feasible to tackle large-scale problems in finite physical system modeling.

Our research underscores the critical role of HPC in advancing the modeling of finite physical systems. The combination of genetic algorithms and exact diagonalization, optimized for HPC environments, offers a powerful method for solving complex optimization problems. This integrated approach has proven effective in refining known parameters and providing new insights into the magnetic properties of molecular nanomagnets and SMMs.

By addressing the computational challenges inherent in these systems, our methodology paves the way for more accurate and insightful studies, ultimately contributing to the advancement of materials science and technology. The findings highlight the importance of precise parameter estimation and the potential for systematic overestimation in previous studies, emphasizing the need for HPC in achieving reliable results.

In conclusion, the use of HPC resources is indispensable for solving optimization problems in the modeling of finite physical systems. Our integrated approach sets a new standard for computational precision and efficiency, offering a robust tool for researchers in the field of molecular magnetism and beyond.

Keywords: Parallelism, High Performance Computing, Single Molecule Magnets, Physical Simulations, Physical System Modelling

Efficient Algorithm for U(N) to U(3) Representation Reduction in Isospin-Adapted Nuclear Structure Calculations

Daniel Langr, Tomáš Dytrych
Faculty of Information Technology,
Czech Technical University in Prague,
Praha, Czech Republic
Nuclear Physics Institute, Czech Academy of Sciences,
Řež, Czech Republic
langrd@fit.cvut.cz, dytrych@ujf.cas.cz

An efficient algorithm for enumerating representations of the unitary group U(3) that occur in a representation of the unitary group U(3) is introduced. The algorithm is applicable to U(N) representations associated with a system of nucleons distributed among the degenerate eigenstates of the selected three-dimensional harmonic oscillator level, where each eigenstate can be occupied by up to 4 particles with different combinations of spin and isospin quantum numbers. A C++ implementation of the proposed algorithm is provided, and its performance evaluation is reported.

Keywords: Computer algorithm, C++, Group representation, Nuclear structure, Three-dimensional harmonic oscillator, U(3) symmetry

New superconducting ScC2H8 ternary hydride at moderate stabilization pressure: ab initio calculations

Izabela Wrona, Radosław Szczęśniak, Artur Durajski

Since Ashcroft introduced the visionary concept of superconductivity in hydrogen [1] and hydrogen-rich materials [2], there have been remarkable advances in both theoretical and experimental research techniques that made it possible to predict and create high-temperature superconductors. Current studies suggest that hydrogen-rich superconductors have significant potential as candidates for achieving room temperature superconductivity. In particular, H3S stands out as the first experimentally confirmed superhydride with a critical temperature above 200 K at high pressure [3, 4]. This finding supports earlier theoretical predictions of its crystal structure and superconducting properties, demonstrating the reliability of first-principles calculations in the study of conventional phonon-mediated hydride superconductors [5, 6]. Following this advance, many hydrogen-based superconductors with high Tc values have been predicted. To date, the binary hydrides of almost all the elements in the periodic table have been studied in detail, theoretically or experimentally. Although binary hydrides exhibit remarkable superconducting properties, they remain stable only under extreme pressure conditions. To overcome that problem, various approaches have been employed to reduce the stability pressure required for hydrogen-based superconductors. Particularly effective strategy is the addition of other elements to existing binary hydrides, leading to the creation of ternary hydrides [7]. Due to their different chemical compositions, ternary hydrides offer a wider range of structural possibilities than binary hydrides. They are highly attractive candidates for superconductivity because they combine the advantages of different elements and induce strong electron-phonon coupling [7–10]. Especially, the introduction of additional light elements, beyond hydrogen, can significantly increase the electron-phonon coupling in ternary hydrides, because of filling the gap in the phonon density of states. In this field, theoretical calculations provide guidance and play an important role in predicting new superconducting materials, showing much lower costs and higher efficiencies than experimental studies, especially at high pressures. Inspired by the recent experimental findings on Fm3m-LaBeH8 [11] and the excellent correlation between theoretical predictions and measured values for this compound [12], we have investigated the superconductivity of ScX2H8 (X=B, C, N, Al, Si, P) under pressures from 10 to 200 GPa. Using first-principles methods

together with the strong-coupling Migdal-Eliashberg approach, we aim to unravel the intricacies of the superconducting behavior in this system. All calculations for the hydrogen-rich systems were conducted using the density functional theory (DFT), as implemented in the Quantum-Espresso package [13, 14]. The generalized gradient approximation (GGA) in the form of the Perdew Burke Ernzerhof (PBE) functional and projector augmented wave (PAW) potentials were used to describe the effect of the electron–ion interactions. Atomic positions and lattice vectors were fully optimized as a function of pressure using the Broyden Fletcher Goldfarb Shanno (BFGS) method [15]. We concentrated on examining the stability and superconductivity of the well-known sodalite-like clathrate structure, which has been experimentally observed and features a high-symmetry Fm3m space group [16]. First, as a precursor for the ScH10 hydride, we have used Fm3m-LaH10 phase of high-temperature superconductor. Then, we have constructed the ScX2H8 structure by replacing two H atoms with two X atoms (X=B, C, N, Al, Si or P). The stability of these compounds has been rigorously evaluated over a range of pressures from 10 GPa to 200 GPa. By performing the phonon calculations we have shown that the initial Fm3m-ScH10 structure is thermodynamically unstable in the whole pressure range. In the case of Fm3m-ScX2H8 systems, only ScC2H8 compound remains stable, as evidenced by no imaginary frequencies within the pressure range of 40 to 200 GPa. Our calculations have shown that the Fm3m-ScC2H8 system exhibits metallic behavior which is visible in its electronic band structure. Furthermore, the projected density of states reveals that near the Fermi level, the total DOS is dominated by Sc atoms. Furthermore, based on the phonon density of states projected onto individual atoms, we have demonstrated which elements have a dominant contribution to the low and high-frequency range. With the information on the electronic and lattice dynamics, we have calculated the electron-phonon Eliashberg spectral functions for ScC2H8 in the whole stable pressure range from 40 to 200 GPa. We were thus able to determine the electron-phonon coupling constant, which increases with pressure, proving the strong-coupling electron-phonon pairing mechanism in ScC2H8. Through the solution of the Eliashberg equations, we have determined the temperature-dependent characteristics of the superconducting energy gap and the critical temperature as a function of pressure. The critical temperature of Fm3m-ScC2H8 reaches a maximal value of 57 K at the pressure of 200 GPa. Our calculations revealed that the $2\triangle(0)/kBTc$ ratio of ScC2H8 significantly exceeds the universal value of 3.53 predicted by the BCS theory, indicating the existence of retardation effects. Moreover we have calculated the free energy difference between the superconducting and normal states and the specific heat difference between the superconducting and normal states. The obtained results facilitate the calculation of the dimensionless parameter relative jump of the specific heat. Structural stability at finite temperatures plays a crucial role in the

performance of high-temperature superconductors. To examine the thermal stability of ScC2H8 at its superconducting critical temperature, an AIMD simulation was employed. The results show minimal fluctuations in total energy during the AIMD steps, demonstrating good thermal stability of ScC2H8. In particular, our results provide evidence for the electron-phonon superconductivity mechanism of ScC2H8 and greatly expand the scope for the study of superhydrides with high critical temperature. These results present a promising approach to reducing the stabilization pressure of hydrogen-based superconductors through chemical doping, potentially opening the way for further research and applications.

Acknowledgements

Artur P. Durajski acknowledges financial support from the National Science Centre (Poland) under Project No. 2022/47/B/ST3/00622. All authors are grateful to the Czestochowa University of Technology – MSK CzestMAN for granting access to the computing infrastructure built-in project no. POIG.02.03.00-00-028/08 "PLATON – Science Services Platform" and POIG.02.03.00-00-110/13 "Deploying high-availability, critical services in Metropolitan Area Networks (MAN-HA)".

[1] N. W. Ashcroft, Metallic hydrogen: a high-temperature superconductor?, Phys. Rev. Lett. 21, 1748 (1968).
[2] N. W. Ashcroft, Hydrogen dominant metallic alloys: high temperature superconductors?, Phys. Rev. Lett. 92, 187002 (2004).
[3] A. P. Drozdov, M. I. Eremets, I. A. Troyan, V. Ksenofontov, and S. I. Shylin, Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system, Nature 525, 73 (2015).
[4] I. Troyan, A. Gavriliuk, R. Ruffer, A. Chumakov, A. Mironovich, I. Lyubutin, D. Perekalin, A. P. Drozdov, and M. I. Eremets, Observation of superconductivity in hydrogen sulfide from nuclear resonant scattering, Science 351, 1303 (2016).
[5] Y. Li, J. Hao, H. Liu, Y. Li, and Y. Ma, The metallization and superconductivity of dense hydrogen sulfide, J. Chem. Phys. 140, 174712 (2014).
[6] D. Duan, Y. Liu, F. Tian, D. Li, X. Huang, Z. Zhao, H. Yu, B. Liu, W. Tian, and T. Cui, Pressure-induced metallization of dense (H2S)2H2 with high-TC superconductivity, Sci. Rep. 4, 6968 (2014).
[7] X. Zhang, Y. Zhao, and G. Yang, Superconducting ternary hydrides under high pressure, Wiley Interdiscip. Rev. Comput. Mol. Sci. 12, e1582 (2022).
[8] S. Di Cataldo, W. von der Linden, and L. Boeri, First-principles search of hot superconductivity in La-X-H ternary hydrides, Npj Comput. Mater. 8, 2 (2022).
[9] P. Song, Z. Hou, K. Nakano, K. Hongo, and R. Maezono, Potential high-TC superconductivity in YCeHx and LaCeHx under pressure, Mater. Today Phys. 28, 100873 (2022).
[10] R. Vocaturo, C. Tresca, G. Ghiringhelli, and G. Profeta, Prediction of ambient-

pressure superconductivity in ternary hydride PdCuHx, J. Appl. Phys. 131, 033903 (2022).
[11] Y. Song, J. Bi, Y. Nakamoto, K. Shimizu, H. Liu, B. Zou, G. Liu, H. Wang, and Y. Ma, Stoichiometric ternary superhydride LaBeH8 as a new template for high-temperature superconductivity at 110 K under 80 GPa, Phys. Rev. Lett. 130, 266001 (2023).
[12] Z. Zhang, T. Cui, M. J. Hutcheon, A. M. Shipley, H. Song, M. Du, V. Z. Kresin, D. Duan, C. J. Pickard, and Y. Yao, Design principles for high-temperature superconductors with a hydrogen-based alloy backbone at moderate pressure, Phys. Rev. Lett. 128, 047001 (2022).
[13] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, Quantum espresso: a modular and open-source software project for quantum simulations of materials, J. Phys. Condens. Matter 21, 395502 (2009).
[14] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. D. Jr, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Kucukbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Ponce, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, Advanced capabilities for materials modelling with Quantum ESPRESSO, J. Phys. Condens. Matter 29, 465901 (2017).
[15] S. R. Billeter, A. Curioni, and W. Andreoni, Efficient linear scaling geometry optimization and transition-state search for direct wavefunction optimization schemes in density functional theory using a plane-wave basis, Comput. Mater. Sci. 27, 437 (2003).
[16] V. S. Minkov, S. L. Budko, F. F. Balakirev, V. B. Prakapenka, S. Chariton, R. J. Husband, H. P. Liermann, and M. I. Eremets, Magnetic field screening in hydrogen-rich high-temperature superconductors, Nature Commun. 13, 3194 (2022).

Keywords: superconductivity, hydrogen-rich superconductors, ternary hydride, DFT, Eliashberg theory, ab initio calculations

Study of the magnetic behaviour of oleic-acid coated Co ferrite nanoparticles: A multiscale modeling approach

Marianna Vasilakaki, Nikolaos Ntallis, Kalliopi Trohidou
nstitute of Nanoscience and Nanotechnology, NCSR 'Demokritos', Aghia Paraskevi, Attiki, Greece
k.trohidou@inn.demokritos.gr

A multi-scale modeling approach is employed for the study of the effect of oleic-acid (OA) coverage on the magnetic behaviour of Co ferrite nanoparticles (CFNs), using high performance computing (HPC). Our study is performed in three different length scales: first we use density functional theory (DFT) calculations to study the magnetic properties of ultra-small OA coated CFNs. Next taking input from the DFT data, we calculate the magnetic characteristics of larger in size OA coated CFNs at an atomic scale. Finally, a mesoscopic modelling approach for interacting assemblies of nanoparticles is employed for the study of the magnetic behaviour of CFNs covered with different percentage of oleic-acid, at finite temperature. The results demonstrate that the DFT magnetic moment and magnetic anisotropy of the nanoparticle decrease with the increase of the percentage of the surfactant. However, in the assembly of CFNs the interplay between the exchange and dipolar inter-particle interactions results in the increase of the magnetic anisotropy and the decrease of the saturation magnetization as the percentage of OA coverage increases, in agreement with experimental findings. The proposed multi-scale computational approach illustrates its ability to handle numerical calculations on complex magnetic interactions of multiple structural components, overcoming computational limitations and to predict optimum parameters for hybrid organic/inorganic nanomaterials for various applications.

Keywords: multiscale modeling, magnetic nanoparticles, DFT calculations, Monte Carlo simulations

Modeling Magnetic Properties of Molecular Nanomagnets Using Genetic Algorithms

Michał Antkowiak
Adam Mickiewicz University
Poznań, Poland
antekm@amu.edu.pl

The advent of high-performance computing (HPC) has significantly transformed the landscape of molecular modeling, enabling researchers to simulate complex systems with unprecedented precision and efficiency. This study explores the use of genetic algorithms (GAs) in conjunction with exact diagonalization methods to model the magnetic properties of molecular nanomagnets, specifically focusing on compounds incorporating rare-earth ions and 3d transition metals. The primary objective is to elucidate the magnetic behaviors of these compounds, including those exhibiting spin crossover phenomena, by leveraging the computational power of HPC to enhance the accuracy and feasibility of such simulations.

Molecular nanomagnets are pivotal in various technological applications, including quantum computing, magnetic storage media, and spintronic devices. The compounds modeled in this study comprise rare-earth ions, specifically three distinct ions, alongside 3d metals such as Nickel (Ni) and Iron (Fe). The unique electronic configurations and strong spin-orbit coupling of rare-earth elements, combined with the diverse magnetic properties of 3d metals, present a rich field for investigation. Additionally, compounds displaying spin crossover, wherein the spin state of a molecule changes due to external stimuli like temperature or pressure, are included to provide a comprehensive understanding of the magnetic phenomena in these systems.

Genetic algorithms, known for their robustness in optimization problems, are employed to explore the parameter space efficiently. These algorithms mimic the process of natural selection by generating a population of potential solutions and iteratively refining them through operations such as selection, crossover, and mutation. This approach is particularly advantageous for the complex parameter landscapes of molecular nanomagnets, where traditional optimization methods may falter due to local minima.

The integration of genetic algorithms with exact diagonalization methods allows for an accurate determination of the energy spectrum and magnetic properties of the compounds. Exact diagonalization, despite its computational intensity, is

crucial for obtaining precise solutions for many-body systems, thereby providing a detailed insight into the magnetic interactions and anisotropies present in the nanomagnets.

The results demonstrate that the proposed methodology successfully captures the magnetic characteristics of the modeled compounds. For the rare-earth ion-containing systems, the simulations reveal intricate details about the interplay between spin-orbit coupling and magnetic anisotropy. In the case of 3d metal compounds, the method accurately describes their magnetic exchange interactions and anisotropic behavior. Furthermore, the study extends to spin crossover compounds, where the genetic algorithm-exact diagonalization framework effectively models the spin transition processes, correlating well with experimental observations.

In conclusion, this research establishes the efficacy of combining genetic algorithms with exact diagonalization in modeling the magnetic properties of molecular nanomagnets. The high computational capabilities afforded by HPC resources are instrumental in handling the complexity and computational demands of these simulations. The insights gained from this study not only advance the understanding of magnetic phenomena in these compounds but also pave the way for the design and optimization of new materials with tailored magnetic properties for advanced technological applications.

Keywords: molecular nanomagnets, numerical simulation, Heisenberg model, genetic algorithms