Workshop on Software Co-Design Actions in European Flagship HPC Codes

INVITED TALK:

Co-design with Proxy-Apps: A match made in heaven?

INVITED SPEAKER: Jens Domke (RIKEN R-CCS)

WEBSITE: https://domke.gitlab.io/

BIO: Jens is a Research Scientist of the High Performance Big Data Research Team at the RIKEN Center for Computational Science (R-CCS), Japan. He received his doctoral degree from the Technische Universität Dresden, Germany, in 2017 for his work on HPC routing algorithms and interconnects. Jens started his career in HPC in 2008, after he and a team of five students of the TU Dresden and Indiana University, won the Student Cluster Competition at SC08. Since then, he published several peer-reviewed journal and conference articles. Jens contributed the DFSSSP and Nue routing algorithms to the subnet manager of InfiniBand, and built the first large-scale HyperX prototype at the Tokyo Institute of Technology. His research focus is on interconnects, topologies, and routing algorithms for HPC systems. Furthermore, he is interested in SDN networks, scheduling algorithms for parallel architectures, and performance evaluation and optimization of parallel applications.

TITLE: Performance Assessment and Energy Efficiency of MaX Codes

SPEAKER: Daniele Cesarini (CINECA)
CoE: MaX

ABSTRACT: In this talk we present the co-design actions put in place by the MaX CoE in the second and conclusion part of the project. In our analysis, we analyzed the performance of applications on selected HPC micro architectures relevant for the EuroHPC context. We focused our investigation on the CPU exploitation by comparing four different micro architectures: Intel Xeon Skylake, ARM ThunderX2, AMD Rome, and IBM Power9. For this co-design activity, the performance and the power efficiency of all MaX codes was extensively analyzed in order to quantify the impact of the MaX's work on the evolution of the codes. We compared the latest version of the MaX flagship codes at the end of the project with a baseline referred to the code version at the beginning of MaX. We will show the evolution in terms of execution time, micro architecture performance, and power efficiency covering the above-mentioned architectures used in several today's supercomputers. This analysis mainly focuses on single-node configuration using general-purpose CPUs, since most of the MaX codes didn't support accelerated systems at the project's start time and because we focused on the micro architecture efficiency of the systems. To conclude, we will present an analysis on one of the MaX flagship codes, Quantum ESPRESSO, when run at scale in order to explore the power efficiency using different MPI and OpenMP configurations.

BIO: Daniele Cesarini received the graduate degree in computer engineering, in 2014, and the PhD degree in electrical engineering, in 2019, both from the University of Bologna, Italy. He is currently an HPC analyst with Cineca HPC department where he works in the area of performance optimization for next-generation HPC architectures. His research interests concern the development of SW-HW co-design strategies and parallel programming support for energy efficient HPC systems.

TITLE: Alya - computational fluid dynamics on exascale GPU hardware for the wind community

SPEAKER: Herbert Owen (BSC)
CoE: EoCoE

ABSTRACT: Alya is an HPC-based multi-physics simulation code developed at BSC that has been designed to run efficiently in parallel supercomputers. Its target domain is engineering, with complex geometries and unstructured meshes. Based on a finite element discretization in space and both explicit and implicit time discretization to solve compressible and incompressible flow problems, solids mechanics, and thermal coupling. Alya is part of the UEABS (Unified European Applications Benchmark Suite) for CPU and GPU. Its scalability has been tested up to 100K cores. This work presents the advances toward Exascale developed within the EoCoE project, the Energy-oriented Center of Excellence. The Centers of Excellence are the European projects in charge of developing codes that can use the exascale computers available shortly. EoCoE focuses on energies that do not contribute to climate change. It deals with five main energy topics: Fusion, Water, Meteorology, Materials, and Wind. For wind, the main objective is improving wind resource assessment over complex terrain with Large Eddy Simulation (LES). The momentum equation is treated explicitly. Therefore the two main kernels for a neutral case are the assembly of a Right Hand Side (RHS) vector and the solution of a linear system for the pressure. For the solution of the linear system or the pressure, we have interfaced Alya with several libraries developed within EoCoE. The best results have been obtained with AGMG and PSCToolkit. This work will present weak scalability results up to 16000M elements on 24k CPU cores using the algebraic multigrid preconditioner within PSCToolkit. Thanks to multigrid, Alya can now obtain correct algorithmic scalability. The RHS vector assembly has been optimized for the GPU to obtain an implementation that is more than 60 times faster than Alya's previous GPU implementation, reaching a floating performance close to 50% of the peak value in an A100 Nvidia GPU. The improvements involved a better use of local memory, specialization, and register lifetime optimization. The specialization includes concentrating exclusively on linear tetrahedral elements for the moment and focusing only on the explicit time discretization. While we had been working on the GPU implementation since 2017, before these improvements, it was not clear if we could obtain an unstructured finite element GPU implementation that was significantly more efficient on GPUs than on CPUs. Now we are confident about the advantages of GPUs for our problems. We have a clear plan to provide the wind community with a GPU code that can efficiently perform LES for complex wind problems, including thermal coupling, canopy, actuator disc models, and forcing from the mesoscale.

BIO: Senior Researcher, Department pf Computer Applications in Science and Engineering Barcelona Supercomputing Center (BSC-CNS). He is involved in the EoCoE European Centre of Excellence.

TITLE: Resources for co-design in the POP CoE

SPEAKER: Xavier Teruel (BSC)
CoE: POP

ABSTRACT: The main objective of the resources for co-design activity within the POP Center of Excellence is to synthesize potential performance problems that may arise when executing HPC applications and to propose software solutions to such problems. The resulting information must be made publicly available for the HPC community (in the shape of an accessible website); including a detailed description that may point to potential co-design activities with other parties (e.g., the compiler or the communication library). The aforementioned problems and solutions are also exemplified using kernels and including their corresponding POP metrics. The Resources for co-design website is a section within the POP website which gathers together a set of aforementioned typical behavioural patterns, potentially resulting in some kind of performance degradation, that POP has identified in our analyses of user applications. For each of these patterns, the site links to the corresponding best-practice(s) that address their performance issues and, in many cases, also provides access to the kernels to allow interested parties to compare the behaviour before and after applying a given best-practice. The site becomes a data base where performance analysts can include their observations and make them available for other actors in the HPC community. The workshop talk will include an introduction of what is our definition of design and co-design and why patterns and best-practices are good candidates to drive the co-design activities. It will continue with a brief description of the main activities carried out within the project concerning the co-design task, and some metrics achieved during the life of the POP2 project. Finally, the presentation will provide an example of how the users may interact with the site.

BIO: Xavier Teruel works as a researcher at the Barcelona Supercomputing Center (BSC) and he has participated in several research projects as well as in the standardization of task-based extensions to the OpenMP programming model. In the context of POP2, he was leading the co- design activities carried out on this project. The main goal was to create the fundamentals of a new data base containing useful resources for co-design.

TITLE: HPC codesign in GROMACS

SPEAKER: Szilárd Páll (KTH)
CoE: BioExcel

ABSTRACT: GROMACS is a widely used molecular dynamics simulation package known for its versatility, performance, and portability, with excellent efficiency and scalability from laptops to supercomputers. This is enabled by state-of-the art parallel algorithms and bottom-up performance optimizations to target all levels of hardware parallelism from SIMD to multicore, NUMA, accelerators and intra- and inter-node. Motivated by the end of Dennard scaling, the increasing need to parallelize, and to make efficient use of microprocessors, the GROMACS engine has evolved through a series of algorithmic and parallelization redesigns. Codesign efforts have been at core of these efforts and this talk will give an overview of these. Fundamental molecular dynamics algorithms have been reformulated to target wide SIMD/SIMT-style architectures combining physics and accuracy-based approach, computer science, and HPC performance engineering. A SIMD abstraction layer and algorithm benchmark mini-app was developed to facilitate codesign and allow quick and relatively easy porting to new SIMD instruction sets without expert knowledge of the algorithms or the application, often in a just hours to days. As early as 2011 an NVIDIA collaboration became an integral part of the GROMACS development. This collaboration had two-way impact: learning from hardware and library engineers provided essential guidance, while our use-cases motivated improvements and features exposed in CUDA like stream priorities, 3D FFTs optimizations, and most recently the distributed cuFFTmp library. The collaboration evolved into an ongoing codesign project with important recent results like: a GPU-resident loop, design and implementation of the direct GPU communication layer, and a distributed GPU-optimized particle mesh Ewald algorithm. In the frame of the Intel CoE collaboration project delivered long-term impact results with a nearly complete SYCL backend of GROMACS, planned to become the new standards-based GPU portability layer with support for all major GPU platforms and the primary means to target AMD and Intel.

BIO: Szilárd Páll is an HPC researcher at the PCD Center for High Performance Computing at KTH Royal Institute of Technology in Stockholm. He has a background in computer science and computational biophysics and has programmed GPU accelerators since 2008. He worked on reformulating key parallel algorithms for modern processor architectures and co-authored the first heterogeneous CPU-GPU parallelization of GROMACS, one of the most widely used HPC simulation software. His recent focus is on efficient asynchronous task scheduling on exascale heterogeneous architectures.

TITLE: Co-Designing a high performance and portable library (QMCKL): one of the major challenges addressed by TREX CoE

SPEAKER: William Jalby (Université de Versailles Saint-Quentin)

CoE: TREX

ABSTRACT: Quantum Monte Carlo (QMC) methods have many application areas, ranging from boosting catalyst design for sustainable fuel production to the simulation of new materials with high critical superconducting temperature. For some large-system applications, QMC methods represent the only possibility to compute high-accuracy properties. However, these methods are extremely CPU intensive and very often require Exascale computing capabilities. Fortunately, their embarrassingly parallel nature facilitates the use of a large number of nodes. However, efficiently using the computational power of recent multicores and accelerators remains a major challenge. To address this issue, the TREX (Targeting Real Chemical Accuracy at the EXascale) CoE (https://trex-coe.eu/) has chosen to develop a high performance portable library QMCKL. This library aims at fulfilling three (classical) goals: productivity, performance and portability. Simultaneously reaching these three goals is rather challenging and we will discuss here how, in QMCKL, we are tackling these issues, namely,

- by making available multiple library versions: a reference one easy to integrate and understand and several optimized versions;
- by making tradeoffs between performance and portability;
- by using high level algorithm descriptions to encompass multiple optimizations common to different architectures;
- by defining a hierarchy of software layers trying to unify different CPU ISA
- by relying on different tools such as auto-tuners to get the “last mile” in performance;
- by taking into account numerical accuracy issues allowing the developer to use lower precision (including FP16 format)

We will demonstrate QMCKL use within QMC=CHEM an important QMC code and we will show performance analysis on different target architectures. These developments are first supported by the use of MAQAO, an advanced performance analysis framework (www.maqao.org) which helps the developer select the most profitable optimization. Numerical accuracy work is supported by the use of VERIFICARLO (https://github.com/verificarlo/verificarlo), a tool for debugging and assessing floating point precision and reproducibility.

BIO: W. Jalby started his career first at INRIA as a researcher then joined University of Illinois (CEDAR project), got appointed Associate Professor at University of RENNES I before joining University of Versailles as a Full Professor. His research interests are focusing on memory system analysis and optimization, compilers, parallelism and performance analysis methods/tools. Most of his research has been carried out in close collaboration with hardware suppliers (Fujitsu, Bull and INTEL), tools developers (JSC, TUD, University of Oregon, CAPS Entreprise) and application developers both from research (CEA, EDF, CNRS) and Industry (ESI, MAGMAsoft, Dassault, GNS, RECOM, ...). Since 2004, he is the director of a joined Lab (ITACA) between CEA DAM and UVSQ focusing on code optimization techniques. In 2010, he got appointed as CTO of the Exascale Computing Research Lab (a joined HPC laboratory founded by CEA, INTEL and UVSQ) and he is leading ECR research activities. He has authored over 100 technical publications in international journals and conferences and directed over 35 PhD Theses. He has been Coordinating UVSQ participation to two CoE’s: POP2 and TREX.

TITLE: Performance analysis and code optimizations for distributed training of autoencoders

SPEAKER: Rakesh Sarma (JSC)
CoE: RAISE

ABSTRACT: CoE RAISE develops innovative AI methods on heterogeneous high-performance computing (HPC) architectures capable of scaling towards Exascale. A framework to apply various machine learning (ML) methods on HPC systems using exceptionally large simulation data is developed. This framework is deployed on various HPC systems in Europe. It is optimized for performance, maintained to adapt to the latest system changes, and documented for universal application. The parallelization of the ML training is based on a distributed data-parallelism (DDP) approach, i.e., it relies on distributing the training dataset to multiple workers, where the trainable parameters of the network are occasionally exchanged between the workers. A single epoch of a reference convolutional autoencoder performing a training on a large computational fluid dynamics (CFD) dataset with 65,000 trainable parameters is estimated to take 28 hours on a single worker, i.e., a graphics processing unit (GPU). As the trained model converges after around 1,000 epochs, the corresponding training would require several months. Applying the DDP approach decreases the computational time for a single epoch to less than 0.5 hours using 512 workers employing the NVIDIA Collective Communications Library (NCCL). Further investigations show that the data loader is a bottleneck of the training. A reduction of the single epoch time by 90% is achieved by optimizing the data loader and memory structure, resulting in an epoch time of 160 seconds. With these optimizations, less than two days are required to fully train the model. A consequence of using the DDP approach is an increase of the mini-batch size when more workers are employed, which affects the training accuracy. Adjusting other training hyperparameters can overcome this issue. Such hyperparameter tuning is, however, costly due to the required repeated computations. It is therefore recommended to further reduce the solution time by further acceleration of the epoch computations.

BIO: Rakesh Sarma works at Forschungszentrum Jülich, contributing primarily to the CoE RAISE project in the aspect of developing parallel and scalable AI methods for HPC applications. He obtained his PhD from Delft University of Technology, Netherlands, in 2018. His doctoral thesis was on the development of Bayesian inference and reduced order modelling methods for prediction of instabilities in aeroelastic structures. Thereafter, he worked at the Dutch National Center for Mathematics and Computer Science in Amsterdam in the domain of ML/AI in space weather and stratified turbulence applications.