Workshop Agenda and speaker's profile

Dr Massimiliano and Ganesan Narayanasamy - Introduction about the workshop 2.30 pm CEST

Dr Jose Moreira, IBM Research Open ISA and Power Innovations 2.45 pm CEST

Dr Sameer Shende , TAU founder and University of Oregon E4S and TAU features and Demo 3.15 PM CEST

Brian Allison , IBM Advanced Accelerator 3.45 PM CEST

Dr Panda , Ohio State University Distributed Deep learning and tools 4.15 PM CEST

Florin Manaila , IBM AI at Scale . 4.45 PM CEST

Conclusion summary - by
Dr Massimiliano and Ganesan Narayanasamy 5.15 PM

Title: Advanced High-Performance Computing Features of the OpenPOWER ISA

Abstract: Power ISA processors have a long history of offering superior features for HPC applications. Well known examples include POWER3, used in the ASCI White supercomputer, various PowerPC processors used in the Blue Gene family of massively parallel computers, and POWER9, present in the leading supercomputers of today, Summit and Sierra. OpenPOWER ISA has enabled open access to many of these features. IBM's most recent contribution to OpenPOWER ISA, in the form of Power ISA Version 3.1, includes the Matrix-Multiply Assist (MMA) instructions. The MMA instructions are designed to deliver additional performance both for classical high-performance computing, in the space of scientific and technical computing, and for the increasingly important space of business analytics. In addition, the Open Memory Interface (OMI), also developed by IBM, opens new levels of memory bandwidth and capacity for the most demanding applications. Our goal is to raise awareness of and interest in these new features, which we believe can lead to further research in processor architecture and programming environments. Some of the most promising application areas include graph algorithms, classical machine learning and deep learning.

Speaker's Profile

Dr JoséE. Moreira is a Distinguished Research Staff Member in the Scalable Systems Department at the Thomas J. Watson Research Center. He received a B.S. degree in physics and B.S. and M.S. degrees in electrical engineering from the University of Sao Paulo, Brazil, in 1987, 1988 and 1990, respectively. He also received a Ph.D. degree in electrical engineering from the University of Illinois at Urbana-Champaign in 1995. Since joining IBM at the Thomas J. Watson Research Center, he has worked on a variety of high-performance computing projects. He was system software architect for the Blue Gene/L supercomputer and chief architect of

the Commercial Scale Out project. He currently leads the IBM Research work on the architecture of POWER processor. He is an author or coauthor of over 100 technical papers and 15 US patents. Dr. Moreira is a Senior Member of the IEEE (Institute of Electrical and Electronics Engineers) and a Distinguished Scientist of the ACM (Association for Computing Machinery).

Title : Distributed Deep learning and Tools

Abstract

This talk will start with an overview of challenges being faced by the AI community to achieve high-performance, scalable and distributed DNN training on Modern HPC systems with both scale-up and scale-out strategies. After that, the talk will focus on a range of solutions being carried out in my group to address these challenges. The solutions will include: 1) MPI-driven Deep Learning, 2) Co-designing Deep Learning Stacks with High-Performance MPI, 3) Out-of-core DNN training, and 4) Hybrid (Data and Model) parallelism. Case studies to accelerate DNN training with popular frameworks like TensorFlow, PyTorch, MXNet and Caffe on modern HPC systems will be presented.

Speaker's Profile:

DK Panda is a Professor and University Distinguished Scholar of Computer Science and Engineering at the Ohio State University. He has published over 450 papers in the area of high-end computing and networking. The MVAPICH2 (High Performance MPI and PGAS over InfiniBand, Omni-Path, iWARP and RoCE) libraries, designed and developed by his research group (http://mvapich.cse.ohio-state.edu), are currently being used by more than 3,090 organizations worldwide (in 89 countries). More than 739,000 downloads of this software have taken place from the project's site. This software is empowering several InfiniBand clusters (including the 3rd, 5th, 8th, 14th, 15th, and 18th ranked ones) in the TOP500 list. The RDMA packages for Apache Spark, Apache Hadoop and Memcached together with OSU HiBD benchmarks from his group (http://hibd.cse.ohio-state.edu) are also publicly available. These libraries are currently being used by more than 320 organizations in 35 countries. More than 36,400 downloads of these libraries have taken place. High-performance and scalable versions of the Caffe and TensorFlow framework are available from https://hidl.cse.ohio-state.edu. Prof. Panda is an IEEE Fellow. More details about Prof. Panda are available at http://www.cse.ohio-state.edu/~panda.

Title : Performance Evaluation using TAU Performance System and E4S

Abstract : The DOE Exascale Computing Project (EC) Software Technology focus area
is developing an HPC software ecosystem that will enable the efficient
and performant execution of exascale applications. Through the

Extreme-scale Scientific Software Stack (E4S), it is developing a comprehensive and coherent software stack that will enable application developers to productively write highly parallel applications that can portably target diverse exascale architectures - including the IBM OpenPOWER with NVIDIA GPU systems. E4S features a broad collection of HPC software packages including the TAU Performance System(R) for performance evaluation of HPC and AI/ML codes. TAU is a versatile profiling and tracing toolkit that supports performance engineering of codes written for CPU and GPUs and has support for most IBM platforms. This talk will give an overview of TAU and E4S and how developers can use these tools to analyze the performance of their codes. TAU supports transparent instrumentation of codes without modifying the application binary. The talk will describe TAU's support for CUDA, OpenACC, pthread, OpenMP, Kokkos, and MPI applications. It will describe TAU's use for Python based frameworks such as Tensorflow and PyTorch. It will cover the use of TAU in E4S containers using Docker and Singularity runtimes under ppc64le. E4S provides both source builds through the Spack platform and a set of containers that feature a broad collection of HPC software packages. E4S exists to accelerate the development, deployment, and use of HPC software, lowering the barriers for HPC users.

Websites:
Extreme-scale Scientific Software Stack (E4S) [https://e4s.io]
TAU Performance System [http://tau.uoregon.edu]

Speaker’s Profile

Dr. Sameer Shende serves as the Director of the Performance Research Laboratory at the University of Oregon and the President and Director of ParaTools, Inc. and ParaTools, SAS in France. He serves as the lead developer of the TAU Performance System, Program Database Toolkit (PDT), HPCLinux, and Extreme-Scale Scientific Software Stack (E4S). His research interests include performance instrumentation, compiler optimizations, measurement, and analysis tools for HPC. He leads the SDK project for the Exascale Computing Project (ECP), in the Programming Models and Runtime (PMR) area to provide ECP Software Technology (ST) products in a containerized environment for HPC. He has served as the the chair of the Performance Measurement, Modeling, and Tools track at SC17 and the co-chair for the technical program at the ICPP 2018 conferences. He received his B.Tech from IIT Bombay in 1991, and his M.S. and Ph.D. from the University of Oregon in 1996 and 2001 respectively.

Presentation Title: AI in IBM

Abstract: The session will present HPC challenges in fuelling machine learning and deep learning into the simulations. Besides, we will present a user-centric view of IBM Watson ML Community Edition and the newly IBM inference system IC922 adoption into AIops of large HPC clusters (from deployment to inference).

Speaker’s Profile

Florin leads enterprise transformation by adopting of High-Performance Computing, Distributed Deep Learning and Emerging Technologies in order to help them to rapidly transform the way businesses operate, solve problems, and gain competitive advantage. He is responsible for performance, availability, and scalability of Cognitive Systems infrastructure. It has created IBM Distributed Deep Learning reference architecture as well as important industry blueprints of applied AI including edge fabric. His client experience covers EU commercial and government civil and defence agencies. Is passionate about in-memory computing and Spiking Neural Networks.

Title : Advanced Accelerator – OpenCAPI

Abstract: The Open Coherent Accelerator Processor Interface (OpenCAPI) is an industry-standard architecture targeted for emerging accelerator solutions and workloads. This session will address these following areas : 1.) The latest technology advancements surround OpenCAPI, 2.) The OpenCAPI strategy as it relates to the other industry acceleration standards. ie Intel's CXL, Gen-Z and CCIX, 3.) The open initiatives surrounding OMI and OpenCAPI 3.0 and GitHub, 4.) Industry Open Source Initiatives around OpenCAPI, 5.) OC-Accel - Our new FPGA programming framework, supporting OpenCAPI 3.0, targeting higher level programming languages such as C, C++ 6.) Interesting Use Cases 7.) China's OpenCAPI Contest

Speaker’s Profile:

Brian Allison is an IBM Senior Technical Staff Engineer and Chief Engineer for External Engagements including CAPI and OpenCAPI. Over his 28 year career at

IBM, Mr. Allison was the Chief Engineer of various industry leading IBM chip sets that spanned cache coherent, memory and node controllers. His expertise is in computer architecture and logic design