International Workshop on Machine Learning Hardware (IWMLH)

https://mlhardware.github.io

Abstract:

Recent years have seen a surge of investment in AI chip companies worldwide. These companies are, however, mostly targeting applications outside of the scientific computing community. As the use of Machine Learning(ML) accelerates in the HPC field itself, there is concern that the scientific community should influence the design of this new specialized hardware. Indeed, scientific computing has a distinctive set of requirements regarding workload type, usage model, and platform administration. How those chips answer those demands will shape the future of their integration within the global scientific computing infrastructure. In this workshop, we propose to let the community and select vendors engage on questions related to the programming models, compiler toolchains, system interfaces, and architecture trade-offs of these chips. It is crucial that such interactions happen early in this field, and ISC is a natural venue for those interactions to happen on a global scale. This proposal follows the outcome of a successful SC19 BoF, where an emphasis on compiler technology emerged. Accordingly, this workshop will feature an invited talk from A.Cohen (Google Research) on the Multi-Level Intermediate Representation(MLIR) project, followed by technical presentations from various industrial participants: Groq, SambaNova, GraphCore, Cerebras, and Preferred Networks.

More on the context and subject matter:

The main goal for this workshop is to foster exchange between ML Hardware (also called AI Chip) companies and their potential HPC users. There are many design questions in this highly heterogeneous chip landscape. We believe that both industrial players and the HPC field would benefit from having early discussions about how these chips can fit scientific HPC workloads. Indeed, it is potentially risky for the HPC community as a whole to let industrial machine learning users dominate the discussion with nascent hardware vendors. HPC ML workloads have different needs, which we hope this workshop will push.

The main takeaway from the Machine-Learning Hardware: Architecture, System Interfaces, and Programming Models SC19 BoF was that the main source of uncertainty and complexity inherent to these chips comes from the software stack and compiler toolchain used to map models to hardware. For this workshop to be successful, it must inform the scientific community and stimulate in-depth exchanges with vendors on these technical aspects. In order to enable these goals, the workshop will start with an invited talk by one of the main actors from the MLIR (https://mlir.llvm.org/) project (Albert Cohen, G Research).

ISC Relevance:

This workshop will serve as a forum to inform the global scientific computing community on ML accelerators and let them engage with vendors. The PC is global and the industrial participant list is a healthy mix of companies based in the US (Groq, SambaNova, Cerebras), UK (Graphcore) and Japan(Preferred Networks). The invited speaker is from an european research center.

Workshop format and framing:

We will focus the session on technical aspects of the hardware and software associated with these chips. The structure of the session will be threefold:

First, a major actor from the MLIR project will give an invited talk (Albert Cohen - Title and abstract to be determined).

Second, we will have speakers from select AI chip companies present their architectures with a focus on HPC use. Cerebras, SambaNova, GraphCore, Groq, and Preferred Networks will participate, and we are discussing with one other company in that space (see the “participants” section below). We have already ran a BoF session with the first four vendors in this list and are confident that they will know how to frame their technical presentation for an HPC audience, and the other(s) companies should have HPC researchers present. We are asking them to describe the architecture and its power/performance tradeoffs, the programming models and compiler toolchains they envision for their platforms, as well as the low-level interfaces (explicit memory discoverability and access, RDMA, etc) they plan to expose (see the “scope” section below).

Third, we will encourage the attendees to give feedback to the speakers about the relevance of these architectures to their scientific HPC workloads. The session organisers will drive informal discussions about how these architectures could benefit HPC. This discussion will encompass both the needs of the workloads themselves as well as the software stacks and programming models future accelerators would have to interact with in the HPC space. The goal is to identify gaps in the current design of ML hardware or its software interfaces, and discuss how the scientific computing community could help close those gaps. In particular, we would like to identify relevant HPC efforts that could provide benchmarks, design feedback, and integration with vendors APIs.

Participants:

Albert Cohen (Invited "keynote" talk). Albert is a research scientist at Google. He has been a research scientist at Inria from 2000 to 2018. He graduated from École Normale Supérieure de Lyon and received his PhD from the University of Versailles in 1999 (awarded two national prizes). He has been a visiting scholar at the University of Illinois, an invited professor at Philips Research, and a visiting scientist at Facebook Artificial Intelligence Research. Albert Cohen works on parallelizing and optimizing compilers, parallel programming languages and systems, and synchronous programming for reactive control systems. He served as the general or program chair of major conferences, including PLDI, PPoPP, HiPEAC, CC, the embedded software track of DAC, and as a member of the editorial board of ACM TACO and IJPP. He coauthored more than 180 peer-reviewed papers and has been the advisor for 26 PhD theses. Several research projects initiated by Albert Cohen resulted in effective transfer to production compilers and programming environments.

- **Groq (**https://groq.com): John Barrus. John Barrus is the Director of Business Development at Groq where the team is delivering a high-performance processor with sub-millisecond response times to accelerate ML workloads. Before Groq, Barrus launched accelerated computing on Google Cloud, introducing both GPU and TPU products. Prior to working as a Sr. Product Manager at Google, Barrus was the VP of Research and Director of the Ricoh Innovations Research Lab. He started Ricoh's EWS business unit and became the VP of Engineering after developing new cloud-enabled tablet technology at the lab. Barrus has started and launched 5 new businesses and raised over $100M in funding for new businesses and products. As a Distinguished Research Scientist, Barrus led groups of scientists and engineers in developing new core technologies, products and businesses worldwide. Barrus has published 25 technical papers and journal articles and has 90 issued U.S. patents. Barrus earned a Ph.D. in Mechanical Engineering from MIT.
- **SambaNova** (https://sambanova.ai): To be announced
- **GraphCore** (https://www.graphcore.ai): Matt Fyles. Matt Fyles is the Vice President of Software at GraphCore. Matt is a a computer scientist with over 20 years' experience in the design, development, delivery and support of software and hardware for the microprocessor market, spanning a wide range of applications from consumer electronics to high performance computing, with a particular focus on parallel processors.
- **Cerebras** (https://www.cerebras.net): Andy Hock. Dr Andy Hock is the head of Product at Cerebras Systems. He and his team are building a new class of computer system to accelerate deep learning, AI, data science and analytics. His current work with Cerebras is directed at addressing the big challenge in compute that he feels AI researchers face today. Previously, Andy was the Product lead for data science and analytics at Google for the "Terra Bella" satellite image processing project.
- **Preferred Networks** (https://preferred.jp): Kei Hiraki, University of Tokyo Emeritus Professor.

Scope:

The workshop will feature the participation of select AI accelerator companies, with discussions centered on the following aspects:

- **Programming models**. Most AI chips are designed to leverage regularity in the dataflow inherent in ML models. This design supposes the use of standard model representation formats and/or custom dataflow graph formats. Those programming models are of high interest to the scientific community. We need to understand which types of ML/scientific applications will be supported by each platform, as well as the associated development and maintenance costs.
- **Compiler toolchain**. ML accelerators often rely on complex compilation technology to map dataflow descriptions to hardware. These compilers can be radically different from existing compiler stacks in that they may solve complex placement and routing problems. The usage model of those compiler toolchains in a scientific computing context is a subject worthy of discussion. Indeed, the features of the compilers and constraints around their use will play a large role in the scientific process itself.
- **System interfaces**. Understanding what low-level system interfaces will be available can help cast light on which usage and administration model to expect. In particular, we'd like participants to discuss expected capabilities in terms of concurrency, partitioning, debugging, power management, and performance characterization.
- **Architecture**. Successful integration of AI chips in the existing scientific computing infrastructure will require a proper understanding of chips in terms of operator optimization and customization, bandwidth, latency, memory management, power demands, and network capabilities.

Expected Outcome:

The PC will publish invited talk and industrial participant slides to the website (https://mlhardware.github.io). The subject matter of this workshop is novel and rich enough for the event to be reconvened yearly at the same venue.

More importantly, we hope that this workshop will nurture in-depth interactions between the AI Chip companies and the scientific computing community at large, and that such discussions will drive forward R&D efforts from both sides.

Advertising:

Industrial vendors are willing to advertise the workshop and its website (https://mlhardware.github.io) through their social media outreach.

Program Committee:

Pete Beckman is the co-director of the Northwestern-Argonne Institute for Science and Engineering. From 2008-2010 he was the director of the Argonne Leadership Computing Facility, where he led the Argonne team working with IBM on the design of Mira, a 10 petaflop Blue Gene/Q. Pete coordinates the collaborative research activities in extreme-scale computing between the US Department of Energy and Japan's ministry of education, science, and technology (MEXT), and leads Argo, an Exascale Computing Project focused on low-level resource management for the OS and runtime. He is the founder and leader of the Waggle project for AI@Edge. The Waggle technology and software framework is being used by the Chicago Array of Things project and is deployed in over 10 cities around the world. Dr. Beckman has a Ph.D. in Computer Science from Indiana University (1993)

Swann Perarnau is an Assistant Computer Scientist at Argonne. He leads the topology, memory and power management efforts for the Argo ECP project. In particular, he is designing low-level system software mechanisms to help applications discover the features and performance of complex heterogeneous hardware, as well as composable abstractions to make the most efficient use of it.

Rosa M. Badia holds a PhD from the UPC (1994). She is the manager of the Workflows and Distributed computing group at the Barcelona Supercomputing Center (BSC). She is a Scientific Researcher at the Spanish National Research Council (CSIC). She graduated on Computer Science at the Facultat d' Informàtica de Barcelona (UPC, 1989). She was lecturing and doing research at the Computer Architecture Department (DAC) at the UPC from 1989 to 2008, where she held an Associate Professor position from 1997 to 2008; she is currently part-time lecturing again at the same department.

Kentaro Sano is the team leader of the Processor Research team at RIKEN. Dr. Kentaro Sano received his Ph.D. from GSIS, Tohoku University, in 2000. Since 2000 until 2005, he had been a Research Associate at Tohoku University. Since 2005 until 2018, he has been an Associate Professor at Tohoku University. He was a visiting researcher at the Department of Computing, Imperial College, London, and Maxeler corporation in 2006 and 2007. Since 2017 until present, he has been a team leader of a processor research team at R-CCS, Riken. His research interests include FPGA-based high-performance reconfigurable computing systems especially for scientific numerical simulations and machine learning, high-level synthesis compilers and tools for reconfigurable custom computing machines, and system architectures for next-generation supercomputing based on the data-flow computing model.

Valentin Reis is a Postdoctoral appointee at Argonne. He leads the effort towards dynamically reconfigurable hardware within the Argo ECP project. In particular, he is interested in software-based approaches to automatically learn the best application performance/power tradeoff point of CPU power scaling. His interests span machine learning, functional programming and infrastructure.