MLSys 2020 Workshop book

Generated Mon Feb 17, 2020

Workshop organizers make last-minute changes to their schedule. Download this document again to get the lastest changes, or use the MLSys mobile application.

Schedule Highlights

March 4, 2020

Ballroom A, **On-Device Intelligence** *Chandra, Warden, Venkatesh, Lin*

Level 1 Room 3, **SARA: Secure and Resilient Autonomy** *Bose, Chandramoorthy, Vega, Swaminathan*

Level 3 Room 10, **Automated Machine Learning For Networks and Distributed Systems** *Arzani, Darvish Rouhani*

Level 3 Room 5, **MLOps Systems** *Dutta, Zaharia, Zhang*

Level 3 Room 6, **Benchmarking Machine Learning Workloads on Emerging Hardware** *St John, Emani*

Level 3 Room 8, **Resource-Constrained Machine Learning (ReCoML 2020)** *Ben Itzhak, Narodytska, Aberger*

Level 3 Room 9, **Software-Hardware Codesign for Machine Learning Workloads** *Gupta, Wohlbier, Low, Vetter, Vassilieva*

On-Device Intelligence

Vikas Chandra, Pete Warden, Ganesh Venkatesh, Yingyan Lin

Ballroom A, Wed Mar 04, 09:00 AM

AI has the potential to transform almost everything around us. It can change the way humans interact with the world by making the objects around them "smart" — capable of constantly learning, adapting, and providing proactive assistance. The beginnings of this trend can already be seen in the new capabilities coming to smartphones (speech assistant, camera night mode) as well as the new class of “smart” devices such as smart watches, smart thermostats, and so on. However, these “smart” devices run much of the computation on the cloud (or a remote host) — costing them transmission power and response latency as well as causing potential privacy concerns. This limits their ability to provide a compelling user experience and realize the true potential of an “AI everywhere” world.

This workshop seeks to accelerate the transition towards a truly “smart” world where the AI capabilities permeate to all devices and sensors. The workshop will focus on how to distribute the AI capabilities across the whole system stack and co-design of edge device capabilities and AI algorithms. It will bring together researchers and practitioners with diverse backgrounds to cover the whole stack from application domains such as computer vision and speech, to the AI and machine learning algorithms that enable them, to the SoC/chip architecture that run them, and finally to the circuits, sensors, and memory technologies needed to build these devices.

Workshop Schedule Highlights

Morning Session: Enabling new experiences on smart devices and agents
Keynote Speaker: [Blaise Aguera y Arcas](https://en.wikipedia.org/wiki/Blaise_Ag%C3%BCera_y_Arcas)
Speaker Bio:
Blaise leads an organization at Google AI working on both basic research and new products. Among the team’s public contributions are MobileNets, Federated Learning, Coral, and many Android and Pixel AI features. They also founded the Artists and Machine Intelligence program, and collaborate extensively with academic researchers in a variety of fields. Until 2014 Blaise was a Distinguished Engineer at Microsoft, where he worked in a variety of roles, from inventor to strategist, and led teams with strengths in interaction design, prototyping, machine vision, augmented reality, wearable computing and graphics. Blaise has given TED talks on Seadragon and Photosynth (2007, 2012), Bing Maps (2010), and machine creativity (2016). In 2008, he was awarded MIT’s TR35 prize.

Afternoon Session: Model, Software and Hardware co-design and optimization
Keynote Speaker: [Diana Marculescu](http://www.ece.utexas.edu/people/faculty/diana-marculescu)
Speaker Bio:
Diana Marculescu is Department Chair, Cockrell Family Chair for Engineering Leadership #5, and Professor, Motorola Regents Chair in Electrical and Computer Engineering #2, at the University of Texas at Austin. Before joining UT Austin in December 2019, she was the David Edward Schramm Professor of Electrical and Computer Engineering, the Founding Director of the College of Engineering Center for Faculty Success (2015-2019) and has served as Associate Department Head for Academic Affairs in Electrical and Computer Engineering (2014-2018), all at Carnegie Mellon University. She received the Dipl.Ing. degree in computer science from the Polytechnic University of Bucharest, Bucharest, Romania (1991), and the Ph.D. degree in computer engineering from the University of Southern California, Los Angeles, CA (1998). Her research interests include energy- and reliability-aware computing, hardware aware machine learning, and computing for sustainability and natural science applications. Diana was a recipient of the National Science Foundation Faculty Career Award (2000-2004), the ACM SIGDA Technical Leadership Award (2003), the Carnegie Institute of Technology George Tallman Ladd Research Award (2004), and several best paper awards. She was an IEEE Circuits and Systems Society Distinguished Lecturer (2004-2005) and the Chair of the Association for Computing Machinery (ACM) Special Interest Group on Design Automation (2005-2009). Diana chaired several conferences and symposia in her area and is currently an Associate Editor for IEEE Transactions on Computers. She was selected as an ELATE Fellow (2013-2014), and is the recipient of an Australian Research Council Future Fellowship (2013-2017), the Marie R. Pistilli Women in EDA Achievement Award (2014), and the Barbara Lazarus Award from Carnegie Mellon University (2018). Diana is a Fellow of both ACM and IEEE.

Important Deadlines
- Submission deadline: Jan 15, 2020
- Paper decision notification: Jan 27, 2020
- Presentation/Poster for accepted submissions: Feb 28, 2020

Schedule

09:00 AM	**Enabling new experiences on smart devices and agents**
11:00 AM	**AutoML for on-device vision by Minxing Tan (Google)**
11:30 AM	**Imitation Learning from Observation by Prof Peter Stone**
02:00 PM	**Model, Software and Hardware Co-optimization**
04:00 PM	**TFLite: deploying models on micro controllers by Nat Jeffries (Google)**
04:30 PM	**How to Evaluate Deep Learning Accelerators by Prof. Vivienne Sze**

SARA: Secure and Resilient Autonomy

Pradip Bose, Nandhini Chandramoorthy, Augusto Vega, Karthik Swaminathan

Level 1 Room 3, Wed Mar 04, 09:00 AM

This workshop will bring classical system architecture and design experts and AI/ML algorithmic experts together in one forum. The goal is to brainstorm about challenges in designing secure and resilient AI-centric systems in general, but with a special focus on autonomous systems (such as self-driving cars and industrial robots) - where safety and security are of paramount value.

The knowledge and expertise of classical mainframe and server architects who are experts in designing ultra-reliable and secure systems will be blended with domain experts in AI - particularly those with an established expertise in developing reliable and secure AI algorithms.

Detailed workshop information, abstract submission instructions, dates: https://sara-workshop.org

Schedule

09:00 AM	**Introduction: Dr. Pradip Bose (IBM Research)**
09:05 AM	**Keynote 1: Dr. Thomas Rondeau (DARPA)**
09:50 AM	**Coffee Break**
10:05 AM	**Perspectives from Industry: Dr. Sanu Matthew (Intel Corp)**
10:25 AM	**Paper 1: Presentation**
10:40 AM	**Perspectives from Academia: Prof. Saibal Mukhopadhyay (Georgia Tech)**
11:00 AM	**Paper 2: Presentation**
11:15 AM	**Paper 3: Presentation**
11:30 AM	**Poster Session**
12:00 PM	**Lunch Break**
01:30 PM	**Poster Session (contd)**
02:00 PM	**Keynote 3: Prof. Xue Lin (Northeastern University)**
02:45 PM	**Paper 4: Presentation**
03:00 PM	**Paper 5: Presentation**
03:15 PM	**Embedded Tutorial: Dr. Pin-Yu Chen (IBM Corp)**
03:45 PM	**Coffee Break**
04:00 PM	**Panel Discussion: Pin-Yu Chen (IBM), Akshay Deshpande (Soothsayer Analytics), Xue Lin (Northeastern University), Sarita Adve (UIUC); Moderator: Dr. Pradip Bose (IBM)**
05:00 PM	**Closing Remarks: Organizers (IBM)**

Abstracts (1):

Abstract 18: **Closing Remarks: Organizers (IBM) in SARA: Secure and Resilient Autonomy**, 05:00 PM

Closing remarks and discussion on special journal issue.

Automated Machine Learning For Networks and Distributed Systems

Behnaz Arzani, Bita Darvish Rouhani

Level 3 Room 10, Wed Mar 04, 09:00 AM

The first workshop on "Towards A Domain-Customized Automated Machine Learning Framework For Networks and Systems" aims at creating a coalition of researchers who aim to build an AutoML platform for network operators. The platform helps network operators bridge the expertise gap when using ML to solve challenging networking problems.

Researchers at this workshop will discuss how we, as a community, can build a framework that: enables users to use ML to solve problems in networked systems without having in-depth ML expertise and that, similarly, enables ML experts to contribute to solving problems in networked systems without having expertise in these domains. We will discuss whether existing AutoML frameworks, as-is can be used by network operators? If not, what needs to change? Can domain customization help? If yes, what are the components of a domain-customized AutoML framework and how are they different from traditional AutoML solutions? What are the important criteria that such a system needs to meet? What are the techniques we can use to build such a framework? What are the collaborations we can initiate across industry and academia to make headway on solving this problem?

Schedule

09:00 AM	**[Coming soon] We will have an updated schedule on the workshop website**

MLOps Systems

Debo Dutta, Matei Zaharia, Ce Zhang

Level 3 Room 5, Wed Mar 04, 09:00 AM

Due to the complexity in putting ML into production, the actual machine learning capability is a small part of a complex system and its lifecycle. This new evolving field is known as MLOps. Informally MLOps typically refers to the collaboration between data scientists and operations engineers (e.g. SRE) to manage the lifecycle of ML within an organization. This space is new and has yet to be explored from a research perspective.

In this workshop we aim to cover research problems in MLOps, including the systems and ML challenges involved in this process. We will also cover the software engineering questions including specification, testing and verification of ML software systems. We will bring together a wide variety of experts from both industry and academia, covering persona ranging from data scientists to machine learning engineers.

Benchmarking Machine Learning Workloads on Emerging Hardware

Tom St John, Murali Emani

Level 3 Room 6, Wed Mar 04, 09:00 AM

With evolving system architectures, hardware and software stacks, diverse machine learning (ML) workloads, and data, it is important to understand how these components interact with each other. Well-defined benchmarking procedures help evaluate and reason the performance gains with ML workload-to-system mappings. We welcome all novel submissions in benchmarking machine learning workloads from all disciplines, such as image and speech recognition, language processing, drug discovery, simulations, and scientific applications. Key problems that we seek to address are: (i) which representative ML benchmarks cater to workloads seen in industry, national labs, and interdisciplinary sciences; (ii) how to characterize the ML workloads based on their interaction with hardware; (iii) which novel aspects of hardware, such as heterogeneity in compute, memory, and networking, will drive their adoption; (iv) performance modeling and projections to next-generation hardware. Along with selected publications, the workshop program will also have experts in these research areas presenting their recent work and potential directions to pursue.

Call for Papers can be found here:
https://memani1.github.io/challenge20/

Paper Submission Deadline: January 15, 2020
Author Notification: January 27, 2020
Camera-Ready Papers Due: February 21, 2020

Schedule

09:00 AM	**Introduction - Tom St. John (Tesla Inc.)**
09:10 AM	**MLPerf Inference Deep Dive - Vijay Janapa Reddi (Harvard University)**
10:00 AM	**Morning Break**
10:30 AM	**Morning Paper Session**
11:20 AM	**Formula One vs. Family Car, or the Need for Broader, Generalizable Benchmarks - Natalia Vassilieva (Cerebras Systems)**
12:00 PM	**Lunch**
02:00 PM	**Benchmarking Science: Datasets and Exascale Infrastructure - Geoffrey Fox (Indiana University)**
02:45 PM	**Afternoon Paper Session**
03:30 PM	**Afternoon Break**
04:00 PM	**Panel - Peter Mattson (Google), Shuaiwen Song (University of Sydney), Gennady Pekhimenko (University of Toronto), Carole-Jean Wu (Facebook, Arizona State University), Elise Jennings (Argonne National Laboratory), Grigori Fursin (CodeReef)**
05:20 PM	**Conclusion - Murali Emani (Argonne National Laboratory)**

Abstracts (2):

Abstract 4: **Morning Paper Session in Benchmarking Machine Learning Workloads on Emerging Hardware**, 10:30 AM

Precious: Resource-Demand Estimation for Embedded Neural Network Accelerators - Stefan Raif (FAU Erlangen-Nürnberg), Benedict Herzog (FAU Erlangen-Nürnberg), Judith Hemp (FAU Erlangen-Nürnberg), Timo Hönig (FAU Erlangen-Nürnberg), Wolfgang Schröder-Preikschat (FAU Erlangen-Nürnberg)

Benchmarking Machine Learning Workloads in Structural Bioinformatics Applications - Heng Ma (Argonne National Laboratory), Austin Clyde (Argonne National Laboratory), Venkatram Vishwanath (Argonne National Laboratory), Debsindhu Bhowmik (Oak Ridge National Laboratory), Arvind Ramanathan (Argonne National Laboratory), Shantenu Jha (Rutgers University, Brookhaven National Laboratory)

Benchmarking Alibaba Deep Learning Applications Using AI Matrix - Wei Zhang (Alibaba Group), Wei Wei (Alibaba Group), Lingjie Xu (Alibaba Group), Lingling Jin (Alibaba Group)

Abstract 8: **Afternoon Paper Session in Benchmarking Machine Learning Workloads on Emerging Hardware**, 02:45 PM

Deep Learning Workload Performance Auto-Optimizer - Connie Yingyu Miao (Intel Corporation), Andrew Yang (Intel Corporation), Michael Anderson (Intel Corporation)

Challenges with Evaluating ML Solutions in Data Centers - Shobhit Kanaujia (Facebook), Wenyin Fu (Facebook), Abhishek Dhanotia

(Facebook)

Benchmarking TinyML Systems: Challenges and Direction - Colby Banbury (Harvard University), Vijay Janapa Reddi (Harvard University), Will Fu (Harvard University), Max Lam (Harvard University), Amin Fazel (Samsung Semiconductor Inc.), Jeremy Holleman (Syntiant, University of North Carolina Charlotte), Xinyuan Huang (Cisco Systems), Robert Hurtado (HurtadoTechnology Inc.), David Kanter (Real World Insights), Anton Lokhmotov (dividiti), David Patterson (University of California Berkeley, Google), Danilo Pau (STMicroelectronics), Jeff Sieracki (Reality AI), Jae-Sun Seo (Arizona State University), Urmish Thakkar (Arm), Marian Verhelst (KU Leuven, Imec), Poonam Yadav (University of York)

Resource-Constrained Machine Learning (ReCoML 2020)

Yaniv Ben Itzhak, Nina Narodytska, Christopher Aberger

Level 3 Room 8, Wed Mar 04, 09:00 AM

The workshop will cover broad aspects that are related to ML over resource-constrained environments, such as Internet-of-Things (IoT) devices, and edge-computing. Resource-constrained ML is challenging due to several reasons: First, current ML models usually have high resource requirements in terms of CPU, memory and I/O. Naive solutions that reduce these resource consumption would result in significant ML performance degradation. Therefore, new ML models and frameworks are required in order to employ ML with reasonable ML performance over resource-constrained environments. Second, resource-constrained environments, such as edge computing and IoT, are usually being used for real-time applications. Hence, the model serving is a critical issue, such that an ML model should respond quickly and accurately while being employed over limited resources. The workshop will specifically include the following topics: model/hardware architectures, models compression, interpretability, use-cases.

The organizers will select papers based on a combination of novelty, quality, interest, and impact.
Topics of interest include, but are not limited to:

- Compression of deep ML model architectures
- Quantized and low-precision neural networks
- Optimization of ML model architectures for resource-constrained environments
- Hardware accelerators for deep ML models
- Explainability of ML models in the context of resource-constrained environments
- ML deployments over resource-constrained environments, e.g. Internet-of-Things (IoT) devices and edge-computing.

Reviewing process: All submissions should include the author’s names and their affiliations. The authors are allowed to post their paper on arXiv or other public forums.

Key dates related to the reviewing process are given below:
Paper submission deadline: January 15, 2020 AoE (at midnight anywhere on earth)
Decision notification: January 27, 2020

We invite research contributions in different formats:
Original research papers (up to 6 pages, not including references)
Position, opinion papers and extended abstracts (up to 4 pages, not including references)

Submission link: link

Dual submission policy: We will not accept any paper which, at the time of submission, is under review for another workshop or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published.
Proceedings: Accepted papers will be published in the form of online proceedings.
Submission format: To prepare your submission to ReCoML 2020, please use the LaTeX style files provided at SML2020style.tar.gz . Submitted papers will be in a 2-column format, each reference must explicitly list all the authors of the paper.

Organizing Committee
Yaniv Ben-Itzhak, VMware Research, ybenitzhak (at) vmware (dot) com
Nina Narodytska, VMware Research, nnarodytska (at) vmware (dot) com
Christopher R. Aberger, Stanford and SambaNova Systems, christopher.aberger (at) sambanovasystems (dot) ai

Software-Hardware Codesign for Machine Learning Workloads

Ritwik Gupta, John Wohlbier, Tze Meng Low, Jeffrey Vetter, Natalia Vassilieva

Level 3 Room 9, Wed Mar 04, 09:00 AM

Machine learning development workflows today involve the siloed design and optimization of task-specific software for a limited number of fixed hardware options. As a result, hardware and software are seen as individual components where the impact of either SW or HW on each other cannot be optimized or assessed jointly. This abstraction leads to computationally inefficient machine learning workloads.

Recently, both software and hardware have taken steps to become more domain specific. Machine learning focused software libraries provide operations and abstractions limited to workload-relevant use cases. Hardware makers have started manufacturing workload-relevant chips in the form of FPGAs, ASICs, and DLAs. However, these efforts are still largely independent of each other, resulting in inefficiencies and less-than-ideal workload performances.

Ideally, hardware and software would be codesigned for a specific ML workload, but investing in a particular hardware design is costly, especially in the face of the rapidly evolving state of ML. This workshop is soliciting extended abstracts that seek to bridge the gap between software and hardware in the areas of model design, model abstractions, model primitives, workload compression, hardware design, hardware optimization for power, data flow optimization, and compiler technologies.

Schedule

08:00 AM	**Welcome, Introduction, Logistics**

08:10 AM	**DARPA**
08:35 AM	**SambaNova**
09:00 AM	**Groq**
09:25 AM	**Graphcore**
09:50 AM	**Break**
10:05 AM	**Cerebras**
10:30 AM	**Oak Ridge National Laboratory**
10:55 AM	**Carnegie Mellon University**
11:20 AM	**University of Washington**
11:45 AM	**Columbia University**
12:10 PM	**Lunch**
02:00 PM	**Facebook**
02:25 PM	**AMD**
02:50 PM	**Xilinx**
03:15 PM	**Break**
03:30 PM	**Intel**
03:55 PM	**Arm**
04:20 PM	**Panel**

Abstracts (13):

Abstract 2: **DARPA in Software-Hardware Codesign for Machine Learning Workloads**, 08:10 AM

Dr. Thomas Rondeau - Program Manager - DARPA

Title:

Abstract: AWAITING DARPA RELEASE

Keywords: software defined hardware, domain specific SoC, reconfigurable hardware, many core heterogeneous specialized accelerators

Bio: Tom Rondeau is a program manager in DARPA's Microsystems Technology Office with a focus on adaptive and reconfigurable radios, improving the development cycle for new signal-processing techniques, and exploring new approaches and applications with the electromagnetic spectrum. Prior to joining DARPA, Tom was the maintainer and lead developer of the GNU Radio project, a visiting researcher with the University of Pennsylvania, and an Adjunct with the IDA Center for Communications Research in Princeton, NJ.

Abstract 3: **SambaNova in Software-Hardware Codesign for Machine Learning Workloads**, 08:35 AM

Dr. Christopher Aberger - Director, Software Engineering - SambaNova Systems

Title:

Abstract: In many applications traditional software development is being replaced by machine learning generated models resulting in accuracy improvements and deployment advantages. This fundamental shift in how we develop software is known as Software 2.0. The continued success of Software 2.0 will require efficient and flexible computer hardware optimized for the dataflow computational graphs at the core of machine learning. In this talk, we will discuss the design of high-performance dataflow computer architectures for machine learning. Our vertically integrated approach to machine learning performance combines new machine learning algorithms, new domain-specific languages, advanced compilation technology and software-defined hardware.

Keywords: dataflow computational graph, domain specific languages, compiler technology, software defined hardware
Bio: Dr. Christopher Aberger is a director of software engineering at SambaNova Systems where he leads the machine learning team. Christopher works on efficient training algorithms for new and emerging hardware architectures. He received his Ph.D. degree in Computer Science from Stanford University where he studied the intersection of graph, database, and machine learning systems; this work received a Best Of award at VLDB in 2016 and an invited TODS article in 2017.

Abstract 4: **Groq in Software-Hardware Codesign for Machine Learning Workloads**, 09:00 AM

Dr. Dennis Abts - Chief Architect - Groq

Title: From Supercomputers to Superchips: Deep Learning One PataOp at a Time

Abstract:

Keywords: tensor streaming architecture, inference, software defined hardware, compiler technology

Bio: Dennis is the Chief Architect at Groq, and is an expert in scalable vector architectures for high performance computing. Previously at Google, he worked on datacenter network topologies for energy-proportional networking and Cray where he was a Sr. Principal Architect on several Top500 massively-parallel supercomputers. Dennis has published over 20 technical papers in areas of memory systems, interconnection networks, and fault-tolerant systems. He holds over two dozen patents spanning 20+ years of experience at Cray and Google. Dennis holds a Ph.D. in Computer Science from the University of Minnesota and is a Senior Member of IEEE and ACM Computer Society.

Abstract 5: **Graphcore in Software-Hardware Codesign for Machine Learning Workloads**, 09:25 AM

Matt Fyles - VP Software - Graphcore

Title: Compiling For Distributed Memory Architectures

Abstract: The Graphcore Intelligence Processing Unit (IPU) is designed for targeting machine learning workloads and supporting the scaling of applications across multiple devices. The IPU architecture is based around massively parallel distributed processing where applications are mapped over thousands of processor cores and operate using a Bulk Synchronous Parallel (BSP) execution model which separates computation from communication. In order to achieve performance from applications mapped onto the IPU the software tool chain has to deal with the complex task of partitioning machine learning computational graphs. In this presentation we discuss how we take a machine learning application and through our software tools partition and schedule the work across the IPU. We also discuss the hardware / software trade-offs that were made to build a processor to execute these workloads.

Keywords: IPU, ML computational graph, bulk synchronous parallel, training, inference

Bio: Matt Fyles is 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. He began his career at STMicroelectronics, Europe's largest semi-conductor company, followed by SuperH, Clearspeed and XMOS. He is currently Vice President of Software at Graphcore, a Bristol-based artificial intelligence hardware and software company. Matt is a graduate of Computer Science from the University of Exeter.

Abstract 7: **Cerebras in Software-Hardware Codesign for Machine Learning Workloads**, 10:05 AM

Dr. Natalia Vassilieva - Technical Product Manager - Cerebras Systems

Title: Accelerating Deep Learning with a purpose-built solution: the Cerebras approach

Abstract: The new era of chip specialization for deep learning is here. Traditional approaches to computing can no longer meet the computational and power requirements of this workload, arguably the most important of our generation. What is the right processor for deep learning? To answer this question, this talk will provide an overview of deep neural nets, discuss computational requirements of different types of models and limitations of existing hardware architectures and scale-out approaches. Then we will discuss Cerebras' approach to meet computational requirements of deep learning with the Cerebras Wafer Scale Engine (WSE) -- the largest computer chip in the world, and the Cerebras Software Platform, co-designed with the WSE. The WSE provides cluster-scale resources on a single chip with full utilization for tensors of any shape -- fat, square and thin, dense and sparse -- enabling researchers to explore novel network architectures and optimization techniques at any batch sizes. Finally, we will discuss potential co-design ideas for new neural net models and learning methods for the WSE.

Keywords: WSE, training, inference, dataflow computational graph

Bio: Natalia Vassilieva is a Technical Product Manager at Cerebras Systems, a computer systems company dedicated to accelerating deep learning. Her focus is machine learning and artificial intelligence, analytics, and application-driven software-hardware optimization and co-design. Most recently before joining Cerebras Natalia has been a Sr. Research Manager at Hewlett Packard Labs, where she led the Software and AI group and served as the head of HP Labs Russia from 2011 till 2015. Prior to HPE, she was an Associate Professor at St. Petersburg State University and worked as a software engineer for different IT companies. Natalia holds a PhD in computer science from St. Petersburg State University.

Abstract 8: **Oak Ridge National Laboratory in Software-Hardware Codesign for Machine Learning Workloads**, 10:30 AM

Dr. Jeffrey Vetter - Future Technologies Group Leader - Oak Ridge National Laboratory

Title:

Abstract:

Bio:

Abstract 9: **Carnegie Mellon University in Software-Hardware Codesign for Machine Learning Workloads**, 10:55 AM

Professor Tze Meng Low - Carnegie Mellon University

Title:

Abstract:

Keywords:

Bio: Tze Meng Low is an Assistant Research Professor with the Department of Electrical and Computer Engineering at Carnegie Mellon University. He graduated from the University of Texas at Austin with an M.S.(C.S) in 2004, and a Ph.D. in Computer Science in 2013. His research focuses on the systematic derivation and implementation of high-performance algorithms through the use of formal methods and analytical models. His goal is to achieve performance portability across both architectures and domains by understanding and capturing the interaction between software algorithms and hardware features through analytical models so as to build better code-generators, and/or software libraries for emerging domains and architectures.

Abstract 10: **University of Washington in Software-Hardware Codesign for Machine Learning Workloads**, 11:20 AM

Professor Michael Taylor - University of Washington

Title:

Abstract:

Bio:

Abstract 11: **Columbia University in Software-Hardware Codesign for Machine Learning Workloads**, 11:45 AM

Professor Luca Carloni - Columbia University

Title: Accelerating Embedded Machine Learning with the Open-Source ESP Infrastructure

Abstract: Recent advances in machine learning (ML) have depended on the continued progress of hardware computing platforms. Future advances will depend even more on the synergistic progress of hardware and software. This is the case particularly for embedded ML applications, where developers must meet performance requirements under tighter resource constraints. The emerging open-source hardware community can play a unique role in supporting embedded ML research. ESP is an open-source research platform to design and program heterogeneous systems-on-chip. With the design automation capabilities of ESP, application developers can synthesize hardware accelerators from models specified in common ML frameworks, integrate these accelerators in a complete system-on-chip, and quickly obtain FPGA-based prototypes to evaluate their design by running embedded ML applications.

Keywords: SoC design, ML frameworks, FPGA

Bio: Luca Carloni is Professor of Computer Science at Columbia University in the City of New York. He holds a Laurea Degree Summa cum Laude in Electronics Engineering from the University of Bologna, Italy, and the MS and PhD degrees in Electrical Engineering and Computer Sciences from the University of California, Berkeley. His research interests include methodologies and tools for system-on-chip platforms with emphasis on heterogeneous computing, intellectual property reuse, design of networks-on-chip, embedded software, and distributed embedded systems. He coauthored over one hundred and fifty refereed papers and is the holder of two patents. Luca received the Faculty Early Career Development (CAREER) Award from the National Science Foundation in 2006, was selected as an Alfred P. Sloan Research Fellow in 2008, and received the ONR Young Investigator Award and the IEEE CEDA Early Career Award in 2010 and 2012, respectively. In 2013 Luca served as general chair of Embedded Systems Week (ESWeek), the premier event covering all aspects of embedded systems and software. Luca is an IEEE Fellow.

Abstract 14: **AMD in Software-Hardware Codesign for Machine Learning Workloads**, 02:25 PM

Mayank Daga - Director, Deep Learning Software - AMD

Title: ROCm: Using Open-Source to Foster Hardware-Software Co-Design for ML

Abstract: The unprecedented success of ML in recent years has led to a burgeoning list of HW accelerators to optimize ML workloads in the marketplace. However, a novel HW platform is only part of the solution; each HW platform requires a complex SW stack to fully utilize the HW capabilities. Lack of availability of a robust and open-source middleware puts the burden on each HW vendor to develop this complex software stack themselves. This inhibits hardware-software co-design among the industry partners while incurring substantial investment dollars and longer time to discovery.

AMD believes that fully open-source middleware coupled with already open-source ML frameworks will power the next wave of AI research and adoption, wherein machines learn to reason and not just perceive. AMD is pioneering the development of the first complete open-source software platform optimized for ML called ROCm. AMD's ROCm ecosystem comprises of kernel-driver, language runtimes and compilers, optimized ML libraries and frameworks. This talk will demonstrate the capabilities of the ROCm ecosystem for ML using GPUs as the target HW platform. In addition, we will also discuss how our learnings from ML workloads have influenced our HW and SW optimizations.

Keywords: ML frameworks, open source, ROCm, GPUs, HW/SW

Bio: Mayank Daga is the Director of ML Software at AMD's Radeon Software Technologies Group. He leads the team developing optimized ML libraries and frameworks for GPUs. Before being enamored by ML, he spent several years optimizing domain specific applications in the realm of high-performance computing and data-science for AMD accelerators including inventing an industry leading sparse matrix-vector multiplication algorithm when announced. Mayank is the author of more than twenty refereed publications and five patents.

Abstract 15: **Xilinx in Software-Hardware Codesign for Machine Learning Workloads**, 02:50 PM

Nick Ni - Director of Product Marketing, AI and Software - Xilinx

Title: Vitis AI: TensorFlow to FPGAs from edge to cloud

Abstract: AI scientists are moving from research (training) using high price, high power, large form factor HPCs to productization (inference). AI inference requires orders of magnitude more horsepower while keeping the price, power, latency, form factor intact, Xilinx adaptable devices are ideal for that. However, the biggest challenge has been the programming model where it required developers to be hardware savvy. In this talk, we will introduce the newly released development environment called Vitis AI, which allows users to directly take their TensorFlow trained models and target Xilinx devices from edge to cloud. Vitis AI consists of a suite of familiar tools for AI scientists: quantizer, pruner, compiler, profiler, runtime, and pre-optimized Deep Learning Processing Units (DPU).

Keywords: ML toolkit, FPGA, DPU

Bio: Nick Ni is the director of product marketing, AI and software at Xilinx. His team's responsibilities include business development, go-to-market plans, ecosystem development, and outbound marketing for Xilinx's artificial intelligence products and software/hardware development tools. Ni joined Xilinx in 2014. Before Xilinx, he held multiple roles in R&D and applications at ATI, AMD, Qualcomm, and Intel, focusing on embedded systems design and high-level synthesis. Ni earned a master's degree in Computer Engineering from the University of Toronto and holds over 10 patents and publications.

Abstract 17: **Intel in Software-Hardware Codesign for Machine Learning Workloads**, 03:30 PM

James Moawad - Technical Solution Specialist - Intel

Title:

Abstract: We propose to show a Deep Learning Inference toolkit (OpenVINO), which provides a common API for inference independent of the underlying compute hardware. The inference engine can operate on CPU or be accelerated with GPU, VPU or FPGA. We will further look into details of an OpenCL based Deep Learning Accelerator running on FPGA and how this is integrated into the software flow. We will conclude with a brief discussion of the potential use of the oneAPI unified programming model could be used for future developments of such hardware agnostic accelerators.

Keywords: Inference, ML toolkit, CPU, GPU, VPU, FPGA

Bio: James Moawad is a Technical Solution Specialist with Intel's Programmable Solutions Group specializing in compute acceleration using Field Programmable Gate Arrays (FPGA). He holds a B.S. in Electrical Engineering from the University of Illinois at

Urbana-Champaign and a M.S. in Electrical and Computer Engineering from Georgia Institute of Technology with a focus on processor architecture. He designed telecommunication systems at Bell Laboratories / Lucent Technologies from 1999 to 2006 utilizing FPGAs and multi-processor arrays. Since 2006, he has worked as a Field Application Engineer helping customers architect systems with FPGA, embedded processors, DSP and various memory solutions including DRAM, solid state drives and high bandwidth memory (HBM).

Abstract 18: **Arm in Software-Hardware Codesign for Machine Learning Workloads**, 03:55 PM

Dr. Kshitij Sudan - Principle Solutions Architect - Arm

Title:

Abstract: Machine learning processing gets a lot of attention due to novel hardware accelerators being developed to speed-up emerging use-cases. The large and rapidly evolving accelerator space for ML processing however is eclipsed in reality by the amount of ML processing that happens on general purpose CPUs. Some estimates rate >80% of ML inference to occur on general-purpose CPUs. The driving factors for on-CPU processing are three fold: 1) Ease of programming, 2) Integration of ML analysis output with business applications, 3) Duty-cycle of ML workloads. In this talk we will first outline the use-cases that are well served by on-CPU ML workload execution followed by how Arm is working to enable more efficient use of general-purpose Arm CPUs for edge-to-cloud processing of ML workloads. Efficient processing requires both hardware and software features to be co-developed – especially since ML algorithms are rapidly evolving. Arm is leveraging this co-design philosophy along with its traditional strength in energy efficient design to make on-CPU ML processing pervasive and easy-to-use.

Keywords: inference, CPU

Bio: Dr. Kshitij Sudan is a Principal Solution Architect in the Infrastructure Business Unit at Arm where he helps build solutions to address market and customer needs. A solution could either be a single piece of Arm IP or a whole platform offering consisting of Arm IP and enabling open-source software stack. His current areas of focus include smart-offload (like SmartNICs), platform security, video encoding, and efficient ML/AI processing. He received his Ph.D. from the University of Utah where his research focused on DRAM-based memory systems. He has been granted two US patents and has multiple applications in the pipeline.