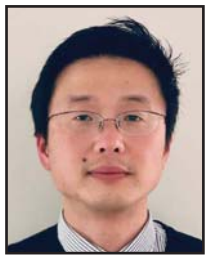
F1: *Striking the Balance Between Energy Efficiency & Flexibility: General-Purpose vs Special-Purpose ML Processors*

Organizers:

Luca Benini
ETHZ
Zurich, Switzerland

Yong Pan Liu
Tsinghua University
Beijing, China



SukHwan Lim
Samsung
Hwaseong, Gyeonggi, Korea

Tanay Karnik
Intel
Hillsboro, OR

Hsie-Chia Chang
National Chiao Tung University
HsinChu, Taiwan

The forum provides a comprehensive full-stack (hardware and software) view of ML acceleration from cloud to edge. The first talk focuses on the main design and benchmarking challenges facing large general-purpose accelerators, including multi-die scaling, and describes strategies for conducting relevant research as the complexity gap between research prototype and product continues to widen. The second talk looks at how to leverage and specialize the open-source RISC-V ISA for edge ML, exploring the trade-offs between different forms of acceleration such as lightweight ISA extensions and tightly-coupled memory accelerators. The third talk details an approach based on a practical unified architecture for ML that can be easily "tailored" to fit in different scenarios ranging from smart watches, smartphones, autonomous cars to intelligent cloud. The fourth talk explores the co-design of hardware and DNN models to achieve state-of-the-art performance for real-time, extremely energy/throughput-constrained inference applications. The fifth talk deals with ML on reconfigurable logic, discussing many examples of forms of specializations implemented on FPGAs and their impact on potential applications, flexibility, performance and efficiency. The sixth talk describes the software complexities for enabling ML APIs for various different types of specialized hardware accelerators (GPU, TPUs, including EdgeTPU). The seventh talk look into how to optimize the training process for sparse and low-precision network models for general platforms as well as next-generation memristor-based ML engines.

Emulating Large Machine-Learning Accelerators with Small Research Chips

Brian Zimmer, *Nvidia, Mountain View, CA*

Market opportunities for machine learning have attracted a wide range of both established companies and startups to develop and tape-out their own accelerators. Many of these new products have unprecedented scale, with multiple reticle-sized chips in leading process nodes and custom interconnect forming enormous computing systems, and even accelerators for embedded devices are performing trillions of operations per second. This talk will discuss some of the main challenges facing large general-purpose accelerators, including multi-die scaling, model storage options, exploiting sparsity, achieving strong scaling, improving utilization, and choosing benchmarks. Furthermore, research into solving these problems is extremely challenging, as most research projects are still confined to tiny multi-project-wafer (MPW) test chips in older technology nodes due to practical constraints such as cost and complexity. This talk will conclude by discussing strategies for small teams to continue to conduct relevant research as the complexity gap between research and product continues to widen.

Brian Zimmer is a Senior Research Scientist with the Circuits Research Group at NVIDIA in Santa Clara, CA. He received the M.S. and Ph.D. degrees in electrical engineering and computer sciences from the University of California at Berkeley, in 2012 and 2015, respectively. His research interests include soft-error resilience, energy-efficient digital design, low-voltage static random-access memory (SRAM) design, machine learning accelerators, productive design methodologies, and variation tolerance.

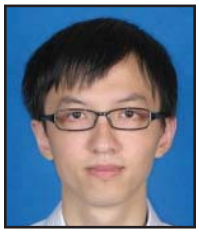
2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9365804

Extending RISC-V Platforms for ML at the Extreme Edge of the IoT

Davide Rossi, *University of Bologna, Bologna, Italy*

The "Internet of Everything" envisions trillions of connected objects loaded with high-bandwidth sensors requiring massive amounts of local signal processing, fusion, pattern extraction and classification. While machine-learning techniques are becoming the de facto standard for near-sensor analytics due to their capability to extract meaningful information from raw sensor data, their computational demand is posing a unique challenge to the small and battery-operated IoT computing systems. In this talk, I will describe the evolution of the Parallel-Ultra-Low-Power (PULP) platform - and open-source RISC-V-based architecture for extreme edge processing - toward an energy-efficient specialized architecture for embedded machine learning, exploring the trade-offs between different forms of acceleration such as lightweight ISA extensions and tightly-coupled memory accelerators.

Davide Rossi received the PhD from the University of Bologna, Italy, in 2012 where he currently holds an assistant professor position. His research interests focus on energy-efficient digital architectures in the domain of heterogeneous and reconfigurable multi- and many-core systems on a chip. This includes architectures, design implementation strategies and runtime support to address performance, energy efficiency, and reliability issues of both high-end embedded systems and ultra-low-power computing platforms targeting the IoT domain. In these fields, he has published more than 100 papers in international peer-reviewed conferences and journals. He is the recipient of the 2019 IEEE TCAD Donald O. Pederson Best Paper Award, 2020 IEEE Transactions on Circuits and Systems Darlington Best Paper Award, and the 2020 IEEE Transactions on Very Large Scale Integration Systems Prize Paper Award.



Unified Ascend: from Tasks to Processor Design

Hu Yuxing and Xia Jing, *HiSilicon-Technologies, Shen Zhen, China*

In this work, we demonstrate a practical unified architecture, called Ascend, to support multiple artificial intelligent applications.

This unified architecture can be easily "tailored" to fit in different scenarios ranging from smart watches, smartphones, autonomous cars to intelligent cloud.

We analyze this work from tasks to processor design, gradually.

First, heterogeneous computing units with optional features are employed to support various tasks, and their datapaths are adapted according to the requirement of computing and data access.

Second, scaling the Ascend architecture from a single core to a cluster containing thousands of cores, involves design efforts, such as memory hierarchy and system-level integration.

Xia Jing
Architect of KunPeng/Ascend, from Turing Business Department, Hisilicon Technologies Co. Limited.

Hu Yuxing
Researcher of Ascend, from Linx Lab, Turing Business, Hisilicon Technologies Co. Limited. (Linx Lab continuously focuses on computing architecture research.)

Co-Designing Hardware and Models for Efficient Neural Network Inference

Paul N. Whatmough, *Arm ML Research, Waltham, MA*

Deep neural networks (DNNs) have quickly become an important workload across all computing applications, including IoT, mobile, automotive, infrastructure and datacenter. However, DNN inference demands an enormous number of arithmetic operations and a large memory footprint. In this talk, I will explore the co-design of hardware and DNN models to achieve state-of-the-art performance for real-time, energy/throughput-constrained inference applications. I will start by discussing some IoT and smart hearing aid applications, and show how DNNs can be optimized to fit inside the miniscule flash memory budgets of the cheapest microcontroller ICs, which can be as small as 2kB. Next, I will discuss a technique for hardware specialization of DNN inference using fixed-weight datapaths in tandem with transfer learning techniques to provide generalization. Then I will cover analog and mixed-signal circuits, which may ultimately be more energy efficient for low-precision inference, but introduce new challenges from the ML model design point of view. Finally, I will wrap up with some forward-looking observations around efficiency and flexibility trends in hardware design for ML.

Paul Whatmough received the Doctorate degree from University College London, UK in 2012. From 2005 to 2008, he was a Research Scientist at Philips/NXP Research Labs, UK, working on hardware architecture and signal processing for software-defined radio in consumer cellular applications. He has been with Arm Research since 2008, working on machine learning (ML), hardware accelerators, digital signal processing (DSP), variation tolerance, supply voltage noise and circuits and systems for emerging IoT applications. Currently, he leads research at Arm ML Research Lab Boston, and is an Associate at Harvard University

 978-1-7281-9549-0/21/$31.00

Balancing Hardware Flexibility and Efficiency for Deep Learning

Michaela Blott, *Xilinx Research in Dublin, Ireland*

Performance scaling and power efficiency with traditional computing architectures becomes increasingly challenging as next-generation technology nodes provide diminishing performance and energy benefits. Semiconductor companies aim to unleash new levels of efficiency through further specialization of compute and memory subsystems for specific application domains. However, this comes at the cost of an increasing diversity of highly customized devices, which can only serve smaller markets.

During this talk, we will discuss many examples of forms of specializations implemented on FPGAs that have been leveraged by the industry for deep learning specifically with their impact on potential applications, flexibility, performance and efficiency.

Michaela Blott is a Distinguished Engineer at Xilinx Research in Dublin, Ireland, where she heads a team of international scientists driving exciting research to define new application domains for Xilinx devices, such as machine learning, in both embedded and hyperscale deployments. She earned her Master's degree from the University of Kaiserslautern in Germany and brings over 25 years of leading-edge computer architecture and advanced FPGA and board design, in research institutions (ETH Zurich and Bell Labs) and development organizations. She is heavily involved with the international research community serving as the technical co-chair of FPL'2018, workshop organizer (H2RC), industry advisor on numerous EU projects, and member of numerous technical program committees (FPL, ISFPGA, DATE, etc.) and most recently received the Women in Tech Award 2019.

Hybrid Digital and Analog Computing for Efficiency and Generality Optimization

Wang Shaodi, *WITIN Tech, Beijing, China*

Neural Networks (NNs) have been widely employed in modern artificial intelligence (AI) systems due to their unprecedented capability in classification, recognition and detection. However, the massive data communication between the processing units and the memory has been proven to be the main bottleneck to improve the efficiency of NN-based hardware. Furthermore, the significant power demand for massive addition and multiplication limits its adoption at the edge devices. In addition, the cost is another major concern for an edge device. Therefore, an edge neural processing chip with simultaneous low power, high performance, and low cost is in urgent need for the fast-growing AI-and-IoT (AIoT) market. In this talk, we will introduce an ultra-low-power neural processing SoC chip in 40nm with computing-in-memory technology. We have designed, fabricated, and tested this chip based on 40nm ESF3 eFlash technology. It solves the data processing and communication bottlenecks in NNs with computing-in-memory technology. Furthermore, it combines classic digital solutions together with the analog computing-in-memory macro to achieve 12-bit high-precision computing. To enable a sub-mW system in AIoT applications, a Risc-V microprocessor with DSP instruction was designed with dynamic-voltage-and-frequency scaling (DVFS) to adapt with various lower-power and real-time computing tasks. The chip supports multiple NNs including DNN, TDNN, and RNN for different applications, e.g., smart voice, and health monitoring.

Wang Shaodi received his B.S. degree from Peking University in 2011 and the Ph.D. degree in electrical engineering from UCLA in 2017. He founded WITINMEM Co. Ltd in 2017, and currently serves as the CEO of WITINMEM, dedicated to developing chips with computing-in-memory technology. He has published 20+ journal and conference papers and applied for over 50 patents. He also served as reviewer and TPC member in several IEEE and ACM journals and conferences.

Bridging the Gap: Software Cost of Hardware Specialization

Jacques Pienaar, *Google, Mountain View, CA*

In accelerator design the need to balance communication and compute is often an important accelerator design consideration. However balancing the programmability and compute of accelerators is often omitted. The term co-design has been with us for 30 years and yet still not always implemented as best practices for chip designs. That can result in a critical design and deployment failure, as there is cost to enable hardware accelerators that are too specific. Overly specialized hardware, with respect to the programming model and domain, increases the need and cost of the development of software layers/tools and in particular the compiler. But these software layers are often areas that are under-invested in, during new hardware accelerator design, which compounds the problem. This talk describes some of the software complexities for enabling ML APIs for various different types of hardware accelerators (GPU, TPU v1, TPU v2 and EdgeTPU). It will describe some of the pains encountered while enabling these new hardware accelerators as well as our approach at Google to address these pain points.

Jacques Pienaar is a Staff Software Engineer at Google, one of the leads of MLIR and TensorFlow Graph Compiler projects. Previously he was the co-developer of Lanai compiler, the open-source GPGPU compiler gpucc. H was also one of the first engineers on XLA and the XLA TPU backend, and TF2XLA developer. His research interests span multiple layers of an ML software stack from programming model down to runtime. Jacques holds a Ph.D. from Purdue university where he was a member of the Integrated Systems Laboratory in the School of Electrical and Computer Engineering.

Efficient Machine Learning: Algorithms-Circuits-Devices Co-design

Hai Helen Li, *Duke University, Durham, NC*

Following technology advances in high-performance computation systems and fast growth of data acquisition, machine learning, and especially deep neural networks (DNNs), made remarkable success in many research areas and applications. Such a success, to a great extent, is enabled by developing large-scale network models that learn from a huge volume of data. The deployment of such a big model, however, is both computation-intensive and memory-intensive. Though the research on hardware acceleration for neural networks has been extensively studied, the progress of hardware development still falls far behind the upscaling of DNN models at the software level. The holistic co-design across algorithm, circuit, and device levels emerges more importantly for execution acceleration, energy efficiency, and design flexibility. In this presentation, we will present our studies on how to optimize the training process for sparse and low-precision network models for general platforms. We will also discuss the memristor-based computing engine designs optimized for DNN inference and training.

Hai "Helen" Li is the Clare Boothe Luce Professor and Associate Chair of the Department of Electrical and Computer Engineering at Duke University. She received her B.S and M.S. degrees from Tsinghua University and a Ph.D. from Purdue University. At Duke, she co-directs the Duke University Center for Computational Evolutionary Intelligence and the NSF IUCRC for Alternative Sustainable and Intelligent Computing (ASIC). Her research interests include neuromorphic circuits and systems for brain-inspired computing, machine learning acceleration and trustworthy AI, conventional and emerging memory design and architecture, and software and hardware co-design. Dr. Li served/serves as the Associate Editor for multiple IEEE and ACM journals. She was the General Chair or Technical Program Chair of multiple IEEE/ACM conferences and Technical Program Committee member of over 30 international conference series. Dr. Li was a Distinguished Lecturer of the IEEE CAS society (2018-2019) and a distinguished speaker of ACM (2017-2020). Dr. Li is a recipient of the NSF Career Award, DARPA Young Faculty Award, TUM-IAS Hans Fischer Fellowship from Germany, ELATE Fellowship, eight best paper awards and another nine best paper nominations. Dr. Li is an IEEE fellow and a distinguished member of the ACM.

 978-1-7281-9549-0/21/$31.00