Short Course: Machine Learning Hardware: Considerations and Accelerator Approaches

2024 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3503-0620-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISSCC49657.2024.10454428

Organizer: *Daniel Friedman*
IBM Thomas J. Watson Research Center
Yorktown Heights, NY

Time:	Topic:
8:00 AM	Breakfast
8:25 AM	Introduction by Chair, ***Daniel Friedman*** *IBM Thomas J. Watson Research Center, Yorktown Heights, NY*
8:30 AM	**Introduction to Machine Learning Applications and Hardware-Aware Optimizations** ***Ranghajaran Venkatesan,*** *Nvidia, Santa Clara, CA*
10:00 AM	Break
10:30 AM	**Architecture and Design Approaches to ML Hardware Acceleration: Performance Compute Environment** ***Leland Chang,*** *IBM Thomas J. Watson Research Center, Yorktown Heights, NY,*
12:15 PM	Lunch
1:20 PM	**Architecture and Design Approaches to ML Hardware Acceleration: Edge and Mobile Environments** ***Marian Verhelst,*** *KU Leuven, Leuven, Belgium*
2:50 PM	Break
3:20 PM	**Emerging ML Accelerator Approaches: In-Memory Computing Architectures** ***Naresh Shanbhag,*** *University of Illinois Urbana-Champaign, Champaign, IL*
4:50 PM	Conclusion

Introduction

The growth in the application of machine learning and artificial intelligence technology to problems across virtually all spheres of endeavor has been and is expected to remain extraordinary. Hardware acceleration for machine learning tasks is a critical vector that has enabled this exceptionally rapid growth. Further accelerator advances are necessary to drive everything from improved efficiency for inference, to support ever-growing network sizes to improvements in support for network training, to enabling broadening of ML deployments across platforms with a wide range of power and performance envelopes. In this short course, we will first present an overview of machine learning and inference, including describing key metrics, frameworks, application areas, and approaches to support model scaling. In the second presentation, we will discuss architectural and design approaches to ML hardware acceleration for applications in high performance compute environments. In the third presentation, we will turn to approaches to mapping ML hardware acceleration to constrained compute footprint contexts as in edge and mobile applications. Finally, we will present a framework for considering a key emerging topic in hardware design for ML acceleration, namely, compute-in-memory approaches.

OUTLINE

SC1:
Introduction to Machine Learning Applications and Hardware-Aware Optimizations

Ranghajaran Venkatesan, *Nvidia, Santa Clara, CA*

Deep neural networks (DNNs) have become a crucial solution for tackling complex challenges in a wide range of fields, such as image recognition, natural language processing, robotics, healthcare, and autonomous driving. The landscape of DNN applications is constantly expanding, driving the ongoing evolution of DNN models. These models come in various architectures, including convolutional neural networks, transformers, diffusion models, and more, each tailored to meet the unique demands of their respective applications. These DNN models vary significantly in size and computational complexity, driving the need for efficient neural-network computing chips. This has led to a growing exploration of hardware and software co-design techniques, balancing energy efficiency and performance without compromising accuracy. This short course offers an introductory exploration of different neural network models, shedding light on their individual characteristics and applications. It also delves into various design strategies, emphasizing the importance of achieving a delicate balance between efficiency, scalability, and adaptability across different neural network paradigms, all while paving the way for the emergence of efficient neural network models and computing architectures.

Rangharajan Venkatesan is a Senior Research Scientist in the ASIC & VLSI Research group in NVIDIA. He received the B.Tech. degree in Electronics and Communication Engineering from the Indian Institute of Technology, Roorkee in 2009 and the Ph.D. degree in Electrical and Computer Engineering from Purdue University in 2014. His research interests are in the areas of low-power VLSI design and computer architecture with particular focus in deep learning accelerators, high-level synthesis, and spintronic memories. He has received Best Paper Awards for his work on deep learning accelerators from the IEEE/ACM Symposium on Microarchitecture (MICRO) and the Journal of Solid-State Circuits (JSSC). His work on spintronic memory design was recognized with the Best Paper Award at the International Symposium on Low Power Electronics and Design (ISLPED), and Best Paper nomination at the Design, Automation and Test in Europe (DATE). His paper titled, “MACACO: Modeling and Analysis of Circuits for Approximate Computing”, received the IEEE/ACM International Conference on Computer-Aided Design (ICCAD) Ten Year Retrospective Most Influential Paper Award in 2021. He has served as a member of the technical program committees of several leading IEEE/ACM conferences including ISSCC, DAC, MICRO, and ISLPED.

SC2:
Architecture and Design Approaches to ML Hardware Acceleration: Performance Compute Environment

Leland Chang, *IBM Thomas J. Watson Research Center, Yorktown Heights, NY*

With the recent explosion in generative AI and large language models, hardware acceleration has become particularly important in high-performance compute environments. In such applications, AI accelerators should address a broad range of AI models and enable workflows spanning model pre-training, fine-tuning, and inference. System-level design and software co-optimization must be considered to balance compute and communication costs, especially with inference workloads driving aggressive latency targets and model size growth driving the use of distributed systems. This talk will discuss these considerations in the context of high-performance system deployments and explore approaches to AI accelerator circuit design as well as research roadmaps to improve both compute efficiency and communication bandwidth.

Leland Chang is a Principal Research Scientist and the Senior Manager of AI Hardware at IBM Research, where he leads a team developing AI hardware accelerators for next-generation server and mainframe products. He has worked across technology, circuits, architecture, and software with key technical contributions to FinFET technologies, SRAM scaling, integrated voltage regulators, and AI accelerators. He received the B. S., M. S., and Ph.D. degrees in EECS from UC Berkeley and has authored 100 technical articles and 135 patents. He is a former memory subcommittee chair of the ISSCC technical program committee.

SC3:
Architecture and Design Approaches to ML Hardware Acceleration: Edge and Mobile Environments

Marian Verhelst, *KU Leuven, Leuven, Belgium*

Various applications demand more and more powerful machine inference in resource-scarce distributed devices, such as phones, watches, glasses, robots or drones. To allow intelligent applications at ultra-low energy and low latency, one needs customized processor architectures optimized for extreme edge applications. This need has resulted in the creation of a wide variety of novel hardware architectures, supported by HW-algorithm co-optimization methods. This talk will zoom into ML processor architectures for the edge, as well as tools for efficient mapping of ML algorithms onto such architectures.

Marian Verhelst is professor at the MICAS laboratories of KU Leuven and a research director at imec. Her research focuses on embedded machine learning, hardware accelerators, HW-algorithm co-design and low-power edge processing. She received a PhD from KU Leuven in 2008, and worked as a research scientist at Intel Labs from 2008 till 2010. Marian currently is a member of the board of directors of tinyML, scientific advisor to multiple startups and active in the TPC's of DATE and ESSCIRC. She enjoys science communication as an IEEE SSCS Distinguished Lecturer, as a regular member of the Nerdland science podcast (Dutch), and as the founding mother of KU Leuven's InnovationLab high school program.

SC4:
Emerging ML Accelerator Approaches: In-Memory Computing Architectures

Naresh Shanbhag, *University of Illinois Urbana-Champaign, Champaign, IL*

In-memory computing (IMC) has emerged as an attractive complement to digital accelerators for enhancing the energy efficiency of machine learning tasks. IMC addresses the energy and latency costs of memory accesses dominating AI workloads by transforming conventional memory accesses into ones that compute functions of data in the memory core. As a result, IMC chips have demonstrated at least an order-of-magnitude reduction in the energy-delay product over equivalent von Neumann architectures at iso-accuracy. IMCs also exhibit a fundamental energy vs. SNR trade-off that designers need to exploit to enhance energy efficiency while meeting task-level accuracy requirements. Since its inception in 2014, IMC design has become an active area of research in the integrated circuits and architecture communities. This talk will provide an overview of IMCs, describe various IMC design principles and architectures, review current trends via data-driven extensive benchmarking of IMC chip prototypes, and identify future opportunities and challenges in deploying IMCs at scale in emerging applications.

Naresh R. Shanbhag is the Jack Kilby Professor of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign. He received his Ph.D. degree from the University of Minnesota (1993) in Electrical Engineering. From 1993 to 1995, he worked at AT&T Bell Laboratories at Murray Hill where he led the design of high-speed transceiver chipsets for very-high-speed digital subscriber line (VDSL), before joining the University of Illinois at Urbana- Champaign in August 1995. He has held visiting faculty appointments at the National Taiwan University (Aug.-Dec. 2007) and Stanford University (Aug.-Dec. 2014). His research focuses on the design of energy-efficient systems for machine learning, communications, and signal processing, spanning algorithms, VLSI architectures, and integrated circuits. He has more than 200 publications in this area, holds thirteen US patents, and is a co-author of two books and multiple book chapters (see https://shanbhag.ece.illinois.edu/ for details). He received the 2018 SIA/SRC University Researcher Award, became an IEEE Fellow in 2006, received the 2010 Richard Newton GSRC Industrial Impact Award, the IEEE Circuits and Systems Society Distinguished Lecturership in 1997, the National Science Foundation CAREER Award in 1996, and multiple best paper awards. In 2000, Dr. Shanbhag co-founded and served as the Chief Technology Officer of the Intersymbol Communications, Inc., which introduced mixed-signal ICs for electronic dispersion compensation of OC-192 optical links, and became a part of Finisar Corporation in 2007. From 2013-17, he was the founding Director of the Systems On Nanoscale Information fabriCs (SONIC) Center, a 5-year multi- university center funded by DARPA and SRC under the STARnet program. He is currently on the leadership teams of the JUMP 2.0 DARPA and SRC funded Centers for Ubiquitous Connectivity (CUbiC) and Codesign of Cognitive Systems (CoCoSys), and the NSF-industry funded Center for Advanced Semiconductor Chips for Accelerated Performance (ASAP).