A Little Goes a Long Way: Building Domain-Specific Chiplets and Emerging Interconnects for Next Era of AI/ML Systems

Vikram Jain

Postdoctoral Researcher
University of California, Berkeley

Monday, March 10
1:30 PM • MSEE 239

ECE Computes Faculty Candidate Search Seminar

Abstract

Recent advancements in deep neural networks (DNNs), especially transformer-based large language models (LLMs), have driven significant progress in artificial intelligence (AI). As demand grows, models expand to trillions of parameters, potentially requiring dedicated nuclear power plants for data centers. While GPUs are commonly used, they are outperformed in energy efficiency by domain-specific accelerators (DSAs). Modern system-on-chip (SoC) designs utilize these DSAs to enable parallel workload execution, known as accelerator-level parallelism (ALP). SoCs need to scale to meet the growing demand but encounter challenges like reticle limits, yield issues, and thermal management. Chipletization—combining multiple chips in one package—offers a solution for improved scalability and composability, leading to what I call chiplet-level parallelism (CLP). Future systems will incorporate various little domain-specific chiplets, enhancing parallel execution. Additionally, technologies like silicon photonics will be vital for scaling these architectures to bridge the gap to warehouse-scale computing. This talk will cover the challenges and optimizations for ALP, CLP, and beyond Moore's architectures. First, I will present my work on enabling energy-efficient heterogeneous SoCs for edge machine learning applications through ALP. I will discuss our design space exploration framework, ZigZag, which was created to allow rapid exploration of hardware architectures for ML accelerators. ZigZag played a crucial role in co-designing an ML accelerator implementation integrated into my two silicon prototypes: TinyVers, an all-digital heterogeneous SoC featuring a RISC-V core and efficient power management for IoT, and Diana, the first hybrid digital and analog ML SoC, utilizing the strengths of both architectures for enhanced energy efficiency. Scaling beyond SoCs, the second part of my talk explores energy-efficient chiplet architectures and CLP. CLP can be seen as a constrained ALP, enabling us to apply many insights from ALP, such as memory management, data orchestration, resource allocation, and more, to chiplets. However, to harness the potential of CLP fully, we need a co-design infrastructure. I will showcase my work on automatic chiplet generation and the universal chiplet interconnect express (UCIe) die-to-die interface standard, which facilitates the creation of a plug-and-play chiplet ecosystem. Additionally, I will present two of my recent silicon prototypes: Cygnus, the first academic RVV1.0 multi-core vector processor chiplet designed for digital signal processing (DSP), and Sirius, the first UCIe-compliant LLM chiplet utilizing a novel quantization scheme. As we enter the age of AI proliferation, domain-specific chiplets will play a significant role in building modular systems for edge and data centers. However, to enhance energy efficiency in warehouse-scale computing, systems-in-package (SiP) must evolve into systems-in-cluster (SiC), connected through emerging silicon photonics and optical networks. A co-design approach that aligns model architecture with hardware specifications is essential for energy-efficient scaling for edge and data centers. In the final section of my talk, I will present a vision for a unified framework focused on partitioning, scheduling, design space exploration, simulation, and hardware generation tailored for scale-up and scale-out architectures. This will enable the development of future energy-efficient and scalable AI/ML systems.

Bio

Vikram Jain is a postdoctoral researcher at the SpeciaLized Computing Ecosystem (SLICE) Lab and the Berkeley Wireless Research Center (BWRC) at the University of California, Berkeley. In addition, he serves as a Lecturer in the Electrical Engineering and Computer Sciences (EECS) department at UC Berkeley. His research focuses on heterogeneous integration and chiplet architectures (2.5D and 3D) for emerging high-performance computing and AI applications. Vikram earned his Ph.D. in energy-efficient heterogeneous systems for embedded machine learning from the MICAS laboratories at KU Leuven, Belgium. He has also been a visiting researcher at the IIS Laboratory at ETH Zurich, where he worked on the design of high- performance networks-on-chip for deep neural network platforms. He has published numerous papers, workshops, and posters in leading conferences and journals, including ISSCC, JSSC, the Symposium on VLSI Technology and Circuits (VLSI), MICRO, HPCA, ISLPED, DAC, ISCAS, DATE, TCAS-I, TVLSI, and TC. Vikram received the Solid-State Circuits Society (SSCS) Predoctoral Achievement Award for his contributions to embedded machine learning hardware design for 2022-2023. He was also awarded the SSCS Student Travel Grant in 2022 and the Lars Pareto Travel Grant in 2019. Moreover, he held a prestigious research fellowship from the Swedish Institute (SI) during his master's program for 2016–2017 and 2017–2018. Vikram also serves as a reviewer for the IEEE Journal of Solid-State Circuits, IEEE Transactions on Very Large-Scale Integration Systems (TVLSI), IEEE Transactions on Circuits and Systems I (TCAS-I), and IEEE Transactions on Computers (TC).

Hosts Mithuna Thottethodi ~ mintuna@purdue.edu and Sumeet Gupta ~ guptask@purdue.edu
Zoom: https://purdue-edu.zoom.us/j/94700508338 ~ Meeting ID: 947 0050 8338

PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering