Electrical & Computer Engineering

S E M I N A R

Louisiana State University

Hardware-Software Co-design of Heterogeneous Architectures to Enable AI at Scale

Gaurav Narang

Arizona State University

Abstract—Today's computing systems struggle to keep pace with the rapidly growing compute demands of modern AI application workloads from edge to cloud. This necessitates orders-of-magnitude improvements in efficiency, which can be achieved by synergistically combining the strengths of machine learning (ML) and emerging compute technologies. This talk spans across design and resource management of von Neumann and heterogeneous in-memory computing (IMC) architectures to address this gap. Specifically, I will present a novel ML-based methodology for dynamic power management that is applicable to both mobile and datacenter-scale systems. This methodology addresses the challenge of maximizing performance and minimizing energy consumption for dynamically changing applications. Next, I will present a framework for the design of heterogeneous IMC architectures to address the challenges of deep neural network training on 3D integrated systems. I will also discuss thermal-aware kernel mapping and runtime reliability-aware computation of different neural networks (CNNs, Graph NNs, and Transformers) on 2.5D and 3D integrated systems. Together, these contributions will demonstrate how ML-enabled design and optimization in heterogeneous architectures can enable energy-efficient, high-performance, and reliable AI inference and training. I will conclude with my vision for future research directions in silicon photonics, brain-inspired computing, and reliability in 2.5D and 3D advanced packaging.

Bio—Gaurav Narang is a Postdoctoral Research Scholar at Arizona State University. He received his Ph.D. in Computer Engineering from Washington State University. Prior to his doctoral studies, he worked at STMicroelectronics (2015-2018) and Synopsys (2018-2021), where he contributed to digital design and logic synthesis for PHY testchips and memory repair (BIST and ECC) logic.

His research focuses on hardware and software co-design for heterogeneous, nonvon Neumann architectures aimed at achieving energy-efficient, high-performance, and reliable acceleration of emerging AI workloads. Dr. Narang's work on ML-driven runtime resource management received Best Paper Award at the 2023 International Symposium on Low Power Electronics and Design (ISLPED) and he earned the Voiland College of Engineering and Architecture, Outstanding Graduate Research Assistant Award in 2025.

When: Thursday, **19 February 2026,** 10:00 - 11:00
Where: Room 3316E Patrick F. Taylor Hall
Info: `https://www.lsu.edu/eng/ece/seminar`
Food: ***Refreshments will be served.***