NSF NANOSCALE SCIENCE AND ENGINEERING GRANTEES CONFERENCE:
NANO AND AI CONVERGENCE
DECEMBER 9-10, 2024

"*The Third Dimension of Technology Scaling*: Co-designing for Direct Functionality Embedding in a Device"

AMIT RANJAN TRIVEDI

Position title: Associate Professor
Institution: University of Illinois at Chicago (UIC)

Bio: Amit Ranjan Trivedi is an associate professor in the department of electrical and computer engineering at the University of Illinois at Chicago (UIC). Trivedi was awarded Sigma Xi best Ph.D. thesis award from Georgia Tech, IEEE Electron Device Society Fellowship, NSF CAREER Award, and best paper award at IEEE AICAS. His research interests include compute-in-memory and neuromorphic technologies. He has more than 100 publications in referred journals and conferences.

Abstract: This talk explores a novel approach to achieving enhanced scalability of AI models within ultra-low-power systems. As AI models continue to grow exponentially in size and complexity to address increasingly diverse use cases, the limitations of transistor scaling have become apparent. While techniques like 3D and heterogeneous integration offer an interim solution by opening a second scaling dimension, the exponential growth of machine learning models necessitates a fundamental rethinking of acceleration strategies. We present an innovative direction: device-based computing, where key functionalities of AI models are directly embedded into devices, such as building circuits in a device and architectures in a circuit, through co-design of hardware and models. Starting from memristors that encapsulate entire multiply-accumulate (MAC) units within a single device to Gaussian transistors enabling reasoning functions, we delve into the underlying currents of these emerging approaches. We present them systematically as a cohesive concept, ultimately revealing a third dimension of scalability. This new paradigm paves the way for achieving higher functionality within constrained area and power budgets, offering transformative pathways for AI acceleration.