THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING SPEAKER SERIES

PRESENTS

# *GreenTPU: Ultra Low Power Hardware AI Accelerator for the Edge*

**Dr. Sanghamitra Roy**
Professor, Department of Electrical and Computer Engineering,
**Utah State University, USA**

Monday**, 15 April 2024,** 9:55 AM

Room: **Zoom** (Meeting ID# 976-269-9678; Passcode: K91Bwy https://zoom.us/j/9762699678?pwd=RUp5ZmN3cHUyQ1FvUExVQjVsc1hVUT09)

## LECTURE ABSTRACT

A paradigm shift in the software-hardware ecosystem is widely attributed to the profound developments in the Artificial Intelligence (AI) domain. AI has enabled a plethora of applications in recent years, many of which were well beyond imagination even a decade ago. A typical use-case scenario involves training a Deep Neural Network (DNN) on a server farm involving powerful CPU-GPU systems, and then using a trained model for inference at the edge. To enable efficient and widespread use of inference, there is an emergence of domain-specific architectures that are able to deliver substantial performance and energy-efficient alternatives to general purpose processors. Among such DNN accelerators, Google Tensor Processing Unit (TPU) has transpired to be the best-in-class, offering more than 15× speedup over the contemporary GPUs. However, the rapid growth in several DNN workloads conspires to escalate the energy consumptions of the TPU-based datacenters. In order to restrict the energy consumption of TPUs, we propose GreenTPU-a low-power near-threshold (NTC) TPU design paradigm. To ensure a high inference accuracy at a low-voltage operation, GreenTPU identifies the patterns in the error-causing activation sequences in the systolic array and prevents further timing errors from the same sequence by intermittently boosting the operating voltage of the specific multiplier-and-accumulator units in the TPU. Compared to a cutting-edge timing error mitigation technique for TPUs, GreenTPU enables 2X–3X higher performance in an NTC TPU, with a minimal loss in the prediction accuracy.

## SPEAKER BIOSKETCH

**Dr. Sanghamitra Roy** is a professor in the department of Electrical and Computer Engineering at Utah State University. She received her Ph.D. degree in Electrical and

Computer Engineering from the University of Wisconsin-Madison. She received her M.S. degree in Computer Engineering from Northwestern University in Dec 2003. Dr. Roy has authored over 85 peer reviewed publications in top tier journals and conferences. She has won several Best Paper Awards/nominations. She received the NSF CAREER Award in 2013. Her research interests are in VLSI circuit design and optimization and exploring reliability aware novel circuit styles and architectures. Dr. Roy was named in the “125 People of Impact”- the list of most influential alumni to graduate from the University of Wisconsin-Madison, in recognition of her ongoing success in academia and research. She is the inventor in 12 issued US patents.