nanoML: Pushing the Limits of Edge AI with Weightless Neural Networks

Abstract: Mainstream artificial neural network models, such as Deep Neural Networks (DNNs) are computation-heavy and energy-hungry. Weightless Neural Networks (WNNs) are natively built with RAM-based neurons and represent an entirely distinct type of neural network computing compared to DNNs. WNNs are extremely low-latency, low-energy, and suitable for efficient, accurate, edge inference. The WNN approach derives an implicit inspiration from the decoding process observed in the dendritic trees of biological neurons, making neurons based on Random Access Memories (RAMs) and/or Lookup Tables (LUTs) ready-to-deploy neuromorphic digital circuits. WNNs are a natural fit for edge AI due to the low area, energy and latency properties offered by them. This talk will describe the state of the art of Weightless Neural Networks, and their applications for edge inferencing.

Biography: Lizy Kurian John is Truchard Foundation Chair in Engineering at the University of Texas at Austin. Her research interests include workload characterization, performance evaluation, and high performance architectures for emerging workloads. She is recipient of many awards including Joe J. King Professional Engineering Achievement Award (2023), and The Pennsylvania State University Outstanding Engineering Alumnus Award (2011). She has authored 3 books and has edited 4 books including a book on Computer Performance Evaluation and Benchmarking. She holds 20 US patents and is an IEEE Fellow (Class of 2009), ACM Fellow, AAAS Fellow and Fellow of the National Academy of Inventors (NAI).