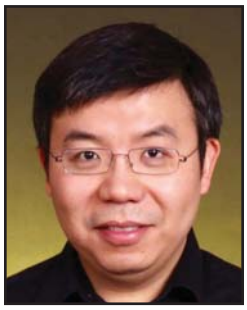
Session 15 Overview:
Compute-in-Memory Processors for Deep Neural Networks
MACHINE LEARNING SUBCOMMITTEE

Session Chair:
Jun Deguchi
Kioxia Corporation, Kawasaki, Japan



Session Co-Chair:
Yongpan Liu
Tsinghua University, Beijing, China

Session Moderator:
Yan Li
Western Digital, Milpitas, CA

Compute-in-memory (CIM) processors for deep neural networks continue to expand their capabilities, and to scale to larger datasets and more complicated models. All four of the papers in this session have integrated CIM into their system setup, and comprehensively evaluate a variety of ML models in high bit precision. The first paper demonstrates a large CIM array with 4.5Mb with bit precision of 1-to-8b. The second paper reduces system energy by using zero skipping, a shared ADC using ping-pong CIM, and digital-predictor-assisted adaptive bit-precision to save power in the ADC. The third paper reduces memory-device footprint by replacing 6T SRAM with 3T plus capacitor. The final paper in the session applies the tensor-train method to decompose and compress neural networks so that they fit within on-chip memory.

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9365855

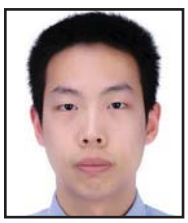
8:30 AM

15.1 A Programmable Neural-Network Inference Accelerator Based on Scalable In-Memory Computing

Hongyang Jia, Princeton University, Princeton, NJ

In Paper 15.1, Princeton University describes a scalable neural-network (NN) inference processor based on a 4×4 array of programmable cores combining precise mixed-signal capacitor-based in-memory-computing (IMC) with digital SIMD near-memory computing, interconnected with a flexible on-chip network. Implemented in 16nm with 1-to-8b configurable precision, their 25mm^2 chip achieves 30TOPS/W in 8b mode.

8:38 AM



15.2 A 2.75-to-75.9TOPS/W Computing-in-Memory NN Processor Supporting Set-Associate Block-Wise Zero Skipping and Ping-Pong CIM with Simultaneous Computation and Weight Updating

Jinshan Yue, Tsinghua University, Beijing, China and Pi2star Technology, Beijing, China

In Paper 15.2, Tsinghua University, Pi2star Technology and National Tsing Hua University describe an energy-efficient computing-in-memory (CIM) neural-network processor. Innovations include a set-associate block-wise zero-skipping (SABZA) and a ping pong-CIM (PP-CIM) architecture using a digital-predictor-assisted adaptive 0/2/4b ADC. Their 65nm, 12mm^2 chip supports the ImageNet dataset (8b activations, 4b weights) with 2.75TOPS/W system energy efficiency, and can reach a peak system energy efficiency of 75.9TOPS/W with 2b activations and 1b weights.

15

8:46 AM

15.3 A 65nm 3T Dynamic Analog RAM-Based Computing-in-Memory Macro and CNN Accelerator with Retention Enhancement, Adaptive Analog Sparsity and 44TOPS/W System Energy Efficiency

Zhengyu Chen, Northwestern University, Evanston, IL

In Paper 15.3, Northwestern University presents a 3T dynamic analog RAM-based computing-in-memory macro and associated CNN accelerator, leading to an effective bit size of only 75% of 6T foundry SRAM. Using special analog sparsity and retention enhancement techniques, their 3.3mm^2 test chip in 65nm technology achieves state-of-art energy efficiency of 217TOPS/W at CIM macro level and 44TOPS/W at system level, for 4b weight/input operation.

8:54 AM

15.4 A 5.99-to-691.1TOPS/W Tensor-Train In-Memory-Computing Processor Using Bit-Level-Sparsity-Based Optimization and Variable-Precision Quantization

Ruiqi Guo, Tsinghua University, Beijing, China

In Paper 15.4, Tsinghua University, University of Electronic Science and Technology of China and National Tsing Hua University present a 3.31×2.71mm^2 computing-in-memory processor in 28nm technology, achieving 5.99-to-691.1TOPS/W energy efficiency, by applying tensor-train decomposition method to compress DNNs to fit within SRAM-CIM, exploiting bit-level sparisty and optimizing activation quantization.