Session 22 - Memory Technology/Emerging Device and Compute Technology - Focus Session: Emerging AI Hardware
Tuesday, December 10, 2:15 p.m.
Continental Ballroom 5
Co-Chairs: C. Petti, Sunrise Memory, Inc.
T-H Hou, National Chiao Tung University

2:20 PM **22.1** Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators (Invited)
Tien-Ju Yang, Vivienne Sze, Massachusetts Institute of Technology

This paper describes various design considerations for deep neural networks that enable them to operate efficiently and accurately on processing-in-memory accelerators. We highlight important properties of these accelerators and the resulting design considerations using experiments conducted on various state-of-the-art deep neural networks with the large-scale ImageNet dataset.

2:45 PM **22.2** Towards 10000TOPS/W DNN Inference with Analog in-Memory Computing – A Circuit Blueprint, Device Options and Requirements (Invited)
Stefan Cosemans, Bram Verhoef, Jonas Doevenspeck, Ioannis Papistas, Francky Catthoor, Peter Debacker, Arindam Mallik, Diederik Verkest, imec, KU Leuven

This paper presents a circuit blueprint for a 10000TOPS/W matrix-vector multiplier for neural network inference based on Analog in-Memory Computing, using pulse-width encoded activations and precharge-discharge summation lines. Three suited device options are discussed: SOT-MRAM, IGZO-based 2T1C DRAM gain cell, and projection PCM with separate write path.

3:10 PM **22.3** The Marriage of Training and Inference for Scaled Deep Learning Analog Hardware (Invited)
Tayfun Gokmen, Malte J. Rasch, Wilfried Haensch, IBM Research AI

Here, we show that for large scale deep neural networks (DNNs) the model's parameters (weights) must come from a training procedure that accounts for hardware induced constraints, such as ADC, DAC and noise, for the inference task to be successful when run on analog hardware composed of crossbar arrays.

3:35 PM **22.4** Can in-memory/Analog Accelerators be a Silver Bullet for Energy-efficient Inference? (Invited)
Jun Deguchi, Daisuke Miyashita, Asuka Maki, S. Sasaki, Kengo Nakata, Fumihiko Tachibana, Kioxia Corporation

Although energy efficiency of in-memory/analog accelerators looks better than that of digital accelerators based on our benchmark, accuracy of in-memory/analog accelerators is usually deteriorated. Then we introduce our proposed quantization technique and a specific hardware architecture. Finally, we discuss whether in-memory/analog accelerators can be a silver bullet for energy-efficient inference.

4:00 PM *COFFEE BREAK*

4:25 PM **22.5** AI Edge Devices Using Computing-In-Memory and Processing-In-Sensor: From System to Device (Invited)
Tzu-Hsiang Hsu, Yen-Cheng Chiu, Wei-Chen Wei, Yun-Chen Lo, Chung-Chuan Lo, Ren-Shuo Liu, Kea-Tiong Tang, Meng-Fan Chang, Chih-Cheng Hsieh, National Tsing Hua University

This paper presents the advanced technologies, including CIS and PIS techniques, for low-power and low-latency AI edge devices. From system perspective, CIM and PIS techniques are effective by reducing power dissipation and data transmission for computations in CNN models. Furthermore, the performance is strongly relied on the corresponding device enhancement.

4:50 PM **22.6** Hybrid Analog-Digital Learning with Differential RRAM Synapses (Invited)
Tifenn Hirtzlin, Marc Bocquet, Maxence Ernoult, Jacques-Olivier Klein, Etienne Nowak, Elisa Vianello, Jean-Michel Portal, Damien Querlioz, University Paris-Sud, Aix-Marseille Université, CEA-Leti

Exploiting analog RRAM for learning is compelling, but raises important challenges. Here, we investigate a learning architecture, based on Binarized Neural Networks, which exploits the analog properties of hafnium-oxide RRAM, but avoids these challenges: it uses exclusively low-overhead digital CMOS, is highly resilient to device imperfections, and shows outstanding endurance.

5:15 PM **22.7** Active Memristor Neurons for Neuromorphic Computing (Invited)
Wei Yi, Kenneth K. Tsang, Stephen K. Lam, Xiwei Bai, Jack A. Crowell, Elias A. Flores, HRL Laboratories, LLC.

Memristors provide a new paradigm to realize biomimetic and scalable neuron and synapse neuromorphic computing primitives capable of efficiently emulating the rich dynamics of biological counterparts. Using VO_2 active memristors, we show that memristor neurons possess most of the known biological neuronal dynamics and all three classes of neuron excitability.

5:40 PM **22.8** Towards Large-Scale Photonic Neural-Network Accelerators (Invited)
Ryan Hamerly, Alex Sludds, Liane Bernstein, Mihika Prabhu, Charles Roques-Carmes, Jacques Carolan, Yoshihisa Yamamoto, Marin Soljacic, Dirk Englund, MIT, NTT Research Inc.

We review leading photonic AI platforms based on beamsplitter mesh networks, weight banks, and photoelectric multiplication. Theoretical performance advantages, as well as practical issues of chip area, input / output, and crosstalk, are considered. We address fundamental and near-term limitations to energy efficiency and investigate bandwidth limitations from temporal crosstalk.