Session Oral 11 (8/8 Thu. 11:30 – 12:30)

Session Topic: Memory/Computing Cooperation & Optimization

Session Chair: Hsie-Chia Chang (National Chiao Tung University)

Room: 6F 茗廳

1. 11:30 – 11:42 (SC11) Memory-contention Aware Warp Scheduler for Computing GPU

   Chien-Ming Chiu, Kuan-Chung Chen, Jhi-Han Jheng, Kuan-Lin Huang, Tsung-Han Tsou, Feng-Ming Hsu, Juin-Ming Lu, and Chung-Ho Chen

   National Cheng Kung University

   We will first introduce a computing GPU which supports both OpenCL and TensorFlow framework. The proposed GPU aims at the deployment of edge AI computing devices. To address the memory contention problem of the GPU, a serious performance bottleneck in resource-limited GPUs, we propose a Memory

Contention Aware Warp Scheduler (MAWS) to strike a dynamic balance between the memory workload requirements and the given memory resources. By measuring the Load/Store Unit (LSU) Stall ratio in a sampling interval and accurately monitoring the variations in memory contention, MAWS finds a suitable warp concurrency that fits the limited memory resources well, and as a result significantly improve the effective throughput.

2. 11:42 – 11:54 (SC12) Establishing Cooperation Between Camera Applications and Flash-Based Storage to Improve JPEG File Reliability

   Yu-Chun Kuo, Chia-Yu Hu, Ruei-Fong Chiu, and Ren-Shuo Liu

   National Tsing Hua University

   NAND flash-based storage such as SD cards and eMMC chips are the most widely used media for Camera Applications. In this work, we propose to establish cooperation between camera applications and flash-based storage to improve the reliability of JPEG files stored in the storage. We conduct realsystem experiments by storing JPEG files on flash chips to evaluate the benefits of our proposed techniques. Experimental results demonstrate that the reliability of JPEG files can be significantly enhanced.

3. 11:54 – 12:06 (SC13) Considerations of Integrating Computing-In-Memory and Processing-In-Sensor into Convolutional Neural Network Accelerators for Low-Power Edge Devices

Kea-Tiong Tang (1), Wei-Chen Wei (1), Zuo-Wei Yeh (1), Tzu-Hsiang Hsu (1), Yen-Cheng Chiu (1), Cheng-Xin Xue (1), Yu-Chun Kuo (1), Tai-Hsing Wen (1), Mon-Shu Ho (2), Chung-Chuan Lo (1), Ren-Shuo Liu (1), Chih-Cheng Hsieh (1), and Meng-Fan Chang (1)

(1)National Tsing Hua University and (2) National Chung Hsin University.

In quest to explore emerging deep learning algorithms at edge devices, developing low-power and low-latency deep learning acceeelerators (DLAs) have become top priority. To achieve this goal, data processing techniques in sensor and memory utilizing the array structure have drawn much attention. Processing-in-sensor (PIS) solutions could reduce data transfer, computingin-memory (CIM) macros could reduce memory access and intermediate data movement. We propose a new architecture to integrate PIS and CIM to realize low-power DLA. The advantages of using these techniques and the challenges from system point-of-view are discussed.

4. 12:06 – 12:18 (SC14) STT-MRAM for Deep Convolutional Neural Network Acceleration

   Chih-Cheng Chang, Chun-Hsien Li, Tian-Sheuan Chang, and Tuo-Hung Hou

   National Chiao Tung University

   Binary STT-MRAM is a highly anticipated embedded non-volatile memory technology in advanced logic nodes < 28 nm. How to enable its in-memory computing (IMC) capability is critical for enhancing AI Edge. Based on the soon-available STT-MRAM, we report the first binary deep convolutional neural network capable of both local and remote learning. Exploiting intrinsic cumulative switching probability, accurate online training of CIFAR-10 color images (~ 90%) is realized using a relaxed endurance spec (switching 20 times) and hybrid digital/IMC design. For offline training, the accuracy loss due to imprecise weight placement can be mitigated using a rapid non-iterative training-with-noise and fine-tuning scheme.

5. 12:18 – 12:30 (SC15) Efficient Design of Multiple Writes for Algorithmic Multi-ported Memory

   Bo-Ya Chen, Bo-En Chen, and Bo-Cheng Lai

   National Chiao Tung University

   This paper proposes REMAP+, a novel design that enables efficient write scheme for algorithmic multi-ported memory, and attains better performance with smaller area. REMAP+ applies the banking structure of memory design and implements the remap table with SRAM cells instead of costly registers. In the remap table, REMAP+ only keeps the most significant bit of write addresses to more efficiently utilize the space in the table. The hash write controller is simplified with the first fit algorithm to handle write conflict with shorter latency. REMAP+ is implemented in a pipeline scheme to further increase the processing throughput. For a 3W1R memory with 16K depth, REMAP+ has attained 22% shorter access latency and 31.3% smaller area when compared with the previous design.