KIAST PIM반도체설계연구센터 '인공지능 PIM 반도체' 특강

1. 일정

1) 일시 : 04월 28일(금) 09:00-12:00

2) 장소 : 온라인 zoom (주소 추후 공지 예정)/한국어 강의

2. Program

Time	Title	Speaker
09:00 ~ 09:10	Opening	Prof. Hoi-Jun Yoo, KAIST
09:10 ~ 09:40	(PIM) Advances and Trends on Compute-in-Memory based Accelerator Designs	Prof. Jae-Sun Seo, Arizona State University
09:40 ~ 10:10	(PIM) Mixed-Signal and Digital Memory-Centric Computing	Prof. BongJin Kim, University of California Santa Barbara
10:10 ~ 10:40	(PIM) iMCU: A 730-μJ/Classification Digital In-Memory Computing-based Microcontroller Unit for Edge TinyML	Prof. MinGoo Seok, Columbia University
10:40 ~ 11:10	(Architecture) Energy Efficient Deep Learning Algorithm - Architecture Co-Optimization	Prof. Hun Seok Kim, University of Michigan
11:10 ~ 11:40	(Architecture) Disrupting Processor Architecture with Non-invasive Near-Data Processing: DNN Training Case Study	Prof. GwangSun Kim, POSTECH
11:40 ~ 12:10	(Interface) Energy Proportional Link	Prof. WooSeok Choi, SNU

3. 접수방법 및 기간

1) 링크 : https://ai-pim.org/ → 홈페이지 상단 인력양성 →
프로그램 선택 → 신청하기 → 이름/E-mail/소속/연락처 입력 제출

2) 기간 : 04월 23일(일) 18:00 까지

4. 특강 포스터

5. 연사별 강의 요약문

1) Advances and Trends on Compute-in-Memory based Accelerator Designs

Jae-Sun Seo, Cornell University

Abstract

Artificial intelligence (AI) algorithms have intensive computation and memory requirements. To efficiently accelerate complex AI algorithms, a number of custom application-specific integrated circuits (ASIC) chip designs have been demonstrated. Many of such AI accelerators have separate memory (e.g. SRAM) and compute engines (e.g. systolic array of processing elements), and thus have shown that accessing memory is the biggest bottleneck for energy-efficient AI inference, loading data from embedded SRAM memory and moving them to where computing actually occurs. Conventional SRAMs require row-by-row access to load the weights and communicate them to physically-separate compute engines, which limits the parallelism and dissipates a large amount of read/write energy for AI work- loads.

To address this limitation, compute-in-memory (CIM) technique has been proposed to embed computation inside the memory architecture, effectively reducing the on-chip memory access and communication cost. These include SRAM, non-volatile memory (NVM), and DRAM/HBM based in-/near-memory computing schemes. Many analog CIM works started to demonstrate macro-level energy and area benefits while being susceptible to variability, and recently digital CIM works also garnered interest due to the elimination of analog-to-digital converters (ADCs) and higher robustness against variability while consuming more area. Using both analog and digital CIM macros, accelerator systems that integrate many CIM macros have been presented in the literature. This talk will present such advances and recent trends on compute-in-memory based accelerator designs, including analog vs. digital CIMs, CIMs that support fixed-point vs. floating-point precision, and trade-offs of SRAM vs. NVM vs. DRAM/HBM based in-/near-memory computing.

2) Mixed-Signal and Digital Memory-Centric Computing

Bongjin Kim

University of California Santa Barbara (UCSB)

Abstract

Recent advancements in the development of memory-centric computing macros and processors enabled the energy-efficient acceleration of deep learning (DL) with lower-precision mixed-signal integer multiply-and-accumulate (MAC) operations. However, their applications were limited to tiny DL inferences on edge devices with lower accuracy requirements. In this talk, we will present our recent research efforts toward robust and reconfigurable memory-centric computing to overcome the limitations of prior mixed-signal compute-in-memory approaches. Besides the processing of DL, we will also introduce our recent research projects on memory-centric computing for solving challenging combinatorial optimization problems.

3) **iMCU: A 730-μJ/Classification Digital In-Memory Computing-based Microcontroller Unit for Edge TinyML**

MinGoo Seok, Columbia University

Abstract

TinyML envisions performing a deep neural network (DNN)-based inference on an edge device, which makes it paramount to create a neural microcontroller unit (MCU). Toward this vision, some of the recent MCUs integrated in-memory computing (IMC) based accelerators. However, they employ analog-mixed-signal (AMS) versions, exhibiting limited robustness over process, voltage, and temperature (PVT) variations. They also employ a large amount of IMC hardware, which increases silicon area and cost. Also, they do not support a practical software dev framework such as TensorFlow Lite for Microcontrollers (TFLite-micro). Because of this, those MCUs did not present the performance for the standard benchmark MLPerf-Tiny, which makes it difficult to evaluate them against the state-of-the-art neural MCUs. In this paper, we present iMCU, the IMC-based MCU in 28nm, which outperforms the current best neural MCU (SiLab's xG24-DK2601B) by 88X in energy-delay product (EDP) while performing MLPerf-Tiny. Also, iMCU integrates a digital version of IMC hardware for maximal robustness. We also optimize the acceleration targets and the computation flow to employ the least amount of IMC hardware yet still enable significant acceleration. As a result, iMCU's total area is only 2.03 mm^2 while integrating 433KB SRAM and 32KB IMC SRAM.

4) Energy Efficient Deep Learning Algorithm – Architecture Co-Optimization

Hun Seok Kim, University of Michigan

Abstract

This talk presents holistic approaches to realize energy-optimized machine-learning (ML) algorithms, VLSI architectures/accelerators and systems. The optimized system integration is a major challenge in machine-learning systems. A truly energy-optimal ML-IoT solution is attainable only by a cross-layer optimization that requires a full characterization of the complete end-to-end system. Addressing this critical technical challenge in emerging ML-IoT applications, a cross-layer interdisciplinary research that spans deep learning algorithms and VLSI hardware architecture will be discussed in this talk.

Recent advances in model pruning have enabled sparsity-aware deep neural network accelerators that improve the energy-efficiency and performance of inference tasks. This talk introduces a novel transform-domain (TD) neural network accelerator in which convolution operations are replaced by element-wise multiplications with sparse-orthogonal weights. It employs an output stationary dataflow coupled with an energy-efficient memory organization to reduce the overhead of sparse-orthogonal TD kernels that are concurrently processed without any conflicts. Weights in the proposed architecture are non-uniformly quantized with bit-sparse canonical-signed-digit representations to reduce multiplications to simple additions. Moreover, for sparse fully-connected layers (FCLs), the proposed scheme introduces column-based-block structured pruning, which is integrated into the same architecture that maintains full multiply-and-accumulate (MAC) array utilization. Compared to prior dense and sparse neural networks accelerators, the proposed architecture can reduce inference energy by 5.1x and 2.4x, and increase performance by 5.2x and 2.1x, respectively, for convolution layers. For sparse FCLs, the proposed architecture can reduce inference energy by 2.4x and increase performance by 2x compared to prior work.

This talk also discusses technical challenges and promises of processing in memory (PIM) for deep learning acceleration will be discussed in the holistic ML-IoT system perspective.

5) Disrupting Processor Architecture with Non-invasive Near-Data Processing: DNN Training Case Study

GwangSun Kim, POSTECH

Abstract

In modern high-performance systems, including GPU systems, the memory bandwidth wall affects overall system performance for many important workloads such as DNN training by limiting the high utilization of the large number of compute cores. While processing-in-memory shows promise in making DRAM's high internal bandwidth available for computation within DRAM, it's also crucial to maximize the utilization of memory bandwidth available to the processor die. In particular, by placing near-data processing (NDP) units near the memory controllers within the processor die and offloading memory-bound operations from the core, compute-bound operations executed on the cores can be overlapped with memory-bound operations executed on the NDP units.

However, realizing an effective NDP mechanism presents three major challenges. First, overlapping compute- and memory-bound operations often cannot be done due to dependencies. For example, in CNNs, memory-bound layers, such as batch normalization, depend on the preceding convolutional layer and prevent overlap. Second, while fine-grained NDP can maximize NDP offloading opportunities, existing approaches require a special instruction executed on the core to generate NDP commands explicitly. Introducing such an instruction to ISA could require extensive changes in the core microarchitecture and core-side software and prevent widespread adoption of NDP. Third, the NDP unit should be flexible enough to support various operations while incurring minimal hardware overhead.

In this talk, I'll introduce a memory access-triggered NDP (mtNDP) architecture that addresses these challenges. The mtNDP is a versatile and non-invasive mechanism that can be deployed to various components of the system, including within a processor die (e.g., near memory controllers) or memory expanders (e.g., within CXL controller). Our case study on mtNDP for DNN training on a multi-GPU system shows that it can achieve a significant speedup of up to 2.7x (47% on average) while reducing energy consumption by up to 41% (38% on average).

6) **Energy Proportional Link**

WooSeok Choi, SNU

Abstract

With the explosive growth of data traffic in the AI and Big Data era, data movement/communication in both high performance computing systems and mobile platforms are starting to consume significant portion of the system power. This continuing trend urgently calls techniques for energy efficient data communication before energy spent on moving data dominates the information processing stack. This talk discusses potential strategies to improve data movement efficiency and presents design examples for energy proportional links (EPLs). EPLs address the limitation of data movement efficiency that improvement on energy efficiency of link building blocks is not fully translated to power savings at system-level, especially in many practical applications where links are only sporadically utilized. Silicon measurements of the energy proportional links will be presented to prove the effectiveness and efficiency.