IEEE JOURNAL ON
EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS

Call for Papers

Communication-aware Designs and Methodologies for Reliable and Adaptable On-Chip AI SubSystems and Accelerators

Guest editors

Kun-Chih (Jimmy) Chen, National Sun Yat-sen University, Taiwan (kcchen@mail.cse.nsysu.edu.tw)
Masoumeh (Azin) Ebrahimi, KTH Royal Institute of Technology, Sweden (mebr@kth.se)
Maurizio Palesi, University of Catania, Italy (maurizio.palesi@dieei.unict.it)
Tim Kogel, Synopsys, Germany (tim.kogel@synopsys.com)

Scope and Purpose

The Big data and Internet of Things (IoTs) trend helps to drive the progress of artificial intelligence (AI) technologies in recent years. The notable benefits of neural network (NN) technologies (*e.g.*, artificial neural network (ANN), spiking neural network (SNN), etc.) are widely applied to many real-world applications, such as speech recognition and image classification, and the resulting accuracies have been well above human-level. Due to its undoubted significance, research works of "AI accelerator/subsystem designs" have drawn lots of attention from both academia and industry.

Due to the massive parallel processing, the performance of the current large-scale artificial neural network is often limited by the massive communication overheads and storage requirements. However, the issues of *interconnection*, *communication, computation synchronized with memory subsystem*, *reliability*, and *flexibility* of AI engines receive less attention. As a result, flexible interconnections and efficient data movement mechanisms for future on-chip AI accelerator are worthy of study, such as:

- **New efficient data movement in contemporary AI subsystems:** Memory access latency and overhead of NN connections already become the performance bottleneck. The memory area already dominates the total silicon cost, and the data access contributes to a large portion of power consumption in accelerators of deep neural networks (DNN). Many researchers tackled this issue by improving on conventional Von Neumann-type architectures, such as *pruning and scheduling*. In recent years, novel Non-Von Neumann architectures, such as *In-memory Computing* and *Near-memory Computing* techniques, had been proposed extensively to accommodate the computing/processing-in-memory issue, which is worthy of investigation in this special issue.
- **Design methodology considering tradeoffs among computing engine (PE arrays) and data movement/storage unit (memory hierarchy) from energy/power/timing point of view:** In conventional AI subsystem designs, most works focus on task scheduling to improve the efficiency of the data movement between the memory and the computing engine. In the emerging AI-on-Chip designs, in addition to timing efficiency, power and energy efficiencies become major concerns.

Consequently, novel and efficient task scheduling and data movement methodologies from energy/power/timing optimization points of view are covered in this special issue as well.

- **"Flexible" and "reliable" communication-aware AI subsystems for future on-chip adaptive learning application:** With predefined functional datapath of dedicated neural network models, the computing flows of the current AI accelerators are usually "fixed" and "non-adaptive." On the other hand, the reliable issue of the SNN design becomes even severe in leading technology nodes because the spike signals are noise sensitive. Therefore, this special issue wants to invite research works of novel computing flows, in consideration of flexible and reliable data movements and communications, for versatile modern AI applications.

Topics of interest

Topics of interest to this special issue include, but are not limited to:

- Challenges of massive memory data access in deep learning
- Near- and In-memory computing techniques for saving data movement
- ANN design based on emerging non-volatile memory devices
- Data movement optimization through task scheduling of the artificial neural network
- Energy- and accuracy-aware pruning and quantization mechanism of neural networks
- Novel interconnection networks for the neural networks (e.g., DNN, RNN, ANN, SNN, etc.)
- Efficient on-chip communication of multicore-based artificial neural network
- Communication/traffic-aware artificial neural network architecture and application
- New topology of on-chip communication for efficient neural network computing
- NoC design for heterogeneous ANN computing
- Reliable and robust computing method and on-chip interconnection for AI computing
- Cross-layer optimization for artificial neural network architecture and application
- Tradeoffs among computing engine and data movement from energy/power/timing point of view
- Flexible computing flow and reconfigurable neural network architecture for on-chip learning applications and Artificial General Intelligence (AGI)
- AI-on-Chip interconnections/designs to meet with the requirements and constraints of emerging application (e.g., 5G, medical applications, Industry 4.0, etc.)
- "Flexible" and "reliable" communication-aware AI subsystems for future on-chip adaptive learning application

Submission procedure

Prospective authors are invited to submit their papers following the instructions provided on the JETCAS website: https://mc.manuscriptcentral.com/jetcas. The submitted manuscripts should not have been previously published nor should they be currently under consideration for publication elsewhere.

Important dates

- Manuscript submissions due 2020-05-20
- First round of reviews completed 2020-06-20
- Revised manuscripts due 2020-07-10
- Second round of reviews completed 2020-07-30
- Final manuscripts due 2020-08-15
- Target publication date 2020-09-30

Request for information

- Kun-Chih (Jimmy) Chen (kcchen@mail.cse.nsysu.edu.tw)