Kyuk-Jae Lee

Seoul National University, South Korea

hyukjae@snu.ac.kr

Biography

Hyuk-Jae Lee received the B.S. and M.S. degrees in Electronics Engineering from Seoul National University, Korea, in 1987 and 1989, respectively, and the Ph.D. degree in Electrical and Computer Engineering from Purdue University, West Lafayette, IN, in 1996. From 1998 to 2001, he worked at the Server and Workstation Chipset Division of Intel Corporation in Hillsboro, Oregon as a Senior Component Design Engineer. From 1996 to 1998, he was on the faculty of the Department of Computer Science of Louisiana Tech University at Ruston, Louisiana. In 2001, he joined the Department of Electrical and Computer Engineering at Seoul National University, Korea, where he is currently working as a Professor and the Department Chair. In 2002, Dr. Lee founded Mamurian Design, Inc., a fabless SoC design house for multimedia applications and served as an Acting CTO until 2012 to develop MPEG-4, H.264 and AMR multimedia processors. From 2013 to 2015, He served as the Director of System Semiconductor Program at the KEIT (Korea Evaluation Institute of Industrial Technology), in charge of managing the research fund from Korean Ministry of Industry, Trade, and Energy. From 2018, he has been serving as the Director of Strategic Industrial Collaboration Program between Seoul National University and Samsung Electronics Device Solution. He is currently a vice president of IEIE (Institute of Electronics and Information Engineers). His research interests are in the areas of computer architecture and SoC design for neural network processing and video applications. He has published more than 90 journal papers and 100 conference papers. He has been serving as an Associate Editor of IEEE Transactions on Circuits and Systems for Video Processing from 2015 and also as a CASS Distinguished Lecture from 2020. He was a guest editor of a special issue for Display Journal in 2015 and is currently a guest editor of a special issue of MDPI.

Lecture 1: Memory Access Optimization for Neural Network Processors

Recently, a number of neural network processors have been developed to efficiently process deep neural networks. Most these processors include a memory system with a large capacity of on-chip SRAMs as well as high-bandwidth off-chip DRAMs. In order to fully utilize the hardware resources of neural processors, efficient access of both on-chip SRAMs and off-chip DRAMs is essential. This tutorial presents traditional and state-of-art optimization techniques for memory access for neural network applications. State-of-art neural processors are introduced and common characteristics of the memory systems are explained. The optimization techniques for the data access of an on-chip SRAM are introduced. The scheduling, parallelization and data allocation of various deep learning algorithms are presented and the pros and cons of optimizations for a given memory system are explained. Additional data optimization for efficient access of an off-chip DRAM is explained in the next. To this end, the basic characteristics of a DRAM organization are introduced and then the data access scheduling for efficient DRAM access is explained. As the last subject of this tutorial, future memory systems are introduced.

Processing-in-Memory (PIM) and Approximate Memory (AM) architecture are briefly introduced and data optimizations for PIM and AM in deep neural processing are presented. A SCM (Storage Class Memory), a potential new memory hierarchy, is introduced and data access techniques for them are also presented.

Lecture 2: New Memory Architecture for Deep Learning Applications

Deep learning applications demand a large amount of data movement between processors and memory devices. To reduce the time and power consumption for data movement, extensive research has been carried out to develop new memory architecture suitable for deep learning. This talk introduces new trends of memory architecture for deep learning applications. Among them, Processing-near-Memory (PNM) and Approximate Memory (AM) architectures attract wide attention. PNM architecture is used to reduce data movement by placing computation near DRAM devices. On the other hand, AM architecture attempts to reduce the precision of deep learning data, and consequently, to reduce memory traffic. AM is especially suitable for deep learning applications of which accuracy may not be degraded significantly even with a loss of data precision. Recent developments in PNM and AM architectures are briefed and then data access optimizations for these memory architectures are explained. A proposal of a new memory architecture combining the two architectures is presented. The new memory architecture is simulated by modifying a GPU simulator and its effectiveness is presented with simulation results.