7.2 Accelerators using novel memory technologies

Date: Wednesday, March 27, 2019
Time: 14:30 - 16:00
Location / **Room:** Room 2

Chair:
Mladen Berekovic, TU Braunschweig, DE

Co-Chair:
Andrea Marongiu, Università di Bologna, IT

The session focuses on accelerating three complex applications. They all use novel combination of memory and computing elements to achieve this goal. The first paper focuses on pattern matching and proposes to use resistive-RAM (RRAM) to accelerate an Automata Processor (AP). The second one focuses on Homomorphic Encryption (HE) and improves the performance and energy consumption using near-data processing on a 3D-stacked memory (Hybrid Memory Cube). The third paper focuses on Inference (DCNN) proposes a 3D-stacked neuromorphic architecture consisting of Processing Elements and multiple DRAM layers.

Time	Label	Presentation Title Authors
14:30	7.2.1	**TIME-DIVISION MULTIPLEXING AUTOMATA PROCESSOR** **Speaker**: Mottaqiallah Taouil, Delft University of Technology, NL **Authors**: Jintao Yu, Hoang Anh Du Nguyen, Muath Abu Lebdeh, Mottaqiallah Taouil and Said Hamdioui, Delft University of Technology, NL ***Abstract*** *Automata Processor (AP) is a special implementation of non-deterministic finite automata that performs pattern matching by exploring parallel state transitions. The implementation typically contains a hierarchical switching network, causing long latency. This paper proposes a methodology to split such a hierarchical switching network into multiple pipelined stages, making it possible to process several input sequences in parallel by using time-division multiplexing. We use a new resistive RAM based AP (instead of known DRAM or SRAM based) to illustrate the potential of our method. The experimental results show that our approach increases the throughput by almost a factor of 2 at a cost of marginal area overhead.* Download Paper (PDF; Only available from the DATE venue WiFi)
15:00	7.2.2	**NEAR-DATA ACCELERATION OF PRIVACY-PRESERVING BIOMARKER SEARCH WITH 3D-STACKED MEMORY** **Speaker**: Alvin Oliver Glova, University of California, Santa Barbara, US **Authors**: Alvin Oliver Glova, Itir Akgun, Shuangchen Li, Xing Hu and Yuan Xie, University of California, Santa Barbara, US ***Abstract*** *Homomorphic encryption is a promising technology for enabling various privacy-preserving applications such as secure biomarker search. However, current implementations are not practical due to large performance overheads. A homomorphic encryption scheme has recently been proposed that allows bitwise comparison without the computationally-intensive multiplication and bootstrapping operations. Even so, this scheme still suffers from memory-bound performance bottleneck due to large ciphertext expansion. In this work, we propose HEGA, a near-data processing architecture that leverages this scheme with 3D-stacked memory to accelerate privacy-preserving biomarker search. We observe that homomorphic encryption-based search, like other emerging applications, can greatly benefit from the large throughput, capacity, and energy savings of 3D-stacked memory-based near-data processing architectures. Our near-data acceleration solution can speed up biomarker search by 6.3x with 5.7x energy savings compared to an 8-core Intel Xeon processor.* Download Paper (PDF; Only available from the DATE venue WiFi)
15:30	7.2.3	**TOWARDS CROSS-PLATFORM INFERENCE ON EDGE DEVICES WITH EMERGING NEUROMORPHIC ARCHITECTURE** **Speaker**: Yi Wang, Shenzhen University, CN **Authors**: Shangyu Wu[1], Yi Wang[1], Amelie Chi Zhou[1], Rui Mao[1], Zili Shao[2] and Tao Li[3] [1]Shenzhen University, CN; [2]The Chinese University of Hong Kong, HK; [3]University of Florida, US ***Abstract*** *Deep convolutional neural networks have become the mainstream solution for many artificial intelligence appli- cations. However, they are still rarely deployed on mobile or edge devices due to the cost of a substantial amount of data movement among limited resources. The emerging processing-in-memory neuromorphic architecture offers a promising direction to accelerate the inference process. The key issue becomes how to effectively allocate the processing of inference between computing and storage resources for mobile edge computing. This paper presents Mobile-I, a resource allocation scheme to accelerate the Inference process on Mobile or edge devices. Mobile-I targets at the emerging 3D neuromorphic architecture to reduce the processing latency among computing resources and fully utilize the limited on-chip storage resources. We formulate the target problem as a resource allocation problem and use a software-based solution to offer the cross-platform deployment across multiple mobile or edge devices. We conduct a set of experiments using realistic workloads that are generated from Intel Movidius neural compute stick. Experimental results show that Mobile-I can effectively reduce the processing latency and improve the utilization of computing resources with negligible overhead in comparison with representative schemes.* Download Paper (PDF; Only available from the DATE venue WiFi)

Time	Label	Presentation Title Authors
16:00	IP3-11, 445	**VISUAL INERTIAL ODOMETRY AT THE EDGE - A HARDWARE-SOFTWARE CO-DESIGN APPROACH FOR ULTRA-LOW LATENCY AND POWER** **Speaker**: Dipan Kumar Mandal, Intel Corporation, IN **Authors**: Dipan Kumar Mandal[1], Srivatsava Jandhyala[1], Om J Omer[1], Gurpreet S Kalsi[1], Biji George[1], Gopi Neela[1], Santhosh Kumar Rethinagiri[1], Sreenivas Subramoney[1], Hong Wong[2], Lance Hacking[2] and Belliappa Kuttanna[2] [1]Intel Corporation, IN; [2]Intel Corporation, US ***Abstract*** *Visual Inertial Odometry (VIO) is used for estimating pose and trajectory of a system and is a foundational requirement in many emerging applications like AR/VR, autonomous navigation in cars, drones and robots. In this paper, we analyze key compute bottlenecks in VIO and present a highly optimized VIO accelerator based on a hardware-software co-design approach. We detail a set of novel micro-architectural techniques that optimize compute, data movement, bandwidth and dynamic power to make it possible to deliver high quality of VIO at ultra-low latency and power required for budget constrained edge devices. By offloading the computation of the critical linear algebra algorithms from the CPU, the accelerator enables high sample rate IMU usage in VIO processing while acceleration of image processing pipe increases precision, robustness and reduces IMU induced drift in final pose estimate. The proposed accelerator requires a small silicon footprint (1.3 mm2 in a 28nm process at 600 MHz), utilizes a modest on-chip shared SRAM (560KB) and achieves 10x speedup over a software-only implementation in terms of image sample-based pose update latency while consuming just 2.2 mW power. In a FPGA implementation, using the EuRoC VIO dataset (VGA 30fps images and 100Hz IMU) the accelerator design achieves pose estimation accuracy (loop closure error) comparable to a software based VIO implementation.* Download Paper (PDF; Only available from the DATE venue WiFi)
16:01	IP3-12, 263	**CAPSACC: AN EFFICIENT HARDWARE ACCELERATOR FOR CAPSULENETS WITH DATA REUSE** **Speaker**: Alberto Marchisio, Vienna University of Technology (TU Wien), AT **Authors**: Alberto Marchisio, Muhammad Abdullah Hanif and Muhammad Shafique, Vienna University of Technology (TU Wien), AT ***Abstract*** *Recently, CapsuleNets have overtaken traditional Deep Convolutional Neural Networks (CNNs), because of their improved generalization ability due to the multi-dimensional capsules, in contrast to the single-dimensional neurons. Consequently, CapsuleNets also require extremely intense matrix computations, making it a gigantic challenge to achieve high performance. In this paper, we propose CapsAcc, the first specialized CMOS-based hardware architecture to perform CapsuleNets inference with high performance and energy efficiency. State-of-the-art convolutional CNN accelerators would not work efficiently for CapsuleNets, as their designs do not account for unique processing nature of CapsuleNets involving multi-dimensional matrix processing, squashing and dynamic routing. Our architecture exploits the massive parallelism by flexibly feeding the data to a specialized systolic array according to the operations required in different layers. It also avoids extensive load and store operations on the on-chip memory, by reusing the data when possible. We synthesized the complete CapsAcc architecture in a 32nm CMOS technology using Synopsys design tools, and evaluated it for the MNIST benchmark (as also done by the original CapsuleNet paper) to ensure consistent and fair comparisons. This work enables highly-efficient CapsuleNets inference on embedded platforms.* Download Paper (PDF; Only available from the DATE venue WiFi)
16:02	IP3-13, 576	**SDCNN: AN EFFICIENT SPARSE DECONVOLUTIONAL NEURAL NETWORK ACCELERATOR ON FPGA** **Speaker**: Suk-Ju Kang, Sogang University, KR **Authors**: Jung-Woo Chang, Keon-Woo Kang and Suk-Ju Kang, Sogang University, KR ***Abstract*** *Generative adversarial networks (GANs) have shown excellent performance in image generation applications. GAN typically uses a new type of neural network called deconvolutional neural network (DCNN). To implement DCNN in hardware, the state-of-the-art DCNN accelerator optimizes the dataflow using DCNN-to-CNN conversion method. However, this method still requires high computational complexity because the number of feature maps is increased when converted from DCNN to CNN. Recently, pruning has been recognized as an effective solution to reduce the high computational complexity and huge network model size. In this paper, we propose a novel sparse DCNN accelerator (SDCNN) combining these approaches on FPGA. First, we propose a novel dataflow suitable for the sparse DCNN acceleration by loop transformation. Then, we introduce a four stage pipeline for generating the SDCNN model. Finally, we propose an efficient architecture based on SDCNN dataflow. Experimental results on DCGAN show that SDCNN achieves up to 2.63 times speedup over the state-of-the-art DCNN accelerator.* Download Paper (PDF; Only available from the DATE venue WiFi)

Time	Label	Presentation Title Authors
16:00		End of session **Coffee Break** in Exhibition Area

Coffee Breaks in the Exhibition Area

On all conference days (Tuesday to Thursday), coffee and tea will be served during the coffee breaks at the below-mentioned times in the exhibition area.

Lunch Breaks (Lunch Area)

On all conference days (Tuesday to Thursday), a seated lunch (lunch buffet) will be offered in the Lunch Area to fully registered conference delegates only. There will be badge control at the entrance to the lunch break area.

Tuesday, March 26, 2019

- Coffee Break 10:30 - 11:30
- Lunch Break 13:00 - 14:30
- Keynote Lecture "Leonardo da Vinci, Humanism and Engineering between Florence and Milan" by Claudio Giorgione in room 1 13:50 - 14:20
- Coffee Break 16:00 - 17:00

Wednesday, March 27, 2019

- Coffee Break 10:00 - 11:00
- Lunch Break 12:30 - 14:30
- Keynote Lecture "Heterogeneous, High Scale Computing in the Era of Intelligent, Cloud-Connected" by David Pellerin, Amazon, US in room 1 13:50 - 14:20
- Coffee Break 16:00 - 17:00

Thursday, March 28, 2019

- Coffee Break 10:00 - 11:00
- University Booth Best Demo Award Presentation at the University Booth 10:30
- Lunch Break 12:30 - 14:00
- Keynote Lecture "A Fundamental Look at Models and Intelligence" by Edward A. Lee, University of California, Berkeley, US in room 1 13:20 - 13:50
- Coffee Break 15:30 - 16:00