Oral S02

Efficient SoC Design & Key Computing Blocks

Date/Venue	**7/31(三) 13:30-15:00 [海棠廳]**
Chairs(s)	**陳春僥 /國立高雄大學電機系** **賴信志 /國立虎尾科技大學**

S02.1 | 13:30-13:41

A 40nm 143.5GOPS/W Arbitrary Kernel Size Applicable Universal Flexible-Oriented Network on Chip (UFONoC) Processor with ARM Compatible Data Transmission Protocol

Kun-Chih (Jimmy) Chen, Yi-Sheng Liao, Hao-Hsiang Peng, Ting-En Kao, Wei-Ren Syu, and Pin-Ching Shen

[1]National Yang Ming Chiao Tung University, Hsinchu City, Taiwan

[2]National Sun Yat-sen University, Kaohsiung City, Taiwan

In recent years, the hardware accelerator for Deep Neuron Network (DNN) has attracted attention due to the increased demand for real time data processing in Artificial intelligence of things (AIoT) applications. With the development of various DNN architectures, the flexible architecture of neural network accelerator design has become more important. This work introduces a highly adaptable DNN accelerator design called Universal Flexible-Oriented Network on Chip (UFONoC). The UFONoC employs a kernel-based weight-wise Neural Network (NN) processing mechanism that can support any size and shape of convolution kernel or feature map in the target DNNs, resulting in greater flexibility. The UFONoC utilizes a hybrid data reuse method for packet transmission in the NoC to reduce memory access. In addition, it adopts the AXI4-stream protocol for data transmission in NoC architecture to minimize data transmission latency. Compared with the prior works, our proposed UFONoC achieves energy savings by 60% to 93% and improves the hardware efficiency by 23% to 742%.

S02.2 | 13:42-13:53

SG-Float: Achieving Memory Access and Computational Power Reduction Using Self-Gating Float in CNN Accelerators

Yu-Yang Liu, Jun-Shen Wu, Ren-Shuo Liu

Department Of Electrical Engineering, National Tsing Hua University

Convolutional neural networks (CNNs) are crucial for enabling the future artificial intelligence world. However, due to its large quantity of data and computation requirements, devices need considerable memory and hardware resources, limiting the implementation of energy-constrained or

hardwareconstrained devices, e.g., IoT devices. In this work, we present self-gating float (SG-Float), algorithm hardware co-design of a novel binary number format, which significantly reduces CNN memory accesses and computational power. First, we propose the novel SG-Float number format that uses the exponent as the indicator to self-gate the man-tissa to zero. Using SG-Float, the relatively small values are approximately represented only by the exponent. As a result, SG-Float can increase the zero proportion of mantissas and reduce mantissa multiplications. Second, we offer an optimization technique to best match SG-Float with CNN accelerators, SGFloat buffering strategy, which reduces the memory accesses of SG-Float. Finally, we apply the SG-Float buffering strategy to floating-point, vector-vector multiplication processing elements (PEs), which NVDLA adopts, in TSMC 40nm technology. Our evaluation results show that SG-Float can achieve up to 37% memory access power reduction with our proposed SG-Float buffering strategy and up to 46% computational power reduction compared with AdaptivFloat with negligible power and area overhead. Furthermore, the inference accuracy loss caused by SG-Float is within 1%.

S02.3 | 13:54-14:05

Low-Complexity VLSI Design of Blind Rotation for Torus Fully Homomorphic Encryption

Tzyy-Shiuan Yang and Ming-Der Shieh
[1]Department of Electrical Engineering National Cheng Kung University
[2]Electrical Engineering Department, National Tsing Hua University
[3]Taiwan Electronic System Design Automation

Torus-based fully homomorphic encryption (TFHE) is notable for its faster and programmable bootstrapping (PBS) algorithm. However, TFHE suffers from its high computational complexity in PBS, especially in the Blind Rotation (BR) process. BR involves multiple iterations of vector-matrix multiplication of polynomials, necessitating the use of Number Theoretic Transform (NTT) for efficient computation. Moreover, algorithm optimization, such as bootstrapping key unrolling, requires more expensive computations of an iteration in BR, leading to an increased number of NTT. To address these issues, this work mainly explores the design of reduced NTT (RNTT), which leverages the sparsity of polynomial multiplications in PBS. We proposed a reconfigurable RNTT architecture with an efficient memory conflict-free addressing algorithm adaptable to different configurations. Experimental results show that the developed RNTT design only requires less than 2% area and latency compared to a highly parallel approach. Overall, it contributes to a 21.76% performance enhancement in the Blind Rotation Core (BRC), which is a low-complexity accelerator developed in this work.

S02.4 | 14:06-14:17

Trojan Horse Detection for RISC-V Cores Using Cross-Auditing

Wei-Po Huang, Shi-Yu Huang, Chi-Kang Chen
[1]Graduate Degree Program of College of Institute of Artificial Intelligence
[2]Innovation, National Yang Ming Chiao Tung University, Taiwan
[3]Institute of Electronics, National Yang Ming Chiao Tung University, Taiwan

In security-critical applications, malicious Trojan Horses embedded in a CPU core could impose great threats on the security of an SoC. In this work, we propose a "Trojan-Horse detection framework" using a cross-auditing scheme. Our framework takes a target RISC-V core, and then pairs it up with another reference RISC-V core to conduct the functional simulation using a set of benchmark programs. The "care outputs" of both cores are compared to reveal the potential Trojan Horses in the target core. A set of well-known Trojan Horses are implanted into an open-source RISC-V core to evaluate the effectiveness of this framework. We found that we can successfully detect almost every implanted Trojan Horse as long as it has been activated and manifested by the benchmark programs.

S02.5 | 14:18-14:29

Optimization of Hardware-Software Co-Design for Underwater Ultrasonic Object Recognition System Based on FPGA

Pin-Hao Tung, Geng-Shi Jeng, Bo-Cheng Lai
[1]Graduate Degree Program of College of Institute of Artificial Intelligence
[2]Innovation, National Yang Ming Chiao Tung University, Taiwan
[3]Institute of Electronics, National Yang Ming Chiao Tung University, Taiwan

With the advancement of technology and rapid development of machine learning, unmanned vehicles have made significant progress, especially in aquatic environments. The vast and unpredictable nature of the ocean necessitates the use of Unmanned Surface Vehicles (USVs) for exploration. USVs utilize underwater sonar devices to detect their surroundings and identify targets using machine learning techniques. However, current research often separates the optimization of sonar imaging from image recognition, lacking an integrated approach. This study focuses on analyzing the various software and hardware design factors and their interrelationships in underwater ultrasonic image recognition tasks. By developing algorithms that consider user requirements, we aim to determine the optimal hardware configurations for these tasks. This integrated approach is expected to improve recognition accuracy, reduce latency, and optimize resource usage, thereby enhancing the efficiency and effectiveness of USVs in complex aquatic environments.

S02.6 | 14:30-14:41

Design and Implementation of a Four-Core Multi-Stream SIMD RISC-V Processor Architecture Supporting SPMD Mechanism for Machine Learning Jih-Ching Chiu, Zhe-You Yan, Hao-Yi Chen, Jiun-Wen Hsiao, Chen-Hao Chao

[1]Department of Electrical Engineering National Sun Yat-sen University Kaohsiung, Taiwan
[2]Department of Electrical Engineering National Sun Yat-sen University Kaohsiung, Taiwan
[3]Department of Electrical Engineering National Sun Yat-sen University Kaohsiung, Taiwan
[4]Department of Electrical Engineering National Sun Yat-sen University Kaohsiung, Taiwan
[5]Chen-Hao Chao Department of Electrical Engineering National Sun Yat-sen University Kaohsiung, Taiwan

Based on the Implanting Machine Learning Accelerator with Multi-Streaming SIMD Mechanism to RISC-V Processor, we have designed and implemented the quad-core multi-stream SIMD RISC processor architecture that supports the SPMD mechanism for machine learning. Additionally, we propose the architecture with an Authorizing Zone (AZ) for SPMD operations efficiently to take the processor addressing, make the thread invocation and keep the mutex operations of the critical sections for the procedure cooperation. Furthermore, we proposed the booting procedure to support the SPMD algorithm models in the lightweight manner for this proposed system architecture. Using the Design Compiler based on 40- nanometer technology, this architecture successfully achieves its functional mechanisms at 250MHz. According to the analysis results of the VGG16 model, the execution speed is 3.89 times faster than that of a single-core SIMD RISC processor, 18.5 times faster than that of an ARM Cortex-M55 processor, 1.76 times faster than an Intel® Core i7-4970K CPU, and approaching the efficiency of GPU computing.

S02.7 | 14:42-14:53

A Sense-Amplifier Flip-Flop with transition completion detection scheme in 16nm for Low- Voltage Applications

Cheng-Hsueh Yang and Jin-Fa Lin
Department of Information and Communication Engineering, Chaoyang University of Technology

A low power sense-amplifier (SA) based flip-flop (FF) with transition completion detection for low voltage applications is proposed. The proposed design integrates the generated detection circuit to indicate the completion of both SA and RS-latch and thus overcoming the operational yield degradation

when keep transistor-count in minimum. Simulation results shown that the minimum VDD of our design is 200mV lower than convention design, which means our design can operate even when VDD is in the subthreshold region.

S02.8 | 14:53-15:04

A Low-cost and High-speed Polynomial Multiplier for NTRU Algorithm

Xin-Han Liu, Shiann-Rong Kuang, Ryan Tso, Yi-Ting Lin, and Pei-Yu Chung
Department of Computer Science and Engineering , National Sun Yat -sen University, Kaohsiung, Taiwan

The NTRU cryptosystem is one of the main alternatives for practical implementations of post-quantum public-key cryptography, and large-degree polynomial multiplication is usually its performance bottleneck. The polynomial multiplication of NTRU is typically achieved by executing two phases iteratively. The first phase involves generating one row of partial products using two's complement circuits and multiplexers based on the coefficient values of the ternary polynomial. The second phase involves accumulating the partial product row with the current partial sum. In this paper, we propose a very simple and low-cost polynomial multiplier to accelerate NTRU's main operation, making it suitable for use in resource-constrained embedded devices. In the proposed multiplier, the add-one operation in the first phase is delayed and integrated into the accumulation phase without significantly complicating the accumulator. FPGA implementation results show that the proposed architecture achieves 15.4% area reduction and 20.0% speedup compared to previous designs.