Poster Session 2

Digital & System

Date/Time	8/2(三) 15:30—17:00
Chair(s)	黃朝宗／國立清華大學 電機工程學系

PD01

Ureka: A Blockchain-based Decentralized Access Management Framework for Securing IOT Devices

Yi-Chun Yang[1], Ren-Song Tsay[1], Borhan Lee[2]
[1]Department of Computer Science, National Tsing Hua University
[2]Institution of Information Systems and Applications, National Tsing Hua University

As the Internet of Things (IoT) continues to grow, there is a pressing need for secure and efficient management of device access rights. This paper presents Ureka, an open system that uses Blockchain technology to manage IoT access rights in a transparent and bidirectional way. Ureka consists of a personal private system, a Blockchain smart contract system, and a ticket module for IoT devices. Each user and IoT device is identified and authenticated through a unique asymmetric key-pair, and access to devices is mediated through smart contracts and tickets. Ureka also implements regulation mechanisms to govern user behavior and prevent rule violations. Our evaluation shows that Ureka is effective in managing IoT access rights and preventing unauthorized access. The contributions of this paper include the Ureka open standard, the implementation of a transparent and bidirectional smart contract and ticket system, and the regulation mechanisms for governing user behavior.

Keywords: Access Control; Blockchain; Decentralization; General Data Protection Regulation (GDPR); Internet of Things (IoT)

PD02

The Polymorphous Quad-Core RISC-V Processor with Atomic Operations and Reconfigurable Caches

Jih-Ching Chiu, Bo-Yu Lai, Chen-Hao Chao, Hao-Yi Chen
Department of Electrical Engineering National Sun Yat-sen University

With the development of technology and the improvement of various computing needs, especially in the application of the Artificial Intelligence of Things (AIoT). It is used to manage and respond to the Edge computing of sensors, to achieve real-time, multi-

tasking and energy-saving computational ability. The prototype concept of multi-core processor architecture with flexible structure management based on superscalar design was proposed by our laboratory in 2011. However, due to the complexity of the overall management and control of the instruction flow and data flow, it is difficult to efficiently implement the processor design for embedded-oriented requirements. Therefore, the concept of a polymorphous processor is proposed. Using a specific switchable execution mode design, combined with an architecture that simplifies the instruction path and shared registers. Finally, this design is implemented by the prototype of quad-core, and the result verifies that the characteristics of this processor meet the requirements. According to the research results, this paper is based on the RISC-V instruction set architecture RV32I. And refer to the embedded polymorphous quad-core processor architecture proposed by our lab. Improve the architecture and add the following designs: 1. Forwarding Center: Used for the transfer of data between the four cores and reduce the time for waiting for the data to be write back to the register file. 2. Atomic Memory Operation: To ensure the correctness of the read and write sequence of data in the critical region (Critical Region) during the execution of multi-core and multi-threading. 3. Reconfigurable Cache: To provide the memory of polymorphous multi-core processor with a flexible architecture, and to maintain high-bandwidth benefits of instruction and data access. After verification, Reconfigurable I-Cache can increase the average instruction fetch speed by about 17.12%. The RISC-V instruction set architecture brings the features of simplified instructions, easy-to-add instructions, and open network resources. This paper uses Verilog to design and implement a synthesizable quad-core processor that can switch multiple processor modes, support atomic operations, and reconfigurable cache with improved performance. Finally, use the compilation tools provided by riscv-toolchain to generate programs for verifying and comparing performance to ensure that the results conform to the design concept. The circuit synthesis uses the TSMC 40nm process, the processor core runs at 333MHz, and the area is about 128,084um^2.

PD03

Super High Speed Transport Layer ASIC for Universal Serial Bus 4.0 Protocol

Guo-Ming Sung, Yu-Di Huang, Bo-Rui Huang, Kuang Chung, Chih-Ping Yu
Department of Electrical Engineering, National Taipei University of Technology

This paper proposes a super high speed transport layer ASIC for Universal Serial Bus 4.0 (USB 4.0) protocol. The proposed ASIC is designed and verified with FPGA develop board and is fabricated in TSMC 0.18-μm CMOS process. The USB 4.0 operates in 40 GB/s and accepts many previous compatible protocols. In USB protocol, several functional layers are considered, such as protocol adapter layer, configuration layer, transport layer,

logical layer, electrical layer, and so on. This study focuses on designing a transceiver ASIC in transport layer, which is used to process the tunneled packet from protocol adapter layer or the control packet from configuration layer. Those packets will be routed and verified with Quality of Service (QoS) to guarantee the packet transmission correctly. After the designed functions have been verified with FPGA board (Intel DE-10), an ASIC is implemented. The simulation results present that the throughput and power consumption are 966.784 Mbps and 154.87 mW, respectively, at the operational frequency of 125/100 MHz and supply voltage of 1.8 V.

PD05

Heart Valve Disease Recognition Using Phonocardiogram Signal Based on A Lightweight Convolution Neural Network

Yen-Ching Chang[1], Szu-Ting Wang[2], Ying-Hsiu Hong[1], Yao-Feng Liang[3], Ming-Hwa Sheu[1], and Shin-Chi Lai[3,4]
[1]Department of Electronics Engineering, National Yunlin University of Science and Technology
[2]Doctor's Program of Smart Industry Technology Research and Design, National Formosa University
[3]Department of Automation Engineering, National Formosa University
[4]Smart Machinery and Intelligent Manufacturing Research Center, National Formosa University

Cardiovascular diseases (CVDs) can be detected early through heart sounds, making auscultation one of the important methods for diagnosing such conditions. In order to assist doctors more effectively in clinical diagnosis, a lightweight system is proposed for automatic recognition of CVDs using Phonocardiogram (PCG). The proposed method employs short-time Fourier transform (STFT) for time-frequency analysis of heart sounds and utilizes maxout residual network (MaxResNet) for disease recognition. The application of the skip connection technique prevents gradient vanishing and accelerates model convergence. The results from the 10-fold cross-validation demonstrate that the proposed method achieves a recognition accuracy of 99% using only 0.16M parameters. In comparison to the approach by Karhade et al., this work achieves a comparable accuracy while significantly reducing parameter usage by 87%. Moreover, the proposed method achieves real-time performance on Raspberry Pi, where it takes only 6 ms to recognize 1.1 seconds of PCG recordings.

PD06

IMPLEMENTATION OF A TILE-GRAINED PIPELINE ARCHITECTURE FOR CNN ACCELERATOR

Yi-Yuan Chen, Chung-Bin Wu, Chun-Tung Kuo
National Chung-Hsing University, Taichung

In this paper, we propose a pipelined neural network accelerator circuit. The accelerator consists of two 256MACs PE arrays, operates at a frequency of 250MHz, and achieves a theoretical throughput of 256GOPS while utilizing relatively fewer FPGA resources. We implement the hardware on the Xilinx Zynq UltraScale+ MPSoC ZCU102 platform. Experimental results demonstrate that using this architecture can reduce DRAM access by nearly 40%.

PD07

Reliable Transmission with Headerless Routing Algorithm in NoC

Kun-Chih (Jimmy) Chen[1] and Ting-En (Nick) Kao[2]
[1] Institute of Electronics, National Yang Ming Chiao Tung University
[2] Department of Computer Science and Engineering, National Sun Yat-sen University

Network-on-Chip (NoC), as a core architecture in System-on-Chip (SoC) or in multi-system applications, may have security vulnerabilities during its development process. This is because its supply chain often relies on different third-party manufacturers or component developers in the global market. This creates potential threats that may come from suspicious intellectual property (IP) developed by third parties. The impact of these malicious IPs on the system can vary greatly. Especially in a multi-core system, there may be a data center responsible for storing confidential data. If there are problems such as data leakage or loss during transmission, the credibility and functional correctness of the multi-core system are often questioned. These two problems are achieved through eavesdropping and tampering attacks by attackers. In conventional methods, most of them encrypt the transmitted data, but in order to preserve the readability of the transmission in the NoC, only the payload field in the packet is encrypted. The result of this is that the source and destination information of the packet are exposed. That is to say, if the attacker is not targeting the data in the packet itself, such protection will be completely ineffective. Therefore, in this paper, we propose a routing mechanism for NoC under encrypted transmission, named Headerless Routing Algorithm. This method allows transmission between routers without relying on header information such as the source and destination. Instead, it utilizes routing calculations on the routers to determine the hop in various directions through lookup table. This approach prevents attackers from correctly identifying the source and destination information and also ensures the integrity of the transmission process, thereby protecting the system's confidentiality and security. The experimental results primarily demonstrate that the inclusion of the headerless technique significantly reduces packet transmission errors in the presence of attacks. The transmission accuracy improves from an initial 81% error rate to a 100% correct transmission rate. This effectively proves the effectiveness and attack tolerance capabilities of this technology.

PD08

Implementation of a Resource-efficient Depthwise Separable Convolution Accelerator

Yuan-Ting Li, Chung-Bin Wu, Chih-Sheng Chiang
National Chung-Hsing University

This paper presents the implementation of a depthwise separable convolution accelerator that is built upon the lightweight approach described in the referenced materials and tailored to the proposed hardware acceleration architecture. The primary focus of this hardware design is to support depthwise-separable convolution operations, with a distinctive feature of simultaneously reducing the utilization of on-chip memory and hardware resources. This design aims to achieve the objectives of low bandwidth consumption and high resource efficiency.

PD09

A VLSI Implementation of the Tiny-YOLOv2 Algorithm

Kai-Lun Lin[1], Yuan-Ho Chen[1,2]
[1]Dept. of Electronics Engineering, Chang Gung University
[2]Dept. of Radiation Oncology, Chang Gung Memorial Hospital

With the development of artificial intelligence, CNN has been widely used in computer vision. ASIC is favored for its low power consumption, high speed, and flexibility in CNN acceleration. In this paper, we designed a target detection system based on Tiny-YOLOv2 for ASIC and used Winograd fast algorithm to accelerate the operation by reducing the use of multipliers. The Tiny-YOLOv2 architecture was also modified to achieve the results on ASIC. The research shows that compared with CPU, the design of this paper is 1.22 times faster per image and greatly reduces power consumption. Therefore, the design in this paper can be well applied to target recognition.

PD10

An Efficient CNN Accelerator with Group of Pictures Mode for Object Detection in Videos

Kuan-Hung Chen[1], Chun-Wei Su[2]*
[1] Department of Electronic Engineering, Feng Chia University
[2] Department of Electronic Engineer, PH.D. program of Electronical and Communications Engineering

Artificial Intelligence (AI) has made great progress in many computer vision tasks. However, while providing high precision and robustness, AI models often require high computation and power consumption. Previous studies have attempted to design architectures with specific models on Field Programmable Gate Array (FPGA) platforms, but the limitation of hardware resources often limits the frame rate. In this paper, we present a novel solution for object detection in videos. We proposed an efficient object detection model with an advanced group of pictures (GoP) mode technique, which were implemented on an FPGA development board. Our design can reach a 5.1 times faster frame rate and still maintain high accuracy level compared to the previous design.

PD11

High-Performance Energy-Efficient Decomposed-CNN Accelerator Chip Design with Encoder-Free Approximate Pooling Prediction

I-Hung, Lai, Jing-Xun, Hu, Jun-Fu, Chen, Huan-Ge, Xu, I-Chyn, Wey
Department of Electrical Engineering and Artificial Intelligence Research Center, Chang Gung University

This paper proposes an accelerator that combines reconfigurable architecture and max-pooling predictions for faster processing. By using approximate pooling predictions, unnecessary MAC (Multiply Accumulate) operations can be terminated early without significant loss in accuracy. Furthermore, a low-cost pooling prediction architecture is designed by reusing the decomposed convolutional structure, which enhances hardware efficiency and reduces design complexity. The accelerator can perform pooling prediction and MAC operations, increasing overall hardware usage. Compared with conventional decomposed convolutional neural network accelerators, the proposed accelerator reduces power consumption by 10.83% and computation time by 45%. Moreover, it reduces the average area by 39.1% with only a 0.74% loss in accuracy compared to traditional pooling prediction accelerators.

PD12

A Pure Sum-of-Product Composed Cosine Distance Alternative for Object Tracking

Kuan-Hung Chen, and Xuan-Fan Lin
Department of Electronic Engineering, Feng Chia University

Object tracking has gained significant progress and has been widely adopted in various fields in recent years. As the number and category of detected targets in the scene increase, enhancing the computational speed of object tracking without compromising accuracy becomes a crucial challenge. Therefore, using hardware design to implement the computational aspects of tracking is a highly advantageous approach, offering superior processing speed and capabilities for handling large-scale tracking tasks. From the literature, we observed that calculating the cosine distance to determine the similarity of appearance features is a computationally intensive task, especially when dealing with multiple tracks. Hence, we improved the cosine distance calculation by eliminating the square root operation required for normalization. This approach offers the advantage of converting the entire computation into pure sum-of-product operations, beneficial to corresponding hardware design. Based on the experimental results, almost all important evaluation measures including MOTA and ID switches in the MOT Challenge MOT16 dataset exhibit consistency with the original scheme.

PD13

A Reconfigurable Double-Group Systolic Array-based Accelerator Architecture for Deep Neural Network Training

Shu-Yen Lin, Kuan-Tsen Kuo, and Hao-Wen Chang
Department of Electrical Engineering, Yuan Ze University

In this work, the double-group systolic array-based accelerator architecture (DGSA) is proposed. To optimize the hardware utilization and energy consumption for DGSA, the schedule diagram for DGSA is optimized to reduce the latency of the neural network. DGSA is implemented in TSMC 90-nm CMOS technology with the 166MHz operating frequency. Compared with the previous design, the power consumption of DGSA is decreased by 32.6%. The area cost is 28.2% lower than the previous design. The latency in ResNet-20 reduces by 13.4% when epoch equals 100.

PD14

Gradient Sparse Update for Deep Neural Network Training Accelerator

I-Hsuan Li, and Tian-Sheuan Chang
Institute of Electronics, National Yang Ming Chiao Tung University

Training accelerators enable us to fine-tune a pre-trained model to deliver higher performance in real-world applications. However, intermediate features and losses cause large memory usage when computing gradients for backpropagation step. To address this issue, we propose a training acceleration with pruning method and gradient sparse update strategy. The pruning method combines structured pruning and unstructured pruning to achieve a sparse model to reduce data and save memory. When performing gradient sparse update strategy, we update the weights of the last few layers and skip the gradient computation of less important layers and sub-tensors. Our study enables ResNet-18 model on CIFAR-10 to achieve 91.15% accuracy by updating 6% of convolution weights and saving 95% of feature memory usage compared to dense update.

PD15

EasyNPU A Verilog HDL Generator for Systolic Array-based Neural Network Accelerator

Shu-Yen Lin, Yu-Shuo Chen, and I-Hao Chen
Department of Electrical Engineering, Yuan Ze University

In the past few years, Deep Neural Networks (DNNs) have become popular in many

applications, and the systolic array (SA) architectures are usually used to complete the DNN operations. The SA architecture is a homogeneous network with many processing element (PE) nodes. This paper proposes a Verilog HDL generator (EasyNPU) for different SA architectures. With the proposed EasyNPU, the users can quickly and flexibly create a SA architecture as long as they provide the PE architecture and configuration files. Three different PE architectures are applied to verify the functionality of EasyNPU. All generated SA designs are verified by TSMC 90-nm CMOS technologies with 100MHz operating frequencies.

PD17

TCB Convolution: Ternary-Coded Binarized Convolutions with Fixed-Point

Zong-Yu Wang, Ting-Yu Chen and Tsung-Chu Huang
Department of Electronics Engineering, National Changhua University of Education

Winograd convolution has almost become a standard for reducing computation complexity for convolutional neural network. However, its efficiency is limited due to the surge in the number of additions. In this paper, we analyze the upper bound of addition count of Winograd convolution with fixed-point filters and an estimated arithmetic weight density, and then proposed a ternary-coded binarized (TCB) convolution that can eliminate all multiplications and reduce more than 99% of addition operations.

PD18

Base on Deep Neural Network Structure Hardware Implementation for Electric Vehicles Battery State Estimation

Muh-Tian Shiue, Yi-Fong Wang, Yang-Chieh Ou, Bing-Jun Liu and Ping-Hao Liu
Department of Electrical Engineering, National Central University

This study proposes a battery model that predicts the Li-ion SOC of electric vehicles using a deep neural network (DNN) structure. Unlike most machine learning-based models, which employ large neural networks with billions of parameters to determine target functions, this study aims to achieve high accuracy rates using smaller DNN models. The study utilizes data collected from the Center for Advanced Life Cycle Engineering (CALCE) and verifies the accuracy of the DNN model using Federal Urban Driving Schedule (FUDS) [1] drive cycle datasets. The study also uses actual electric vehicle battery data from Argonne National Laboratory for testing the DNN model. This approach makes the hardware implementation of deep neural network estimation more efficient. The study develops and evaluates a series of DNN models with varying numbers of neurons and finds that increasing the number of neurons does not reduce

the error rate, but it increases costs. Additionally, the study reduces the complexity of the neural network algorithm circuit using battery measurement circuit accuracy analysis [2]. The objectives of this study are to demonstrate that DNN models can accurately estimate the SOC of Li-ion batteries with sufficient battery data, identify the optimal number of neurons for SOC estimation, and implement the testing part of the DNN model in hardware.