**Oral S06**
**Efficient Hardware Design for AI Acceleration**

| Date/Venue | 7/31(三) 15:30-17:00 [薔薇廳] |
|---|---|
| Chairs(s) | 陳坤志 /國立陽明交通大學<br>黃柏蒼 /陽明交通大學國際半導體產業學院 |

### S06.1 | 15:30-15:41

**Design and Implementation of a Hardware-Friendly Replacement for the Softmax Function**

Meng-Hsun Hsieh, Xuan-Hong Li, Chi-Yao Liang, and Juinn-Dar Huang
[1]Department of Electronics and Electrical Engineering & Institute of Electronics
[2]National Yang Ming Chiao Tung University, Hsinchu, Taiwan

The Softmax function plays an essential role in most machine learning algorithms. Conventional realization of Softmax necessitates computationally intensive exponential operations and divisions, thereby posing inevitable challenges in developing low-cost hardware implementations. This paper proposes a promising hardware-friendly substitute, Squaremax, which eliminates complex exponential operations. The function definition is straightforward and can thus be efficiently implemented in both software and hardware. Experimental results show that Squaremax consistently achieves comparable or superior accuracy over several popular models. Besides, this paper also presents an efficient hardware architecture design of Squaremax. It demands no functional units for exponential and logarithmic operations, and is even lookup table (LUT) free. It adopts a flexible 16-bit fixed-point Q format for I/O to better preserve the output precision, which leads to higher model accuracy. Moreover, it yields significant improvement in speed, area, and power, as well as achieves exceptional area and power efficiency of 664 G/mm2 and 1396 G/W in a 40nm process. Therefore, hardware-friendly Squaremax is a very promising substitute for complex Softmax in both software and hardware for deep learning applications, and the proposed hardware architecture design and efficient LUT-free implementation do achieve an outstanding improvement in speed, area, and power.

### S06.2 | 15:42-15:53

**Bit-Serial Cache: Exploiting Input Bit Vector Repetition to Accelerate Bit-Serial Inference**

Kai-Jui Chen, Yun-Chen Lo, and Ren-Shuo Liu
Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan

Bit-serial computation has demonstrated superiority in processing precision-varying DNNs by slicing multi-bit vectors into multiple single-bit vectors and computing the inner product using multiple steps

of shift-and-adds. In this paper, we identify that performing real-world DNNs inference with bit-serial computation exhibits high input bit vector locality, where up to 85.7% of non-zero input bit vectors, as well as their associated computation, are previously-seen and previously-done ones. We propose Bit-Serial Cache to transfer the identified locality into performance and energy efficiency gains. The key design strategy is to store recently-computed partial sums of input bit vectors to a cache and utilize cache accesses to replace redundant computations. In addition to the bit-serial computation architecture, we also present: 1) request clustering and 2) interleaved scheduling, to further enhance the performance and energy efficiency. Our experiments using six popular DNNs (in both 8-b and 4- b) show that Bit-Serial Cache speeds up DNN inference by up to 2.72×, 1.82×, and 4.03×, energy efficiency by 3.19×, 3.29×, and 2.82×, area efficiency by 1.35×, 1.24×, and 2.76× over state-ofthe-art Loom, DPRed Loom, and Laconic.

### S06.3 | 15:54-16:05

**Enhancing Finite State Machine Design Automation with Large Language Models and Prompt Engineering Techniques**

Qun-Kai Lin, Cheng Hsu, and Tian-Sheuan Chang
Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Large Language Models (LLMs) have attracted considerable attention in recent years due to their remarkable compatibility with Hardware Description Language (HDL) design. In this paper, we examine the performance of three major LLMs, Claude 3 Opus, ChatGPT-4, and ChatGPT-4o, in designing finite state machines (FSMs). By utilizing the instructional content provided by HDLBits, we evaluate the stability, limitations, and potential approaches for improving the success rates of these models. Furthermore, we explore the impact of using the prompt-refining method, To-do-Oriented Prompting (TOP) Patch, on the success rate of these LLM models in various FSM design scenarios. The results show that the systematic format prompt method and the novel prompt refinement method have the potential to be applied to other domains beyond HDL design automation, considering its possible integration with other prompt engineering techniques in the future.

### S06.4 | 16:06-16:17

**VESTA: A Versatile SNN-Based Transformer Accelerator with Unified PEs for Multiple Computational Layers**

Ching-Yao Chen, Meng-Chieh Chen, and Tian-Sheuan Chang
Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Spiking Neural Networks (SNNs) and transformers represent two powerful paradigms in neural computation, known for their low power consumption and ability to capture feature dependencies, respectively. However, transformer architectures typically involve multiple types of computational layers, including linear layers for MLP modules and classification heads, convolution layers for tokenizers, and dot product computations for self-attention mechanisms. These diverse operations pose significant challenges for hardware accelerator design, and to our knowledge, there is not yet a hardware solution that leverages spike-form data from SNNs for transformer architectures. In this paper, we introduce VESTA, a novel hardware design that synergizes these technologies, presenting unified Processing Elements (PEs) capable of efficiently performing all three types of computations crucial to transformer structures. VESTA uniquely benefits from the spike-form outputs of the Spike Neuron Layers [1], simplifying multiplication operations by reducing them from handling two 8-bit integers to handling one 8-bit integer and a binary spike. This reduction enables the use of multiplexers in the PE module, significantly enhancing computational efficiency while maintaining the low-power advantage of SNNs. Experimental results show that the core area of VESTA, excluding SRAM, is 0.21mm2 . It operates at 500MHz and is capable of real-time image classification at 30 fps.

### S06.5 | 16:18-16:29

## FM-P2L: An Algorithm Hardware Co-design of Fixed-Point MSBs with Power-of-2 LSBs in CNN Accelerators

Bo-Zhi Tsai, Jun-Shen Wu and Ren-Shuo Liu
Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan

Implementing Convolutional neural networks (CNNs) can be challenging due to the high requirements of memory storage and computational resources. To alleviate these deficiencies, this work presents a novel algorithm hardware co-design centered on a new number format, fixed-point MSBs with power-of-2 LSBs (FM-P2L), which can significantly reduce the requirement of computational resources of the CNN accelerators by trading negligible accuracy loss. First, we propose the novel FM-P2L number format that uses fixed-point to represent MSBs and power-of-2 to represent LSBs, which can reduce the computation complexity of multiplications. The optimal bitwidths of MSBs and LSBs are determined from our proposed algorithm with the capability to preserve accuracy. Second, we propose a novel multiplier that best matches FM-P2L with CNN accelerators, which can significantly reduce the area and power of the computational units. Finally, to evaluate the benefits of FM-P2L, we compared FM-P2L with fixed-point and low-bitwidith floating-point on a weight-stationary based systolic array accelerator with vector-vector multiplication PEs. Our evaluation results demonstrate that FM-P2L can achieve up to 44% computing power and 50% area reduction compared with fixed-point and up to 55% computing power and 65% area reduction compared with floating-point, while maintaining negligible inference accuracy loss on the state-of-the-art CNN models.

### S06.6 | 16:30-16:42

**Efficient Implementation of Transformer Inference via a Tiled-Based Architecture on an FPGA***

Ling-Chi Yang, Chi-Jui Chen, Trung Le, Bo-Cheng Lai, Scott Hauck, Shih-Chieh Hsu
[1]National Yang Ming Chiao Tung University HsinChu, Taiwan
[2]National Yang Ming Chiao Tung University HsinChu, Taiwan
[3]Electrical and Computer Engineering University of Washington Washington, USA
[4]National Yang Ming Chiao Tung University HsinChu, Taiwan
[5]Electrical and Computer Engineering University of Washington Washington
[6]Physics University of Washington Washington, USA

This paper presents an ultra-low-latency implementation of a machine learning inference algorithm called a ”Transformer”. In this research, we utilized the FlashAttention-2 algorithm on an FPGA, which is a device that has greater on-chip memory resources compared to a GPU. This involved transitioning from row-wise to tile-wise data accesses and using smaller tiles to create a more fine-grained pipeline. To address the challenge of low efficiency on dataflow architecture due to limited memory ports and data conflicts, we implemented a set of ping-pong buffers that allow interleaved access without stalling the computation of the attention mechanism. Our proposed Dataflow architecture demonstrates a significant increase in power efficiency, achieving improvements of 61% to 321% over existing FPGA-based transformer accelerators.

### S06.7 | 16:43-16:54

**MoblieNets Accelerator with Reconfigurable Dataflow and Merged-Layers Computation**

Shen-Fu Hsiao, Bo-Ching Tsai, Bao-Qi He, and Yu Kuo
Department of Computer Science and Engineering National Sun Yat-sen University Kaohsiung, Taiwan

MobileNets are popular light-weight deep neural network (DNN) models for resource-limited embedded systems. MobileNets adopt depthwise separable convolution (DSC) composed of depthwise convolution (DWC) and pointwise convolution (PWC) to reduce the complexity in standard convolutional neural network (CNN) models. We present an efficient DNN accelerator design with run-time reconfigurable dataflow with proper selection of hardware parallelism types in order to maximize hardware utilization efficiency for different convolution operations that might have significant differences of computation complexity. Furthermore, contiguous layers of operations in MobileNets are merged to reduce external memory accesses. Our design considers both data communication (between internal and external memory) and data computation time (inside the accelerator) to minimize the idle

time of processing elements (PEs). Implementation results show that the proposed merged-layers design reduces up to 31% execution time and 58% energy consumption compared with the counterpart without merged-layers.

### S06.8 | 16:55-17:06

## Fully Convolution Based Denoise Autoencoder AI Accelerator for ECG Arrhythmia Classification

Yu-Chun Wu, Shuenn-Yuh Lee, and Ju-Yi Chen
Department of Electronic Engineering, National Cheng Kung University , Tainan, Taiwan

The electrocardiogram(ECG) is an efficient indicator for arrhythmia detection. This paper presents an arrhythmia classification system for the identification of cardiovascular diseases. Classification criteria following the international standards set by the Association for the Advancement of Medical Instrumentation (AAMI), arrhythmic diseases are classified into five major categories. This paper proposes an AI algorithm combining noise reduction and disease classification. Noise reduction is based on a fully convolution-based denoise autoencoder (DAE) model which can be applied to reconstruct the clean data. The denoising performance, signal-to-noise ratio (SNR), can achieve 20.19 dB. For 24- hour ECG data, integrating filtering into the hardware, as opposed to the filter in the preprocessing stage, can save 20% of the data preprocessing time. Moreover, the prediction accuracy of the model is improved by transferring learning the encoder of DAE fuses with the R-peak interval features model. To test the robustness of the model, testing data are randomly selected in the MIT-BIH database and AHA database before training the model. Finally, the accuracy of the model can achieve 97.8% in blind testing. Furthermore, this paper provides an artificial intelligence (AI) accelerator further with Taiwan Semiconductor Manufacturing Company (TSMC) 180-nm, CMOS technology to demonstrate the proposed method, which uses the dynamic zero-gating process element to save up to 55.8% power consumption. The chip consumes 14 KB static random-access memory (SRAM), 2 FIFO data buffers, and 8 process elements, each with 3 MACs. The total power consumption is 0.533 mW and the accelerator can inference in only 55 ms.