Oral A1
Data Converters and Clocking Circuits

8 月 4 日（三）14:00-15:30
Session Chair: 陳佳宏教授 | 國立陽明交通大學 電機工程學系
林鴻文教授 | 元智大學 電機工程學系

A1.1 14:00-14:20
[Invited Paper]
Design and Implementation of a Fully Integrated Switched-Capacitor Voltage Regulator for IoT applications

Bing-Chen Wu and Tsung-Te Liu
National Taiwan University

In this paper, a fully integrated switched-capacitor dc–dc voltage regulator (SCVR) for energy-constrained IoT devices is presented. A regulation algorithm is proposed to mitigate the voltage variation caused by the loading transitions. Several circuit techniques are introduced to further enhance the SCVR current density and operating range. Experimental results show that the proposed SCVR achieves a faster load-transient response and a more efficient operation when compared to the state-of-the-art SCVR designs.

A1.2 14:20-14:30
An All-Digital Spread-Spectrum Clock Generator with Dual Phase-Rotating Compensation for Multi-Standard Devices

Hsueh-Hua Hsiang, Chun-Yao Chang and Kuo-Hsing Cheng
Department of Electrical Engineering, National Central University

This paper presents an all-digital spread-spectrum clock generator (SSCG) with dual phase-rotating compensation technique. Conventional delta-sigma modulator (DSM) based SSCG implements the spread-spectrum operation with an average divisor ratio, which causes instantaneous timing error and degrades the signal quality. To prevent the instantaneous timing error issues, the proposed all-digital SSCG adopts the phase compensation technique to realize a truly fractional frequency synthesizer for the spread-spectrum operation. In addition, the proposed phase compensation technique is designed to apply for multi-standard/frequency without electromagnetic interference (EMI) performance degradation. The experimental results show that under the operating frequency of 6 GHz, 3 GHz and 1.5 GHz, the EMI reduction is 22.44 dB, 20.48 dB and 18.83 dB, respectively. The maximum difference of EMI reduction between the operating frequency of 6 GHz and 1.5 GHz is 3.61 dB. The power consumption is 9.7 mW at the highest operating frequency.
A 10.7b 300MS/s Digital-slope-assisted SARADC in 65nm CMOS

Chun-Chieh Peng, Hou-Chung Chou, and Ta-Shun Chu
National Tsing-Hua University

A 10.7-bit 300MS/s digital-slope-assisted SARADC is proposed. The proposed SARADC requires 3 comparison cycles in each conversion. These three comparisons provide 1-bit, 5.7-bit and 4-bit resolution, respectively. The proposed hybrid ADC is fabricated in 65nm CMOS technology. The peak DNL and INL are +0.54 / -0.51 LSB and +0.62 / -0.84 LSB, respectively. The measured SNDR and SFDR at Nyquist input are 60.72dB and 70.05dB, respectively. With the power consumption of 6.2mW, the corresponding Walden FoM is 23.3 fJ/conversion-step.

A 12-ENOB 2nd-order Noise Shaping SAR ADC with Voltage-Time-Voltage Converter

Chih-Cheng Chen and Chih-Cheng Hsieh
Department of Electrical Engineering, National Tsing Hua University

This paper presents a 2nd-order noise shaping SAR ADC with the Voltage-Time-Voltage (V-T-V) converter. Using the proposed V-T-V converter, the overall design demonstrates robust performance and does not require any gain calibration for the noise transfer function, which reduces the circuit complexity and power consumption. The prototyped ADC was fabricated in 90nm CMOS process with an active area of 0.039mm2. At 1V supply and a sampling rate of 10MS/s, it achieves a SNDR of 73.8dB over 625kHz bandwidth. The resultant Schreier figure of merit (FoM) is 173.2dB, and the Walden FoM is 14.2fJ/conversion-step.

Design of Current-Sensing VCO-Based 2nd-Order CTDSM for Sensor Readout Applications

Shih-Shuo Chang, Hao-Yun Lee, and Shuenn-Yuh Lee
Department of Electrical Engineering, National Cheng-Kung University

This paper presents a current-sensing voltage-controlled oscillator (VCO)-based continuous-time delta-sigma modulator (CTDSM) for current readout application. The proposed VCO-based CTDSM can act as an amplifier-less current readout circuit which has the capability of direct quantization to the sensor current in signal acquisition. Second order noise shaping is achieved by using an extra capacitor as a current integrator and a dual-VCO structure as a phase integrator. The degeneration R-DAC technique is adopted as the feedback path to the second stage in a cascaded integrators with distributed feedback (CIFB) structure, which can dramatically reduce both the circuit complexity and
the power consumption of the feedback network. Simulation results show that the proposed current-sensing VCO-based CTDSM can achieve a signal-to-noise-and-distortion ratio (SNDR) of 90 dB in 200 Hz bandwidth with a high power efficiency of 0.9 while consuming only 59.8 μW, which is suitable for sensor applications such as the electrochemistry acquisition.

**A1.6 15:00-15:10**

**An Ultra-low Power 12bit SAR ADC with Linearly Adjustable Sampling Rate and Power Consumption for Always-on Sensors**

Chiao-Teng Chung(Jordan), Chih-Cheng Lu and Wei-Shu Rih
Industrial Technology Research Institute

Abstract—This paper presents an ultra-low power (ULP) 12-bit asynchronous successive-approximation register (SAR) analog-to-digital converter (ADC) with adjustable sampling rate and the corresponding power consumption. It can be used to optimize system power consumption by reduce sampling frequency. The 12-bit ADC was fabricated using a 180-nm CMOS technology. It consumes 828nW with a 1.2-V supply at 16kSPS. The measured SNDR and SFDR are 67.5 and 88dB, respectively. The ENOB is maintained at 10.93 bits, equivalent to a figure-of-merit of 25.8fJ/conversion step. The ADC sampling frequency can be scaled from 1MHz to 0.1Hz with ENOB rating of 10.93 bits and power consuming from 50uW to 0.01uW. The voltage supply can be scaled from 1.2-V supply to 1.8-V.

**A1.7 15:10-15:20**

**A 2.7 Gb/s Clock-Embedded, ±10% Spread-Spectrum Modulation Depth, Multiplexed-DLL-Based CDR Circuit**

Yen-Kuei Lu and Ching-Yuan Yang
Department of Electrical Engineering, National Chung Hsing University

In this paper, a spread spectrum signal of 200 kHz modulation frequency and ±10% modulation depth with a triangular-shape frequency modulation for 2.7 Gb/s data transmission is applied, and the multiplexed-delay-locked loop (M-DLL) has bandwidth of 16.3 MHz and phase margin of 78˚ for simulation. Implemented in 90-nm CMOS process, the chip active area of proposed M-DLL-based CDR is 0.118 mm². With a clock-embedded PRBS (EmPRBS), the simulated peak-to-peak jitter of retimed data and recovery clock is 6.64 ps and 4.05 ps, respectively. For the spread spectrum modulated data, the simulated maximum phase error is 0.19 UI.
A 10 GHz Dual-Loop PLL with Active Noise Cancellation Achieving 12dB Spur and 29% Noise Reduction

Yu-Sian Lu, Cheng-Lung Lee, and Wei-Zen Chen
Institute of Electronics, National Yang Ming Chiao Tung University

This paper presents a 10 GHz dual-loop PLL with active noise/interference cancellation. To achieve wide range and low noise operation, a sampling-based phase detector (SPD) is adopted for the inband noise suppression. Besides, an active noise cancellation (ANC) loop with zoom-in phase detector (zPD) is proposed to counteract the interference outside the PLL loop bandwidth. An experimental prototype has been fabricated in a TSMC 40 nm CMOS technology. By activating the ANC, the spurious tones can be reduced by 12 dB when a 260 MHz interference is injected. The integrated jitter from 1kHz to 260 MHz can be reduced from 413.7 fs to 293.21 fs, which corresponds to 29% noise reduction. It achieves adaptive noise suppression covering from 220-336 MHz, which is 60 X larger than its loop bandwidth. The proposed PLL demonstrates active noise cancellation at 3X higher frequencies compared to the prior art.
Oral D1
AI Accelerator

8 月 4 日（三）14:00-15:30
Session Chair: 張添烜教授 | 國立陽明交通大學 電子研究所

D1.1 14:00-14:20
[Invited Paper]
Real-time Super Resolution CNN Accelerator with Constant Kernel Size Winograd Convolution
Po-Wei Yen, Yu-Sheng Lin, Chia-Yang Chang, Shao-Yi Chien
National Taiwan University

This paper presents a super-resolution CNN designed for real-time hardware processing and the associated hardware architecture. Previous networks typically contain numerous layers, various kernel sizes and deconvolution layers, making it hard for hardware implementation. In this paper, we present a CNN only consisting of 3x3 convolution, replacing the deconvolution by pixel shuffling. Such regularity of kernel size enables us to employ Winograd convolution to implement the whole network. The proposed architecture achieves output resolution of 1920x1080 (FHD) at 60 fps while working at a clock frequency of 200 MHz. It also outperforms other 12K-parameter networks in image quality.

D1.2 14:20-14:40
An Energy-Efficient CycleGAN Accelerator for Edge AI Devices
Yi-Yen Hsieh¹, Yu-Chi Lee¹, Chia-Hsiang Yang¹,²
¹Graduate Institute of Electronics Engineering, National Taiwan University
²Department of Electrical Engineering, National Taiwan University

Cycle-consistent generative adversarial networks (CycleGANs) have been commonly used for unsupervised-learning applications, especially for image-to-image translation. A CycleGAN has more complex dataflow since it features two generator-discriminator pairs.

Massive external memory access also results in a long latency for both training and inference. Data structure for transposed convolution also needs to be tailored. This paper presents the first dedicated CycleGAN accelerator for energy-constrained mobile applications. The numbers of external and internal memory accesses are reduced by 98.3% and 68.3% through spatial data reuse, input feature map reuse, and local data reuse. The computational complexity is reduced by 79.4% by skipping zeros in the transposed convolutional layers. An architecture with two processing cores is proposed to improve the utilization by 2x. Designed in a 40-nm CMOS technology, the proposed CycleGAN accelerator dissipates 445 mW at 227 MHz from a 0.9-V supply. It achieves a 38x higher throughput-to-area ratio and 127x higher energy efficiency than a GPU.
D1.3 14:40-14:50

VSA-Reconfigurable-Vectorwise-Spiking-Neural

Hong-Han Lien¹, Chung-Wei Hsu², and Tian-Sheuan Chang²
¹AI Graduate Program, ²Institute of Electronics, Nat’l Chiao Tung University

Spiking neural networks (SNNs) that enable lowpower design on edge devices have recently attracted significant research. However, the temporal characteristic of SNNs causes high latency, high bandwidth and high energy consumption for the hardware. In this work, we propose a binary weight spiking model with IF-based Batch Normalization for small time steps and low hardware cost when direct training with input encoding layer and spatio-temporal back propagation (STBP). In addition, we propose a vectorwise hardware accelerator that is reconfigurable for different models, inference time steps and even supports the encoding layer to receive multi-bit input. The required memory bandwidth is further reduced by two-layer fusion mechanism. The implementation result shows competitive accuracy on the MNIST and CIFAR-10 datasets with only 8 time steps, and achieves power efficiency of 25.9 TOPS/W. Index Terms—Spiking neural network, deep learning accelerators

D1.4 14:50-15:00

Modularized Expanding Calculation Unit Design and Implement for AI Accelerator

Yu-Kuan Hsiao, Chung-Bin Wu
National Chung-Hsing University

This accelerator is a general convolutional network accelerator, suitable for object detection networks such as the Yolo series. In view of the large amount of computation required by the convolutional network, its computing speed depends on the number of Process Elements. However, there are different considerations in terms of the number of Process Elements, unit area, and power consumption in response to different usage requirements. For this reason, this paper proposes a set of Process Elements architectures that can be modularized and expanded, and the number of Process Elements required can be expanded for different needs.

D1.5 15:00-15:10

Low-Power Machine Learning Chip Design with Zero-Shot Learning

I-Chyn Wey, Shih-Wei Yang
Graduate Institute of Electrical Engineering, Chang Gung University

In this paper, a low-power machine learning classifier chip is proposed, it’s applied to human activity recognition system, research and analysis based on the demands of
wearable systems or edge-side computing. In feature extraction operations, the data segmentation method is proposed to simplify the operation. The feature processing circuit is subjected to motion analysis and simplifies standard deviation. The power of the feature vector extraction circuit can be reduced to 24% of the original algorithm, and the correct recognition rate can still be maintained above 97%. The classifier circuit of SVM in this paper subtracts the multiplication of α values and uses the improved kernel operation. In this way, the chip power can be reduced to 22.9% of the original algorithm. The RF classifier circuit of this paper fixes the depth of the decision tree. Compared with the RF design in [5], the proposed RF can reduce the power to 39% while the accuracy is maintained. In order to overcome the learning difficulties of new movement patterns, Zero-Shot learning RF design is proposed to achieve the ability that machine learning couldn’t self-learn new movement classifications in the past, while sacrificing with only 1.03% accuracy degradation.

**A Performance Aware Reconfigurable Accelerator for Quantized Light-weight Neural Network**

Hung-Yi Chiang, Chi-Wei Hsu, Yu-Guang Chen, Tsung-Han Hsieh, Jing-Yang Jou
Department of Electrical Engineering, National Central University

Convolution neural network (CNN) has been widely applied in the fields of computer vision tasks. However, standard neural networks require a lot of operations and parameters, this is a challenge for embedded devices. MobileNets, a novel CNN which adopts depthwise separable convolution to replace the standard convolution has substantially reduced operations and parameters with only limited loss in accuracy. There are mainly two different calculation methods in MobileNets, pointwise and depthwise. If the same accelerator is used to calculate these two different operations, it will be wasted due to different operation parameters. In addition, there are some methods for neural network quantization, which reduce the bit width to reduce computing energy and parameters. If the same precision hardware is used to calculate quantized operation, the maximum benefit cannot be achieved. Therefore, A novel architecture which can effectively calculate quantized MobileNets is proposed in this paper.
A Hierarchical-Based Reconfigurable Process Element Design for Quantized Convolutional Neural Networks

Chi-Wei Hsu, Hung-Yi Chiang, Yu-Guang Chen, Tsung-Han Hsieh, Jing-Yang Jou
Department of Electrical Engineering, National Central University

Convolutional Neural Network (CNN) is one of the most popular deep neural networks for visual imagery analysis. When applying CNN to various applications, data size and accuracy are the two major concerns to perform efficient and effective computations. In conventional CNN models, 32bits data are frequently used to maintain high accuracy. However, performing a bunch of 32bits multiply-and-accumulate (MAC) operations causes significant competing efforts as well as power consumptions. Therefore, recently researchers develop various methods to reduce data size and speed up calculations. Quantization is one of the techniques which reduces the number of bit of the data as well as the computational complexity at the cost of accuracy loss. To provide better computation effort and accuracy trade-off, different bit number may be applied to different layers within a CNN model. Therefore, a flexible processing element (PE) which can support operations of different bit numbers is in demand. In this paper, we propose a hierarchy-based reconfigurable processing element (PE) structure that can support 8bits x 8bits, 8bits x 4bits, 4bits x 4bits and 2bits x 2bits operations. The structure we propose applies the concept of hierarchical structure that can avoid the redundant hardware in the design. To improve the calculation speed, our 8bits x 8bits PE applies two stage pipelines. The experimental results with 90nm technology show that in 2bits x 2bits PE, we can save the area by 57.5% to 60% compared to a precision-scalable accelerator. In the 8bits x 8bits PE, the two-stage pipelines can maintain almost the same calculation speed of the 4bits x 4 bits PE.
A Compression Algorithm for Topological Quantum Error-corrected Circuits Based on Bridging

Chen-Hao Hsu¹, Wan-Hsuan Lin², Wei-Hsiang Tseng¹, Yao-Wen Chang²
¹Graduate Institute of Electronics Engineering, National Taiwan University
²Department of Electrical Engineering, National Taiwan University

Topological quantum error correction (TQEC) is promising for scalable fault-tolerant quantum computation. The required resource of a TQEC circuit can be modeled as its space-time volume of a three-dimensional geometric description. Implementing a quantum algorithm with a reasonable physical qubit number and computation time is challenging for large-scale complex problems. Therefore, it is desirable to minimize the space-time volume for large-scale TQEC circuits. Previous work proposed bridge compression, which can significantly compress a TQEC circuit, but it was performed manually. This paper presents the first automated tool that can perform bridge compression on a large-scale TQEC circuit. Our proposed algorithm applies the bridge compression technique to compactify TQEC circuits with modularization. Besides, we offer a time-ordering-aware 2.5D placement for compacting TQEC circuits and satisfying time-ordered measurement constraints. On the other hand, we suggest friend net-aware routing to effectively reduce the required routing resource under topological deformation. Compared with the state-of-the-art work, experimental results show that our proposed algorithm can averagely reduce space-time volumes by 83%.

Routing Topology and Time-Division Multiplexing Co-Optimization for Multi-FPGA Systems

Tung-Wei Lin¹, Wei-Chen Tai¹, and Iris Hui-Ru Jiang¹,²
¹Department of Electrical Engineering, National Taiwan University
²Graduate Institute of Electronics Engineering, National Taiwan University

Time-division multiplexing (TDM) is widely used to overcome bandwidth limitations and thus enhances profitability in multi-FPGA systems due to the shortage of I/O pins in an FPGA. However, multiplexed signals induce significant delays. To evaluate timing degradation, nets with similar criticalities are often grouped to form NetGroups. In this
paper, we propose a framework concerning routing topology and time-division multiplexing co-optimization for multi-FPGA systems. The proposed framework first generates high-quality topologies considering NetGroup criticalities. Then, inspired by column generation, TDM ratio assignment is solved optimally by Lagrangian relaxation. Experimental results show that our approach outperforms the top three entries of ICCAD 2019 CAD Contest. Moreover, our TDM ratio assignment algorithm can further improve the results of the top three winners to almost as good as ours.

E1.3 14:40-15:00

Thermal-aware Optical-electrical Co-routing Algorithm for Networks-on-Chip

Yu-Sheng Lu¹, Kuan-Cheng Chen¹, Yu-Ling Hsu², and Yao-Wen Chang¹,²
¹Graduate Institute of Electronics Engineering, National Taiwan University
²Department of Electrical Engineering, National Taiwan University

The optical interconnection is a promising solution for on-chip signal communication in modern system-on-chip (SoC) and heterogeneous integration designs, providing large bandwidth and high-speed transmission with low power consumption. Previous works do not simultaneously handle two main issues for on-chip optical-electrical (O-E) co-design: the thermal impact during O-E routing, and the trade-offs among power consumption, wirelength, and congestion. As a result, the thermal-induced band shift might incur transmission malfunction, the power consumption estimation is inaccurate, and thus only suboptimal results are obtained. To remedy these disadvantages, we present a thermal-aware optical-electrical routing co-design flow to minimize power consumption, thermal impact, and wirelength. Experimental results based on the ISPD 2019 contest benchmarks show that our co-design flow significantly outperforms state-of-the-art works in power consumption, thermal impact, and wirelength.

E1.4 15:00-15:10

A Complete PCB Routing Methodology with Concurrent Hierarchical Routing

Shih-Ting Lin¹, Hung-Hsiao Wang¹, Chia-Yu Kuo¹, Yolo Chen², and Yih-Lang Li¹
¹Department of Computer Science, National Yang Ming Chiao Tung University
²Wistron NeWeb Corporation

Trends in high pin density and an increasing number of routing layers complicate printed circuit board (PCB) routing, which is categorized as escape and area routing. Traditional escape routing research has focused on escape routing but has not considered the quality of area routing among chip components at the same time. In this work, we propose a complete PCB routing methodology, including simultaneous escape routing (SER), post-SER refinement, and gridless area routing. The SER completes the layer assignment of all nets and produces an escape order ensuring suitable escape and area routing on each layer. Length-matching constraints and differential pair routing are
satisfied in each stage of the routing flow. The experiment results indicate that the proposed PCB routing method can complete routings for seven commercial PCB designs, whereas the commercial PCB tool cannot complete any of them.

E1.5 15:10-15:20

Standard Cells with Inverted Inputs in 7nm Technology

Tung-Chun Wu and Rung-Bin Lin
Yuan Ze University

This paper presents a method to determine which inverted input cells (IICs) should be included into a cell library designed with a 7nm process technology. Standards cells with some inverted inputs could not only reduce the number of cells in a design but also possibly eliminate the area overhead (diffusion breaks) required for isolating the inverters and the standard cells without inverted inputs. Our experimental results show that with a 7nm standard cell library containing about 50 IICs derived from only 17 master cells could achieve up to 12% and on average 4.8% area saving per circuit without worsening timing performance.

E1.6 15:20-15:30

On-the-Fly Pin Length Reduction of 7nm Standard Cell Designs

Chin-Kai Yang and Rung-Bin Lin
Yuan Ze University

This article presents an on-the-fly pin length reduction method that minimizes pin length of a standard cell design with 7nm process technology. The method replaces standard cells of an already-routed design with their layout counterparts tailored for pin length reduction. Experimental results show that our method can on average reduce total pin length by 26.04% and total wire length (including pin length) by 7.58% without via count increase. Compared with the previous work that reported a reduction of 12.1% and 3.34% respectively, our approach is viable and effective.
Oral A2
RF/mm-Wave and Energy Harvesting Techniques
8月4日（三）15:45-17:15
Session Chair: 彭朋瑞教授 | 國立清華大學 電機工程學系
蔡政翰教授 | 國立臺灣師範大學 電機工程學系

A2.1 15:45-16:05
[Invited Paper]
Design of a Low-Jitter Area-Efficient Fractional Output Divider Using Replica-DTC-Free Calibration Technique
Chun-Yu Lin, Yu-Ting Hung, Tun-Ju Wang, and Tsung-Hsien Lin
National Taiwan University

A fractional output divider (FOD) generates an output signal whose frequency is a fractional division of the input frequency. The equivalent fractional division is realized by dithering the divide ratios. In an FOD, the quantization error due to dithering is compensated by incorporating a digital-to-time converter (DTC). Prior works adopt additional replica DTCs to facilitate the DTC gain calibration. However, extra DTCs lead to larger chip area and higher power consumption. In this work, an FOD with replica-DTC-free background calibration is proposed, and good performance is demonstrated.

A2.2 16:05-16:25
A Sub-THz Frequency Doubler with Quadrature Outputs in 40-nm CMOS
Ho-Chun Chang, Hong-Shen Chen, Yung-Shun Huang, and Jenny Yi-Chun Liu
Institute of Electronics Engineering, National Tsing Hua University

A sub-THz frequency doubler with quadrature outputs is demonstrated in a 40-nm CMOS technology. The proposed scheme facilitates a local frequency generator operating near the maximum oscillation frequency of the technology. An impedance transformation network provides high impedance at the 2nd harmonic frequency, rejects the desired signal from feeding back and enhances the fundamental rejection ratio (FRR). The quadrature outputs can be adopted in quadrature phase modulation transceivers. The proposed doubler shows a conversion gain of -11.3 dB and -12.3 dB at the quadrature outputs, with a frequency range from 170 GHz to 184 GHz. The maximum output power is -8 dBm. The FRR better than 40 dB guarantees the output signal quality.
A 28-GHz CMOS Transmitter for Silicon Photonics mmWave-over-Fiber 5G Communications

Wen-Chieh Huang, Hong-Shen Chen, Yu-Wei Cheng, Chia Chan Tsai, Yung-Shun Huang, Jenny Yi-Chun Liu
Institute of Electronics Engineering, National Tsing Hua University

A 28-GHz transmitter is demonstrated in a standard 90-nm CMOS technology for millimeter-wave-over-fiber 5G mobile communications. The first stage of the transmitter is a low-noise amplifier performing good in-band noise compared to traditional transimpedance amplifiers. A power amplifier as the output stage provides sufficient output power for indoor wireless transmission. A 2.2 Gb/s on-off keying modulation signal is tested in this work with an eye diagram. In additional to electrical measurement, the transmitter is packaged with a photodetector and verified with an optical signal source to demonstrate the optical to electric signal transmission path. The proposed transmitter exhibits high gain of 36.6 dB, an output power of over 16 dBm, and a low noise figure performance of 4.5 dB, facilitating a high data rate connection for 5G communication systems.

A Wireline Termination Embedded Energy Harvesting System with 300uW Extracted

Yu-Hong Yang and Tai-Cheng Lee
Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University

A 1-Gb/s wireline energy harvesting system with 3-phase rotating and 5 time-interleaving switched-capacitor is proposed to achieve the low power requirement. By Z0-mimicking at the termination and energy harvesting from the wireline incoming signal, the energy harvesting system extracts 300uW net power from termination and provides 1.42mW throughput by mixing the power from the supplier, which is fabricated in a 40-nm CMOS technology.

A W-Band Low Insertion Loss CMOS Wilkinson Power Divider

Ruo-Hsuan Gao and Chien-Nan Kuo
Institute of Electronics, National Yang Ming Chiao Tung University

An on-chip Wilkinson power divider is realized in 90 nm CMOS for operation in the W-band frequency. Being different from the conventional design, the proposed divider is miniaturized by using transmission lines of one-eighth wavelength with U-shape folded
ground. The core circuit occupies only 0.03 mm². The measured insertion loss less than 1.7 dB over the frequency range from 70 GHz to 110 GHz.

**A2.6 16:55-17:05**

**A 10-11 GHz FMCW Chirp Generator with Background Noise Cancellation**

Kai-Teng Lo, Pei-En Li and Wei-Zen Chen  
Institute of Electronics, National Yang Ming Chiao Tung University

This paper presents a 10-11 GHz delta sigma-PLL based chirp generator for frequency modulated continuous waveform (FMCW) radars. A DTC-free noise-cancellation sampling phase detector (NC-SPD), which consists of a novel sampling PD and a DAC, is proposed to compensate the quantization noise during frequency sweeping. To reduce the frequency error of the FMCW signal, the DAC is background calibrated to continuously track the NC-SPD gain. By the proposed techniques, the FM error is reduced by more than 32%. The FMCW modulator generates a 10-11 GHz chirp signal with a frequency error ratio of 0.017-0.028 (%). The experimental prototype is implemented in TSMC 40 nm CMOS process and consumes 27.67 mW from a 1.2 V supply.

**A2.7 17:05-17:15**

**An Active Inductor for IoT SoC Applications**

Yeong-Lin Lai and Chun-Yi Zheng  
Department of Mechatronics Engineering, National Changhua University of Education

In this paper, we propose a gyrator-based active inductor with a folded circuit structure for Internet of Things (IoT) system-on-chip (SoC) applications. Metal oxide semiconductor field effect transistors (MOSFETs) form gyrators and capacitors, which produce inductive properties. This creates an active inductor based on a gyrator. Negative resistance compensation technology is used to compensate the losses caused by transistors and bias circuits. Inductance values from 14.81 to 50.44 nH are shown. Folded active inductors achieve the characteristics of high quality factor and wide inductance range.
D2.1  15:45-16:05

[Invited Paper]
Low-Power and Efficient MLP Accelerator Supporting Structured Pruning, Sparse Activations and Asymmetric Quantization for Edge Computing

Wei-Chen Lin, Ya-Chu Chang, and Juinn-Dar Huang
National Yang Ming Chiao Tung University

Multilayer perceptron (MLP) is one of the most popular neural network architectures broadly used for regression, classification, recommendation systems, and natural language processing (NLP) today. In this paper, we propose an efficient and low-power MLP accelerator architecture optimized for edge computing. The accelerator has three key features. First, it aligns with a novel structured weight pruning algorithm that merely needs minimal hardware support. Second, it takes advantage of activation sparsity for power minimization. Third, it supports asymmetric quantization on both weights and activations to boost the model accuracy especially when those values are in low-precision formats. Furthermore, the number of PEs is determined based on the available external memory bandwidth to ensure the high PE utilization, which avoids area and energy wastes. Experiment results show that the proposed MLP accelerator with only 8 MACs operates at 1.6GHz using the TSMC 40nm technology, delivers 899GOPS equivalent computing power after structured weight pruning on a well-known image classification model, and achieves an equivalent energy efficiency of 9.7TOPS/W, while the model accuracy loss is less than 0.3% with the help of asymmetric quantization.

D2.2  16:05-16:25

An AI-edge Real-Time Affective Computing Platform with Embedded AI SoC Design

Wei-Chih Li, Cheng-Jie Yang, Bo-Ting Liu, and Wai-Chi Fang
Department of Electronics Engineering & Institute of Electronics, National Yang Ming Chiao Tung University

This paper proposes an AI-edge real-time affective computing platform that integrates an AI System-on-Chip (SoC) design and multimodal signal processing systems composed of electroencephalogram (EEG), electrocardiogram (ECG), and photoplethysmogram (PPG) signals. To extract the emotional features of the EEG, ECG, and PPG signals, we used a short-time Fourier transform (STFT) for the EEG signal and direct extraction using the raw
signals for the ECG and PPG signals. The long-term recurrent convolution networks (LRCN) classifier was implemented in an AI SoC design and divided emotions into three classes: happy, angry, and sad. The proposed LRCN classifier reached an average accuracy of 77.41% for subject-independent evaluation. The platform consists of wearable physiological sensors and multimodal signal processors integrated with the LRCN SoC design. The core area and total power consumption of the LRCN chip were 1.23 x 1.24 mm² and 53.2 mW, respectively. The on-chip training processing time and real-time classification processing time are 5.5 us and 1.9 us per sample. The proposed platform displays the classification results of emotion calculation on the graphical user interface (GUI) every one second for real-time emotion monitoring.

D2.3 16:25-16:35
A ResNet-GRU-Based Lip Reading Model for Day-to-Day Conversations in Chinese Mandarin

Wen-Jie Cai, Shanq-Jang Ruan
Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology

We propose a lip reading recognition model to predict daily Mandarin conversation and collect a new Daily Mandarin Conversation Lip Reading (DMCLR) dataset, consisting of 1,000 videos from 100 daily conversations spoken by 10 speakers. Our model consists of a spatiotemporal convolution layer, a ResNet-18 network, and a 3-layer Bi-GRU network. This model is able to reach 94% of accuracy in the DMCLR dataset. We also prove that the most advanced model in Mandarin is effective once the number of the target sentences is restricted. Such advancement makes it possible for lip reading applications to become practical in real life. Additionally, We are able to improve accuracy from 85.3% to 87.5% and 41.4% to 51.7% on the two most enormous public available lip reading datasets, LRW (English) and LRW-1000 (Mandarin), respectively. The results show that our method achieves state-of-the-art performance on these two challenging datasets.

D2.4 16:35-16:45
3D LiDAR Decoder Design

Sheng-Bi Wang, You-Sheng Xiao, Wei-Zhe Yan, Min-Hua Lu, Hao-An Hsieh, Yu-Cheng Fan
Department of Electronic Engineering, National Taipei University of Technology

In recent years, object detection has been applied in many fields. In one of the most well-known fields, autonomous driving system, the key skill lies in the breakthrough of LiDAR (Light Detection and Ranging). As the core component of object detection, LiDAR can instantly gather a large amount of three-dimensional (3D) depth information with high accuracy, which allows users to rebuild three-dimensional environment. Through the information provided by LiDAR, autonomous driving system can identify all kinds of
circumstances nearby and to choose a much safer route. The mainstream of applying LiDAR skill in autonomous system is the product of Velodyne, HDL-64. Based on HDL-64, this paper aims to decode the data packets which transform the information of original packet into the point cloud data of X, Y, Z. This will allow users to make use of decoding information in more research and applications such as rebuilding three-dimensional environment, making deeper object detection and object classification.

D2.5 16:45-16:55

Tensor Compressive Sensing Processor for Single-Pixel Terahertz Imaging Systems

Wei-Chieh Wang¹, Shang-Hua Yang² and Yuan-Hao Huang¹
¹Institute of Communications Engineering and Department of Electrical Engineering
²Institute of Electronics Engineering, National Tsing-Hua University,

Terahertz (THz) imaging recently becomes an emerging technology for object detection in security applications or defect detection of micro circuitry due to its high-penetration and low-power properties. Because THz detectors are very expensive, single-pixel camera based on compressed sensing technique was proposed to realize the cost-efficient imaging system. The primary challenge of a single-pixel THz camera is the extremely high computational complexity of the image reconstruction algorithm. This paper proposes a tensor-based compressive sensing model and low-complexity image reconstruction algorithm and hardware architecture for THz single-pixel imaging systems.

D2.6 16:55-17:05

Applied to Hierarchical Navigation Based on Density Detection and Pedestrian Flow Detection

Wei-Zhi Zhang and Wei-Liang Lin
National Chung Hsing University

This article is based on the YOLOv4 object detection method, and builds a crowd density detector suitable for this research content through a rigorous crowd training data set, so as to obtain density information. It also uses the FairMOT-based object tracking method to identify and obtain through each frame information on the direction of people flow for each object.

Hierarchical navigation is proposed, which is divided into wide area and narrow area. The wide area considers the density and the flow of people as the path planning, while the narrow area tends to avoid obstacles in the field, and implement this in the Unity virtual environment, and implement hierarchical navigation in the National Chung Hsing University internship store.
Memory-aware Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Underwater Image Enhancement

Meng-Hua Li, Shiann-Rong Kuang, Shao-Heng Hung, Jian-Zhi Chen, and Xiang Yu Chen
Department of Computer Science and Engineering, National Sun Yat-sen University

Contrast Limited Adaptive Histogram Equalization (CLAHE) is a widely adopted technique to improve the visibility of images. However, high memory usage is a big challenge for applying CLAHE to the real-time image processing. In this paper, a memory-aware implementation of CLAHE is proposed to enhance the contrast of underwater turbid images in real time. In the proposed architecture, the computation of CLAHE is merged and the on-chip memory is heavy shared to significantly reduce the memory demand while maintaining the high performance. Experimental results show that the proposed design achieves the same or higher throughput and requires much less on-chip memory than previous designs while maintaining the comparable quality of resulted underwater images.
Oral E2

DFT and Test for Emerging Technologies

8月4日（三）15:45-17:15
Session Chair: 謝東佑教授 | 國立中山大學 電機工程學系

E2.1 15:45-16:05

[Invited Paper]

Prediction of Test Pattern Count and Test Data Volume for Scan Architectures Under Different Input Channel Configurations

Fong-Jyun Tsai¹, Chong-Siao Ye¹, Kuen-Jong Lee¹, Shi-Xuan Zheng¹, Yu Huang², Wu-Tung Cheng², Mark Kassab², Janusz Rajski², Chen Wang², Justyna Zawada², Sudhakar M. Reddy³

¹Department of Electrical Engineering, National Cheng Kung University
²Mentor, A Siemens Business Wilsonville, OR, USA
³Dept. of Electrical & Computer Eng. University of Iowa, Iowa City, USA

As the complexity of industrial integrated circuits continue to increase rapidly, test data compression has now become a de facto technology for large designs to reduce the overall test cost. During the design for test (DFT) planning, it is critical to understand the impact of using different numbers of input/output test channels on test coverage, test cycles, and test data volume. In this paper, two approaches to predict the test pattern counts and test data volumes with different input channel counts are presented, one with the compression tool able to generate channel-scaling patterns and the other without this capability. The results can be used to determine the scan test configuration that results in the smallest or near smallest test data volume. Experiments on industrial circuits show that the average error rates of pattern count prediction for most circuits are less than 10% for both approaches. The error rates of the predicted smallest data volumes are all less than 3.5%. The total ATPG run time can be reduced by a factor of more than 10X compared to the currently used trial-and-error approach.

E2.2 16:05-16:25

A Memory Built-In Peer-Repair Architecture for Mesh-Connected Processor Array

Chih-Hsuan Tung¹, Pai-Yu Tan¹, and Cheng-Wen Wu¹²

¹Department of Electrical Engineering, National Tsing Hua University
²Department of Electrical Engineering, National Cheng Kung University

Driven by artificial intelligence (AI) applications in recent years, the mesh-connected processor array has become a popular high-performance AI computing architecture. In addition to processor cores, however, it is well known that memories in the system play a critical role in performance and power consumption, so normally the embedded
memories occupy more than 2/3 of the overall area of these chips, which in turn dominate the yield and reliability of the computing chips. Memory built-in self-repair (MBISR) has been considered a feasible solution for test and repair of embedded memories in the respective processor cores. However, so far MBISR does not take advantage of the regular topology of the mesh-connected processor array. In this paper, we propose a memory built-in peer-repair (MBIPR) architecture that enables the processor core to share its spare memories with the neighboring cores in the mesh-connected array. Experimental results show that the repair rate of the proposed MBIPR outperforms that of the original MBISR. Compared with MBISR, MBIPR increases the spare utilization and the lifetime by 2.1 to 8 times, with only about 0.2-0.9% higher area overhead.

**E2.3**

**Determine Optimum Test Compression Configuration via Pattern Count Estimations for a Wide Range of Input/Output Channel Combinations**

Shi-Xuan Zheng, Chong-Siao Ye, and Kuen-Jong Lee
Department of Electrical Engineering, National Cheng Kung University

On-chip test compression hardware has become a pragmatic technology to cut down the overall test costs. Determining the input and output channel counts of test compression hardware that results in minimum test data volume is a critical issue to reduce test costs. In this paper, efficient methods to estimate test pattern counts for a large range of input/output counts are developed. These methods require only two ATPG runs for the estimations of up to hundreds of configurations. The results can then be utilized to estimate the test data volume for each input/output configuration. The configuration with the estimated lowest test data volume thus can be determined. The pattern count results of variable configurations for a design can also be used to determine the best suitable configuration when the design is to be embedded in an SoC system.

**E2.4**

**CORONA: A k-COnnected RObust Interconnection Network Generation Algorithm**

Hsin-I Wu, Ren-Song Tsay, Fong-Yuan Chang, Kuan-Fu Lu
Department of Computer Science, National Tsing Hua University

In the paper, we propose a k-connected robust interconnection network (k-RN) for open fault tolerance and design a k-RN generation algorithm to handle large interconnected networks. In a k-RN, any vertex pair has at least k disjoint connections and can tolerate k-1 opens. To realize k-RN in current large networks, we design an effective O(n log n)-time algorithm, Corona, with a novel approach for total connection length reduction. Experiments in IC design application show that Corona can practically handle large nets with 100k pins in a few minutes each. The quality in terms of total wire length of Corona is within 1.6% on average to an O(n^3)-time approximation algorithm, and within 4.7% to exact optimum solutions.
Fault Models and Test Algorithms for Memristor-Based Spiking Neural Network

Hsueh-Hung Cheng¹, Kuan-Wei Hou¹, and Cheng-Wen Wu¹,²
¹Department of Electrical Engineering, National Tsing Hua University
²Department of Electrical Engineering, National Cheng Kung University

Memristor-based computing architectures for deep neural network (DNN) and spiking neural network (SNN) have been explored in recent years, for improving the performance and energy efficiency of AI computing. However, memristor cells are susceptible to manufacturing defects and process variations, which will result in various faults and failures in the system. In this paper, we propose a test method for memristor-based SNN hardware. We first perform circuit-level simulations of the SNN hardware, taking process variations into account. All feasible open and short defects and the memristor program fault are respectively injected and compared with the fault-free circuit. From the analysis result, we conclude that the injected defects and faults can be covered by two simple fault models, i.e., the Slow Integration Fault (SIF) and Fast Integration Fault (FIF). We also develop a fault simulator that can be used to simulate March-like test algorithms for the proposed defects and fault models, based on the memristor-based SNN circuit. We propose a March-like test algorithm, called March-SNN. Testing is done by applying the specified input test patterns (input spikes) according to the algorithm and observing the output test responses (output spikes), with existing operations of the SNN circuit to reduce additional hardware overhead. Experimental results show that March-SNN covers 100% of the proposed defects and fault models.

An Error-Tolerability Enhancement Method of Videos for Object Detection Applications

Tong-Yu Hsieh and Jun-Tsung Wu
Department of Electrical Engineering, National Sun Yat-sen University

This paper presents the first one error-tolerability enhancement method of videos for object detection applications. Different from the previous work that focuses on the human visual system’s perception to video errors, we take machine’s tolerability of errors into consideration in this work. The development of a run-time video re-encoding method for object detection applications and its effectiveness in error-tolerability enhancement is investigated. To the best of our knowledge, this issue has never been addressed in the literature. In our work we employ YOLOv3 as a case study of object detection. Errors are injected into a benchmark video to generate more than 100,000 erroneous videos, and then YOLOv3 is employed to define the quality (acceptability) of these videos. The results demonstrate that error-tolerability of videos can be significantly enhanced by the proposed method with higher efficiency for object detection (computer vision). Better video re-encoding efficiency (smaller video size) is achieved and lower implementation cost is incurred. These are mainly because that the error-tolerability of machines is larger than that of humans.
Oral A3
Low Power Oscillators and Sensors

8 月 5 日（四）13:45-14:45
Session Chair: 鄭光偉教授 | 國立成功大學 電機工程學系

A3.1 13:45-14:05
A ±20-ppm -50°C-105°C 1-μA 32.768-kHz Clock Generator with an AT-Cut-Crystal-Assisted Background Calibration

Chun-Yu Lin and Tsung-Hsien Lin
Graduate Institute of Electronics Engineering and Department of Electrical Engineering, National Taiwan University

This paper describes a 32.768kHz frequency source with AT-Cut crystal calibration technique, which consumes only 0.364mm2 and 1μA power consumption. With the proposed AT-cut XO-assisted background calibration and compensation technique at extreme temperatures, the implemented 32.768kHz clock generator achieves low-power consumption with good frequency stability (<20ppm) over a wide temperature range (-50~105 degree C) without complex temperature trimming.

A3.2 14:05-14:15
A 1.17-nW/kHz Frequency-Locked On-Chip Wake-Up Timer in 0.18-μm CMOS

Bo-Sheng Li, Sheng-Kai Chang, Zhi-Ting Tsai, and Kuang-Wei Cheng
National Cheng Kung University

This work presents a resistive frequency-locked loop on-chip oscillator with a double chopper stabilization technique to improve the temperature stability and long-term stability. A prototype device is fabricated in 0.18-μm CMOS technology and exhibits a 24.7 ppm/°C temperature stability and 2.73 ppm long-term stability while consuming just 293 nW under an oscillation frequency of 250 kHz. The use of the double chopper stabilization technique effectively eliminates the TC-sensitive non-idealities, including the current mismatch and offset voltage of the amplifier. Moreover, the low-frequency flicker noise is also mitigated; resulting in a 16X improvement in the long-term stability.
A Low Power Reconfigurable Hybrid Frequency-Locked Loop For On-Chip Detection of the DNA Hybridization

Chao-Yuan Chen, Kuang-Hung Hsu and Tsung-Heng Tsai
Department of Electrical Engineering and Advanced Institute of Manufacturing with High-tech Innovation, National Chung Cheng University

This paper presents a high area and power efficiency Reconfigurable MEMS Poly Phase Filter Hybrid Frequency-Locked Loop (RMPPF-HFLL) for on-chip DNA impedance detection. The Proposed RMPPF converts the DNA impedance into a time-domain delay through the RC charging and discharging. The readout circuit can detect this delay by the high precision dynamic comparator and control the digital voltage control oscillator (DVCO) to track by the proposed hybrid control algorithm. The 6bits digital codes will be analyzed through the SAR algorithm with digital mode and the residual voltage is tracked continuously through analog mode. For the purpose of wide DNA impedance sensing range, this paper presents a reconfigurable PPF structure which can dynamically adjust the resolution and the range of the capacitance by switching the resistor array of the PPF without changing the specifications of the oscillator. Fabricated in 0.18um standard MEMS CMOS process, this paper demonstrates the closed-loop time-domain DNA sensing scheme. The front-end PPF sensing method can reduce the low-frequency noise introduced by electrode without the need for additional high speed chopper switches. Besides, the reconfigurable PPF achieve up to 4x dynamic sensing range which can apply for different DNA sensing environment. Finally, the novel Hybrid frequency-locked loop combined with PPF analog front-end achieve less than 1ms settling time, the total area only occupy 0.685um^2 and only consumes 14.46uW which is suited the biomedical application.

A 950-pW, 39-pJ/Conversion Leakage-Based Temperature-to-Digital Converter With 43mk Resolution

Cheng-Ze Shao, Ying-Jie Huang, and Yu-Te Liao
ECE Department, National Yang Ming Chiao Tung University

This paper presents a power-efficient temperature-to-digital convertor using a leakage-based ring oscillator. The reversely-biased transistors are added between the supply rails and the core circuit to limit power consumption and enhance temperature-to-current conversion gain. A design of temperature-insensitive sampling clock signal results in better conversion linearity/accuracy. The design was fabricated in a 180-nm CMOS process with an active chip area of 0.04 mm^2. The design can achieve a resolution of 43 mK at 20 oC in 100 measurements and a temperature inaccuracy of -1.6/+2.1 oC over a temperature range of 20-80 oC. At a 0.55 V supply, the power consumption of the whole system is 950 pW, the conversion time is 40.8 ms, and the resulted energy efficiency is 39 pJ/conversion.
A 180-nm CMOS 80-MHz Low-Voltage MEMS Oscillator Using Frequency-Locked Loop

Tsung-Ying Chen and Ching-Yuan Yung
National Chung Hsing University

This work presents a MEMS oscillator using the frequency-locked loop circuit with an auto-zeroing technology. The temperature stability for the oscillator is about 97.49 ppm/°C from 0°C to 100°C and line sensitivity is 47.18 %/V from 0.81V to 1V. The on-chip MEMS oscillator can generate a corresponding frequency with different gravitational acceleration. The proposed auto-zeroing technique can effectively improve the variation of output frequency for PVT effects.
Oral D3

Emerging AI Algorithm and System

8 月 5 日（四）13:45-14:45
Session Chair: 劉志尉教授 | 國立陽明交通大學 電子工程研究所

D3.1 13:45-14:05

[Invited Paper]
Fed-ELM: Extreme Learning Machine System for Federated Learning

Yi-Ta Chen, Yu-Chuan Chuang, and An-Yeu (Andy) Wu
Graduate Institute of Electronics Engineering, National Taiwan University

Federated learning (FL) is a privacy-preserving learning framework, which collaboratively learns a centralized model across edge devices. Each device trains an independent model with its local dataset and only uploads model parameters to mitigate privacy concerns. However, most FL works focus on deep neural networks (DNNs), whose intensive computation hinders FL from the practical realization on resource-limited edge devices. In this paper, we exploit the high energy efficiency properties of extreme learning machine (ELM) and propose a federated ELM system (Fed-ELM) to meet the demand for FL scenarios. Despite non-independent and identically distributed (non-IID) and imbalanced data across edge devices, Fed-ELM still achieves comparable performance to a centralized ELM trained with IID and balanced data. Compared with DNNs under the same transmission cost, the proposed Fed-ELM outperforms FederatedAveraging NN (Fed-NN) by 2.3% accuracy, and the fine-tuning process is 7%-33% less time-consuming.

D3.2 14:05-14:15

An Efficient Frequency Domain Training Framework for Robust Convolutional Neural Networks

Ming-Cheng Chen, Yo-Charn Li, Chih-Wei Liu
Department of Electronics Engineering, National Yang Ming Chiao Tung University

In recent years, with the rapid development of deep learning, many theories and techniques have been proposed. Among them, Convolutional Neural Networks (CNNs) have the most extensive applications. However, with the gradual increase in network complexity, high-dimensional convolution operations have higher computational complexity, resulting in low network training efficiency. Therefore, from the perspective of accelerating network training, this paper proposes a training framework that transforms convolutional neural networks into the frequency domain. This training framework improves the FFT-based acceleration method by avoiding multiple transformations between the spatial domain and the frequency domain, thereby saving a lot of computations. In addition, in terms of network architecture design, we analyze and
integrate the spatial domain concept proposed in [1]–[3] into the frequency-domain network architecture. Then with the characteristics of spectral value distribution, the training method is improved to enhance the robustness of the model. According to the experimental results, the computational complexity required to train the classic CNN model AlexNet in this framework saves nearly 51.9%-91.8% of the computation when comparing with the FFT-based acceleration method. The amount of memory and parameter storage has also been significantly reduced by 56.2%. Under a certain degree of noise interference, the model trained by the frequency domain framework is more robust than the traditional spatial-domain CNN.

**D3.3**  
*14:15-14:25*  
**Automated LPI Radar Classification using Fractional Fourier Transform**  
Jia-Yu Wu, Meng-Chiao Wu, Chih-Wei Liu  
*Department of Electronics Engineering, National Yang Ming Chiao Tung University*

Radar plays a key component for enemy detection in the military application. In addition to tracing surrounding systems, our side must have an ability to detect and classify the radar signal from the enemy, especially in the electronic warfare (EW). Due to the improvement of the digital modulation in radar waveforms, traditional pulse radars have been rapidly replaced by Low Probability of Intercept (LPI) radars. Thus, a fast and efficient LPI radar waveform intercept receiver under ultra-low signal to noise ratio (SNR) is the important function in the future. In this work, we will focus on the classification of twelve common LPI radar waveforms. Different from the previous works using the Wigner-Ville Distribution (WVD) or Choi-Williams Distribution (CWD) as the spectrum analysis, this work chose Fractional Fourier Transform (FrFT). The feature extraction based on FrFT analysis is proposed. Moreover, we also use deep learning model as our classifier which consists of 1D Convolution Neural Network (1D CNN) and Long short-term memory (LSTM). According to simulation results, the classification accuracy of the proposed system is 94.5% at the SNR of -10 dB. Compared with previous works, our algorithm has 4 dB improvement on the same accuracy.

**D3.4**  
*14:25-14:35*  
**Evaluation of Feature Labeling for ML-based AMD Detector**  
Yao-Wen Yu¹, Cheng-Hung Lin¹,², Cheng-Kai Lu³, Jia-Kang Wang¹,⁴, Tzu-Lun Huang¹,⁴  
¹ Department of Electrical Engineering, Yuan Ze University  
² Biomedical Engineering Research Center, Yuan Ze University  
³ Department of Electrical and Electronic Engineering, Universiti Teknologi PETRONAS  
⁴ Department of Ophthalmology, Far Eastern Memorial Hospital

Today’s automated detectors of Age-related macular degeneration (AMD) on optical coherence tomography (OCT) volumes with the support vector machine (SVM) are widely researched in the field of ophthalmology. However, the OCT volumes are three-dimensional (3D) data information composed of several two-dimensional OCT
images capture from subject eyeballs. Therefore, two feature labeling methods, the slice-chain labeling method and the slice-threshold labeling method, are investigated for the 3D OCT volume in this paper. The two labeling methods are evaluated in this paper because they influence detection accuracy for the SVM-based AMD automated detector and the number of features stored in the memory of SVM hardware. From the experiment results, the slice-threshold labeling method achieves a high detection accuracy of 96.36% with 35.34% features saved in the memory of SVM hardware compared with the slice-threshold labeling method.

**D3.5 14:35-14:45**

**Clothing Classifier with Two-Stage Based Lite CNN Model for Intelligent Autonomous Movers**

Hua-Luen Chen¹, Chi-Chun Lai¹, Jie-Min Lin¹, Kuan-Hung Chen², Yin-Tsung Hwang¹, and Chih-Peng Fan¹

¹Department of Electrical Engineering, National Chung Hsing University
²Department of Electronic Engineering, Feng Chia University

For an intelligent autonomous mover, in addition to providing the functions of object detection and collision avoidance, clothing classification has acquired more attentions recently due to its important role in user-friendly use in surrounding-crowds environments. In this study, a two-stage processing based lite convolutional neural network (CNN) architecture is developed for clothes classification, where the pedestrians' boxes inferred by the first stage detection are used as the inputs of the clothes classification by the second stage process. By the proposed second-stage inference model, the Top-1 recognition accuracy of clothes classification is closed to 96% with the image datasets collected in supermarket. Compared with the previously related designs, the proposed approach performs better recognition accuracy of clothes classification for the application of autonomous movers.
Low-Complexity MSE-based Alternating Optimization Design for Intelligent Reflecting Surface (IRS)-Assisted MIMO Communication

Chi-Wei Chen, Sin-Sheng Wong, Chieh-Fang Teng, and An-Yeu Wu
Graduate Institute of Electrical Engineering, National Taiwan University

Recently, intelligent reflecting surface (IRS) has emerged as a promising cost-efficient technology to enhance communication performance by reconfiguring and adding diversity to the propagation environment. However, the IRS realized by the unit-modulus phase shifters makes the optimization problem non-convex, which makes the prior works suffer from high complexity. In this article, we consider a single-user IRS-assisted multiple-input-multiple-output (MIMO) system, aiming for maximizing spectral efficiency. To deal with the complicated non-convex problem, we exploit the relationship between the mean-square-error (MSE) and the spectral efficiency, and reformulate the MSE as the surrogate objective. By deriving a closed-form non-coupling solution, we propose a low-complexity and low-latency MSE-based alternating optimization (MSE-AO). In addition, it can be executed in parallel for further speedup of the IRS computation. Simulation results show that the proposed MSE-AO can reduce 8.2 times multiplications while with only 1% performance degradation compared with prior state-of-the-art algorithm.
Low-energy, high hardware-efficiency fixed-width multipliers for approximate computing have drawn a lot of research interests. This paper utilized the evolutionary algorithms (EA) to automatically generate the hardware-efficient compensation circuit for the fixed-width modified Booth multiplier (FWBM). The proposed evolutionary compensation circuit (EvCC) directly evolves through the truncation part to maximize its accuracy. Take an 8-bit FWBM as an example. The proposed EvCC only has one 2-input XOR gate with an SNR of 34.5, which is much greater than that of the direct truncation method. This paper discusses the potential and feasibility of the proposed automatically-generated EvCC, which can provide the low hardware-cost compensation circuit architecture for FWBM without the exhaustive simulation.

A Geometrical Mean Decomposition Processor design for Wireless MIMO Hybrid Precoding

Kuan-Ting Chen, Yu-Hsian Lin and Yin-Tsung Hwang
Department of Electrical Engineering, National Chung Hsing University

In a closed-loop wireless MIMO communication system, precoding is performed in the transmitting site to pre-compensate channel fading and to alleviate the design complexity of the receiver. Among various linear precoding schemes, singular value decomposition (SVD) is the most popular one. However, for massive MIMO systems, a hybrid approach combining digital precoding and analog beamforming is considered a more cost effective approach. In this paper, a geometrical mean decomposition (GMD) pre-processing module is developed on top of the SVD based hybrid precoding design to enhance the BER performance. A hierarchical computing scheme is developed to reduce the algorithm complexity. In particular, all basic computations are converted into Givens rotations, which can be implemented efficiently in hardware by using a CORDIC scheme. The algorithm to hardware mapping is conducted and resultant architecture consists of two CORDIC processor clusters supporting a maximum degree of computing parallelism. In addition, lifetime analysis technique is applied to minimize the required number of CORDIC processors. The design is realized in a TSMC 40nm CLN40G process technology and can operate up to 333MHz. It suggests a roughly 1.6dB performance enhancement over the SVD based design. The GMD pre-processing module consumes about 127k logic gates and accounts for 25% circuit complexity of the overall hybrid precoding processor design.
A Two-Dimensional FFT/IFFT Processor with Customized Floating-Point Datapath for High-Resolution Synthetic Aperture Radar

Hung-Yuan Chin\(^1\), Pei-Yun Tsai\(^1\), and Szy-Yuan Lee\(^2\)
\(^1\)Department of Electrical Engineering, National Central University
\(^2\)National Space Program Office, National Applied Research Laboratory

A 2D-FFT/IFFT processor is implemented to support 8192-, 16384-, and 32768-point range FFT/IFFT and 8192-point azimuth FFT/IFFT. Azimuth decomposition is adopted to facilitate burst access to external memory. Besides, normal-order input for azimuth FFT and bit-reversed input for azimuth IFFT are designed to reduce the latency and the requirement of re-ordering buffers. We also derive the equation to describe the control logic for look-up tables of twiddle factors in normal-order azimuth FFT and bit-reversed order azimuth IFFT so as to decrease ROM tables. A customized floating-point data-path is utilized to cover the large dynamic range of radar signals in different transformed domains. With 10-bit mantissa and 6-bit exponent, the design 2D-FFT/IFFT processor achieves SQNR of more than 48dB for one transformation and about 38 dB for successive 2D-FFT/IFFT operations. The maximum operating frequency is 111MHz of our 2D-FFT/IFFT processor realized by Xilinx ultrascale VU37P HBM FPGA.

A 0.3V 10T SRAM with Pulse Control Based Read-Assist and Write Data-Aware Schemes for IoT Applications

Kuo-Hung Yang\(^1\), Chang-Ming Tsai\(^1\), Yi-Hsuan Hung\(^1\), S M Salahuddin Morsalin\(^1\), Ming-Hwa Sheu\(^1\), and Jin-Fa Lin\(^2\)
\(^1\)Department of Electronic Engineering, National Yunlin University of Science and Technology
\(^2\)Department of Information and Communication Engineering, Chaoyang University of Technology

A novel 10-transistor (10T) PNN low-power bit-cell Static Random-Access Memory (SRAM) design for sub-threshold voltage applications is proposed in this paper. Features of the proposed design include: (a) pulse control read-assist provides dynamic read decoupling technique to eliminate read interference, (b) write data-aware schemes has been used to cut off the pull-down path, and (c) write capability has been improved through additional write current. The designed architecture of 1-Kb SRAM macros (32 rows * 32 columns) has been implemented based on the TSMC-40nm GP process technology. At 300mV and 10MHz operating frequency, the read and write power consumption is 4.15μW and 3.82μW, while the average energy consumption is only 0.42pJ.
Low Power Approximate Booth Multiplier with Novel Encoding and Pre-Encoded Mechanism

Chun-Huo Hsiao, Chien-Ho Wang, Yen-Jen Chang
Department of Computer Science and Engineering, National Chung Hsing University

Approximate computing is a popular issue which is used to design power-efficient circuits with low complexity by trading accuracy for energy. In this paper, approximate Booth multipliers are designed based on approximate radix-4 modified Booth encoding (MBE) architecture, because the radix-4 Booth algorithm can reduce the number of partial products by half, it is widely used to improve the performance of multiplier. Our design contains a feature that involves both reducing the logic expression complexity of the Booth partial product generator and can effectively reduce the power of Booth multiplier dissipated in the redundant activities by disabling the Booth encoders and decoders. The proposed approximate multipliers are demonstrated to have better performance than existing approximate Booth multipliers in terms of delay and power consumption. Compared to the state-of-the-art designs, the simulation results using HSPICE based on TSMC 40nm technology show that in the 16-bit case the proposed pre-encoded multiplier can reduce the dynamic and static power by 15% and 30%, respectively.
E3.1 13:45-14:05

[Invited Paper]
SFO: A Scalable Approach to Fanout-Bounded Logic Synthesis for Emerging Technologies

Jie-Hong Roand Jiang
National Taiwan University

Fanouts are an essential element for signal cloning to achieve logic sharing, but can be a very limited resource in certain emerging technologies, such as quantum circuits, superconducting electronic circuits, photonic integrated circuits, and biological circuits. Although fanout synthesis has been intensively studied for high performance circuit synthesis, prior methods often treat fanout as a soft constraint for critical path optimization or target on specific high-fanout nets such as clock and reset signals. They are not particularly suited for circuit synthesis of these emerging technologies. By treating fanouts as first class citizens, the problem of fanout-bounded logic synthesis was posed as a challenge in the 2019 IWLS Programming Contest. In this paper, we present our winning method, which achieved the overall best quality in the competition, based on fanout load redistribution among existing or expanded equivalent signals.

E3.2 14:05-14:15

Power Domain Layer Assignment in Package Substrate Design

Yu-Sheng Qin, Xiao-Yu Wang, Yi-Yu Liu
Department of Computer Science and Information Engineering, National Taiwan University

both the chip size and design complexity are inevitably increasing. In addition, the number of power domains is drastically increased owing to the demands of power efficiency in various functional modules. Thus, reliable power delivery has become one critical issue in package substrate design. In this paper, we are the first work to automate power domain layer assignment in package substrate design. First, seed selection step selects critical power domains with larger area conflict and assigns these critical domains to separated metal layers. After that, remaining power domains are assigned to proper metal layers taking power pin distribution into account. Experimental results demonstrate the effectiveness of our layer assignment algorithm for the follow-up polygon layout partition.
Wire-bond Package Finger Placement with Minimal Distance

Yu-En Lin, Che-Hsu Lin, Yi-Yu Liu
Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology

Abstract— Semiconductor packaging is the final stage of fabrication, which encapsulates one or more integrated circuit chips to avoid physical damage and to provide interconnections to outside world. There are a variety of packaging design styles, such as wire-bond, flip-chip, 3-D TSV, and so on. In this work, we propose a framework to handle the finger placement problem in wire-bond package design style. First, a minimum-cost maximum-flow-based global finger placement algorithm is used to derive initial finger locations with overall minimum distance. After that, legalization steps are proposed taking bond-wire crossing constraint and pad row sequence constraint into consideration. Finally, incremental finger placement refinement algorithm is adopted to generate final layout with minimal displacement compared to the initial global placement. With the proposed framework, legalized finger locations are determined in polynomial time for package substrate layout engineer to start with.

Deep Global-Local Shape Feature Learning for Novel Layout Patterns Detection

Hao-Chiang Shao¹, Hsing-Lei Ping², Kuo-shiuian Chen², Weng-Tai Su², Chia-Wen Lin²,³, Shao-Yun Fang⁴, Pin-Yian Tsai⁴, and Yan-Hsiu Liu⁴
¹ Dept. Statistics and Information Science, Fu Jen Catholic University
² Dept. Electrical Engineering, National Tsing Hua University
³ Dept. Electrical Engineering, National Taiwan University of Science and Technology
⁴ United Microelectronics Corporation

The fact that it is too complicated to model the non-linear shape distortion of the fabrication result of a designed IC pattern urges the development of learning-based pre-simulation models. Such models are usually driven by pairwise training samples, each consisting of a layout pattern and a reference contour image after one certain fabrication step. However, it is expensive and time-consuming to collect reference contour images of layouts for training and fine-tuning such pre-simulation models via the IC fabrication process. Therefore, we propose a deep learning-based layout novelty detection algorithm with a Glocal (global-local) novelty score. The proposed algorithm can act as an active learning oracle, based on which users can find a reduced amount of layouts worthy enough to be fabricated for acquiring the ground-truth circuit contours of their IC products. Inspired by residual-based and classification-based novelty detection models, we also devise a layout novelty detection method that can assess the potential novelty of a layout by exploiting two subnetworks, an autoencoder and a pretrained layout-to-SEM prediction model. The former subnetwork characterizes the global structural similarity between the given layout and training samples, and the latter can derive an attention-guided latent code depicting the local deformation.
E3.5

SRAF Insertion by Variational Adversarial Active Learning and Clustering with Data Point Retrieval

Sean Shang-En Tseng¹, Iris Hui-Ru Jiang¹, and James P. Shiely²
¹Graduate Institute of Electronics Engineering, National Taiwan University
²Synopsys, Inc., Hillsboro, OR 97124, United States

As the feature size keeps shrinking in the modern semiconductor manufacturing process, subresolution assist feature (SRAF) insertion is one promising resolution enhancement technique that can improve the printability and lithographic process window of target patterns. State-of-the-art works resort to machine learning to reduce runtime but require abundant training samples to generalize the trained models and achieve high performance. Nevertheless, in advanced lithography, we may have a huge solution space of SRAF insertion but few labeled training samples. Therefore, in this work, we address SRAF insertion from a data efficiency perspective. We separate sample selection from SRAF probability learning and train a variational autoencoder and an adversarial network to discriminate between unlabeled and labeled data effectively. Second, we devise a region-based concentric circle area sampling representation to avoid information loss during feature extraction. Third, we determine the final placement of SRAFs by a novel clustering method based on retrieved data points. Experimental results show that, compared with state-of-the-art works, by using 40% training samples, our framework can achieve comparable or even better process variation bands and edge placement errors.
A4.1 14:45-15:05

An 8-MHz 0.677W/cm² Power Density Constant On-Time Controlled Buck Converter with Charge-Pump based Q-De-Sense Ramp Achieving 55mV/1A Constant Voltage Drop

Chieh-Ju Tsai, Shih-Chieh Hsu, Ching-Jan Chen
Graduate Institute of Electricals Engineering, National Taiwan University

In this paper, a Q-De-Sense Ramp compensation scheme adopted in a ripple-based constant on-time controlled buck converter has been presented. The proposed control scheme is not only simple but also suitable for wide input, output range and against different circuit operating condition. As a result, the prototype design operated at an 8-MHz high switching frequency achieving 0.677W/cm² power density and an 0.062 mm² controller area fabricated in a 0.18μm CMOS process. Furthermore, the small-signal analysis based on describing function approach is used to analyze the transient behavior. The small-signal analysis reveals the transient response is near invariant for the converter operating at different circuit condition. The experimental results show that 55mV/1A almost constant voltage drop from VIN = 3.3V to VO = 0.6 – 1.3V. And the peak efficiency is 87% while a 0.33μH small inductor is used.

A4.2 15:05-15:15

An Improved-Dead-Beat-Controlled Buck Converter with New Current-Sensing Techniques

Hsun-Ming Chan, Yuh-Shyan Hwang
Department of Electronic Engineering, National Taipei University of Technology

This paper presents an improved-dead-beat-controlled buck converter with new current-sensing techniques. This circuit uses the new current-sensing circuit to achieve fast transient response. The proposed converter is with current load range 0.1A~ 0.5A. The buck converter is fabricated in TSMC 0.35μm 2P4M CMOS process. The total area of the chip is 1.5 mm x 1.5 mm. The measurement results show that the load transient responses are 2μs and 2.2μs when the load current are light to heavy and heavy to light, respectively. The maximum efficiency is 91.3% in the load current is 200mA and output voltage is 2V.
A4.3 15:15-15:25

An Adaptive On-Time Buck Converter with New Current-Sensing and Phase-Frequency-Locked Techniques

Zong-Ben Jia, Jiann-Jong Chen
Department of Electronic Engineering, National Taipei University of Technology

This paper presents an adaptive on-time buck converter with new current-sensing and phase-frequency-locked techniques. The proposed converter achieves fast transient response and a wide output voltage range. The proposed buck converter is fabricated with TSMC 0.35 μm 2P4M CMOS technology, and the chip area is 1.4366 mm × 1.4534 mm. The measured results show that the output voltage is 1.8V, the load current changes from 50mA to 500mA, and the transient response from 50mA to 500mA is 2μs and 2.4μs, respectively. When the load current is 300mA, the maximum power efficiency is 90.65%.

A4.4 15:25-15:35

Design of a Switched-Capacitor DC-DC Converter in 180nm Standard CMOS Process

Zhi-Yun Hsu, Jun-Wan Wu, Po-Han Chen, Hao-Chung Cheng, and Po-Hung Chen
Institute of Electronics, National Yang Ming Chiao Tung University

This paper presents a fully integrated step-down switched-capacitor (SC) DC-DC converter, which using 180nm CMOS technology for system-on-chip applications. There are two control mechanisms use in this converter. One is switch array modulation (SAM) technique, which is used to regulate the output voltage, and the other is voltage ripple modulation (VRM) technique, which is used to adjust the output voltage ripple when the load current varied. Due to these mechanisms, the converter is able to generate a regulated output voltage with low voltage ripple. The measurement results show that the converter achieves peak power conversion efficiency of 74.1%, and within the load range of 1.4mA to 5 mA, the converter maintains all power conversion efficiency over 70%. Besides, when the load current is in the range of 0.1mA to 5 mA the output voltage ripple is controlled below 63.4 mV.
A Novel Pulse-Based Multiplier for Maximum Power Point Tracking Circuit

Jian-Zhou Yan, Wei-Han Pan, Hung-Hsien Wu, Tien Hsu, and Chia-Ling Wei  
Department of Electrical Engineering, National Cheng Kung University

A novel pulse-based multiplier for maximum power point tracking (MPPT) circuit used in power converters is proposed. The MPPT circuit adopts perturb-and-observe (P&O) algorithm. The proposed P&O MPPT circuit can accurately detect the maximum power point of the PV module without using power-hungry microcontroller or inaccurate analog multiplier. Instead, a high-accuracy pulse-based multiplier is proposed. Its input signals are analog while its output signal is digital. Moreover, the new MPPT controller based on the proposed multiplier is also presented. The chip was fabricated using a 0.18μm 1P6M mixed-signal CMOS process. According to the measured results, the available power from the chosen PV module is 0.1-3mW, the maximum tracking efficiency of the proposed MPPT controller is 99.3%, and the maximum total efficiency of the proposed harvesting chip is 94.7%.
Oral D5
Multicore and Biomedical Systems

8 月 5 日（四）14:45-15:45
Session Chair: 陳坤志教授 | 國立中山大學 資訊工程學系

D5.1 14:45-15:05

VLSI Implementation of the Abnormal Heart Beat Detector using CNN Accelerator

Pei-Jung Chang¹, Hsin-Tung Hua¹, Yuan-Ho Chen²
¹ Dept. of Electronics Engineering, Chang Gung University
² Dept. of Electronics Engineering, Chang Gung University; Dept. of Radiation Oncology, Chang Gung Memorial Hospital

The heart is the most sophisticated organ of human beings. “Emergency” is often caught off guard. At this time, if the heart rate is monitored, artificial intelligence can quickly determine the disease to achieve early detection and timely treatment. From the MIT-BIH database, six types of ECG signals are extracted for discrimination. Use a convolutional neural network to realize fast identification, and analyze the relationship between each layer of the training module to achieve the best accuracy. At this time, the accuracy is 97.57%. In order to improve the versatility of the circuit and achieve the purpose of repeated use circuit modularization, such as PE Unit, activation function module, etc., control each circuit module so that the circuit has the best calculation effect, at this time The accuracy of hardware discrimination is 96.83%. This article uses TSMC 0.18μm COMS to realize this chip, which achieves an operating frequency of 50MHz, a chip area of 1.419mm², and maximum power consumption of 2.808mW.

D5.2 15:05-15:15

Fourier-Domain OCT Imaging Processor with Improved Memory Efficiency

Song-Nien Tang and Chi-Ho Lou
Department of Information and Computer Engineering, Chung Yuan Christian University

This paper presents a display processing unit for the image formation of the Fourier-domain optical coherence tomography (FDOCT) system. Using the proposed design, the FDOCT imaging operations involving the re-sampler, the real-valued fast Fourier transform (RFFT) and the image display processing can be efficiently performed using a hardware processor, allowing high-frame-rate OCT image display with the direct generation of gray-scale image data. Compared to previous works, the proposed design improved the memory efficiency in terms of the RAM size, the control complexity and the logarithm table size for the re-sampler, the RFFT and the display processing units, respectively. The presented FDOCT imaging processor was verified using an FPGA-SoC
platform through the software-hardware cooperation with accesses of the raw data obtained from the FDOCT front-end equipment.

**D5.3** 15:15-15:25

**PASPO1-3 Algorithms for On-Display Fingerprint Sensors with FPGA Design**

*Huei-Ming Wu and Pei-Yung Hsiao*

*Department of Electrical Engineering, National University of Kaohsiung*

This paper uses the latest two types of on-display fingerprint, ODF, sensors as input images to create three progressive alignment algorithms, PASPO1-3, based on a single-pixel orientation, SPO. We strengthen the performance of alignment function by adopting the SPO distribution with classification by various orientational angular accuracies in the overlapped area of a couple of partial fingerprints from ODF sensors. Besides, for speeding up the computing time, we design a parallel array of hardware module for those averaging differences of SPO, AvD, and other parameters. A faster and more accurate integrated scheme of the partial fingerprint alignment with FPGA accelerator for fingerprint recognition is therefore successfully accomplished.

**D5.4** 15:25-15:35

**Entropy-based Thermal Sensor Placement for the Temperature-aware Multi-core System by Using Simulated Annealing**

*Kun-Chih (Jimmy) Chen and Chia-Hsin (Dolly) Chen*

*Department of Computer Science and Engineering, National Sun Yat-sen University*

Because of the high design complexity in multi-core system, it suffers from serious thermal issue caused by the large variety of workload. To monitor the heat phenomenon, the cost-efficient way is to place number-limited thermal sensors on the multi-core system. Consequently, many researches proposed a lot of methods to find the proper location for the thermal sensor placement, such as greedy algorithm, statistical model analysis, genetic algorithm, etc. However, due to the time-varying characteristics of the temperature behavior on the system, the temperature distribution usually changes along with time. Besides, the temperature distribution also depends on the target application on the multi-core system, which worsen the difficulty to find the proper locations for the thermal sensor placement. The improper locations for the thermal sensor placement results in large reconstruction error of the temperature distribution while using the sensing information from the placed thermal sensors. To solve the disadvantages of the state-of-the-art, we propose an entropy-based sensor placement method, which applies the probability model to find the entropy-based objective function and optimize it to achieve the goal eventually. In addition, a simulated annealing method is proposed to prevent from the regional best solution, which further improve the efficiency of the proposed approach. The experimental results show that we can reduce 47.01% average full-system temperature reconstruction error compared with the previous CS-based
sensor placement approaches. In addition, compared with the traditional non-SA method, the proposed approach can further reduce the average error and reduce it by 5.18%.

**D5.5 15:35-15:45**

**A Lego-based Convolutional Neural Network Design Methodology with NoC Interconnection**

*Kun-Chih (Jimmy) Chen, Yi-Sheng Liao, and Cheng-Kang Tsai*

*Department of Computer Science and Engineering, National Sun Yat-sen University*

The Convolutional Neural Network (CNN) has been shown its superiority to solve the problems of classification and recognition in recent years. However, the CNN hardware implementation is challenged due to the high computational complexity and high diverse dataflow according to different CNN models. To mitigate the design challenge, many researches focus on specific CNN models or layers, including dedicated dataflow. However, the dedicated designs for specific CNN models or layers limit the design flexibility. Because each different CNN models involve similar computing functions with proper permutations, we propose a novel Lego-based convolutional Neural Network on Chip (CNNoC) design methodology in this work. We predefined some common neural computing units, such as multiply-accumulation, pooling, etc., called Neu-Lego blocks. Several identical Neu-Lego blocks will be grouped and assigned to one processing element, called Neu-Lego PE. To increase the design flexibility, we further adopt the high flexible Network-on-Chip (NoC) interconnection to connect each involved Neu-Lego PE to construct different CNN models. To further implement the large-scale CNN models by following the proposed design methodology, we propose a dynamic mapping method to increase the reusability of each Neu-Lego PE. Because of the flexible design methodology, the proposed approach can leverage the different CNN model implementation. Compared with the conventional approaches, the proposed approach can improve 84% to 544% throughput according to different involved CNN models and mapping algorithms. Besides, the corresponding hardware is implemented to validate the proposed design methodology with the reusable Neu-Lego PEs, which helps to reduce the area overhead by 51% to 94%.
Oral T1

Emerging Technologies

8 月 5 日（四）14:45-15:45
Session Chair: 夏勤勤教授 | 長庚大學工學院

T1.1 14:45-14:55

Edge AI Computing for Vehicle Detection and Class Counting Intelligent System

S M Salahuddin Morsalin¹, Jia-He Lin¹, Ming-Hwa Sheu¹, Jia-Xiang Zheng¹, Cheng-Jian Lin²
¹ Department of Electronic Engineering, National Yunlin University of Science and Technology
² College of Intelligence, National Taichung University of Science and Technology

The Squeeze-and-Activation Choice deep neural network model has been proposed for vehicle detection and Class Counting with parameter optimization. The SAC blocks are incorporated into the convolutional layers of the SSD model that highlights the important features and decline the unimportant features. During the testing phase, the speed has improved through the filtering out and decreasing of unimportant features by using the scale choice mechanism. Besides, the Fuzzy Sigmoid Function is projected to increase the activation that can change the activation interval of the fuzzy sigmoid function through Fuzzy Logic. Furthermore, SAC blocks have avoided many unusable parameters in the saturation interval and improved the performance effectively. The SAC-SSD model has solved many convolutional worthless parameters problem while doing operations and improved speed performances. Finally, the SAC-SSD deep neural network model has built with passing detection lines and class counting windows to implement for the vehicle class counting. As a result, the model tested FPS on edge AI and it reached a real-time processing speed of 38.4 and an accuracy rate of up to 95.1%.

T1.2 14:55-15:05

A High-linearity, 166MS/s Conversion Rate Time-to-Digital Converter for LiDAR Application

Chan-Yuan He, Li-Jung Chou, Hsi-Hao Huang, and Chen-Yi Lee
Institute of Electronics, National Chiao Tung University

In this paper, a tapped delay line (TDL)-based time-to-digital converter (TDC) is presented. The architecture is composed of a multi-phase clock generator, two tapped delay line-based fine interpolators and other circuits. Instead of using a Phase-locked Loop (PLL), a Frequency-locked Loop (FLL) with digital correction is utilized to reduce jitter and maintain PVT-insensitivity. As a result, effects on time sampling caused by noisy reference clock can be minimized. Also, the high linearity is achieved by employing two
fine interpolators on START and STOP signals apart. The proposed TDC chip fabricated in 180 nm CMOS process achieves differential non-linearity (DNL) of 0.05 rms LSB, integral non-linearity (INL) lower than 0.88 LSB and single-shot precision of 147 ps, making it very suitable for automotive LiDAR applications.

T1.3 15:05-15:15

Enhanced Stability of Cryo-CMOS Subthreshold SRAM Cell

Chang-Ju Liu$^1$ and Vita Pi-Ho Hu$^{1,2}$  
$^1$Department of Electrical Engineering, National Central University  
$^2$Department of Electrical Engineering, National Taiwan University

In this work, we analyzed the static noise margin (SNM) of cryo-CMOS SRAM cell operated at superthreshold (Vdd = 0.75 V) and subthreshold (Vdd = 0.2 V) regions. The impact of temperature and write-assist (WA) technique on the ultra-thin-body (UTB) SOI SRAM cell has been examined. UTB SOI MOSFET at 77K shows a steep subthreshold slope and reduced leakage, which enables the UTB SOI SRAM cell to operate at low Vdd. Our results show that at Vdd = 0.2V, compared to 300K, UTB SOI SRAM with WA at 77K shows 39.6%, 32.2%, and 11% improvements in read, hold, and write SNM, respectively. Therefore, cryo-CMOS SRAM enabling low Vdd operation and improving stability is a highly promising solution to improve both performance and power efficiency.

T1.4 15:15-15:25

PQ-HDC: Projection-based Quantization Scheme for Flexible and Efficient Hyperdimensional Computing

Chi-Tse Huang, Cheng-Yang Chang, Yu-Chuan Chuang, and An-Yeu (Andy) Wu  
Graduate Institute of Electronics Engineering, National Taiwan University

Brain-inspired Hyperdimensional (HD) computing is an emerging technique for low-power/energy designs in many machine learning tasks. Recent works further exploit the low-cost quantized (bipolarized or ternarized) HD model and report dramatic improvements in energy efficiency. However, the quantization loss of HD models leads to a severe drop in classification accuracy. This paper proposes a projection-based quantization framework for HD computing (PQ-HDC) to achieve a flexible and efficient trade-off between accuracy and efficiency. While previous works exploit thresholding-quantization schemes, the proposed PQ-HDC progressively reduces quantization loss using a linear combination of bipolarized HD models. Furthermore, PQ-HDC allows quantization with flexible bit-width while preserving the computational efficiency of the Hamming distance computation. Experimental results on the benchmark dataset demonstrate that PQ-HDC achieves a 2.82% improvement in accuracy over the state-of-the-art method.
A Charge-Redistribution Based Computing-in-Memory Edge AI Accelerator for CNNs

Jun-Hui Fu and Soon-Jyh Chang
Department of Electrical Engineering, National Cheng Kung University

This paper presents a charge-redistribution based computing-in-memory (CIM) edge AI accelerator in 40nm CMOS. This CIM macro adopts 8T static random access memory (SRAM) with a read-decoupled port to avoid write-disturbing. A weighted capacitor switching technique is proposed for creating 2-bit weighted charge on the computing bit-line. The proposed technique achieves better linearity than current-domain computation owing to highly linear and robust metal-oxide-metal (MOM) capacitors. A low multiply-accumulate (MAC) value skipping technique is also adopted to enhance the speed and diminish the power consumption. A 64 × 64 b 8T CIM macro with 1-8b inputs, 8b weights achieves energy efficiency of 15.7 TOPS/W in post-layout simulation and covers an area of 0.44mm².

A Novel Action Recognition Network for Basketball Foul Detection

Cheng-Hung Lin, Min-Yen Tsaiy, Po-Yung Chou
Dept. Electrical Engineering, National Taiwan Normal University

In recent years, deep neural networks for action recognition has attracted extensive attention because of its wide range of applications such as anomaly behavior detection in smart surveillance system. Among the proposed deep learning models, 3DCNN works very well in the action classification of large data sets, including UCF-101, HMDB-51, and Kinetics. However, for the classification of fine-grained actions, current action recognition models still need improvement. The finegrained action means that the difference from the normal action is very small, and the time of occurrence is extremely short and difficult to distinguish. For example, in the basketball game, the foul action is a kind of fine-grained actions. Foul action recognition is very challenging because fouls in basketball games are always instantaneous and very similar to normal actions. In this paper, we propose a lightweight fine-grained action recognition model for basketball foul detection. Compared with other action recognition models such as two-stream model, 3DCNN, our proposed network has a better effect on this subtle classification task, and is lighter in parameters. The visualized foul feature distribution is concentrated in a few frames that supports our initial hypothesis that fouls always happen instantaneously. Finally, the output of this research can be used to assist in training basketball referees.
Oral E4
Design Analysis and Verification for Digital Designs and Cloud
8月5日（四）14:45-15:45
Session Chair：陳勇志教授 | 元智大學 資訊工程學系

E4.1 14:45-15:05
Compatible Equivalence Checking of X-Valued Circuits
Yu-Neng Wang\textsuperscript{2}, Yun-Rong Luo\textsuperscript{2}, Po-Chun Chien\textsuperscript{1}, Ping-Lun Wang\textsuperscript{2}, Hao-Ren Wang\textsuperscript{1}, Wan-Hsuan Lin\textsuperscript{2}, Jie-Hong Roland Jiang\textsuperscript{1,2} and Chung-Yang Ric Huang\textsuperscript{1,2}
\textsuperscript{1}Graduate Institute of Electronics Engineering, \textsuperscript{2}Department of Electrical Engineering, National Taiwan University

The X-value arises in various contexts of system design. It often represents an unknown value or a don’t-care value depending on the application. Verification of X-valued circuits is a crucial task but relatively unaddressed. The challenge of equivalence checking for X-valued circuits, named compatible equivalence checking, is posed in the 2020 ICCAD CAD Contest. In this paper, we present our winning method based on X-value preserving dual-rail encoding and incremental identification of compatible equivalence relation. Experimental results demonstrate the effectiveness of the proposed techniques and the outperformance of our approach in solving more cases than the commercial tool and the other teams among the top 3 of the contest.

E4.2 15:05-15:15
MPFD: Modified Parcel-Fit Decreasing for Deploying EDA Services On Public Cloud
Willy Y.-W. Hsu, Jiun-Cheng Tsai, Aaron C.-W. Liang, Charles H.-P. Wen
Dept. Electrical and Computer Engineering, National Yang Ming Chiao Tung University

Cloud computing is becoming a pervasive technology in many fields. However, the electronic-design-automation (EDA) field has not yet developed a working prototype that employs the commercial tools as services and deploys them on the cloud. Accordingly, we aimed at the core problems on private EDA clouds in the prior work and gave an integrated solution for processing requests. The issues have been migrated to the public cloud since the private cloud has fixed and limited resources, whereas the public cloud offers highly scalable resources. However, we further improve the overall cloud performance by resource consolidation (i.e. reallocate services for switching on/off servers) on basis of operating expanse (OPEX). The proposed service-deployment solution, named Modified Parcel-Fit Decreasing (MPFD), effectively builds a public EDA cloud and make 15\%, 7\%, and 4\% more profits than the conventional and two other
E4.3 15:15-15:25

Designing A Compact Convolutional Neural Network Processor on Embedded FPGAs

Yin-Chun Ling, Hsu-Hsun Chin, Hsin-I Wu, Ren-Song Tsay, Jui-Hung Kao
Department of Computer Science and Electronic Engineering, National Tsing Hua University

Designing a FPGA-based Convolutional Neural Network (CNN) processor needs to consider multiple aspects, such as the feasibility of limited configurable resource, external memory latency and scheduling between memory and computation units. Addressing these issues, we elaborate a systematic design approach that allow fast deployment, which includes the parameterized computation and memory unit, which can be configured based on the target platform, and an evaluation approach for searching the optimal setting sets. To evaluate the proposed approach, we performed object detection task, YOLOv2, on PYNQ-Z1 and achieved 48.23 GOPs throughputs as well as 0.611 seconds execution time. This is 42.38 and 12.8 times faster than the same inference on CPU and GPU and is 2.36 times faster than other FPGA implementations. Additionally, our created evaluation model is only 5-22% apart from the implementation result, which is 60% less than previous work.

E4.4 15:25-15:35

SlewFTA: Slew-Bounded Functional Timing Analysis

Allen Z.-H. Tsai, Aaron C.-W. Liang, Charles H.-P. Wen
Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University

Nowadays, timing analysis is a critical step in the modern VLSI design flow. Compared to static timing analysis (STA), functional timing analysis (FTA) can not only derive a critical delay closer to the true delay of the circuit but also consider its function to generate one input pattern to induce such delay. However, same as STA, FTA always selects the worst input slew of a cell during propagation until reaching the primary outputs. In reality, this is overly pessimistic for the final delay. Not always the worst slews should be propagated during derivation of the critical path considering the Boolean function of the circuit in FTA. Therefore, a slew-bounded functional timing analysis considering slew propagation (named SlewFTA) is proposed in this paper. To overcome the problem, binary search and slew shrinking are combined in FTA to consider the propagation of slews more realistically. Experimental results indicate that compared to FTA, after bounding the propagation of slews, SlewFTA further reduces the final delay of 12 benchmark circuits by 6.38% on average (10.46% for the best case). As a result, SlewFTA is proven more
Due to machine learning advancements and broad synthesis and verification applications, circuit learning gains significant attention and is posed as a challenge in recent ICCAD and IWLS contests. Decision-tree-based learning plays a crucial role in state-of-the-art methods. However, tree learning may be ineffective due to its structural restriction. This work proposes a graph learning scheme that allows a compact representation of the underlying learning task. Compared with prior work, our method is more scalable to complicated circuit learning tasks and offers mechanisms for a training procedure to trade-off between circuit size and accuracy. Experimental results show the superiority of our approach to prior work on various circuits under learning in terms of accuracy, training time, and resulting circuit sizes.
**Oral A5**

**Power IC (II)**

8 月 6 日 (五) 10:45-11:45

Session Chair: 陳厚銘教授 | 國立虎尾科技大學 電機工程系

**A5.1** 10:45-11:05

[Invited Paper]

**Fully Integrated GaN-on-Silicon Gate Driver and GaN Switch with Temperature-compensated Fast Turn-on Technique for Achieving Switching Frequency of 50MHz and Slew Rate of 118.3V/ns**

National Yang Ming Chiao Tung University

In this paper, a monolithically integrated driver fabricated by 12V depletion mode GaN (dGaN) and enhanced mode GaN (eGaN) driver is proposed. The proposed driver features an internal temperature-compensated (T-compensated) controller to drive an integrated 650V eGaN power switch. Due to temperature-compensated characteristics, a slew-rate enhancement driver can be well-controlled by the fast turn-on (FTO) technique which is supplied by an on-chip regulator with reference voltage circuit. Therefore, the Miller plateau voltage can be tracked correctly by the proposed controller so that the switching frequency can be raised up to 50MHz and the dVDS/dt slew rate can reach 118.3V/ns for high efficiency and high switching operation.

**A5.2** 11:05-11:15

**Active Output Ripple Elimination Technique for Buck Converter**

Pang-Jung Liu, Chen-Yu Liao, Xing-Yu Chen, and Ting-Yi Liao
Department of Electrical Engineering, National Taipei University of Technology

An active output ripple elimination (AORE) technique is presented in this paper for achieving low-EMI buck converter. The buck converter delivers the dc output current while an AORE circuit is in charge of generating an ac current to precisely eliminate the ripple of inductor current. Thus, the output ripple of the buck converter can be decreased significantly. A dynamic phase lock circuit and a proposed control scheme are adopted to fulfill exact ripple cancellation and guarantee the converter stability. The experimental results show that the output voltage ripple can be decreased from 200mV to 20 mV. Moreover, the effectiveness of ripple cancellation technique is independent of load current and duty cycle.
A5.3 11:15-11:25

SystemVerilog Behavior Models of Digitally Controlled Power Converter ICs for Mixed-Level Design Methodology

Wei-Ting Yeh, Kai-Yu Hu, Chung-Lun Chang, Chun-Yu Chen and Chien-Hung Tsai
Department of Electrical Engineering, National Cheng Kung University

SystemVerilog behavior models of power converter integrated circuits for mixed-level design methodology was proposed. When SystemVerilog is used to establish behavior models for circuits, the process of transistor-level design can be omitted to prevent prolonging the redesign time. Moreover, models created in this manner can sustain simulation accuracy while reducing the simulation time by 57% compared with models conventionally created using Verilog-AMS. In this study, 0.18-μm CMOS fabrication was used to verify the effectiveness of the proposed SystemVerilog model and mixed-level design in the creation of a digitally controlled buck converter integrated circuit.

A5.4 11:25-11:35

An adaptive biasing Capless LDO with enhanced PSRR for Piezoelectric Vibration Energy-Harvesting system

Yong-Zheng Wang, Ching-Yuan Yang
Department of Electrical Engineering, National Chung Hsing University

In this paper, we propose a high PSRR LDO (low dropout regulator) and fast settling time for piezoelectric vibration energy-harvesting system. The LDO was implemented in 0.35-μm CMOS process. Compared to the power supply rejection ratio of the conventional LDO, the simulation shows the PSRR improvement of the proposed LDO is about 35db.

A5.5 11:35-11:45

Class-E² Converter Design for Wireless Power Transfer Application

Yen-Hsiang Chang, Kuan-Te Wu and Heng-Ming Hsu
Department of Electrical Engineering, National Chung-Hsing University

A magnetic resonance wireless power transfer (WPT) system operated at 13.56MHz is proposed in the paper. To increase the transfer efficiency, the class-E topology is implemented in both of the transceiver and receiver. A parallel-tuned push-pull Class-E amplifier is proposed in transceiver design. A Class-E rectifier is used in receiver design to increase the output power. The coupling coil is considered as a load of the transceiver for real WPT system. The simulation shows the WPT system delivers 37W output power with the efficiency of 74%.
A6.1 10:45-10:55

A 4-CH Current-Mode Functional Electrical Stimulator with Adjustable Biphasic Current and Dual-Shape Waveform Selection for Deep Brain Stimulation

Yi-Ting Yeh, Yu-Hsuan Pai, Huang-Hsiang Chang, Yi-Ching Lu, Sheng-Yu Peng
National Taiwan University of Science and Technology

A 4-channel current-mode functional electrical stimulator for deep brain stimulation is proposed and fabricated in TSMC 0.18 μm CMOS process. In this system, stimulation current amplitude ranges from 10 A up to 2.56mA in each channel. A current digital-to-analog converter (DAC) with binary to thermometer decoder is designed to provide programmable biphasic current waveform. For efficiency improvement and different pulse width requirements, pulse shape and decaying exponential shape selection systems are also implemented here. Active-recharge compensator and passive recharge circuit are used to eliminate residual charge in the tissue. According to the characteristic of push-pull amplifier, self-charge compensator can provide proper compensation current more efficiently. Finally, use animal experiments to verify that the chip is safe and can effectively suppress epilepsy symptoms.

A6.2 10:55-11:05

A reconfigurable CMOS-MEMS accelerometer with on-chip readout circuit

Cheng-Yen Lin¹,², Hao-Chiao Hong³,⁴ and Yi Chiu³,⁴

¹Institute of Biomedical Engineering, National Chiao Tung University
²Institute of Biomedical Engineering, National Yang Ming Chiao Tung University
³Dept. Electrical and Computer Engr., National Chiao Tung University
⁴Dept. Electrical and Computer Engr., National Yang Ming Chiao Tung University

This paper reports a reconfigurable capacitive accelerometer fabricated by a commercial 0.35 μm CMOS process. The MEMS structure is composed of a spring, a proof mass and sensing electrodes. In order to achieve reconfigurable and multi-bandwidth functions, a bias voltage is applied to change the resonant frequency to obtain a wide bandwidth. The measurement results show the sensitivity is 12.1 mV/g. The noise floor is 0.14 g/√Hz. The bandwidth tuning range is 18.4%.
A6.3  11:05-11:15

A 4.85nJ/Conversion Sub-Threshold Temperature-to-Digital Converter from -20°C to 120°C

Yi-Lun Tseng, Yong-Yu Lin, Ching-Hsiang Chang and Hongchin Lin  
Dept. of Electrical Engineering, National Chung Hsing University

A 4.85nJ/conversion temperature-to-digital converter was designed and fabricated in 180 nm CMOS technology using the proposed current reduction scheme and the comparators biased by current sources. Its inaccuracy is +1.4/-0.7°C for temperature from -20°C to 120°C at a supply of 0.65 V. The power consumption is 17.6 nW and the conversion time is 275.7 ms at 25°C.

A6.4  11:15-11:25

A Tunable Gate Leakage Current Based Nano-Ampere Constant Current Generator with Low Temperature Coefficients

Ching-Hsiang Chang, Yi-Lun Tseng, Yong-Yu Lin and Hongchin Lin  
Dept. of Electrical Engineering, National Chung Hsing University

A tunable 1.34 nA constant current generator employing gate leakage of a PMOSFET and a 308 mV sub-nW voltage reference was designed and fabricated using 90nm CMOS technology. The temperature coefficients (TC) of the current reference generator at a supply of 0.75 V are 53 ppm/°C and 394 ppm/°C from 0°C to 120°C are shown by simulation and measurement, respectively.

A6.5  11:25-11:35

Analog Front End for Neural Signal Recording with CMRR Boosting Pseudo Right-Leg Driven and Stimulation Artifact Countermeasure

Ming-Yao Li, Ying-Rong Su, and Chung-Chih Hung  
Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University

This paper integrates the Pseudo Right-Leg Driven (Pseudo RLD), analog preamplifier and the analog-to-digital converter (ADC) together to achieve high CMRR and minimize the effect of stimulation. The three circuits together comprise a complete analog front-end (AFE) capture system which converts analog signal into the digital form and effectively provide a solution so that good signal quality can be obtained even when the stimulus signal arrives. The analog front-end adopts a modular approach to increase future tunability. The analog front-end amplifier consists of an AC-coupled fully differential low-noise amplifier, a variable gain amplifier, and a low-pass filter. The proposed Pseudo RLD circuit effectively boosts CMRR. The ADC is a binary weighted switchback monotonic successive approximation analog-to-digital converter (SAR ADC) architecture. The input referred noise (IRN) of the analog front-end amplifier is less than 2uV, with 40/60/80dB
adjustable gain and 0.5~100/1000Hz adjustable bandwidth. The resolution of the ADC is 10-bit, and the accuracy of the overall AFE exceeds 9 bits ENOB. The total system power consumption is less than 20uW.

**A6.6 11:35-11:45**

**A Measurement System Based on Extended-Gate Field-Effect Transistor for Biomedical Applications**

Yung-Yu Chen¹, Po-Yu Kuo¹, Jung-Chuan Chou¹, Chih-Hsien Lai¹, Yu-Hsun Nien², Yu-Hao Huang¹, Zhi-Xuan Kang², Wei-Hao Lai¹, Chun-Hung Chang¹, and Kun-Tse Lee¹

¹Graduate School of Electronic Engineering, National Yunlin University of Science and Technology  
²Graduate School of Chemical and Materials Engineering, National Yunlin University of Science and Technology

In this work, an extended-gate field-effect transistor (EGFET) applied to resistive divider integrated with instrumentation amplifier (UGFPCIA) was proposed for biomedical applications. The EGFET was designed based on arrayed RuO2 uric acid sensing area to detect the uric acid. The EGFET resistive divider can achieve an effective measurement system due to simplicity, convenience, and low cost. The uric acid concentration which covered the human body (2 mg/dL to 7 mg/dL) was applied in the measurement. The experimental results indicated that the EGFET resistive divider had a good uric acid sensitivity of 12.69 mV/(mg/dL), high linearity of 0.997.
Oral D6

Hardware-Software Co-design for AI

8 月 6 日（五）10:45-11:45
Session Chair: 呂仁碩教授 | 國立清華大學 電機工程學系

D6.1 10:45-10:55

A Pipeline-Based Scheduler for Optimizing Latency of Convolution Neural Network Inference over Heterogeneous Multicore Systems

Hsin-I Wu, Da-Yi Guo, Hsu-Hsun Chin, Ren-Song Tsay, Yong-Xuan Chen
National Tsing Hua University

Parallelization is a common design practice for throughput improvement on multicore systems. However, the existing operating systems’ schedulers for CNN inference essentially divide the computational tasks of each convolution layer onto different CPU cores and cause significant inter-core feature-map data movement. Therefore, the overall performance is often degraded. In this paper, we propose a pipeline-based scheduler for convolution neural network inference parallelization with minimal feature-map data movement requirements. The experimental results show that our approach can achieve 73% performance improvement on throughput compared to the existing multi-thread scheduler.

D6.2 10:55-11:05

TCBNN: Ternary-Coded Binarized Neural Network with Per-Neuron Single-Arithmetic-Fault Tolerance

Cheng-Di Tsai, Ting-Yu Chen, Chun-Yen Tsai, Po-Sheng Chang and Tsung-Chu Huang
Department of Electronics Engineering, National Changhua University of Education

Acceleration and power reduction are two critical issues of artificial neural networks, but will be nonsense without reliability. Although input-side deep layers have been shown to possess considerable self-healing, arithmetic faults in shallow decision layers may still cause unimaginable catastrophe. In this paper we propose a novel structure of deep neural network in testing/application stage that converts the trained weights to optimized ternary-coded binary. An arithmetic fault simulator is applied to test the impact to decide how many shallow layers to be product-encoded. The weights of the first encoded layer are multiplied by prime number A prior to transferring to ternary-coded structure. Two types of optimized AN-code decoders are separately proposed for the middle and last shallow layers. Within each layer, a constant-shift sub-layer with almost-free cost is inserted to transfer all multipliers to carry save adders. Except the activation functions of the last layer are simplified to lightweight-slope piecewise lines for calculation by an adder only, the others are designed as rectified linear units for passing the products without more multipliers and dividers. Our evaluations prove the synapse counts will be
less than any regular-weight-quantized BNNs, and from experimental results for a neuron-based block, the MTBF can be improved up to 126 times in the proposed infection-rate model.

D6.3 11:05-11:15

8-Bit Flexible Floating-Point Format for Accurate and Memory-Efficient Inference

Cheng-Wei Huang, Tim-Wei Chen, Yu-Da Zhu, and Juinn-Dar Huang
Institute of Electronics, National Yang Ming Chiao Tung University

Modern deep neural network (DNN) models generally require a huge amount of weight and activation values to achieve good inference outcomes. Those data inevitably demand a massive off-chip memory capacity/bandwidth, and the situation gets even worse if they are represented in high-precision floating-point formats. Effort has been made for representing those data in different 8-bit floating-point formats, nevertheless, a notable accuracy loss is still unavoidable. In this paper we introduce an extremely flexible 8-bit floating-point (FFP8) format whose defining factors – the bit width of exponent/fraction field, the exponent bias, and even the presence of the sign bit – are all configurable. We also present a methodology to properly determine those factors so that the accuracy of model inference can be maximized. The foundation of this methodology is based on a key observation – both the maximum magnitude and the value distribution are quite dissimilar between weights and activations in most DNN models. Experimental results demonstrate that the proposed FFP8 format achieves an extremely low accuracy loss of 0.1% ~ 0.3% for several representative image classification models even without the need of model retraining. Besides, it is easy to turn a classical floating-point processing unit into an FFP8-compliant one, and the extra hardware cost is minor.

D6.4 11:15-11:25

Value-Aware Error Detection and Correction for SRAM Buffers in Low-Bitwidth, Floating-Point CNN Accelerators

Min-Wei Chu, Jun-Shen Wu, Chi-En Wang, Ren-Shuo Liu
Department of Electrical Engineering, National Tsing Hua University

Low-power CNN accelerators are a key technique to enable the future artificial intelligence world. Dynamic voltage scaling (DVS) and dynamic voltage frequency scaling (DVFS) are essential lowpower strategies for low-power CNN accelerators, but they are bottlenecked by on-chip SRAM, which can exhibit stuck-at (SA) faults at a rate as high as 0.1% when the supply voltage is lowered to, e.g., 0.5 V. Although this issue has been studied in CPU cache design, their solutions are tailored for CPUs instead of CNN accelerators, which inevitably incur unnecessary design complexity and SRAM capacity overhead.
To address this issue, we analyze the impacts of SA faults in different SRAM positions and different SA types, i.e., stuck-at-one (SA1) and stuck-at-zero (SA0), for our targeting low-bitwidth, floatingpoint (LBFP) CNN accelerators. The analysis results lead us to the error detecting and correcting mechanisms that prioritize fixing SA1 appearing at SRAM positions where the exponent bits of LBFP are stored. The evaluation results show that our proposed mechanisms can help to push the voltage scaling limit down to a voltage level with 0.1% SA faults (e.g., 0.5 V).

Variational Channel Noise Pruning and Mixed-Precision Quantization

Wan-Ting Chang, Chih-Hung Kuo and Li-Chun Fang
Department of Electrical Engineering, National Cheng Kung University

Model compression is one of the most effective ways to perform applications of neural networks on mobile devices. In this work, we propose a compression framework of pruning and mixed-precision quantization based on channel noise. Our method uses the variational inference technique to optimize Bayesian Deep Neural Networks, which can be more reasonable and stable for compression. We conduct our experiment on CIFAR10 with VGG16. For the pruning scheme, we achieve 7.2x parameter saving and 1.72x MAC compression without noticeable loss in accuracy on CIFAR-10 with VGG16. Further quantizing the network, there is 62.8x parameter saving at the loss of 0.64% accuracy.

An Efficient Implementation of Convolutional Neural Network with CLIP-Q Quantization on FPGA

Wei Cheng, Ing-Chao Lin, and Yung-Yang Shih
Department of Computer Science and Information Engineering, National Cheng Kung University

To reduce the storage requirement for convolutional neural network (CNN), Compression Learning by InParallel Pruning-Quantization (CLIP-Q) was proposed. CLIP-Q reduces a vast amount of storage requirement by using few bits to represent a weight while keeping accuracy close to the full precision model. This feature makes CLIP-Q suitable for a hardware CNN accelerator. The original version of CLIP-Q uses full precision for data operations and input storage. In this work, we implemented a CNN accelerator that takes advantage of CLIP-Q, and 8-bit fixed-point inputs and 2-bit fixed-point weights are used. Our accelerator maintains the same model accuracy of full precision CNN in CIFAR-10 and CIFAR-100 dataset. The proposed CNN accelerator reaches higher computing and energy efficiency than others with a 5x5 reconfigurable convolutional array.
Oral D7

AI Architecture Design in Multimedia Application

8月6日（五）10:45-11:45
Session Chair: 林光浩教授 | 国立虎尾科技大学 電機工程系
李佩君教授 | 国立暨南国际大学 電機工程学系

D7.1 10:45-10:55

Defense Against Adversarial Attacks using FFT

Pin-Han Lin, Shao-Yu Fu, Chih-Wei Liu
Department of Electronics Engineering, National Yang Ming Chiao Tung University

It is pointed out that if the intentionally-crafted, human imperceptible perturbations are added to the input image, and these images with adversarial perturbations have high confidence to fool Convolutional Neural Network (CNN). In this paper, we proposed using the Fast Fourier Transform (FFT) on image and removing some unimportant value via Binary Mask such that perturbation can be removed. In addition, we use the concept of filter bank to improve the problem of requiring multiple sets of Threshold for different adversarial attacks. Furthermore, since the adversarial perturbations are generated through training, we can effectively improve the accuracy of CNN by adding noise to destroy the structure of adversarial perturbations. Our defense mechanism is a part of image preprocessing, and it have no need to training, which can save lots of time. Our defense mechanism can effectively defense against whitebox attack, the Top-1 average accuracy can achieve over 60%, and black-box attack, the Top-1 accuracy can achieve 77%. Compare with state-of-the-art model-agnostic mechanisms, our defense mechanism has better accuracy for white-box attack and black-box attack and has less computational complexity.

D7.2 10:55-11:05

A CNN-GRU Speech Enhancement Architecture based on Noise Classification

Yong-Zhen Chen, Shang-Jang Ruan
Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology

We propose a noise classification based speech enhancement neural network for hearing aids. A noise classification model is trained to determine the noise type of the input noisy speech, and then the noise type information is sent to the speech enhancement model to learn respective noise reduction. The proposed speech enhancement neural network implements the denoising autoencoder (DAE) concept over the convolutional neural network - gated recurrent unit (CNN-GRU) architecture to reduce the background noises. Experimental results show that the proposed method not only achieved a better
performance in speech quality and intelligibility but also improved the memory storage requirement than other methods. The model size is reduced by 85% than other methods.

**D7.3** 11:05-11:15

**Environmental Sound Classifier using Inception-Dense Model on Smartphone**

Po-Jung Ting, Shann-Jang Ruan  
Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology

In recent years, environmental sound classification (ESC) is an important and challenging problem. In contrast to speech, sound events have noiselike nature and may be produced by a wide variety of sources. In this paper, we propose to use the Inception-Dense model for ESC tasks. Our network architecture uses stacked Inception-Dense blocks to extract multilevel feature maps, and utilizing depth-wise separable convolution minimizes model complexity. In this experiment, the proposed model uses separately three different input sizes to explore the length of sound affects model performance. Furthermore, we apply mixup and data augmentation to ESC tasks. The experiment was conducted on the UrbanSound8K dataset. Our experimental results demonstrated that our ESC system had achieved 83.02% on UrbanSound8K.

**D7.4** 11:15-11:25

**Design and Validation of a Digital Performance-Power Management Mechanism with Dual-Voltage Adjustment Technique**

Ching-Hwa Cheng\(^1\), Jiun-In Guo\(^2\)  
\(^1\)Dept. of ECE, Feng-Chia University, \(^2\)Dept. of EE Nat, Chiao-Tung University

Voltage adjustment is the easiest way to reduce power consumption. Our design uses dual voltages to reduce power consumption without performance degradation. A built-in digitalized power management (DPM) is proposed to design a dual-Vdd performance-power adjustable system. This flexible voltage assignment allows the chip performance and power can be dynamically adjusted during circuit operation. To support the DPM mechanisms, the power-switch circuit is developed to support multiple performance-power operation modes for digital designs. The proposed mechanism allows the chip performance and power consumption can be flexibly adjusted during circuit operation. Dual-Vdd technology is used to design a performance-power multi-mode video decoder (MMVD). The MMVD test chip proves that it can efficiently reduce power consumption without causing circuit delay time and die’s area increases. The test chip is successfully validated after system integration and obtains about 24% reduction in power consumption, which is better than that from the same design using a single supply voltage.
VLSI Design Based on Least Square Estimation Method for Back-Mapping of Barrel Distortion Correction

Kun-Ze Su¹, He-Sheng Chou¹, Jr-Yu Lin¹, Shih-Yao Ke¹, Chiung-An Chen², Ting-Lan Lin³, Wen-Shen Lo⁴ and Shih-Lun Chen¹,⁴
¹Department of Electronic Engineering, Chung Yuan Christian University
²Department of Electrical Engineering, Ming Chi University of Technology
³Department of Electronic Engineering, National Taipei University of Technology
⁴Center for Internet of Things and Intelligent Cloud, Chung Yuan Christian University

This paper proposes a real-time correction VLSI implementation of back mapping for barrel distortion. The mathematical model is based on an inverse mapping expansion polynomials approximated by odd-order polynomials. In order to reduce the calculation process, the polynomial is determined to be a fourth-degree polynomial using two coefficients by estimating the total error value after correction. The finite state machine is used to control data transmission to reduce the demands of chip pads. Although the number of bits of the coefficients is increased by two times compared with the previous circuit design, it increased the accuracy of the results.

Monocular Wide-angle Vision Obstacle Detection and Recognition System for Rear-View Camera Applications

Szu-Hong Wang¹, Wei-Chien, Yuan², Tsan-Hung Chung², Shih-Chang Hsia²
¹Bachelor Program in Interdisciplinary Studies, National Yunlin University of Science & Technology
²Department of Electronic Engineering, National Yunlin University of Science & Technology

Whether it is fatigue driving or accident caused by inattentive driving, which provides feedback to driver’s surroundings while driving and alerts driver's attention in a short time. This paper was proposed an obstacle detection and recognition system based on monocular wide-angle vision images. First, the wide-angle lens image is corrected for barrel distortion. Then, various obstacle detection algorithms are combined to obtain the location and distance of the obstacle and further identify the obstacle class by CNN recognition. Finally, the system was implemented on the Google Coral Dev board platform for inference testing. The experimental results reveal that Xception model has 95.2% recognition rate and 20fps real-time application.