1b-16b Variable Bit Precision DNN Processor for Emotional HRI System in Mobile Devices

Chang Hyeon Kim[1], Jin Mook Lee, Sang Hoon Kang, Sang Yeob Kim, Dong Seok Im and Hoi-Jun Yoo[a]

Korea Advanced Institute of Science and Technology, School of Electrical Engineering (KAIST)

E-mail : [1]ch.kim@kaist.ac.kr

***Abstract* - We propose an energy-efficient DNN processor with the proposed look-up-table-based processing engine (LPE) and near-zero skipper. A CNN-based facial emotion recognition model and an RNN-based emotional dialogue generation model are integrated for the natural human-robot interaction (HRI) system, and it is evaluated by the proposed processor. LPE supports 1 to 16 bit variable weight bit precision, and it achieves 57.6% and 28.5% lower energy consumption than the conventional multiplier-accumulator (MAC) units in 1-16 bit weight precision. Furthermore, the near-zero skipper reduces 36% of MAC operations and consumes 28% lower energy consumption in facial emotion recognition tasks. Implemented in 65 nm CMOS process, the proposed processor occupies 1784×1784 μm² areas and dissipates 0.28 mW and 34.4 mW at 1 frame-per-second (fps) and 30 fps facial emotion recognition tasks.**

***Keywords*— Deep learning, Deep learning ASIC, Deep neural network, Emotion recognition, Mobile deep learning**

I. INTRODUCTION

The recent development of deep learning technology enables machines to recognize the human user's emotion accurately with facial expressions [1] and dialogues [2]. Recent studies [3-5] try to adopt emotion recognition into mobile devices, which allows machine to understand user's intend and enables more natural human-robot interaction. For natural human-robot interaction (HRI), we newly introduce an emotional HRI system for mobile devices, as shown in Fig. 1. It consists of three steps, face detection/alignments for face region-of-interest (RoI) generation, CNN-based facial emotion recognition (FER), and RNN-based emotional dialogue generation (EDG). FER generates the user's emotion and it is fed to EDG with the user's speech as an input. EDG performs natural language processing and outputs different dialogues corresponding to the user's emotional state. We optimized weight bit-precision of both FER's and EDG's DNN models in order to realize real-time operation. The optimized CNN model is quantized to 16 bit fixed point weights in the input layer and 1 bit fixed point weights in the other layers, achieving <1% accuracy degradation compared to 32 bit floating point weights. Besides, the EDG model is trained using the reinforcement learning method which uses 4 bit fixed point weights.

In this paper, an energy-efficient DNN processor has been integrated, which supports variable weight bit precisions in order to combine two different DNN models into a battery-powered mobile device. We propose variable weight bit precision DNN processor with an energy-efficient look-up-table based processing engine (LPE). It fully supports 1 to 16 bit weight precision with binary multiplication optimized LPE and bit-serial shifter. It consumes 57.6% and 28.5 % less energy consumption compared with a conventional multiplier. Furthermore, it supports near-zero skipping, which reduces the average 36% multiplier-accumulator (MAC) operation and 28% energy consumption in facial emotion recognition tasks.

The rest of this paper is organized as follows. In Section II, the overall architecture and the processing details of building blocks are described. Section III shows the implementation results and measurement results. Section VI provides a conclusion.

II. DETAILED BUILDING BLOCKS

A. Overall Architecture

Fig. 2 shows the overall architecture of the proposed DNN processor. It consists of the preprocessing core, and the LUT-based PE core (LP Core), which are connected with a network-on-chip interface for communications. The pre-processing core performs face detection and alignment for face RoI generation, and the face RoI is transferred to LPE core. Two 6KB output memories (OMEMs) and 36KB weight memory (WMEM) are integrated into the LP core. When the activation of DNN is fed into the activation buffer, the Near-zero detector bit-shifts activation data and skips the MAC operation of the activations which are smaller than the predefined threshold value. Four LPE clusters perform DNN processing of the different input channels. Each LP cluster consists of four LPEs, and each LPE includes 8-entry look-up-table of four input activations and generates twelve different output data using twelve 8-to-1 multiplexers. The detailed explanation of LPE is described later. Twelve outputs of four LPs and LP clusters are accumulated by twelve 4-way add/sub trees. The bit-serial shifter performs

a. Corresponding author; hjyoo@kaist.ac.kr

Manuscript Received May. 17, 2020, Accepted Jun. 18, 2020

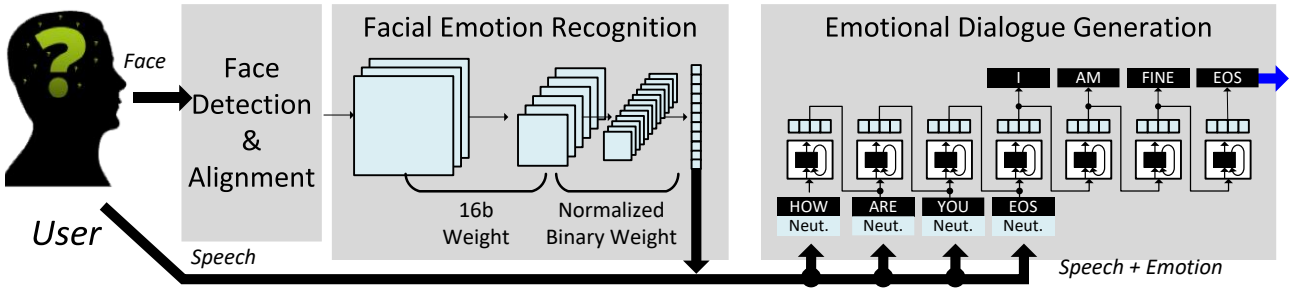


Fig. 1. Emotional HRI System with Facial Emotion Recognition and Emotional Dialogue Generation Model

bit-serial multiplication for weight-bit scalable DNN processing. The outputs are stored into OMEMs, and ping-pong accumulation engine accumulates the partial-summations of DNN.

B. LUT-based Processing Engine (LPE)

Fig. 3 explains the binary weight multiplication of LPE. The outputs of LPEs are accumulated and shifted by the bit-serial shifter for bit-serial multiplication. In the case of 1 bit weight, 1-weights and 0-weights represent addition and substitution of activations, respectively. When LPE fetches new input activations, the LPE updates the LUT with all the combinations of the bit-serial multiplications between four input activations and binary weights (A-step: LUT update). Since each input activation is multiplied with output channel size (C_O) and kernel size (k) of weights, the entire Co×k binary weight multiplications can be obtained by simple indexing of the LUT (B-step: Calculation). To further reduce the number of entries of the LUT, we exploit the characteristic of 2's complement. The total of 16 combinations can be divided into two parts, as shown in Fig. 3. The physical LUT contains 8-entries and stores multiplication results of weight's MSB equals one. The logical LUT represents the multiplication results of weight's MSB equals zero, and is the full inversion of the physical LUT that can be replaced by the half entries, eight 16 bit registers, an inverter, and a multiplexer. Moreover, LPE cluster controller reconfigures the data path of 4-way add/sub trees. It calculates and stores the pre-calculated LUT entries during the A-step, and accumulates and generates the indexed outputs of four LPEs at B-step.

In Fig. 4, we simulated and analyzed the energy consumption of LPE by the number of input activations per LUT size (E) at 200 MHz, 1.1 V operating condition. The normalized energy consumption is measured with the same external memory bandwidth and the same throughput condition. When the E becomes larger, the accumulation power of LUT outputs becomes smaller, but the entry size of LUT becomes exponentially larger. In our simulation, the optimal size of E is four that we designed LUT with 8-entries. In addition, because it is more efficient to index a single LUT with multiple weights, we measured energy consumption by

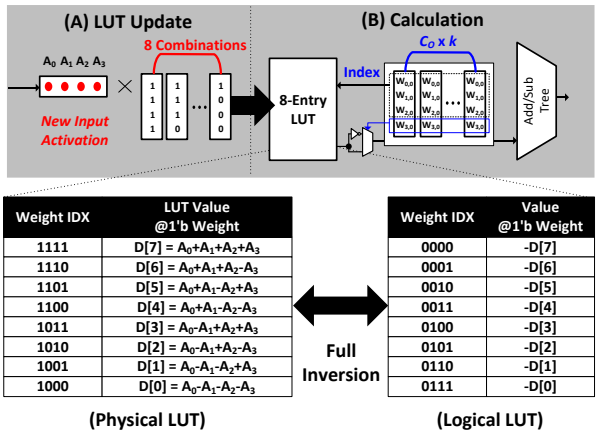


Fig. 3. Detailed Binary Weight Multiplication of LPE

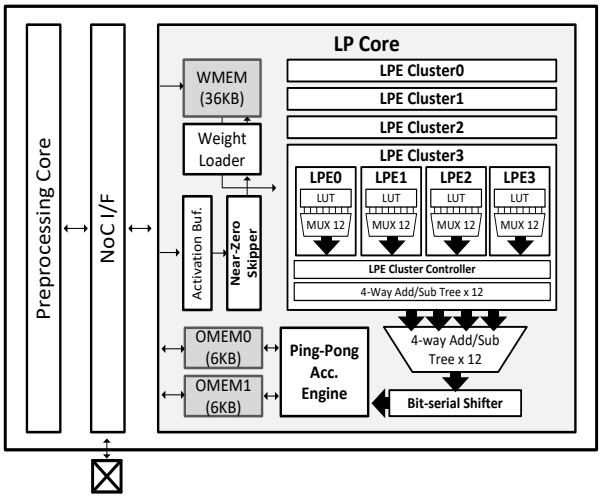


Fig. 2. Overall Architecture

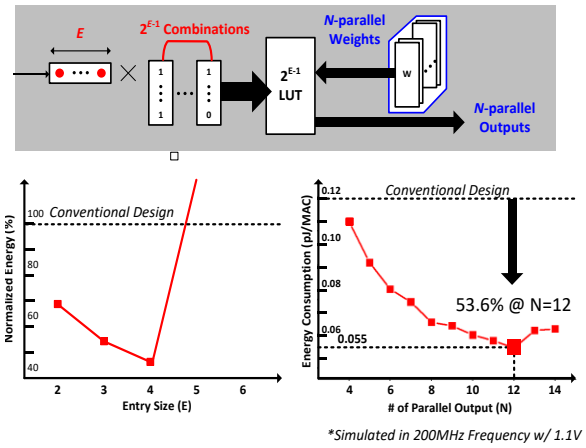
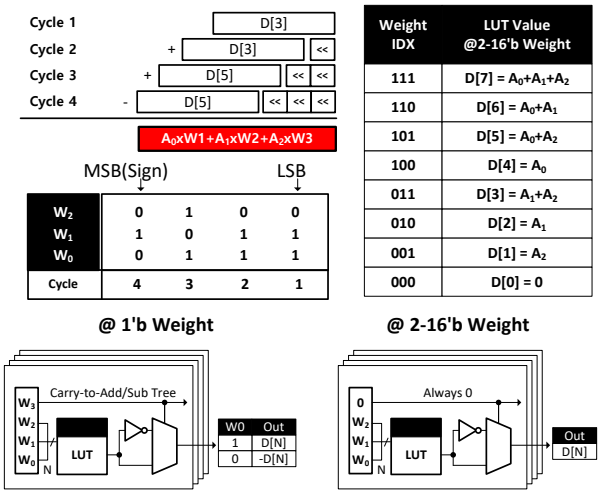


Fig. 4. Energy Consumption of LPE with various number of entry size and parallel outputs.



Weight IDX	LUT Value @2-16'b Weight
111	D[7] = A_0+A_1+A_2
110	D[6] = A_0+A_1
101	D[5] = A_0+A_2
100	D[4] = A_0
011	D[3] = A_1+A_2
010	D[2] = A_1
001	D[1] = A_2
000	D[0] = 0

Fig. 5. 2b-to16b bit-serial multiplication using LPE results

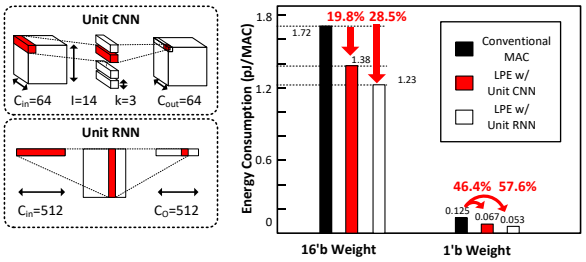


Fig. 6. Energy Consumption Comparison with Conventional MAC and LPE in CNN and RNN

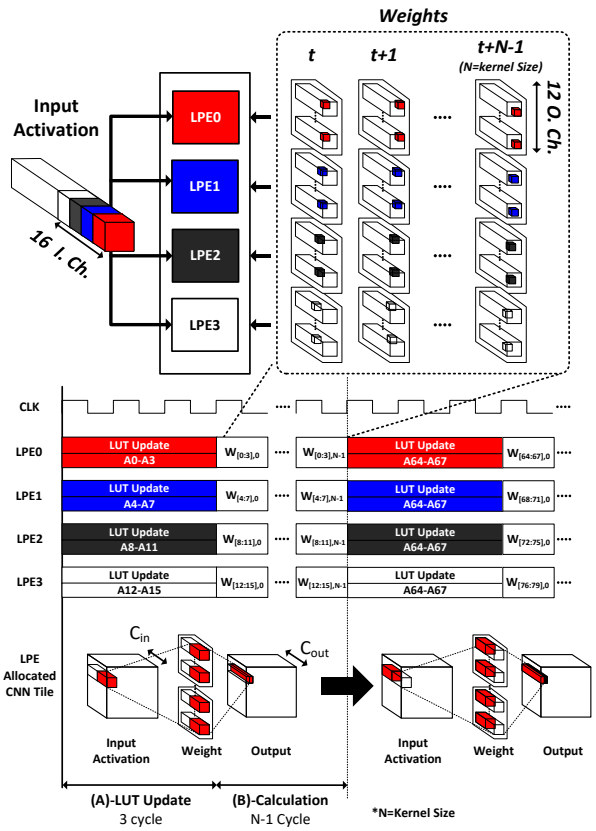


Fig. 7. Spatial and Temporal Data Mapping of LPEs

the number of parallel binary weight multiplications of LPE (N). Finally, we determined the optimal size of E and N as four and twelve. Therefore, twelve multiplexers are integrated into an LPE, and an LPE cluster accumulates 4×12 parallel outputs with twelve 4-way add/sub tree.

With optimization of LPE, the bit-serial multiplication method is adopted in order to realize the variable weight bit precision of DNN. Fig. 5 explains the PE configuration at N-bit weights. It sequentially calculates the addition and substitution of input activations with multiplication results of binary weights from the LSB to the MSB for N-cycles. During the 2-16 bit multiplication, LPE controller reconfigures the LPE that LUT stores all combinations of three input activations as shown in Fig. 5. Then, Fig. 6 describes the energy consumption of LPE during the CNN and RNN operation. As a result, LPE unit reduces the energy consumption of 19.8% in CNN and 28.5% in RNN compared with the conventional MAC units.

Fig. 7 shows the spatial and temporal data mapping of LPEs for CNN processing. In beginning, an LPE cluster fetches 16 input activations, which are the pixels of the same coordinate but 16 different input-channels. Then, each LPE updates LUT with 4 input activations (A-step). It takes 3-cycles. After the LUT update step, LPEs fetch weights and outputs the values of LUT by weights. At every cycle, LPE fetches 4×12 weight, which are the kernels of the 4 different input channels, 12 different output channels, but the same coordinate. So, each LPE outputs 12 LUT values at every cycle, and the 4-way add/sub-tree in LPE cluster accumulates the 4×12 outputs to 12 output. It takes kernel-size cycles. After kernel-size cycles, there are no kernels to multiply with input activations in LPE, that LPE fetches new input activation in the same coordinate but 16 different input-channels and repeat the above procedure. What if the input activations within the same coordinate are fully processed, the LPE fetches the new input activations in the next coordinate.

The near-zero skipper is explained in Fig. 8. While the input buffer fetches new input activations, near-zero-skipper monitors input buffer and blocks the activations, smaller than the threshold value. This threshold value is programmable that can be varied through different layers of DNN. It simply bit-shift the data in input-buffer to the right, and what if the shifted data is 0, it blocks the activations. In the case of negative numbers, it inverses full-bits of the data before bit-shift. Moreover, it skips the corresponding weights of the blocked input activations. While processing the FER

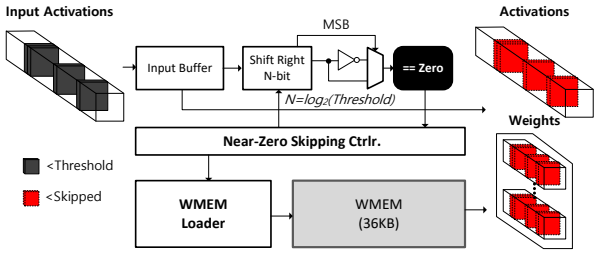


Fig. 8. Near-zero Skipping Method

normalized binary weight CNN with FER2013 dataset [5], the average skipping ratio is 36% with the threshold of 4 that the overall power consumption is reduced by 28%.

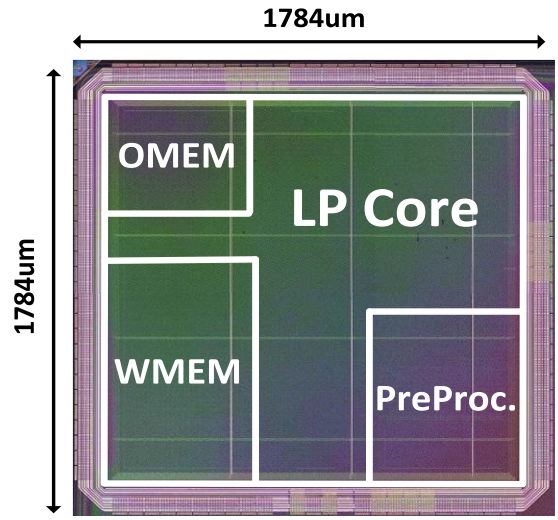


Fig. 9. Chip Micrograph

III. IMPLEMENTATION AND MEASUREMENT RESULTS

Fig. 9 shows the chip micrograph. The proposed processor is fabricated with 65 nm 1P8M Logic CMOS process with 1784×1784 μm^2. It operates from 0.67-1.1V supply voltage with 5-200 MHz clock frequency range. In the case of facial expression recognition with weight-bit precision optimized CNN, it consumes 0.28 mW and 34.4mW at 1fps and 30fps, respectively. For the emotional dialogue generation tasks, it generates different outputs due to the user's emotional states. It consumes 56 mW and takes 1.36-second latency with four layered 8 bit weight GRU model. The peak CNN power efficiency is measured as 13.6 TOPS/W at 0.66V, 5 MHz operating condition, and peak RNN/FC power efficiency is measured as 15.7 TOPS/W at the same condition with binary weight bit precision.

Fig. 10 describes the results of facial emotion recognition with FER2013 dataset and the performance summary of the proposed processor. Fig. 11 illustrates the performance of the processor during testing the emotional dialogue generation using RNN model. As a result, the proposed processor successfully demonstrates both facial emotion recognition and emotional dialogue tasks by accelerating the CNN and RNN model efficiently.

III. CONCLUSIONS

We propose an energy-efficient DNN processor with a LUT-based processing engine and near-zero skipper. A CNN-based facial emotion recognition and an RNN-based emotional dialogue generation model is integrated for the natural HRI system and tested with the proposed processor. LPE supports 1-16 bit variable weight bit precision with and 57.6% and 28.5% lower energy consumption than conventional MAC arithmetic units for 1 bit and 16 bit weight precision. Also, the near-zero skipper reduces 36% of MAC operation and consumes 28% lower energy consumption for facial emotion recognition tasks. Implemented in 65nm CMOS process, the proposed processor occupies 1784×1784 μm^2 areas and dissipates 0.28 mW and 34.4 mW at 1fps and 30fps in CNN-based facial emotion recognition task with face detection and face alignment. Furthermore, for the RNN-based emotional dialogue generation task, it consumes 56mW with 1.36-second latency. In conclusion, the 1b-to16b fully variable weight bit precision low-power DNN processor for <100mW natural HRI system is successfully realized for mobile devices.

Happy

Neutral

Facial Expression Recognition: 71.8% FER2013										
Layers	C1	C2	C3	C4	C5	C6	C7	C8	FC1	FC2
Kernel	7x7	3x3								
CH. In	3	64	64	128	256	256	256	256	1024	1024
CH. Out	64	64	128	128	256	256	256	256	1024	7
W. Bit.	16	1 (Binary)								

Performance

Precision	Feature – 16b integer
	Weight – 16b (input layer), 1b (others)
Avg. Skip Ratio	36% @ Threshold=4
Power	0.28mW @ 5MHz, 0.66V (1 fps)
	34.4mW @ 100MHz, 1.1V (30 fps)

Fig. 10. Facial emotion recognition test with FER2013 dataset and performance summary

Dialogue 1.

U Speech	Excuse me.	
U Emotion	Neutral	Fear
Response	Yes. What can I do for you ?	Yes. May I help you ?

Dialogue 2.

U Speech	I am so sad	
U Emotion	Sad	Happy
Response	What did you find ?	What did you say ?

Emotional Interaction RNN Model

# of Layers	*Words / Sentence*	*Word Dim.*	*Emotion Dim.*	*Size of Params*	*# of OPs*
4	*10 words*	*512*	*7*	*61.65 MB*	*0.26 G*

Performance

Precision	Feature – 16b integer
	Weight – 4 integer
Latency	1.36 seconds @ 200MHz, 1.1V
Power	56 mW @ 200MHz, 1.1V

Fig. 11. Emotional dialogue generation test and performance summary

ACKNOWLEDGMENT

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning under Grant NRF-2018R1A2A1A19023099.

REFERENCES

[1] Li, Shan, and Weihong Deng. "Deep facial expression recognition: A survey." arXiv preprint arXiv:1804.08348 (2018).

[2] Hazarika, Devamanyu, et al. "Conversational memory network for emotion recognition in dyadic dialogue videos." Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018.

[3] Jiang, Shiqi, et al. "Memento: An Emotion-driven Lifelogging System with Wearables." ACM Transactions on Sensor Networks (TOSN) 15.1 (2019): 8.

[4] Suja, P., and Shikha Tripathi. "Real-time emotion recognition from facial images using Raspberry Pi II." 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2016.

[5] C. Pierre-Luc, et al., “FER-2013 face database,” Universitde Montral, 2013.

Chang Hyeon Kim (S’16) received the B.S. and M.S. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2014 and 2016, respectively, where he is currently pursuing the Ph.D. degree.

His current research interests include low-power system-on-chip design, especially focused on parallel processor for artificial intelligence and machine learning algorithms.

Jin Mook Lee (S′15) received the B.S. degrees in electrical engineering from Hanyang University, Seoul, South Korea, in 2014, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2016, where he is currently pursuing the Ph.D. degree.

His current research interests include energy-efficient deep learning inference/training accelerator ASIC design, embedded deep learning platform design and verification, embedded system development with FPGA programming, and deep learning algorithm for sequence recognition.

Sang Hoon Kang (S′16) received the B.S. and M.S. degrees with the School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2016. He is currently pursuing the Ph.D. degree with the School of Electrical Engineering.

His current research interests include low-power vision system-on-chip design and deep learning processor design.

Sang Yeob Kim (S’18) received the B.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2018, where he is currently pursuing the M.S. degree.

His current research interests include low-power system-on-chip design, deep neural network accelerators, and machine learning algorithms for deep learning.

Dong Seok Im (S’18) received the B.S. degree in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2018. Currently, he is working toward the M.S. degree in electrical engineering at the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea.

His current research interests include energy-efficient deep learning SoC design & intelligent vision system.

Hoi Jun Yoo (Fellow, IEEE) graduated from the Electronic Department, Seoul National University, Seoul, South Korea, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1985 and 1988, respectively.

He has served as a member for the Executive Committee of ISSCC, Symposium on VLSI, and A-SSCC, the TPC Chair for the A-SSCC 2008 and ISWC 2010, the IEEE Distinguished Lecturer from 2010 to 2011, the Far East Chair for the ISSCC from 2011 to 2012, the Technology Direction Sub-Committee Chair for the ISSCC in 2013, the TPC Vice Chair for the ISSCC in 2014, and the TPC Chair for the ISSCC in 2015.