Final Program

April 17, 2024 (Japan Standard Time)

13:00-13:20 **Session I: Welcome & Opening Remarks**
Chairs: Yuki Kobayashi (NEC), Hiroki Matsutani (Keio Univ.)

Fumio Arakawa (Chair of the Organizing Committee)
Tadao Nakamura (Chair of the Steering Committee)
Gabriel Southern (Chair of IEEE/CS TCMM)
Tetsuya Kawanishi (President of IEICE/ES)

13:20-14:10 **Session II: Keynote Presentation 1**
Chairs: Takanori Ueda (IBM Japan), Hiroki Matsutani (Keio Univ.)

13:20-14:10 **Accelerating AI with Analog In-Memory Computing**
Stefano Ambrogio (IBM Research)
Abstract: The last decade has witnessed the pervasive spread of AI in a variety of domains, from image and video recognition and classification to speech and text transcription and generation. In general, we have observed a relentless run towards larger models with huge number of parameters. This has led to a dramatic increase in the computational workload, with the necessity of several CPUs and GPUs to train and inference neural networks. Therefore, improvements in the hardware have become more and more essential. To accommodate for improved performance, in-memory computing provides a very interesting solution. While digital computing cores are limited by the data bandwidth between memory and processor, computation in the memory avoids the weight transfer, increasing power efficiency and speed. The talk will describe a general overview, highlighting our own 14-nm chip, based on 34 crossbar arrays of Phase-Change Memory technology, with a total of around 35 million devices. We demonstrate the efficiency of such architecture in a selection of MLPerf networks, demonstrating that Analog-AI can provide superior power performance with respect to digital cores, with comparable accuracy. Then, we provide guidelines towards the next steps in the development of reliable and efficient Analog-AI chips, with specific focus on the architectural constraints and opportunities that are required to implement larger and improved Deep Neural Networks.

14:10-14:30 **Break**

14:30-15:20 **Session III: Keynote Presentation 2**
Chairs: Tohru Ishihara (Nagoya Univ.), Hiroe Iwasaki (TUAT)

14:30-15:20 **Energy-Efficient Heterogeneous Photonics for Next Generation AI and Hardware Accelerators**
Stanley Cheung (Hewlett Packard Enterprise)
Abstract: As Moore's law, Dennard scaling, and the Von-Neumann bottleneck continually push the boundaries of conventional computing, there arises a need to seek alternative architectures, systems, and devices. On the other hand, AI/ML applications in mega-data centers and high-performance post-exa-scale computing will require an interconnect solution that scales in bandwidth and energy efficiency. In this talk, I will first introduce our unique heterogeneous III-V quantum-dot-on-Si platform for high-

performance optical interconnects and post-exa-scale computing systems. Key devices and architectures will be discussed that significantly improve bandwidth-density/energy-efficiency metrics by orders of magnitude. This lays the groundwork for exploring photonic in-memory neuromorphic computing leveraging our recent breakthroughs in non-volatile photonic devices: photonic memristors & photonic charge-trap memory.

15:20-15:40 **Break**

15:40-16:50 **Session IV: FPGAs and Reconfigurable Processors**

Chairs: Kotaro Shimamura (Hitachi), Ryuichi Sakamoto (Tokyo tech.)

15:40-16:05 **MRCA: Multi-grained Reconfigurable Cryptographic Accelerator for Diverse Security Requirements**

Pham Hoai Luan[1], Hai Hau Nguyen[2], Vu Trung Duong Le[1], Thi Diem Tran[2], Tuan Hai Vu[1],Thi Hong Tran[3], and Yasuhiko Nakashima[1] ([1]Nara Institute of Science and Technology, Nara, Japan, [2]University of Information Technology - VNUHCM, Vietnam, [3]Osaka Metropolitan University, Osaka, Japan)

16:05-16:20 **SLMLET: A RISC-V Processor SoC with Tightly-Coupled Area-Efficient eFPGA Blocks**

Takuya Kojima[1], Yosuke Yanai[2], Hayate Okuhara[3], Hideharu Amano[2], Morihiro Kuga[4], and Masahiro Iida[4] ([1]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan, [2]Department of Information and Computer Science, Keio University, Yokohama, Japan, [3]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, [4]Department of Computer Science and Electrical Engineering, Kumamoto University, Kumamoto, Japan)

16:20-16:35 **FPGA-based Low Power Acceleration of HARK Sound Source Localization**

Zirui Lin[1], Katsutoshi Itoyama[1,2], Kazuhiro Nakadai[1], and Hideharu Amano[3] ([1]Dept. of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, Tokyo, Japan, [2]Honda Research Institute Japan Co., Ltd., Saitama, Japan, [3]Dept. of Information and Computer Science, Keio University, Kanagawa, Japan)

16:35-16:50 **ISP Parameter Optimization and FPGA Implementation for Object Detection in Low-Light Conditions**

Kento Mishima[1], Naoya Niwa[1], Kazutoshi Wakabayashi[2], and Hiroe Iwasaki[1] ([1]Tokyo University of Agriculture and Technology, Tokyo, Japan, [2]Systems Design Lab, Tokyo, Japan)

16:50-17:00 **Break**

17:00-17:40 **Session V: Special Invited Presentation**

Chairs: Fumio Arakawa (Univ. of Tokyo), Yuki Kobayashi (NEC)

17:00-17:40 **Achieving the most energy-efficient compute fabric for ML and HPC using of thousands of RISC-V cores**

Dave Ditzel (Esperanto Technologies)

Abstract: Esperanto Technologies will describe high performance compute fabrics for machine learning and high performance computing. The fabric is based on clusters of 32 custom designed RISC-V processors called ET-Minions, each of which as an attached vector unit and attached tensor unit. This fabric and software to drive it was implemented in 7nm with over a thousand ET-Minions on a single die. Future implementations will utlize a chiplet based approach to be able to grow the compute

fabric across multiple die. Chiplet designs for 4nm and 2nm are discussed, allowing for up to 4096 ET-Minions and performance levels that could exceed high end GPUs. The implementation approach is focused on energy effiiency though low-voltage techniques, and and performance per watt projections for different data types are presented. The conclusion is that an array of general purpose processsors based on RISC-V can provide one of the most energy-efficient compute fabrics.

April 18, 2024 (Japan Standard Time)

9:00-9:50 **Session VI: Keynote Presentation 3**

Chairs: Yukinori Sato (Toyohashi Univ. of Tech.), Yoshio Hirose (Fujitsu)

9:00-9:50 **Processing-in-Memory: from Technology to Products**

Kyomin Sohn (Samsung Electronics)

Abstract: The traditional computing architecture represented by Von Neuman maintains a simple memory hierarchy that is still in use to this day. However, the strong demand for computing power, which began with big data and AI applications, is evolving from a new memory hierarchy. In particular, the emergence of large language models in generative AI, requires higher bandwidth and higher capacity of memory. This talk provides an explanation of the key concepts of in-memory computing, which is referred as CIM (compute-in-memory) or PIM (processing-in-memory). CIM technology enables a memory array as a processing unit using inherent feature of multiplication between wordline and bitline. In contrast, PIM technology utilizes internal memory bandwidth by allocating processing units near memory array and activating them simultaneosly. The concept of PIM technology is already proven by HBM2-PIM and GDDR6-AiM from the major DRAM vendors. It is the time to apply this technology to the mass-produced DRAM products. In the system and application having low operational intensity of data, PIM technology looks very attractive to overcome the limitation. However, there are obstacles to apply PIM technology to the conventional DRAM directly. The challenges will be discussed and the several suggestions will be given. From the PIM technology to the DRAM products with PIM technology, this journey will go on.

9:50-10:10 **Break**

10:10-11:00 **Session VII: SLAM Processors**

Chairs: Yukihiro Sasagawa (Socionext), Shinya Takamaeda-Yamazaki (Univ. of Tokyo)

10:10-10:35 **A Low-power and Real-time Semantic LiDAR SLAM processor with Point Neural Network Segmentation and kNN Acceleration for Mobile Robots**
Jueun Jung[1], Seungbin Kim[1], Bokyoung Seo[1], Wuyoung Jang[1], Sangho Lee[1], Jeongmin Shin[1], Donghyeon Han[2], and Kyuho Jason Lee[1] ([1]Ulsan National Institute of Science and Technology (UNIST), Ulsan, Republic of Korea, [2]Massachusetts Institute of Technology, Cambridge, MA, USA)

10:35-11:00 **A Low-power and Real-time Neural-Rendering Dense SLAM Processor with 3-Level Hierarchical Sparsity Exploitation**
Gwangtae Park, Seokchan Song, Haoyang Sang, Dongseok Im, Donghyeon Han, Sangyeob Kim, Hongseok Lee, and Hoi-Jun Yoo (School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea)

11:00-12:00 **Session VIII: Poster Lightning Talk**

Chair: Koji Hashimoto (Fukuoka Univ.)

Poster 1 **Defragmentation-based Efficient Allocation on On-Chip Scratchpad Memory for Lightweighted Microncontrollers**
Gihyeon Jeon and Daejin Park (School of Electronic and Electrical Engineering, Kyungpook National University)

Poster 2 **Exploring the Effectiveness of GPU Time-Slicing under Gang Scheduling in HPC systems**
Akane Koto, Jun Kato, and Masahiro Miwa (Fujitsu Ltd., Kawasaki, Japan)

Poster 3 **Implementation of Batch Matrix Multiplication for Large Language Model Training on A64FX CPUs**
Hiroki Tokura, Takumi Honda, Kentaro Kawakami, Masafumi Yamazaki, Sameer Deshmukh, and Koichi Shirahata (Fujitsu Limited, Kawasaki, Japan)

Poster 4 **Accelerating Elementary Functions Computation Using a CGRA : A Case Study on Exponential Function**
Toshihiro Shimizu, Hiroki Tokura, Yousuke Nakamura, and Naoto Fukumoto (Fujitsu Limited, Kawasaki, JAPAN)

Poster 5 **A Method for Comparing Node Performance for the Detection of Malfunctioning Nodes**
Mari Yamaoka, Akira Hirai, and Akihiko Kasagi (Computing Laboratory, Fujitsu Limited. Kawasaki, Japan)

Poster 6 **Approximate Random Weight Generator & CiM Integration for Neuromorphic Computing**
Naoko Misawa, Shunsuke Koshino, Ruhui Liu, Chihiro Matsui, and Ken Takeuchi (Department of Electrical Engineering and Information Systems, Graduate School of Engineering The University of Tokyo, Tokyo, Japan)

Poster 7 **Compact Edge Vision Transformer Design for Non-volatile Computation-in-Memory**
Tao Wang, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi (Department of Electrical Engineering and Information Systems, Graduate School of Engineering The University of Tokyo, Tokyo, Japan)

Poster 8 **Real-Time Ray Tracing Acceleration Using Multi-FPGA Parallel Rendering**

Ji Young Kim[1], Yun Ho Han[1] and Woo Chan Park[1,2] ([1]Sejong University, Republic of Korea, [2]Exarion Inc., Republic of Korea)

Poster 9 **A Framework for Real-time Sound Tracing Based on an FPGA**

Eunjae Kim[1], Sukwon Choi[2] and Woo-Chan Park[1,2] ([1]Exarion Inc., Republic of Korea, [2] Sejong University, Republic of Korea)

Poster 10 **Performance Evaluation of Sound Tracing Based on an FPGA**

Sukwon Choi[1], Jungwoong So[1], Uijun Kim[1] and Woo-Chan Park[1,2] ([1]Sejong University, Republic of Korea, [2]Exarion Inc., Republic of Korea)

Poster 11 **Energy-Efficient Core Configuration Policies for Heterogeneous Multi-Core Systems**
Yifan Jin[1], Jiaheng Liu[2], Keichi Takahashi[2], Yoichi Shimomura[2], and Hiroyuki Takizawa[2] ([1]Graduate School of Information Sciences, Tohoku University, Sendai, Japan, [2]Cyberscience Center, Tohoku University, Sendai, Japan)

12:00-13:00 Lunch

13:00-14:30 Session IX: Special Invited Lecture 1

Chair: Jun Shiomi (Osaka Univ.)

13:00-14:30 **Navigating Aging Realities: Integrating Reliability into Cutting-Edge Computing Systems**
Yu-Guang Chen (National Central Univ., Taiwan)

Abstract: As CMOS technology undergoes further scaling down, the emergence of the aging effect poses a significant threat to the lifetime reliability of computing systems, with the potential to induce performance degradation or timing failures. Effectively addressing these challenges necessitates a comprehensive understanding of how aging effects impact the outcomes of modern computing systems, prompting the development of methodologies dedicated to aging detection, mitigation, and tolerance. This presentation aims to provide a concise overview of major aging effects and their root causes. Subsequently, a deeper exploration will unfold, focusing on two pivotal aspects: (1) Aging-aware, energy-efficient task deployment for heterogeneous multicore systems, and (2) Aging-aware SRAM-based Computing-In-Memory architecture specifically tailored for multiply-accumulate operations. Throughout the talk, I will showcase innovative concepts devised by our research team to confront these challenges and elaborate on the encountered implementation difficulties. The overarching goal is to furnish the audience with a foundational background in the design of reliable computing systems and to inspire additional researchers to contribute to this dynamic and evolving field.

14:30-14:50 Break

14:50-16:20 Session X: Special Invited Lecture 2

Chair: Jun Shiomi (Osaka Univ.)

14:50-16:20 **Radiation-hardened circuit design for space application**

SinNyoung Kim (IMEC)

Abstract: Space projects are being divided into two distinct categories. The first category consists of projects driven by private funding, focusing on cost, time to market and volume as new business models. The second category comprises the traditional projects led by national or international space agencies. There is not only a growing demand of Commercial Off-The-Shelf (COTS) products for projects in the first category but also an increasing importance of System-on-Chip (SoC) designs with radiation-hardening for projects in the second category, particularly in light of moon exploration efforts. As more and more countries join the race for moon exploration, such as Artemis project, highly reliable microelectronic systems to enable communication, rover control, environmental observation on the moon, and other tasks becomes essential for lunar missions. Consequently, the radiation-hardened SoCs are required to ensure high reliability of systems on the moon that endure the high energy and probability of particle hits. This talk will introduce the design of radiation-hardened circuits, which are one of the basic elements for radiation-hardened SoCs, ensuring the successful operation of your chips on the lunar surface.

16:20-16:40 **Break**

16:40-18:00 **Session XI: Panel Discussion**

Organizer and Moderator: Yasuhiko Nakashima (NAIST)

16:40-18:00 **Exploring the Potentials, Limitations, and Challenges of PiM (Processing-in-Memory) and CiM (Computation-in-Memory)**
Panelists:
Stefano Ambrogio (IBM Research)
Kyomin Sohn (Samsung Electronics)
Hoi-Jun Yoo (KAIST)
Yu-Guang Chen (National Central Univ.)

Abstract: This panel discusses Processing-in-Memory (PiM) and Computation-in-Memory (CiM). The main topics include the potential, limitations, and challenges of each computational technique from their respective standpoints. We will continue the discussion with comments and questions from the audience to help us understand future directions.

April 19, 2024 (Japan Standard Time)

9:00-9:50 **Session XII: Keynote Presentation 4**

Chairs: Teruaki Sakata (TIER IV), Ryusuke Egawa (Tokyo Denki Univ.)

9:00-9:50 **Intel Foundry Advanced Packaging and Test: Enabling Disaggregation in AI and PC**

Chunqing Peng (Intel)

Abstract: Intel Foundry advanced packaging and Test solution are well positioned to enable next generation of AI and PC packaging. Future of packaging are relying on disaggregation in all front. Intel's broad profolio range from FCBGA 2D pkg to 2.5D EMIB package to 3D Foveros and 3.5D CoEMIB technology. This Presentation will share Intel's Technology Roadmap and example in both High-performance compute AI Chip and PC compute chip.

9:50-10:10 **Break**

10:10-11:00 **Session XIII: Keynote Presentation 5**

Chairs: Yuki Kobayash (NEC), Shunsuke Sasaki (Toshiba)

10:10-11:00 **Hot AI by COOL SoCs**

Hoi-Jun Yoo (KAIST)

Abstract: In the current landscape of computing, the prevalence of AI applications on mobile devices has emphasized the critical importance of designing energy-efficient System-on-Chip (SoC) systems to curtail power consumption. Recognizing this, we present a comprehensive suite of low-power design techniques tailored to address the intricacies of various AI applications. One of the key pillars of our approach lies in harnessing the inherent sparsity within Convolutional Neural Networks (CNN). By strategically leveraging sparsity, we can intelligently skip unnecessary operations during the inference process. Our innovative design includes specific strategies like the Single Zero Skipping Logic, Dual Zero Skipping Logic, and Triple Zero Skipping Logic. These mechanisms collectively contribute to achieving a state-of-the-art level of energy efficiency, setting a new standard in the field. Beyond CNN optimization, we introduce Spiking Neural Networks (SNN) to increase sparsity within the input feature map. This nuanced incorporation enhances our ability to tailor the SoC design to the specific characteristics of AI workloads, further contributing to gains in energy efficiency. Moreover, we explore the synergies between CNN and SNN, presenting an approach that capitalizes on the strengths of both architectures for high energy efficiency. The culmination of these advancements results in the development of a highly energy-efficient SoC, proficient in processing a myriad of AI applications with remarkable power efficiency. Our design techniques extend beyond conventional applications, achieving the state-of-the-art energy efficiency in specialized domains such as deep reinforcement learning, 3D rendering utilizing Neural Radiance Fields (Nerf), and natural language processing with Large Language Models (LLM). In summary, our multifaceted approach to SoC design not only addresses the pressing need for energy efficiency in the realm of neural network processing, but also pushes the boundaries of AI applications that can be processed, making significant progress toward sustainable high-performance computing.

11:00-11:20 **Break**

11:20-12:15 **Session XIV: CNN and GCN**

Chairs: Koji Hashimoto (Fukuoka Univ), Ryohei Kobayashi (Univ. of Tsukuba)

11:20-11:45 **Bit-Separable Radix-4 Booth Multiplier for Power-Efficient CNN Accelerator**

Seunghyun Park and Daejin Park (School of Electronic and Electrical Engineering, Kyungpook National University)

11:45-12:00 **Power-Efficient Acceleration of GCNs on Coarse-Grained Linear Arrays**

Dohyun Kim, Koki Asahina, Yirong Kan, Renyuan Zhang, Yasuhiko Nakashima (Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Japan)

12:00-12:15 **A 22 nm 10 TOPS Mixed-Precision Neural Network SoC for Image Processing with Energy-Efficient Dilated Convolution Support**
Simon Friedrich[1], Robert Wittig[1], Emil Matus[1], Darius Grantz[2], Martin Zeller[2], Jens Benndorf[2], Gerhard Fettweis[1] ([1]Vodafone Chair for Mobile Communications Systems, Technical University Dresden, Dresden, Germany, [2]Dream Chip Technologies GmbH, Garbsen, Germany)

12:15-12:25 **Session XV: ASPIRE Session**

Chair: Hideharu Amano (Keio Univ.)

12:15-12:25 **About the ASPIRE funding program and current open calls**

Dai Minowa (Japan Science and Technology Agency)

12:25-13:25 **Lunch**

13:25-14:45 **Poster Break**

14:45-15:50 **Session XVI: Parallel and Distributed Computing**

Chairs: Hiroyuki Takizawa (Tohoku Univ.), Teppei Hirotsu (Denso)

14:45-15:10 **Template-Based Automatic Library Function Generation with Halide for Compute-Intensive Simulink Models**
Qi Li and Masato Edahiro (Nagoya University, Japan)

15:10-15:35 **Branch Divergence-Aware Flexible Approximating Technique on GPUs**

Reoma Matsuo, Yuya Degawa, Hidetsugu Irie, Shuichi Sakai and Ryota Shioya (The University of Tokyo, Tokyo, Japan)

15:35-15:50 **A Microservice Scheduler for Heterogeneous Resources on Edge-Cloud Computing Continuum**
Daiki Saito, Siyi Hu and Yukinori Sato (Toyohashi University of Technology, Toyohashi, Japan)

15:50-16:10 Break

16:10-17:15 Session XVII: Accelerators

Chairs: Takeshi Ohkawa (Kumamoto Univ.), Ryuichi Sakamoto (Tokyo tech.)

16:10-16:35 **NoPIM: Functional Network-on-Chip Architecture for Scalable High-Density Processing-in-Memory-based Accelerator**
Sangjin Kim, Zhiyong Li, Soyeon Um, Wooyoung Jo, Sangwoo Ha, Sangyeob Kim and Hoi-Jun Yoo (School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST))

16:35-17:00 **A Low-Power Neural Graphics System for Instant 3D Modeling and Real-Time Rendering on Mobile AR/VR Devices**
Junha Ryu[1], Hankyul Kwon[1], Wonhoon Park[1], Zhiyong Li[1], Beomseok Kwon[1], Donghyeon Han[2], Dongseok Im[1], Sangyeob Kim[1], Hyungnam Joo[1], Minsung Kim[1], and Hoi-Jun Yoo[1] ([1]School of EE, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea, [2]Dept. of EECS, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts)

17:00-17:15 **Reinforcement Learning Hardware Accelerator using Cache-Based for Optimized Q-Table Selection**
Muhammad Sulthan Mazaya[1,3], Eko Mursito Budi[1], Infall Syafalni[2,3,4], Nana Sutisna[2,3,4], and Trio Adiono[2,3] ([1]Faculty of Industrial Technology, Bandung Institute of Technology, Indonesia, [2]School of Electrical Engineering and Informatics, Bandung Institute of Technology, Indonesia, [3]University Center of Excellence on Microelectronics, Bandung Institute of Technology, Indonesia, [4]Interuniversity Microelectronics Centre (IMEC), Leuven, Belgium)

17:15-17:40 Session XVIII: Poster Award and Closing Remarks

Chairs: Yasutaka Wada (Meisei Univ.), Ryusuke Egawa (Tokyo Denki Univ.)