Final Program

April 16, 2025 (Japan Standard Time)

13:00-13:20 **Session I: Welcome & Opening Remarks**
Chair: Yuki Kobayashi (NEC)

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

13:20-14:10 **Session II: Keynote Presentation 1**
Chairs: Tohru Ishihara (Nagoya University), Yuki Kobayashi (NEC)

13:20-14:10 **A Content-Addressable Engine for Associative Processing**
Jose Martinez (Cornell University)
Abstract: A content-addressable parallel processor, or associative processor for short, is a computing and storage architecture from the 1970s based on content-addressable memories, where sequencing bulk search and update memory operations is the primary means to manipulate in situ many operands in parallel, without employing arithmetic circuits. Our research group has been investigating the potential of this computing paradigm in the context of modern microarchitectures, with the goal of providing a processing-in-memory abstraction that is highly data-parallel and programmable.
In this talk, I will present some results to date on our Content-Addressable Processing Engine (CAPE), an associative processor architecture that employs CMOS-based push-rule SRAM/CAM logic to carry out data-parallel arithmetic and logic operations, abstracted as very long vector instructions that can be expressed using a RISC-V ISA with standard vector extensions for high programmability. I will describe the basic CAPE architecture and show some promising simulation-based results on a diverse set of data-parallel benchmarks. I will explain how CAPE can be part of a tiled multicore architecture that co-exists and cooperates with CPU cores and on-chip caches. I will also describe our co-design effort in the context of analytical databases.

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

14:30-15:10 **Session III: Invited Presentation**
Chairs: Takanori Ueda (IBM Japan), Shunsuke Sasaki (Toshiba)

14:30-15:10 **Device-Algorithm Co-optimization for Analog In-Memory Computing**
Sangbum Kim (Seoul National University)
Abstract: Analog in-memory deep learning is a computing architecture that aims to improve the efficiency of deep learning algorithms by performing calculations in memory rather than transferring data back and forth between memory and processors. This can greatly reduce the energy consumption and latency of deep learning algorithms, making them more efficient and faster.
However, there are still challenges that need to be solved before analog in-memory deep learning can be used for real-world applications. For example, analog memory devices can have significant variability and noise, which can impact the accuracy of calculations. Additionally, the lack of weight update symmetry and linearity impedes the acceleration

of on-chip training operation that is the most computationally expensive in deep learning.
In this talk, I will discuss some of recent device-algorithm co-optimization studies. In the first example, an array of capacitor-based synaptic cells is co-optimized with the Tiki-Taka algorithm that was introduced to mitigate the non-ideal characteristics of synaptic cells storing analog weights. Yet, the detailed practical implementation of the algorithm has not been demonstrated. The ultralow leakage current of IGZO TFT can be utilized to implement a novel 6T1C synaptic cell that can efficiently implement Tiki-Taka algorithm.
In the second example, we demonstrate that neuromorphic hardware based on phase change memory can efficiently implement the Boltzmann Machine on Spiking Neural Networks (sRB). The 6T2R sBM chip not only tolerates noise in devices but also capitalizes on this noise thanks to its stochastic nature.
These examples suggest that the efficient implementation of neuromorphic in-memory computing systems is feasible by pairing synaptic devices with optimal algorithms to mitigate the non-ideal characteristics of these devices. Concurrently, synaptic and neuronal devices must be optimized to meet a new set of requirements imposed by these innovative algorithms.

15:10-15:30 **Break**

15:30-16:45 **Session IV: Security**

Chairs: Ryuichi Sakamoto (Institute of Science Tokyo), Kotaro Shimamura (Hitachi)

15:30-15:55 **UniCrypt: Universal Crypto Engine with Optimized Resource Sharing for Security Applications**
Vu Trung Duong Le, Hoai Luan Pham, Tuan Hai Vu, Van Duy Tran, and Yasuhiko Nakashima (Nara Institute of Science and Technology, Nara, Japan)

15:55-16:20 **A High-Performance Hardware Design for Polynomial Multiplication in the CRYSTALS-Kyber Algorithm**
Hsueh Chen†, Hsiu-Wei Chen†, Shih-Hsu Huang† and Po-Yuan Chen‡ (†Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan, Taiwan ‡System Circuit Testing Technology Department, Electronic and Optoelectronic System Research Laboratories, Industrial Technology Research Institute, Tainan, Taiwan)

16:20-16:45 **Accelerating Elliptic Curve Point Additions on Versal AI Engine for Multi-scalar Multiplication**
Ayumi Ohno, Kotaro Shimamura, Shinya Takamaeda-Yamazaki (The University of Tokyo, Tokyo, Japan)

16:45-17:35 **Session V: Keynote Presentation 2**

Chairs: Ryusuke Egawa (Tokyo Denki Univ.), Hiroe Iwasaki (Tokyo Univ. of Agriculture and Technology)

16:45-17:35 **Advanced Package Substrate Technology for Heterogenous Integration**
Sriram Dattaguru (Intel)
Abstract: The talk focuses on the critical role of package substrates in the context of Moore's Law and the semiconductor industry's evolving needs. Historically, Moore’s Law, which predicted the doubling of transistors in an integrated circuit every two years, has been sustained by improvements in chip scaling and architecture. However, as chip scaling becomes more challenging and costly, package substrates, which traditionally acted as a space transformer between the chip and the external world are becoming more crucial as product differentiators. Over the last several decades,

package substrates have evolved from lead-frames, ceramic, and currently organic materials.
The age of heterogeneous integration is here and driving the need to find advanced packaging solutions that resolve reticle limitations and enable moving higher functionality onto the package. Intel continues to drive leadership in advanced packaging with Embedded Multi-die Interconnect Bridge (EMIB) technology and glass core substrates. EMIB technology have been developed and is high volume production, to enable large form factor high-performance computing (HPC) packages. Roadmap is in place to scale bump pitch and enable through-silicon via (TSV) EMIB bridges. Glass core substrates will enable ultra large form substrates with scalable performance and design rule offerings beyond organic substrates. Both EMIB and glass core substrates technologies will be key to enable scalable chiplet heterogeneous integration for AI and HPC system on package.

April 17, 2025 (Japan Standard Time)

09:00-09:50 **Session VI: Memory**

Chairs: Ryohei Kobayashi (Institute of Science Tokyo), Masayuki Sato (Tohoku Univ.)

09:00-09:25 **Hybrid ReRAM-NMC & SRAM-CiM Matrix Multiplication for Large Language Model**
Tao Wang, Daqi Lin, Kenshin Yamauchi, Naoko Misawa, Chihiro Matsui, and Ken Takeuchi (Dept. of Electrical Engineering and Information Systems, The University of Tokyo, Tokyo, Japan)

09:25-09:50 **A Compute-in-Memory Ascon Implementation Based on a Novel 11T SRAM Processing Macro**
Mark Lee[1], Chris Clark[2], and Saibal Mukhopadhyay[1] ([1]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA [2]Georgia Tech Research Institute, Atlanta, Georgia, USA)

09:50-10:10 **Break**

10:10-11:00 **Session VII: Keynote Presentation 3**

Chairs: Fumio Arakawa (Univ. of Tokyo), Teruaki Sakata (TIER IV)

10:10-11:00 **Open Source RiscV CPU and AI for Edge Applications**

Jim Keller (Tenstorrent)

Abstract: This talk will cover Tenstorrent's solutions in RiscV, IP, software and hardware to bring open source, buildable compute to a wide variety of low power applications.

11:00-12:00 **Poster Session**

Chair: Koji Hashimoto (Fukuoka Univ.)

Poster 1 **Optimizing Computed Tomography Reconstruction with Mixed Precision on Nvidia Jetson Devices**
Xuetao Chen, Amelie Chi Zhou*, Du Wu †, Peng Chen †, Emil Vatai †, Jens Domke †, Mohamed Wahib † (*Hong Kong Baptist University, Hong Kong †RIKEN-CCS, Japan)*

Poster 2 **A Long-Distance Asynchronous Unit for Inter-Domain Connectivity in Network-on-Chip**
Ju-Pyo Lee[1], Sung Yang[2] and Hyun-Gyu Kim[1] ([1]NoC Team, R&D Center, Openedges Technology Inc. Seoul, Korea. [2]PHY Team, R&D Center Openedges Technology Inc. Seoul, Korea)

Poster 3 **A Data Compression Method for DL-SCAs using Denoising Autoencoders**

Masaki Morita, Takuya Kojima, Haruto Ishii, Hideki Takase and Hiroshi Nakamura (Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan)

Poster 4 **A Machine Learning Technique for IR Drop Prediction in VLSI Design**

Chia-Wei Wang and Wei-Kai Cheng (Department of information and computer engineering, Chung Yuan Christian University, Chung-Li City, Taiwan)

Poster 5 **The Bypass Controller for Online Multiplayer Computer Games**

Junya Nagatomi and Hayato Nomura (National Institute of Technology, Akashi College, Akashi, Japan)

Poster 6 **A Machine Learning-Based Power Prediction Framework for DCIM Circuits**

Chih-Li Hsiao[†], Hsiu-Wei Chen[†], Shih-Hsu Huang[†] and Jin-Fu Li[‡] ([†]Department of Electronic Engineering, Chung Yuan Christian University, Taoyuan, Taiwan [‡]Department of Electrical Engineering, National Central University, Taoyuan, Taiwan)

Poster 7 **22-nm Gain Cell DRAM for Cryogenic Operation**

Tomoki Iwase, Shigeru Yamamoto and Kazutoshi Kobayashi (Kyoto Institute of Technology, Kyoto, Japan)

Poster 8 **Tracing Time-Series Percentiles with an FPGA-Based Implementation of Heap Sort**
Riku Kawachi and Yukinori Sato (Toyohashi University of Technology, Toyohashi, Japan)

Poster 9 **Quantitative Analysis of Execution Time in a Homomorphic Encryption**

Hikaru OKAMOTO[1], Yuri KAGOTANI[1], PHAM Hoai Luan[2] and Hideo Akaike[1] ([1]The University of Electro-Communications, Tokyo, Japan [2]Nara Institute of Science and Technology, Nara, Japan)

Poster 10 **An Acceleration of Homomorphic Encryption for GPUs focusing on Polynomial-graph**
Ai Nozaki, Takuya Kojima, Hiroshi Nakamura, and Hideki Takase (The University of Tokyo, Tokyo, Japan)

Poster 11 **Low-Power Booth Multiplier Design for CNN Applications**

Ci-Wun Jhong, Yen-Jen Chang and Hao-Wei Lu (Department of Computer Science and Engineering, National Chung Hsing University, Taiwan)

Poster 12 **Low-switching Data Bus Architecture for Video Encoders**

Tetsuya Aoyama, Naoya Niwa and Hiroe Iwasaki (Tokyo University of Agriculture and Technology, Tokyo, Japan)

Poster 13 **Design and Implementation of Secure Memory Systems with XLS High-Level Synthesis**

Yoshiki Ozaki and Shinya Takamaeda-Yamazaki (The University of Tokyo, Tokyo, Japan)

Poster 14 **Pattern-Aware Encoding Scheme for Parameter Compression of Binary Neural Networks**
Babak Golbabaei[*], Yirong Kan[*], Renyuan Zhang[†], Yasuhiko Nakashima[*] ([*]Division of Information Science, Nara Institute of Science and Technology, Japan [†]School of Information Science and Engineering, Yunnan University, China)*

Poster 15 **Low Power CNN Accelerator Memory Interface with Small Footprint Memory Access**
Hoseong Kim and Daejin Park (School of Electronic and Electrical Engineering, Kyungpook National University, Korea)

Poster 16 **Preliminary Evaluation of Image Recognition Performance and Power Consumption of Edge Devices with AI Accelerators**
Keiichi Sato and Yasutaka Wada (School of Information Science, Meisei University, Tokyo, JAPAN)

12:00-13:00 Lunch

13:00-14:30 **Special Invited Lecture 1**

Chair: Jun Shiomi (Osaka Univ.)

13:00-14:30 **Reliability and Efficiency in Deep Learning Processing Systems**

Alex Orailoglu (UC San Diego)

Abstract: Artificial intelligence techniques driven by deep learning have experienced significant progress in the past decade. The usage of deep learning methods has increased dramatically in practical application domains such as autonomous driving, healthcare, and robotics, where the utmost hardware resource efficiency, as well as strict hardware safety and reliability requirements, are often imposed. The increasing computational cost of deep learning models has been traditionally tackled through model compression and domain-specific accelerator design. As the cost of conventional fault tolerance methods is often prohibitive in consumer electronics, the question of functional safety and reliability for deep learning hardware is still in its infancy. This talk outlines a novel approach to deliver dramatic boosts in hardware safety, reliability, and resource efficiency through a synergistic co-design paradigm.
We start off by reviewing the unique algorithmic characteristics of deep neural networks, including plasticity in the design process, resiliency to small numerical perturbations, and their inherent redundancy, as well as the unique micro-architectural properties of deep learning accelerators such as regularity. The advocated approaches reshape deep neural networks and enhance deep neural network accelerators strategically by prioritizing the overall functional correctness and minimizing the associated costs through the statistical nature of deep neural networks. Experimental results demonstrate that deep neural networks equipped with the proposed techniques can maintain accuracy gracefully, even at extreme rates of hardware errors. As a result, the described methodology can embed strong safety and reliability characteristics in mission-critical deep learning applications at a negligible cost.
The proposed strategies further offer promising avenues for handling the micro-architectural challenges of deep neural network accelerators and boosting resource

efficiency through the synergistic co-design of deep neural networks and hardware micro-architectures. Practical data analysis techniques coupled with a novel feature elimination algorithm identify a minimal set of computation units that capture the information content of the layer and squash the rest. Linear transformations on the subsequent layer ensure accuracy retention despite the removal of a significant portion of the computation units. We further demonstrate that novel complementary sparsity patterns can offer utmost expressiveness levels with inherent hardware exploitable regularity. A novel dynamic training method converts the expressiveness of such sparsity configurations into highly accurate and compact sparse neural networks.

14:30-14:50 **Break**

14:50-16:20 **Special Invited Lecture 2**

Chair: Jun Shiomi (Osaka Univ.)

14:50-16:20 **Next-Generation Quantum Computing: A Computer Architect's Perspective**

Jangwoo Kim (Seoul National University)

Abstract: Quantum computer is the next-generation computing paradigm. Therefore, many researchers are actively working in various domains in quantum computer (e.g,. qubit manufacturing, qubit interface control processor, programming and compiler, application). And, as the real-world quantum computer applications require millions of qubits, we are moving from Noisy Intermediate-Scale Quantum (NISQ) to fault-tolerant quantum computers (FTQC).
In this talk, I first introduce the key challenges in developing a scalable and reliable quantum computer in the FTQC era. Next, I will introduce my research work covering quantum computer modeling, quantum control processor, quantum interface methods, distributed quantum computer, and reliable quantum computer. By integrating these outcomes, we have been contributing to realizing the real-world quantum computers in the FTQC era.

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

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

Organizer and Moderator: Tohru Ishihara (Nagoya University)

16:40-18:00 **Sustainable AI: Emerging Architectures, Devices, and Quantum Computing Towards Future Computing**

Panelists:
Jangwoo Kim (Seoul National University)
Kazutoshi Kobayashi (Kyoto Institute of Technology)
Jose Martinez (Cornell University)

Abstract: This panel will discuss sustainable AI and quantum computing systems. Key topics will include energy efficiency, scalability, and reliability of each computing technology from each panelist's perspective. The discussion will continue with comments and questions from the audience to help us understand future directions.

April 18, 2025 (Japan Standard Time)

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

Chairs: Yuki Kobayashi (NEC), Yasushi Inoguchi (JAIST)

9:00-9:50 **The Challenges of Delivering Power to and Cooling the Cerebras Wafer-Scale Engine**
Jean-Philippe Fricker (Cerebras Systems, Inc.)

Abstract: As AI workloads push the boundaries of computational power, traditional chip architectures struggle to keep pace. Nowhere is this more evident than in wafer-scale computing, where delivering power and managing thermals become critical engineering challenges. This keynote explores the evolving landscape of high-performance computing and the infrastructure bottlenecks limiting future progress.
We begin by examining the ever-growing compute demands and why traditional datacenter infrastructure is struggling to keep up. We'll analyze the impact of system density on both power delivery and cooling, highlighting the inefficiencies of conventional approaches. Next, we'll look at how past innovations—such as water cooling—are making a resurgence as viable solutions.
Finally, we'll explore how Cerebras has tackled these challenges head-on, leveraging novel architectural and cooling innovations to unlock unprecedented performance. We'll compare this approach with conventional solutions to understand why wafer-scale integration represents a paradigm shift in AI computing. Join us for an in-depth look at the future of high-performance computing and what it takes to meet the growing demands of AI inference and training while overcoming power and cooling challenges.

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

10:10-11:00 **Session X: Reconfigurable logic and system**

Chairs: Ryohei Kobayashi (Institute of Science Tokyo), Tomohiro Ueno (RIKEN)

10:10-10:35 **Proposal for Non-Volatilization of eFPGA Core**

Keizo Hiraga[1,2], Kenshu Seto[2], Kazuhiro Bessho[1] and Masahiro Iida[2] ([1]Sony Semiconductor Solutions Corporation, Atsugi, Japan [2]Kumamoto University, Kumamoto, Japan)

10:35-11:00 **A Power-Efficient Reconfigurable Hybrid CNN-SNN Accelerator for High Performance AI Applications**
Heuijee Yun and Daejin Park (School of Electronic and Electrical Engineering Kyungpook National University, Daegu, Korea)

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

11:20-12:15 **Session XI: Image processing**

Chairs: Thiem Van Chu (Institute of Science Tokyo), Teppei Hirotsu (DENSO)

11:20-11:45 **GUPA: Group-Wise Uniform Pruning Accelerator for Depthwise Separable Convolution**
Yi Chen, Malte Wabnitz, Jie Lou, Christian Lanius and Tobias Gemmeke (Chair of Integrated Digital Systems and Circuit Design, RWTH Aachen University, Germany)

11:45-12:00 **HOPE: An Efficient Accelerator with Head-wise Overlap Processing for Sparse Attention in Vision Transformer**
Jihyeon Heo, Soomin Rho, Kwangrae Kim and Ki-Seok Chung (Department of Electronic Engineering, Hanyang University, Seoul, Korea)

12:00-12:15 **Exploring the Versal AI Engine for 3D Gaussian Splatting**

Kotaro Shimamura, Ayumi Ohno and Shinya Takamaeda-Yamazaki (The University of Tokyo, Tokyo, Japan)

12:15-13:15 Lunch

13:15-14:35 Poster Break

14:35-15:25 Session XII: Keynote Presentation 5

Chairs: Ryohei Kobayashi (Institute of Science Tokyo), Yasutaka Wada (Meiji Gakuin Univ.)

14:35-15:25 **Specialized Hardware and Open-Source Tools for Scientific Computing and Instruments**
Kazutomo Yoshii (Argonne National Lab.)

Abstract: High-performance computing (HPC) faces critical challenges as transistor scaling slows, limiting further gains in computational performance and energy efficiency. Scientific instrumentation, meanwhile, faces a different obstacle: rapidly increasing data rates. Instruments such as advanced X-ray detectors generate terabytes of data per second, making it impractical to transmit raw data downstream. On-chip data processing and reduction at the source are now essential to address this bottleneck.
Data movement, rather than computation, has become the dominant factor limiting system performance across both HPC and scientific instruments. The performance gap between processors and memory exacerbates inefficiencies, leaving many data-intensive workloads unable to fully utilize processing capabilities. To optimize system performance, strategies such as data compression and reduction within hardware are increasingly necessary. Data flow and streaming computing paradigms can significantly improve data handling in both HPC and scientific instruments by facilitating efficient, continuous data transfer.
Specialized hardware accelerators offer a promising solution by enhancing both performance and energy efficiency across scientific domains. These accelerators also have the potential to revolutionize scientific instruments by enabling real-time data handling at the source. However, hardware specialization requires expertise in design, verification, integration, and sharable resources such as open-source hardware libraries -- all of which remain scarce globally.
Open-source ecosystems, featuring tools such as Chisel, Verilator, FireSim, Chipyard, Mosaic, and OpenROAD, stimulate collaboration and community-driven prototyping activities by enabling the sharing of research ideas and innovations. These tools, along with open standards like RISC-V, enhance accessibility to hardware innovation,

particularly for professionals from software backgrounds. Cultivating strong open-source collaborations could pave the road for both scientific computing and instrumentation.

15:25-15:45 **Break**

15:45-16:50 **Session XIII: Circuit design and its environment**

Chairs: Ryuichi Sakamoto (Institute of Science Tokyo), Naoya Niwa (Tokyo Univ. of Agriculture and Tech.)

15:45-16:10 **RustSFQ: A Domain-Specific Language for SFQ Circuit Design**

Mebuki Oishi, Sun Tanaka and Shinya Takamaeda-Yamazaki (The University of Tokyo, Tokyo, Japan)

16:10-16:35 **Evaluation of Trade-off between Compression Ratio and Hardware Cost for Adaptive Bandwidth Compression Hardware Platform**
Tomohiro Ueno[1], Kaito Kitazume[2], Masato Kiyama[3], Kazutomo Yoshii[4], Kento Sato[1], Norihisa Fujita[5], Ryohei Kobayashi[6], Taisuke Boku[5] and Kentaro Sano[1]([1]Center for Computational Science, Riken, Kobe, Japan [2]Degree Programs in Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan [3]Faculty of Advanced Science and Technology, Kumamoto University, Kumamoto, Japan [4]Mathematics and Computer Science, Argonne National Laboratory, Lemont, IL, USA [5]Center for Computational Sciences, University of Tsukuba, Tsukuba, Japan [6]Supercomputing Research Center, Institute of Integrated Research, Institute of Science Tokyo, Kanagawa, Japan)

16:35-16:50 **Can the Agile-chip platform carve out a niche between ASICs and FPGAs?**

Hideharu Amano, Atsutake Kosuge, Hirofumi Sumi, Naonobu Shimamoto, Yukinori Ochiai,Yurie Inoue, Tohru Mogami, Yoshio Mita, Makoto Ikeda (System Design Lab, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan)

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

17:00-17:50 **Session XIV: Efficient AI models**

Chairs: Ryohei Kobayashi (Institute of Science Tokyo), Yukihiro Sasagawa (Socionext)

17:00-17:25 **TTF-GNN: Memory-Efficient GNNs via Tensor Train Decomposition and Network Folding**
Hiroaki Ito, Jiale Yan, Kazushi Kawamura, Masato Motomura, Thiem Van Chu and Daichi Fujiki (Institute of Science Tokyo, Tokyo, Japan)

17:25-17:50 **A Lightweight Transformer Model With Dynamic Sparse Mask for Neural Machine Translation**
Nastaran Asadi[1], Babak Golbabaei[1], Yirong Kan[1], Renyuan Zhang[1,2], Yasuhiko Nakashima[1] ([1]Division of Information Science, Nara Institute of Science and Technology, Japan [2]School of Information Science and Engineering, Yunnan University, China)

17:50-18:15 **Session XV: Poster Award and Closing Remarks**

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