Final Program

April 19, 2023 (Japan Standard Time)

13:00-14:30 **Special Invited Lecture 1**
Co-chairs: Takatsugu Ono (Kyushu Univ.), Shotaro Shintani (NSITEXE)

13:00-14:30 **Vortex: An open-source RISC-V based GPGPU accelerator**
Hyesoon Kim (Georgia Institute of Tech, USA)

Abstract: Vortex is an open source Hardware and Software project to support GPGPU based on RISC-V ISA extensions. Currently Vortex supports OpenCL/CUDA and it runs on FPGA. The vortex platform is highly customizable and scalable with a complete open source compiler, driver and runtime software stack to enable research in GPU architectures/compiler/run-time systems. In this talk, I will present the vortex architecture/software stack.

14:30-15:00 **Break**

15:00-16:30 **Special Invited Lecture 2**
Co-chairs: Takatsugu Ono (Kyushu Univ.), Takumi Uezono (Hitachi)

15:00-16:30 **The Parameter and Chip Wars: Shifting the Focus from Model-centric to Data-centric AI**
Vijay Janapa Reddi (Harvard Univ., USA)

Abstract: In recent years, deep learning has revolutionized the field of artificial intelligence by providing a powerful tool for solving complex problems across various domains, from computer vision to natural language processing. Traditionally, deep learning has focused on developing complex models to solve challenging problems, resulting in intense competition in the form of parameter and chip wars to create more powerful hardware. Furthermore, the dramatic scaling at the individual model level has had significant ramifications at the system level, requiring management of the growing complexity surrounding AI systems. Despite these advancements, recent research has highlighted the significant impact of data quality and quantity on model capabilities and performance, revealing that improving data quality often leads to better results than designing more complex models. This finding has prompted a shift towards a data-centric approach, emphasizing the acquisition of high-quality data and the design of effective data engineering pipelines. This talk delves into the challenges and directions presented by the parameter and chip wars in deep learning, including recent developments in hardware and algorithms, and it suggests that a data-centric approach for systems may be a more viable approach to offset the scaling challenges posed by the parameter and chip wars.

16:30-17:00 **Break**

17:00-17:50 **Special Session I**

17:00-17:50 **Keynote Presentation 1**
Co-chairs: Yoshio Hirose (Fujitsu), Yasuo Unekawa (Toshiba Electronic Devices & Storage)

Next Generation Cryogenic Superconductor Computing ~ From Classic to Quantum ~

Koji Inoue (Kyushu Univ.)

Abstract: Moore's Law, doubling the number of transistors in a chip every two years, has so far contributed to the evolution of computer systems. Unfortunately, we cannot expect sustainable transistor shrinking anymore, marking the beginning of the so-called post-Moore era. Therefore, it has become essential to explore emerging devices, and superconductor single-flux-quantum (SFQ) logic that operates in a 4.2-kelvin environment is a promising candidate. Josephson junctions (JJs) are used as switching elements in SFQ logic to compose a superconductor ring (SFQ ring) that can store (or trap) and transfer a single magnetic flux quantum. It fundamentally operates with the voltage pulse-driven nature that makes it possible to achieve extremely low-latency and low-energy JJ switching. This talk shares the history of our SFQ Research, e.g., revisiting microarchitecture and demonstrating over 30 GHz microprocessors, AI accelerator designs, and recently targeting quantum computers. Then, the role of computer architecture for such emerging device computing is discussed.

April 20, 2023 (Japan Standard Time)

9:30-9:40 **Session I**

9:30-9:40 **Welcome and Opening Remarks**
Co-chairs: Yuki Kobayashi (NEC), Hiroki Matsutani (Keio Univ.)

Makoto Ikeda	*Chair of the Organizing Committee*
Tadao Nakamura	*Chair of the Steering Committee*
Jose Renau	*Chair of IEEE/CS TCMM*
Minoru Fujishima	*President of IEICE/ES*

9:40-10:30 **Session II**

9:40-10:30 **Keynote Presentation 2**
Co-chairs: Yukinori Sato (Toyohashi Univ. of Technology), Satoshi Kametani (Sony Semiconductor Solutions)

Compute Express Link (CXL): Shaping the compute landscape
Debendra Das Sharma (Intel, USA)

Abstract: High-performance workloads demand heterogeneous processing, tiered memory architecture, infrastructure accelerators such as SmartNICs, and infrastructure processing units to meet the demands of the emerging compute landscape. Applications such as artificial intelligence, machine learning, data analytics, 5G, automotive, and high-performance computing are driving significant changes within cloud computing, intelligent edge and client computing infrastructure. Interconnect is a key pillar in this evolving computational landscape. The recent advent of Compute Express Link (CXL), a new open standard for cache-coherent interconnect, with its memory and coherency semantics has made it possible to pool computational and memory resources at the rack level using low-latency, higher-throughput, and memory-coherent access mechanisms. CXL is adopting networking features such as multi-host connectivity, pooled memory, persistence flows, and fabric managers while keeping its low-latency load-store semantics intact. CXL is evolving to provide efficient access mechanisms across multiple nodes with advanced atomics, acceleration, SmartNICs, persistent memory support, etc. In this talk we will explore how synergistic evolution across load-store interconnects and fabrics can benefit the compute infrastructure of the future.

10:30-10:40 **Break**

10:40-11:55 **Session III: Processor**
Co-chairs: Hiroyuki Takizawa (Tohoku Univ.), Yuetsu Kodama (RIKEN)

10:40-11:05 **Lookup Table Modular Reduction: A Low-Latency Modular Reduction for Fast ECC Processor**
Anawin Opasatian, Makoto Ikeda (Univ. of Tokyo)

11:05-11:30 **Dual Vector Load for Improved Pipelining in Vector Processors**
Viktor Razilov[1], Juncen Zhong[1], Emil Matus[1], Gerhard Fettweis[1,2] ([1]Technische Univ. Dresden, [2]Barkhausen Institut, Germany)

11:30-11:55 **Cachet: A High-Performance Joint-Subtree Integrity Verification for Secure**

Non-Volatile Memory
Tatsuya Kubo, Shinya Takamaeda-Yamazaki (Univ. of Tokyo)

11:55-12:00 **Break**

12:00-12:30 **Session IV: Poster Short Speeches**
Chair: Koji Hashimoto (Fukuoka Univ.)

Poster 1 **Low-Power Parallel Lane Detection Unit for Lightweight Automotive Processors**
Heuijee Yun, Daejin Park (Kyungpook National Univ., Korea)

Poster 2 **Optimized Deep MLP for Tensor Train-based Inference Engine**
Jiale Yan, Masato Motomura (Tokyo Inst. of Technology)

Poster 3 **Performance Improvement of SpMV using Sector Cache on A64FX**
Toshiyuki Ichiba, Akihiko Kasagi (Fujitsu)

Poster 4 **Quantifying the Effects of Copious 3D-Stacked Cache on HPC Workloads**
Emil Vatai[1], Jens Domke[1], Balazs Gerofi[2], Yuetsu Kodama[1], Mohamed Wahib[1], Artur Podobas[3], Sparsh Mittal[4], Miquel Peric`as[5], Lingqi Zhang[6], Peng Chen[7], Aleksandr Drozd[1], Satoshi Matsuoka[1] ([1]RIKEN, [2]Intel, USA, [3]KTH Royal Inst. of Technology, Sweden, [4]Indian Inst. of Technology, India, 5Chalmers Univ. of Technology, Sweden, [6]Tokyo Inst. of Technology, [7]National Inst. of Advanced Industrial Science and Technology)

Poster 5 **Persistent-Memory-Based Acceleration of Memory-Intensive Deep Learning Workloads**
Jeongha Lee, Soyeon Park, Seokmin Kwon, Hyokyung Bahn (Ewha Univ., Korea)

Poster 6 **An efficient methods for collecting performance information on a large-scale computing environment**
Mari Yamaoka, Akira Hirai, Akihiko Kasagi (Fujitsu)

Poster 7 **Single-Ended Write 10T SRAM Cell Design for In-Memory Computing**
Wei-Ting Chang, Yen-Jen Chang (National Chung Hsing Univ., Taiwan)

Poster 8 **A Prototype Design of Real-Time Encrypted Malicious Traffic Detection based on Hardware Implementation**
Zhenguo Hu[1], Hirokazu Hasegawa[2], Yukiko Yamaguchi[1], Hajime Shimada[1] ([1]Nagoya Univ., [2]National Inst. of Informatics)

Poster 9 **Design-space Exploration of CGRA for HPC**
Boma Adhi[1], Emanuele Del Sozzo[1], Johannes Pfau[2,1], Carlos Cortes[1], Tomohiro Ueno[1], Kentaro Sano[1] ([1]RIKEN, [2]Karlsruhe Inst. of Technology, Germany)

12:30-14:00 **Poster and Lunch Time Break**

14:00-14:50 **Session V: AI / ML (1)**
Co-chairs: Shinya Takamaeda-Yamazaki (Univ. of Tokyo), Shunsuke Sasaki (Toshiba)

14:00-14:25 **COOL-NPU: Complementary Online Learning Neural Processing Unit with CNN-SNN Heterogeneous Core and Event-driven Backpropagation**
Sangyeob Kim, Soyeon Kim, Seongyon Hong, Sangjin Kim, Donghyeon Han, Jiwon Choi, Hoi-Jun Yoo (KAIST, Korea)

14:25-14:50 **A Low-power Neural 3D Rendering Processor with Bio-inspired Visual Perception Core and Hybrid DNN Acceleration**
Donghyeon Han, Junha Ryu, Sangyeob Kim, Sangjin Kim, Jongjun Park, Hoi-Jun Yoo (KAIST, Korea)

14:50-16:10 Poster Break

16:10-17:40 Session VI: Panel Discussions

Topics: "The Past, Present, and Future of COOL Chips"
Organizer and Moderator: Fumio Arakawa (Univ. of Tokyo)
Panelists: Tadao Nakamura (Keio Univ.)
Hiroaki Kobayashi (Tohoku Univ.)
Hideharu Amano (Keio Univ.)
Kunio Uchiyama (AIST)
Makoto Ikeda (Univ. of Tokyo)

April 21, 2023 (Japan Standard Time)

9:30-10:20 **Session VII**

9:30-10:20 **Keynote Presentation 3**
Co-chairs: Tohru Ishihara (Nagoya Univ.), Akihiko Hashiguchi (Sony Semiconductor Solutions)

Sustainability and Fleet Manageability Innovations with 4th Gen Intel Xeon Processor
Arijit Biswas, Pankaj Kumar (Intel, USA)

Abstract: System architecture approach along with architecture innovation of Intel Xeon processors to lower customers' energy use emissions, also known as scope 2 emissions, due to their energy efficiency achieved through power management, platform monitoring technology and design of built-in accelerators. These accelerators are designed for today's in-demand workloads and deliver significant performance per watt advantage. Another innovation is the Optimized Power Mode feature that, when enabled, provides significant energy savings while only minimally impacting performance. This session will also highlight new telemetry features of Intel Xeon processors to enable companies to better monitor and control electricity consumption and carbon emissions. And platform Monitoring Technology for better telemetry fleet management through exposure of CPU core temperature, power consumption, package C state residency and error information telemetry to both in-band and out-of-band agents.

10:20-10:30 **Break**

10:30-11:20 **Session VIII**

10:30-11:20 **Keynote Presentation 4**
Co-chairs: Yuki Kobayashi (NEC), Teruaki Sakata (TIER IV)

RISC-V Robust Ecosystem
Mark Himelstein (RISC-V International, USA)

Abstract: This talk will discuss the state of RISC-V and its software ecosystem. RISC-V members have already shipped in excess of 10 billion cores for profit so what are they and their customers running on RISC-V? From EDA tools to firmware to operating systems to runtime infrastructure and applications, the RISC-V ecosystem provides both open source and commercial products for implementers and customers to take advantage of in order to enable solutions and inevitably success.

11:20-11:30 **Break**

11:30-12:00 **Session IX: System Design and Implementation (1)**
Co-chairs: Ryohei Kobayashi (Univ. of Tsukuba), Yuichiro Shibata (Nagasaki Univ.)

11:30-11:45 **Low power implementation of Geometric High-order Decorrelation-based Source Separation on an FPGA board**
Ziquan Qin[1], Kaijie Wei[1], Hideharu Amano[1], Kazuhiro Nakadai[2] ([1]Keio Univ.,

[2]Tokyo Inst. of Technology)

11:45-12:00 **FPGA Emulation of Through-Silicon-Via (TSV) Dataflow Network for 3D Standard Chip Stacking System**
Takeshi Ohkawa, Masahiro Aoyagi (Kumamoto Univ.)

12:00-13:30 Poster and Lunch Time Break

13:30-14:20 Session X

13:30-14:20 **Keynote Presentation 5**
Co-chairs: Takuya Nakaike (IBM), Yasushi Inoguchi (JAIST)

AI Software stack: enabling co-optimizations on Deep Learning frameworks
Kazuaki Ishizaki (IBM, Japan)

Abstract: Deep neural networks are becoming popular since they improve the accuracy of machine learning tasks in multiple domains among image, object detection, language, speech, conversation, code, and others. These improvements are enabled by increasing the number of parameters and computational operations. AI accelerator is critical to achieving high accuracy in these tasks since they require a lot of computational resources for training and inference. In this talk, I will review co-optimizations among hardware, algorithms, and software to achieve high performance per watt in an AI accelerator. The algorithms to maintain the same level of accuracy in low-precision arithmetic can simplify the hardware design and implementation. That hardware can achieve high performance at good power efficiency. I will present how the AI software stack can enable these optimizations during the compilation from Deep Learning frameworks to AI accelerators.

14:20-15:20 Poster Break

15:20-16:10 Session XI: AI / ML (2)
Co-chairs: Yukihiro Sasagawa (Socionext), Teppei Hirotsu (NSITEXE)

15:20-15:45 **A Real-Time Keyword Spotting System Based on an End-To-End Binary Convolutional Neural Network in FPGA**
Jinsung Yoon, Donghyun Lee, Neungyun Kim, Su-Jung Lee, Gil-Ho Kwak, Tae-Hwan Kim (Korea Aerospace Univ., Korea)

15:45-16:10 **Flexibly Controllable Dynamic Cooling Methods for Solid-State Annealing Processors to Improve Combinatorial Optimization Performance**
Genta Inoue, Daiki Okonogi, Thiem Van Chu, Jaehoon Yu, Masato Motomura, Kazushi Kawamura (Tokyo Inst. of Technology)

16:10-16:20 Break

16:20-17:10 Session XII: System Design and Implementation (2)
Co-chairs: Kotaro Shimamura (Hitachi), Ryuichi Sakamoto (Tokyo Inst. of Technology)

16:20-16:45 **A 2.41-µW/MHz, 437-PE/mm^2 CGRA in 22 nm FD-SOI With RISC-Like Code Generation**
Tobias Kaiser, Friedel Gerfers (Technische Univ. Berlin, German)

16:45-17:10 **MazeCov-Q: An Efficient Maze-Based Reinforcement Learning Accelerator for Coverage**
Infall Syafalni, Mohamad Imam Firdaus, Andi Ilmy, Nana Sutisna, Trio Adiono

(Bandung Inst. of Technology, Indonesia)

17:10-17:20 **Poster Award and Closing Remarks**
Program Committee Co-chairs: Ryusuke Egawa (Tokyo Denki Univ.), Yasutaka Wada (Meisei Univ.)