# Final Program

## April 14, 2021 (Japan Standard Time)

**8:30-10:00** **Special Invited Lecture 1**
*Co-chairs: Takatsugu Ono (Kyushu Univ.), Chikafumi Takahashi (NSITEXE)*

8:30-10:00 **Improving Fidelity of NISQ Machines with Intelligent Software**
*Moinuddin Qureshi (Georgia Institute of Technology, USA)*

**Abstract:** Quantum computing promises exponential speedups for an important class of problems. While quantum computers with few dozens of qubits have been demonstrated, these machines suffer from a high rate of gate errors. Such machines are operated in the *Noisy Intermediate Scale Quantum (NISQ)* mode of computing where the output of the machine can be erroneous. In this talk, I will discuss some of our recent work that aims to improve the reliability of NISQ computers by developing software techniques to mitigate hardware errors. Our first work exploits the variability in the error rates of qubits to steer more operations towards qubits with lower error rates and avoid qubits that are error-prone. Our second work looks at executing different versions of the programs each crafted to cause diverse mistakes so that the machine becomes less vulnerable to correlated errors. Our third work looks at exploiting the state-dependent bias in measurement errors (state 1 is more error-prone than state 0) and dynamically flips the state of the qubit to perform the measurement in the stronger state. We perform our evaluations on real quantum machines from IBM and demonstrate significant improvement in the overall system reliability. Finally, I will also briefly discuss the hardware aspect of designing large-scale quantum computers, including cryogenic processor and cryogenic memory system.

**10:00-10:30** **Break**

**10:30-12:00** **Special Invited Lecture 2**
*Co-chairs: Takatsugu Ono (Kyushu Univ.), Shotaro Shintani (NSITEXE)*

10:30-12:00 **Processor Hardware Security**
*Jakub Szefer (Yale Univ., USA)*

**Abstract:** As the amount of sensitive information processed by computers constantly increases, there is a need to continue to harden the processors, and the whole computer systems. Among the possible threats, the variety of remote attacks are of importance since they do not require attacker to be physically near the target system, they only require that attacker and victim are executing on the same system, such as by being co-located on same server in a public cloud computing data center. At the same time, there is ever-expanding use of machine learning and other algorithms that process sensitive information in the cloud data centers. Both data, as well as the algorithms, e.g. the specific machine learning architectures or models, can be targets of attacks. This opens up the various algorithms to variety of hardware-rooted side and covert channel attacks, which continue to pose threat to our privacy and security. Meanwhile, considering only performance or security is not enough, and the processor designers need to be mindful of the power consumption and energy usage of their systems. In this talk, we will first cover various remote timing and power related information leaks to give background of

the existing threats. We will then cover variety of transient execution attacks, which work with the covert channels, and can further undermine system security. As examples of specific threats, attacks on machine learning algorithms from literature will be reviewed. The talk will next cover various defenses, such as secure caches or secure TLBs that aim to protect from the threats. In addition, the talk will touch upon power and energy issues, and especially the need to better understand the performance-power-security trade off in design of processors. How to design high performance, low power, and secure systems is a research challenge that hopefully this talk can motivate academics and researchers to explore more.

**12:00-13:00** **Lunch Time Break**

**13:00-14:30** **Special Session I: Panel Discussions**
*Co-chairs: Hiroki Matsutani (Keio Univ.), Yutaka Uematsu (Hitachi)*

**Topics: "Hot" Techs for "Cool" AI Computing: Do We have Enough Tricks?**

*Organizer and Moderator: Masato Motomura (Tokyo Institute of Technology)*
*Panelists: Yusuke Doi (Preferred Networks)*
*Avi Baum (Hailo, Israel)*
*Art Swift (Esperanto Technologies, USA)*
*Mitsuhisa Sato (Riken)*

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

**15:00-15:50** **Special Session II**

15:00-15:50 **Keynote Presentation 1**
*Co-chairs: Fumio Arakawa (Univ. of Tokyo), Yasuo Unekawa (Toshiba Electronic Devices & Storage)*

**Why Preferred Networks Made MN-Core?**
*Yusuke Doi (Preferred Networks)*

**Abstract:** At Preferred Networks, we use deep learning as the core of our technology to contribute to various customers, including those in the manufacturing, biotechnology, and healthcare industries. As efficient computation is a critical differentiator in this field, we are also working on high-efficiency computation using MN-Core, an ASIC that we made. Preferred Networks, which initially started with software and algorithm technology as its core, decided to create MN-Core because of how to utilize the power of software in hardware and the economic aspect of computational optimization. In this talk, I will introduce MN-Core backgrounds and targets and the industrial impacts achieved by vertical integration from software to hardware.

**15:50-16:00** **Break**

**16:00-18:00** **Poster Short Speeches**
*Co-chairs: Koji Hashimoto (Fukuoka Univ.), Takumi Uezono (Hitachi)*

Poster 1 **Parallel Implementation of CNN on PYNQ Cluster**
*Yasuyu Fukushima, Kensuke Iizuka, Hideharu Amano (Keio Univ.)*

Poster 2 **Just-In-Time Machine Code Translator for Deep Learning Processing on Supercomputer Fugaku**
*Kentaro Kawakami, Kouji Kurihara, Masafumi Yamazaki, Takumi Honda, Naoto Fukumoto (Fujitsu La*bs.)

Poster 3 **A Toolkit for Power Behavior Analysis of HLS-Designed FPGA Circuits**
*Qidi Zhang, Xiangbo Kong, Hiroyuki Tomiyama (Ritsumeikan Univ.)*

Poster 4 **Multi-FPGA board design using CyberWorkBench, a high-level synthesis tool**
*Hiroaki Suzuki[1], Wataru Takahashi[2], Kazutoshi Wakabayashi[3], Hideharu Amano[1] ([1]Keio Univ., [2]NEC, [3]Univ. of Tokyo)*

Poster 5 **Design and Evaluation of High-performance SHA-3 System on Chip for Society 5.0**
*Tri Dung Phan, Thi Hong Tran, Yasuhiko Nakashima (Nara Institute of Technology)*

Poster 6 **Low-power convolutional neural network accelerator for edge computing**
*Yasuhiro Nakahara[1], Masato Kiyama[1], Motoki Amagasaki[1], Qian Zhao[2], Masahiro Iida[1] ([1]Kumamoto Univ., [2]Kyushu Institute of Technology)*

Poster 7 **Test of less configuration memory FPGA**
*Yuya Nakazato[1], Motoki Amagasaki[1], Qian Zhao[2], Masahiro Iida[1], Morihiro Kuga[1] ([1]Kumamoto Univ., [2]Kyushu Institute of Technology)*

Poster 8 **Evaluation of Narrow Bit-Width Variation for Training Neural Networks**
*Tomoya Akabe[1], Mutsumi Kimura[2], Yasuhiko Nakashima[1] ([1]Nara Institute of Technology, [2]Ryukoku Univ.)*

Poster 9 **A Case Study of DNN Pre-Process with System-Level Design on FPGA**
*Akiyoshi Tanaka[1], Ryota Yamamoto[1], Shinji Ito[3], Shinya Honda[1,2], Masato Edahiro[1] (Nagoya Univ.[1], Nanzan Univ.[2], System-I[3])*

Poster 10 **Evaluation of Neural Network Based Scan Matching for SLAM SoC Implementations**
*Keisuke Sugiura, Hiroki Matsutani (Keio Univ.)*

Poster 11 **Software management system of the PYNQ cluster**
*Takumi Inage, Kazuei Hironaka, Kensuke Iizuka, Hideharu Amano (Keio Univ.)*

Poster 12 **A Conflict-Aware Capacity Control Mechanism for Deep Cache Hierarchy**
*Jiaheng Liu[1], Ryusuke Egawa[2], Mulya Agung[1], Hiroyuki Takizawa[1] ([1]Tohoku Univ., [2]Tokyo Denki Univ.)*

Poster 13 **Optimal placement of coherence directories using memory networks**
*Yuki Kameyama[1], Yoshiya Shikama[1], Naoya Niwa[1], Michihiro Koibuchi[2], Hideharu Amano[1] ([1]Keio Univ., [2]National Institute of Informatics)*

Poster 14 **An Online Trace-Driven Cache Simulator for ARM-Based Supercomputers**
*Kazuki Chugo, Yukinori Sato (Toyohashi Univ. of Technology)*

## April 15, 2021 (Japan Standard Time)

**9:00-9:10** **Session I**

9:00-9:10 **Welcome and Opening Remarks**
*Co-chairs: Yuki Kobayashi (NEC), Takuya Nakaike (IBM)*

| | |
|---|---|
| *Kunio Uchiyama* | *Chair of the Organizing Committee* |
| *Tadao Nakamura* | *Chair of the Steering Committee* |
| *Jose Renau* | *Chair of IEEE/CS TCMM* |
| *Hiroyuki Tsuda* | *President of IEICE/ES* |

**9:10-10:00** **Session II**

9:10-10:00 **Keynote Presentation 2**
*Co-chairs: Yuki Kobayashi (NEC), Takuya Nakaike (IBM)*

**"虎穴に入らずんば虎子を得ず"<High Risk, high return /No Risk, no return>: Domain-specific Processors make for Cool Solutions**
*Avi Baum (Hailo, Israel)*

**Abstract:** In recent years, domain specific architectures are thriving. One main reason that fuels this trend is the prolific domain of machine learning. In this talk I will briefly survey some of the main approaches and a glimpse into theoretical aspects that underlie their suggested benefit. I will share some observations on present and future developments in the field and share my subjective view on about the possible implications on compute architectures.

**10:00-10:10** **Break**

**10:10-11:00** **Session III: Application Specific Systems with FPGAs**
*Co-chairs: Ryohei Kobayashi (Univ. of Tsukuba), Yasutaka Wada (Meisei Univ.)*

10:10-10:35 **Hybrid Network of Packet Switching and STDM in a Multi-FPGA System**
*Tomoki Shimizu, Kohei Ito, Kensuke Iizuka, Kazuei Hironaka, Hideharu Amano (Keio Univ.)*

10:35-11:00 **High Performance Multicore SHA-256 Accelerator using Fully Parallel Computation and Local Memory**
*Van Dai Phan, Hoai Luan Pham, Thi Hong Tran, Yasuhiko Nakashima (Nara Institute of Science and Technology)*

**11:00-11:10** **Break**

**11:10-12:00** **Session IV**

11:10-12:00 **Keynote Presentation 3**
*Co-chairs: Teruaki Sakata (Hitachi), Masato Suzuki (Socionext)*

**High-Efficiency Inferencing for Scalable Machine Learning**
*Art Swift (Esperanto Technologies, USA)*

**Abstract:** The extraordinary market demand for large-scale machine learning

solutions requires more than GPUs, FPGAs, or large multiplier arrays. These approaches deliver high performance, but at high costs: high power consumption, prohibitively complicated programming models, and unacceptable inflexibility. Esperanto Technologies CEO Art Swift will describe the architectural approach and design methodology for the company's first supercomputer-on-chip solution for ML inferencing acceleration. The ET-SoC-1 combines the traditional flexibility and programmability of CPU cores with the high efficiency of autonomous tensor processing to deliver unmatched system-level efficiency and all-layer ML acceleration. Every element of Esperanto's integrated solution represents best-in-class technology: the simplicity of the RISC-V instruction set, proprietary instruction-set extensions for machine learning, an on-chip mesh interconnect, a uniquely optimized memory hierarchy, state of the art process technology, and custom low-voltage circuits. In this way, Esperanto delivers more performance per watt than existing products without compromising flexibility.

**12:00-13:00** **Lunch Time Break**

**13:00-13:50** **Session V: Deep Learning Acceleration I**
*Co-chairs: Yuichiro Shibata (Nagasaki Univ.), Kazushi Kawamura (Tokyo Institute of Technology)*

13:00-13:25 **An Energy-efficient Deep Neural Network Training Processor with Bit-slice-level Reconfigurability and Sparsity Exploitation**
*Donghyeon Han, Dongseok Im, Gwangtae Park, Youngwoo Kim, Seokchan Song, Juhyoung Lee, Hoi Jun Yoo (KAIST, Korea)*

13:25-13:50 **In Search of the Performance- and Energy-Efficient CNN Accelerators**
*Stanislav Sedukhin[1], Yoichi Tomioka[1], Kohei Yamamoto[2] ([1]The Univ. of Aizu, [2]Oki Electric Industry)*

**13:50-14:10** **Break**

**14:10-15:00** **Session VI**

14:10-15:00 **Keynote Presentation 4**
*Co-chairs: Yuetsu Kodama(Riken), Yasushi Inoguchi (JAIST)*

**Codesign and System of the Supercomputer "Fugaku"**
*Mitsuhisa Sato (Riken)*

**Abstract:** We have been carrying out the FLAGSHIP 2020 Project to develop the Japanese next-generation flagship supercomputer, "Fugaku". We have designed an original manycore processor based on Armv8 instruction sets with the Scalable Vector Extension (SVE), an A64FX processor, as well as a system including interconnect and a storage subsystem with the industry partner, Fujitsu. The "co-design" of the system and applications is a key to making it power efficient and high performance. We determined many architectural parameters by reflecting an analysis of a set of target applications provided by applications teams. As a result, the system has been proven to be a very power-efficient system, and it is confirmed that the performance of some target applications using the whole system is more than 100 times the performance of the K computer. In this talk, the pragmatic practice of our co-design effort for "Fugaku" and its performance will be presented as well as an overview of system software.

**15:00-15:10** **Break**

**15:10-16:00** **Session VII: High Performance Processors**
*Co-chairs: Ryusuke Egawa (Tokyo Denki Univ.), Hiroyuki Takizawa (Tohoku Univ.)*

15:10-15:35 **Power/Performance/Area Evaluations for Next-Generation HPC Processors using the A64FX Chip**
*Eishi Arima[1], Yuetsu Kodama[2], Tetsuya Odajima[2], Miwako Tsuji[2], Mitsuhisa Sato[2]*
*([1]Univ. of Tokyo, [2]Riken)*

15:35-16:00 **A Timing Aware Connectivity Optimization Technique for Improving Energy Efficiency of High-Performance CPUs**
*Ayan Datta, Karanvir Singh, Arpita Dutta, Kousik Debnath (Intel, India)*

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

**16:30-18:30** **Online Social Hour**

## April 16, 2021 (Japan Standard Time)

**9:00-9:40** **Session VIII**

9:00-9:40 **Invited Presentation 1**
*Co-chairs: Yukinori Sato (Toyohashi Univ. of Technology), Kunio Uchiyama (AIST)*

**Architectural Challenges in the Era of New Technologies and Extreme Heterogeneity**
*Anastasiia Butko (Lawrence Berkeley National Laboratory)*

**Abstract:** As the end of the Moore's Law is approaching, we enter the era of new technologies and extreme heterogeneity. Novel architectures bring new challenges in their adoption and integration into larger systems. For example, adopting quantum accelerators hinges on building a classical control hardware pipeline that is scalable, extensible, and provides a real-time response. The physical nature of quantum devices creates non-trivial architectural challenges for control hardware that cannot be solved with the existing approaches. In this talk, we address the architectural challenges related to the adoption of novel accelerators and how these challenges can be addressed with the open-source hardware trends.

**9:40-10:00** **Break**

**10:00-10:50** **Session IX: Memory**
*Co-chairs: Kotaro Shimamura (Hitachi), Masanori Muroyama (Tohoku Institute of Technology)*

10:00-10:25 **A Metadata Prefetching Mechanism for Hybrid Memory Architectures**
*Shunsuke Tsukada, Hikaru Takayashiki, Masayuki Sato, Kazuhiko Komatsu, Hiroaki Kobayashi (Tohoku Univ.)*

10:25-10:50 **Nonvolatile SRAM Using Fishbone-in-Cage Capacitor in a 180 nm Standard CMOS Process for Zero-standby and Instant-powerup Embedded Memory on IoT**
*Takaki Urabe[1], Hiroyuki Ochi[2], Kazutoshi Kobayashi[1] ([1]Kyoto Institute of Technology, [2]Ritsumeikan Univ.)*

**10:50-11:10** **Break**

**11:10-11:50** **Session X**

11:10-11:50 **Invited Presentation 2**
*Co-chairs: Akihiro Hashiguchi (Sony), Yoshio Hirose (Fujitsu)*

**The CMOS image sensor Advance in key technology and the Introduction of Next-generation image sensor**
*Akito Kuwabara (Sony Semiconductor Solutions)*

**Abstract:** The CMOS image sensor is widely used not only in video cameras, digital still cameras and smartphones and security cameras, but also in-vehicles and medical, because its productivity and performance has been improved through the development of basic semiconductor technology and stacked structure

technology. In particular, the stacked CMOS image sensor has made it possible to mount various processing circuits on sensor edge and has expanded possibilities of the CMOS image sensor. For example, Intelligent Vision Sensor equipped with CNN（Convolutional Neural Network） processing on sensor edge enables high-speed edge AI processing and extraction of only the necessary data(Metadata), which, when using cloud AI processing, reduces data transmission latency, power consumption and communication costs, and protects privacy and confidential information. In this presentation, we will explain the CMOS image sensor advance in key technology and the features of Intelligent Vision Sensor using stacked structure technology and its architecture.

**11:50-13:20** **Lunch Time Break**

**13:20-14:10** **Session XI: Deep Learning Acceleration II**
*Co-chair: Shunsuke Sasaki (Toshiba Electronic & Devices Storage), Ryuichi Sakamoto (Univ. of Tokyo)*

13:20-13:45 **LSFQ: A Low Precision Full Integer Quantization for High-Performance FPGA-based CNN Acceleration**
*Zhenshan Bao, Kang Zhan, Wenbo Zhang, Junnan Guo (Beijing Univ. of Technology, China)*

13:45-14:10 **Training Low-Latency Spiking Neural Network through Knowledge Distillation**
*Sugahara Takuya, Renyuan Zhang, Yasuhiko Nakashima (Nara Institute of Science and Technology)*

**14:10-14:20** **Poster Award and Closing Remarks**
*Makoto Ikeda, Program Committee Co-chair (Univ. of Tokyo)*