Final Program

April 17, 2019 *Main Hall(7th Floor), Yokohama Joho Bunka Center*
(Yokohama Media & Communications Center)

13:30-15:00 **Special Invited Lecture 1**
Chair: Tohru Ishihara (Nagoya Univ.)

13:30-15:00 **Design, Compilation, and Acceleration for Deep Neural Networks in IoT Applications**
Deming Chen (Univ. of Illinois at Urbana-Champaign, USA)

Abstract: Many new IoT (Internet of Things) applications are driven by the fast creation, adaptation, and enhancement of various types of Deep Neural Networks (DNNs). DNNs are computation intensive. Without efficient hardware implementations of DNNs, these promising IoT applications will not be practically realizable. In this talk, we will analyze several challenges facing the AI and IoT community for mapping DNNs to hardware accelerators. Especially, we will evaluate FPGA's potential role for accelerating DNNs for both the cloud and edge devices. Although FPGAs can provide desirable customized hardware solutions, they are difficult to program and optimize. We will present a series of effective design techniques for implementing DNNs on FPGAs with high performance and energy efficiency. These include automated hardware/software co-design, the use of configurable DNN IPs, resource allocation across DNN layers, smart pipeline scheduling, Winograd and FFT techniques, and DNN reduction and re-training. We showcase several design solutions including Long-term Recurrent Convolution Network (LRCN) for video captioning, bidirectional LSTM for machine translation, and Inception module (GoogleNet) for face recognition. We will also present some of our recent work on developing new DNN models and data structures for achieving higher accuracy for several interesting applications such as crowd counting, music synthesis, and smart sound.

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

15:30-17:00 **Special Invited Lecture 2**
Chair: Tohru Ishihara (Nagoya Univ.)

15:30-17:00 **Low Power Design: Facts, Myths, and Misunderstandings**
Youngsoo Shin (KAIST, Korea)

Abstract: Low-power design is a standard practice these days, so quite often people do not pay much attention on the details. This may create misconception and misunderstanding. An example is a simple statement "Our design consumes 100mW". Power consumption cannot be declared as a single concrete number because of complications in the estimation or measurement process with inherent inaccuracies and uncertainties. Another popular example is "Our design consumes lowest power of 100mW". After understanding that a chip is typically designed with huge amount of margins, it is easy to see that the limit is far beyond. A few of these details of low power design, which often go unnoticed but deserve careful consideration, are addressed in this talk.

17:00-17:15 **Break**

17:15-17:55 **Invited Presentation 1**
Co-chairs: Yoshio Hirose (Fujitsu), Yukinori Sato (Toyohashi Univ. of Technology)

17:15-17:55 **DLU and Domain Specific Computing**
Takumi Maruyama (Fujitsu)

Abstract: Fujitsu recently started development of Domain specific processor such as DLU (Deep Learning Unit) as well as DA (Digital Annealer), in addition to conventional server processors which Fujitsu has developed for many years such as GS processors for mainframe, SPARC processors for UNIX server, and SPARC/ARM processors for HPC. In this talk, what is common and what is unique in domain specific processors compared with conventional general processors will be explained through DLU and DA as examples, to discuss about the future of computer architecture.

April 18, 2019 *Main Hall(7th Floor), Yokohama Joho Bunka Center*
(Yokohama Media & Communications Center)

9:30-9:50 **Session I**

9:30-9:50 **Welcome and Opening Remarks**
Co-chairs: Yuki Kobayashi (NEC), Hiroki Matsutani (Keio Univ.)

Hideharu Amano	*Chair of the Organizing Committee*
Allen J. Baum	*Chair of IEEE/CS TCMM*
Akihiko Kasukawa	*President of IEICE/ES*

9:50-10:40 **Session II**

9:50-10:40 **Keynote Presentation 1**
Co-chairs: Masato Suzuki (Socionext), Kunio Uchiyama (Hitachi)

GPU: A true AI Cool-Chip with High Performance/Power Efficiency and Full-Programmability
Toru Baji (NVIDIA)

Abstract: In the early days of processor LSIs, CPU performance increased almost 1.5 times / year thanks to the Moor's Law. However, around year 2010, due to the leakage current, too complex CPU architecture and Amdahl's Law limitation this rate becomes 1.1 times / year. Today where Moor's Law is at its end, and the performance increase is reported to be almost 3% / year. On the other hand, parallel processing dedicated GPU continues to grow its performance with the rate of 1.5 times / year, and even with the Moor's Law ending, it still continue to grow its performance by architecture evolution and various built-in accelerators. Pascal GPU came with Tensor Core which accelerates the AI matrix multiplication 12 times. And the latest Turing GPU further added an RT Core accelerator to increase the most complex 3D computer graphics (ray tracing) more than 10 times. GPU delivers around one to two order of magnitude performance advantage over the CPU and now is the most widely used processor accelerator in AI and Supercomputing. On the other hand, due to its local data processing architecture and careful circuit/layout design, the performance/power efficiency is almost one order of magnitude better than the CPU. The AI application demands for extraordinary high performance and full programmability to meet the rapidly evolving algorithms. Now it might be said that only the GPU can meet these contradictory requirements. In this talk, GPU basic technology which realize this high performance and high power efficiency will be introduced. Also the applications to Supercomputing, AI and ML and advanced autonomous driving will be reported.

10:40-11:00 **Break**

11:00-11:50 **Session III:**

11:00-11:50 **Keynote Presentation 2**
Co-chairs: Takuya Nakaike (IBM), Hiroki Matsutani (Keio Univ.)

Quantum Computing at IBM – from hardware to software
Patryk Gumann (IBM, USA)

Abstract: In quantum computing, information is processed in a fundamentally

different way than classical computing: by taking advantage of quantum phenomena such as entanglement and interference, a number of quantum algorithms have been, theoretically proven to offer a speedup over classical algorithms. However, the difficulty lies in controlling and operating the quantum bits, or qubits, that make up a quantum computer. All physical qubits are notoriously fragile and sensitive to any fluctuations in their environment, which makes the promise of quantum computing difficult to realize. Fortunately, due to tremendous material research and development, deeper understanding of underlaying decoherence mechanisms as well as smart microwave and cryogenic engineering we are getting closer to building a first NISQ type of quantum computers. At IBM we focus on parallel efforts to improve the qubits and controls to ultimately achieve fault tolerance while also searching for near-term applications that do not require error correction. Recent results in quantum chemistry and on error mitigation techniques show that quantum computing may be able to tackle some of its most anticipated applications before full universal fault-tolerant quantum computers are realized.

11:50-12:00 **Break**

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

Poster 1 **A Deep Learning Tennis Ball Collection Robot and its Future Implementation on FPGA**
Shenshen Gu (Shanghai Univ., China)

Poster 2 **Performance Evaluation of Fine-grained Virtual FPGA Based on SLM Architecture**
Theingi Myint[1], Motoki Amagasaki[1], Qian Zhao[2], Masahiro Iida[1], Morihiro Kuga[1] ([1]Kumamoto Univ., [2]Kyushu Institute of Technology)

Poster 3 **FiC-RNN: A Framework for Accelerating Deep Recurrent Neural Networks on Multiple FPGAs**
Yuxi Sun, Akram Ben Ahmed, Hideharu Amano (Keio Univ.)

Poster 4 **Implementation of Burrows Wheeler Transform on multi-FPGA system: Flow-in-Cloud**
M M Imdad Ullah, Kazusa Musha, Akram Ben Ahmed, Hideharu Amano (Keio Univ.)

Poster 5 **Congestion-aware Collective Communications on Multi-Layer Full-Mesh Topology**
Masahiro Miwa, Toshihiro Shimizu, Ryuichi Sekizawa, Yoshiyasu Doi (Fujitsu Labs.)

Poster 6 **A Machine Learning-based CPU Temperature Prediction Model Considering Environmental Features**
Seung Hun Choi, Seon Young Kim, Young Geun Kim, Sung Woo Chung (Korea Univ., Korea)

Poster 7 **Design Optimization Methodology for FPGA-Based Accelerator with Multiple Users**
Shuhei Yoshida[1], Yuta Ukon[1], Koji Yamazaki[2], Koyo Nitta[1] ([1]NTT, [2]NTT Advanced Technology)

Poster 8 **Design exploration of PRNG for spiking neural network processors**
Atsuya Okazaki, Masatoshi Ishii, Junka Okazawa (IBM)

Poster 9 **Partial Reconfiguration Technique for a Multi-board FPGA System FiCSW**
Miho Yamakura, Kazuei Hironaka, Keita Azegami, Kazusa Musha, Hideharu Amano (Keio Univ.)

Poster 10 **FPGA SoC implementation of a subroutine for Hall Thruster Simulation in OpenCL**
Manfred Orsztynowicz[1], Hiroyuki Noda[1], Takaaki Miyajima[2], Naoyuki Fujita[3], Hideharu Amano[1] ([1]Keio Univ., [2]RIKEN, [3]JAXA)

Poster 11 **A Low Power Radix-4 Booth Multiplier with Precise Operand Exchange Technique**
Yu-Cheng Cheng, Yen-Yuan Wang, Yen-Jen Chang (National Chung Hsing Univ., Taiwan)

Poster 12 **Workload-based Dynamic SCM Capacity Management of SCM/NAND Flash Hybrid Storage**
Chihiro Matsui, Ken Takeuchi (Chuo Univ.)

Poster 13 **Parallel Task Scheduler for High Performance Video Codec Accelerator**
Kwang-Hyun Choi, Hyun Eun, Jee-Hyun Jun, Ju-Ock Lee, Seok-Hee Kim, Hyun-Gyu Kim (Chips & Media, Korea)

Poster 14 **Efficient Flash Memory-Access Power Reduction Techniques for IoT-Driven Rare-Event Logging Application**
Jisu Kwon, Jeonghun Cho, Daejin Park (Kyungpook National Univ., Korea)

Poster 15 **The Framework for Image Processing Aiming to Autonomous Car Using FPGA**
Koki Honda, Wei Kaijie, Hideharu Amano (Keio Univ.)

Poster 16 **Real Chip Performance Evaluation of Inductive Coupling TCI IP**
Hideto Kayashima, Takuya Kojima, Hayate Okuhara, Tsunaaki Shidei, Hideharu Amano (Keio Univ.)

Poster 17 **Implementation of ART algorithm with Xilinx SDAccel**
Yasuaki Okamoto, Hideharu Amano (Keio Univ.)

Poster 18 **An Energy Optimization Method for Hybrid In-Memory Checkpointing**
Muhammad Alfian Amrizal, Mulya Agung, Ryusuke Egawa, Hiroyuki Takizawa (Tohoku Univ.)

Poster 19 **Implementation of Training Phase of Convolutional Neural Networks on Multi FPGA**
Aoi Hiruma, Hideharu Amano, Kazusa Musha, Yugo Yamauchi (Keio Univ.)

Poster 20 **Designing an FPGA Accelerator with Optimization and Specialization for On-Board DRAM**
Kenta Sato, Yukinori Sato (Toyohashi Univ. of Technology)

Poster 21 **Preliminary Discussion of a Time Stride Prefetching**
Hayato Nomura, Tomoki Nakamura, Toru Koizumi, Hidetsugu Irie, Shuichi Sakai (Univ. of Tokyo)

Poster 22 **Mu-Law as Dynamic-Range Parameter Representation for DNN**
Charles Choo[1], Byung-Joo Kim[2] ([1]San Jose State Univ., USA, [2]Mando Innovation Silicon Valley, USA)

12:30-13:50 Lunch Time Break

13:50-14:00 **Poster Open: 7th floor poster show room**

14:00-14:50 **Session V**

14:00-14:50 **Keynote Presentation 3**

Co-chairs: Yuki Kobayashi (NEC), Yasuo Unekawa (Toshiba Electronic Devices & Storage)

Vector Engine Processor of NEC's Brand-New Supercomputer SX-Aurora TSUBASA

Yoshihiro Konno (NEC)

Abstract: NEC has released the latest vector supercomputer, SX-Aurora TSUBASA in 2018. It features superior sustained performance, especially for memory-intensive scientific applications and it inherits DNA over 30 years by the SX Series from the SX-1/2 to the SX-ACE with their specialized vector processors. The system architecture of SX-Aurora TSUBASA is drastically changed from its predecessors of the SX series. The SX-Aurora TSUBASA mainly consists of a vector host (VH) and one or more vector engines (VEs). VH is a standard x86/Linux server, which provides Linux OS functions, and VE OS developed by NEC runs on the Linux to control VEs. VE is implemented as a PCI Express (PCIe) card equipped with the newly developed vector processor, and is connected to VH. By this architecture change the SX-Aurora TSUBASA became able to be applied from the dedicated large scale supercomputer to desktop server. In the presentation at Cool Chips, we will elaborate the design of the VE processor, including its vector architecture such as configurations of vector pipelines and registers, execution sequences of vector operations, configuration of the memory network, and power control and fault tolerance mechanisms. We will also show overall configurations of the SX-Aurora TSUBASA system, VE card implementation including cooling, sustained performance in wider area of benchmark programs and some use cases.

14:50-15:40 **Break (Poster Open: 7th floor poster show room)**

15:40-16:55 **Session VI: Cool Software**

Chair: Hiroyuki Takizawa (Tohoku Univ.), Yasutaka Wada (Meisei Univ.)

15:40-16:05 **The Impacts of Locality and Memory Congestion-aware Thread Mapping on Energy Consumption of Modern NUMA Systems**

Mulya Agung, Muhammad Alfian Amrizal, Ryusuke Egawa, Hiroyuki Takizawa (Tohoku Univ.)

16:05-16:30 **Hybrid Access in Storage-class Memory-aware Low Power Virtual Memory System**

Yusuke Shirota, Satoshi Shirai, Tatsunori Kanai (Toshiba)

16:30-16:55 **A Compiler for Deep Neural Network Accelerators to Generate Optimized Code for a Wide Range of Data Parameters from a Hand-crafted Computation Kernel**

Eri Ogawa, Kazuaki Ishizaki, Hiroshi Inoue, Swagath Venkataramani, Jungwook Choi, Wei Wang, Vijayalakshmi Srinivasan, Moriyoshi Ohara, Kailash Gopalakrishnan (IBM)

16:55-17:10 **Break**

17:10-18:40 **Session VII: Panel Discussions**

Topics: "Where will the computer architecture go?"

Organizer & Moderator: Yasunori Kimura (JST)
Panelists: Toru Baji (NVIDIA)
Deming Chen (Univ. of Illinois at Urbana-Champaign, USA)
Takumi Maruyama (Fujitsu)
Sanjay Patel (Wave Computing, USA)

18:40-19:00 **Break**

19:00-21:00 **Banquet**

April 19, 2019 *Main Hall(7th Floor), Yokohama Joho Bunka Center*
(Yokohama Media & Communications Center)

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

9:30-10:20 **Keynote Presentation 4**
Co-chairs: Chikafumi Takahashi (Desno), Hideharu Amano (Keio Univ.)

A Unified Platform for Processing at the Edge
Sanjay Patel (Wave Computing, USA)

Abstract: There is a need in AI to move intelligence to the edge to circumvent challenges that include latency, security, and bandwidth for AI use cases at the edge. This trend will accelerate with the continued proliferation of IoT. Utilization of cloud-base, AI-as-a-service to address edge use cases can become exceedingly expensive. AI processing on a remote server system can be unreliable where connectivity is poor. Processing AI at the edge comes with its own challenges, achieving desired performance at low-power with constrained memory and processing resources. This talk discusses the ramifications of moving AI processing to the edge for not only inference but also training, focusing on the potential of CPU-centric edge-AI Architectures with acceleration assists, and the distinction of data formats for inference and training.

10:20-10:50 **Break (Poster Open: 7th floor poster show room)**

10:50-11:45 **Session IX: Machine Learning**
Co-chairs: Shunsuke Sasaki (Toshiba Electronic Devices & Storage), Kazushi Kawamura (Waseda Univ.)

10:50-11:15 **Post Training Weight Compression with Distribution-based Filter-wise Quantization Step**
Shinichi Sasaki, Asuka Maki, Daisuke Miyashita, Jun Deguchi (Toshiba Memory)

11:15-11:30 **Performance and Cost Evaluations of Online Sequential Learning and Unsupervised Anomaly Detection Core**
Tomoya Itsubo, Mineto Tsukada, Hiroki Matsutani (Keio Univ.)

11:30-11:45 **Implementing a large application(LSTM) on the multi-FPGA system: Flow-in-Cloud**
Yugo Yamauchi, Kazusa Musha, Hideharu Amano (Keio Univ.)

11:45-12:35 **Session X: Packet Processing**
Chair: Hajime Shimada (Nagoya Univ.), Kotaro Shimamura (Hitachi)

11:45-12:10 **Multi-Level Packet Processing Caches**
Kyosuke Tanaka, Hayato Yamaki, Shinobu Miwa, Hiroki Honda (Univ. of Electro-Communications)

12:10-12:35 **Key-value Store Chip Design for Low Power Consumption**
Yuta Tokusashi, Hiroki Matsutani, Hideharu Amano (Keio Univ.)

12:35-14:00 **Lunch Time Break**

14:00-14:55 **Session XI: Cache and Memory Systems**
Co-chairs: Yasutaka Wada (Meisei Univ.), Ryusuke Egawa (Tohoku Univ.)

14:00-14:25 **Cache-Aware Dynamic Classification and Scheduling for Linux**
Ravi Theja Gollapudi, Gokturk Yuksek, Kanad Ghose (State Univ. of New York at Binghamton, USA)

14:25-14:40 **Perceptron-based Cache Bypassing for Way-Adaptable Caches**
Masayuki Sato, Yongcheng Chen, Haruya Kikuchi, Kazuhiko Komatsu, Hiroaki Kobayashi (Tohoku Univ.)

14:40-14:55 **Statistical Access Interval Prediction for Tightly Coupled Memory Systems**
Robert Wittig, Mattis Hasler, Emil Matus, Gerhard Fettweis (Technical Univ. Dresden, Germany)

14:55-15:25 **Break (Poster Open: 7th floor poster show room)**

15:25-16:40 **Session XII: Signal Processing**
Chair: Makoto Ikeda (Univ. of Tokyo)

15:25-15:50 **Inter-Frame Smart-Accumulation Technique for Long-Range and High-Pixel Resolution LiDAR**
Ken Tanabe[1], Hiroshi Kubota[1], Akihide Sai[2], Nobu Matsumoto[1] ([1]Toshiba Electronic Devices & Storage, [2]Toshiba)

15:50-16:15 **Low Delay 4K 120fps HEVC Decoder with Parallel Processing Architecture**
Ken Nakamura, Yuya Omori, Daisuke Kobayashi, Tatsuya Osawa, Takayuki Onishi, Koyo Nitta, Hiroe Iwasaki, Atsushi Shimizu (NTT)

16:15-16:40 **Low Power Speaker Identification using Look Up-free Gaussian Mixture Model in CMOS**
Alberto Gianelli[1], Nick Iliev[1], Shamma Nasrin[1], Mariagrazia Graziano[2], Amit Ranjan Trivedi[1] ([1]Univ. of Illinois at Chicago, USA,, [2]Politecnico di Torino, Itary)

16:40-17:00 **Break**

17:00-17:40 **Session XIII**

17:00-17:40 **Invited Presentation 2**
Co-chairs: Yasushi Inoguchi (JAIST), Koyo Nitta (NTT)

A64FX High Performance Processor Architecture and its Design Challenges
Shuji Yamamura (Fujitsu)

Abstract: In the last year Fujitsu introduced A64FX at Hot Chips 30. A64FX is the latest Fujitsu's HPC processor which is designed for Post-K supercomputer. Fujitsu is developing Post-K supercomputer as a successor to K computer with RIKEN. A64FX is the latest Fujitsu's HPC processor based on our own microarchitecture, as used in our SPARC64 and mainframe processor development. In this talk, A64FX microarchitecture will be explained and also cache memory system and on-chip network which can provide coherent and high throughput memory to many cores on the chip. We will also introduce our steady development effort in combination of the front-end and the back-end implementations.

17:40-18:00 **Poster Award and Closing Remark**
Makoto Ikeda, Program Committee Co-chair (Univ. of Tokyo)