公開

研究終了報告書

「Single-Wire Driven Efficient Approximate Computing Technologies」

研究期間： 2018 年 10 月～2022 年 3 月
研 究 者： 張 任遠

1． 研究のねらい

This project aims at exploring the **greatest feature, benefit and proper up/down tier technologies for various approximate computing (AC) mechanisms** instead of straightforwardly creating a new or hybrid of several existing AC schemes.

[Identify benefit] Along with the boom of AI, many AC technologies have been proved feasible and profitable with their respective gain-loss maps. As the foundation of this project, it is expected to identify the greatest benefit of entire AC scope. At the kick-off phase of this project, the interconnection reduction (due to single wire driven) is predicted as the principle benefit of AC. The propriety of this statement is verified, hopefully.

[Principle feature/challenge] The flexibility/re-configurability of is assumed as the principle feature/challenge of AC hardware. The milestone of this project will lie on developing the highly flexible platform for general purpose applications by single wire driven AC technologies.

[up/down tier technologies] Regardless of concrete type of AC mechanisms, the ultra-massive core computing architecture must be constructed by AC hardware. Thanks to two above, much more efficient data-traffic scheme is feasible. Consequently, proper architectures and reasonable physics devices are necessary. As the target of this project, it is expected to **develop full-tier technical chain for various single-wire driven AC platforms, which cover the architecture, mechanism, and device levels.** Furthermore, the new scheme of AC is expected to create from the hybrid data representations by existing forms such as analog, stochastic, spiking and else. At last, it is impossible to conclude the "best" AC technology in real-world. However, it is feasible and important to develop the optimum architecture and physics device matching the specific AC mechanism. Concluding this project, several tier-maps should be identified.

２．　研究成果

（１）概要

In this project, the new schemes of analog, stochastic, and spike coding based computational circuits were developed with rich benefits over energy consumption, hardware scale, and inter-connection. During the exploring phase, several new mechanisms of single-wire driven approximate computing (AC) were created by the hybrid of above. Both of computational efficiency and quality are improved with different features as: (a) hybrid of analog and stochastic appears higher accuracy and speed than plain analog and stochastic computing circuits, respectively; (b) hybrid of stochastic and spike coding offers higher quality in the machine learning applications than the spiking neural network; (c) hybrid of analog spike coding enable our novel architecture (see below) in large scale neural networks.

On the basis of above mechanisms, the corresponding proper computing architectures were developed, which escape from the main-stream of many core platforms such as Von-Neumann type, CGRA, or systolic array. The programmable analog calculation units (ACUs) along with novel analog RAM were invented for tensor computing architectures with ultra-small size (high parallelism, namely). The in-memory computing architecture for stochastic computing was proposed for neural network implementations with world-leading cost-quality trade-off map. As the principle progress on the architecture level, we developed a zero-redundancy elastic architecture “DiaNet” for general purpose AC acceleration; even four generations of DiaNet series were released with the superior re-configure efficiency.

On the physics level, not only CMOS devices but also new materials were implemented for respective single-wire driven computing technologies. The Mem-Capacitor based circuits with analog data representation and the memrister based stochastic computing circuits appears higher efficiency than all of similar CMOS implementations.

Concluding this project, the optimum technical tier-maps were clarified as: (a) [arithmetic app.s] + [Von-Neumann-Like tensor computing architecture] + [analog computing] + [CMOS]; (b) [NN app.s] + [in-memory computing architecture] + [stochastic mechanism] + [memrister]; (c) [general app.s] + [DiaNet] + [analog/spiking] + [CMOS].

（２）詳細

Research theme A: computing mechanism driven by single wire

In this scope, mechanism/circuit level explorations of approximate computing technologies, which include analog, stochastic, and spike based computing and their hybrid, were conducted as follows:

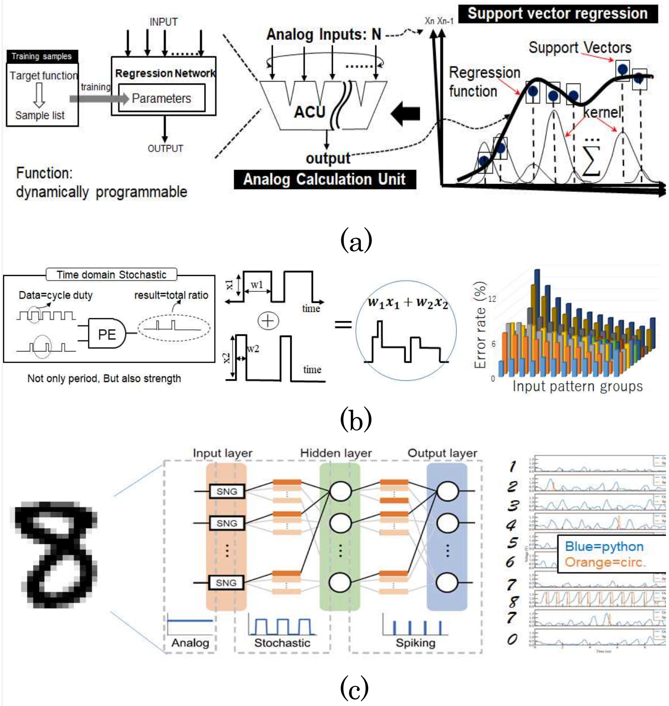

Fig. 1. (a) Programmable analog computing units; (b) analog/stochastic hybrid computing mechanism; (c) analog/stochastic/spike hybrid coding

[analog computing] The programmable analog calculation unit (see Fig. 1(a)) for approximate computing is developed along with its dynamic and static memory circuits [j-3]. Arbitrary complex functions with two operands can be carried out in continuous time by a circuitry with 600 MOS transistors. A power-gating technology is suggested for ACU to eliminate the static power consumption. For proof-of-concept, several example functions are demonstrated. From circuit simulation results, the proposed ACU correctly calculates all the exampled functions with the average error less than 1.7%. We also illustrate the robustness of ACU against temperature and process variations. By applying the pro-posed analog DRAM and Hexadecimal flip-flop, the calculation results of ACU are available in three modes: original analog signal, Hexadecimal, and four-bit digital signal. [analog-stochastic hybrid computing] The multi-domain stochastic computing circuit is proposed for polynomial functions (see Fig. 1(b)) [c-4]. The multiply operation is implemented by analog current pulses where the strength and duty cycle represent two variables. The summation is performed by simply accumulating the current pulses. Employing the original Neuron-MOS based TBSC module and current mirrors, entire MDSC is designed and simulated in a 0:18um CMOS technology. A toy-example is demonstrated for MAC operations. From the circuit simulation results, an average of calculation accuracy of 95.3% is achieved. The performances over data range, circuit complexity, power consumption, and latency are all improved in contrast to the state-of-art TBSC implementations. [analog-digital-spike hybrid computing] By using a set of time-based stochastic computing (TBSC) circuits, the stochastic numbers (SNs) in continuous time-domain are directly fed into the input layer of the spiking neural networks (SNNs) without any additional spike-coding mechanism (see Fig. 1(c)). The analog circuits behaving as synapses and neurons are designed to fit the TBSC coding and generate spikes for the rest of layers. Implementing the exampled pattern recognition tasks, the MNIST recognition accuracy loss is below 4% compared to well-trained artificial neural networks (ANNs). The average firing energy is 0.94pJ per spike, which is 0.5x of state-of-art of low power SNN implementations. The energy consumption of MNIST is estimated as 0.88uJ per classification.

Research theme B: developing new architecture for approximate computing

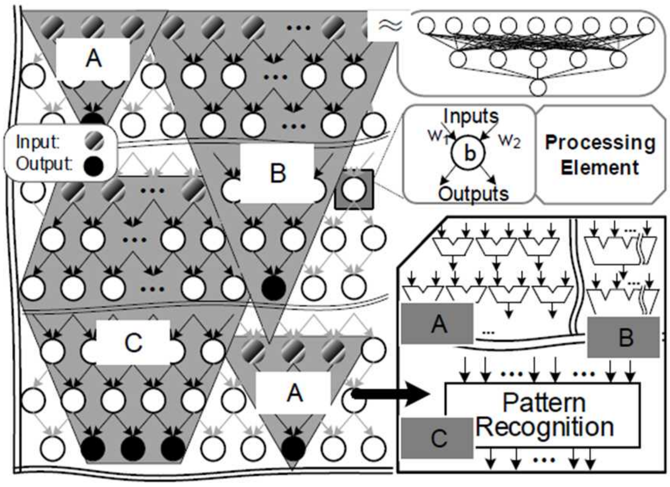

Fig. 2 DiaNet topology along with tensor computer by it

[DiaNet] We were developing an elastic architecture of tensor computation which is named ``DiaNet". This proposal is the reverse of boom of AI application. we build a special NN pre-silicon; then, make it perform arbitrary behaviors including computations or recognitions. Figure 2 illustrates the topology of DiaNet and its motivation. The prototype of DiaNet is seen in [j-1, j-2], where the ALU-like functions are implemented by regression. More complicated tasks such as vector computations or pattern recognitions are also feasible. Since the topology of DiaNet is an ultra-sparse NN, the computation in each processing element (PE) is quite simple and symmetric anywhere. Therefore, the data representation and calculation fundamental is irrelevant to the architecture theoretically. Any specific mechanism such as conventional digital, analog, spiking, stochastic, and even quantum could be applied. For instance, the analog implementation of DiaNet achieves 9-input arbitrary calculation with 96% accuracy and power of 19uW by only 720 Transistors [c-5].

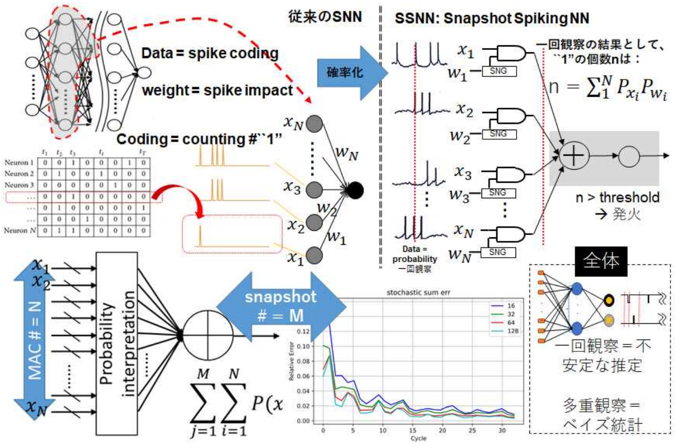

Fig. 3 Flash computing architecture

[Flash computing architecture] The concept of flash computing is seen in Fig. 3. It is realized by behaviors through the novel coding mechanisms and data representations. Differently from the stochastic computing or spiking neural network, the data is carried by the probability of spike at arbitrary timing point. The probabilities interpretations are entangled by massive synapses coupling. Only simplest logic gates and capacitors are necessary to perform entire NN. For observing the data, both of snapshot mode and accumulation mode are easily archived, where snapshot offers quick inference and accumulation gives a Bayes-like statistics feature. In this sense, the operation could be done in an ultra-fast and efficient manner. Meanwhile, the precision and speed are freely tunable. So far, the CMOS implementation of flash computing achieves 128-item MAC with 98% accuracy in two clock cycles by only two traditional ALU-comparable circuit scale.

Research theme C: applying non-CMOS device in single-wire driven computing

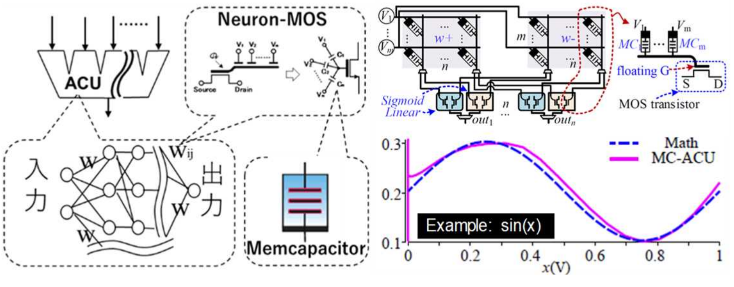

Fig. 4 Applying mem-capacitor in ACU

[Mem-Capacitor] To retrieve any specific function approximately, the regression algorithm through neural network (NN) is realized by a compact analog circuitry. The mem-capacitor (MC) technology is associated to Neuron-MOS structure, which couples multiple capacitors on the floating gate of a MOS transistor to achieve multiply-accumulation (MC) operation as Fig. 4. In this manner, each synapse of NN is emulated by only one MC device and the weight is post-fabrication programmable due to the memristive characteristics of MC. For proof-of-concept, the approximate calculation unit (ACU) for arbitrary two-operand computations is achieved with 406 devices. From circuit simulation results, all the example functions are retrieved with the maximum inaccuracy of 7.8%. Compared to state-of-art works of approximate computing, the energy- and device-count savings reach 63.3% and 84.2%, respectively.

３． 今後の展開

Three seasons are concerned in the future scenario (see Fig. 5). Firstly, the quantum-spike coding methodology will be explored. We are going to start from some toy-examples such as conventional neural networks. Two schemes including one-shot observation and statistic observation are verified to perform regression and pattern recognition. Then, we will migrate this coding methodology into our DiaNet. Secondly and simultaneously, more series of DiaNet (so far, till version 3.0) and flash computing architecture are expected to evolve. As soon as above techniques ready, our last task is to offer practical architecture solutions. We might migrate some existing tensor computing structures such as systolic ring by partitioning the DiaNet into reasonable pieces. As the further step rooting on this project, it is expected to develop the cool super computer towards almost-zero-carbon society. Considering the cooling power (very important but always ignored in computer research evaluation), the entire energy is hopefully reduced by combining CMOS devices and other cold device such as quantum or super conductor. By then, the non-deterministic computing technologies from this project will play a key role.

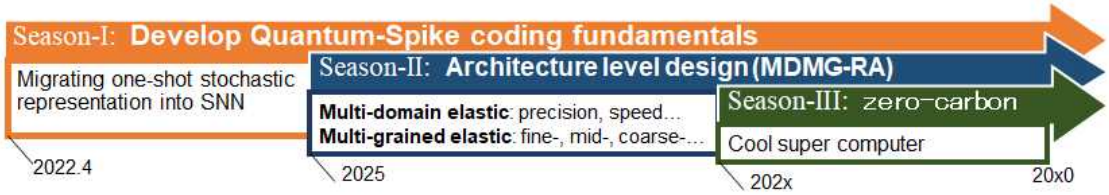
Season-I: Develop Quantum-Spike coding fundamentals
Migrating one-shot stochastic representation into SNN
Season-II: Architecture level design (MDMG-RA)
Multi-domain elastic: precision, speed…
Multi-grained elastic: fine-, mid-, coarse-…
Season-III: zero-carbon
Cool super computer
2022.4
2025
202x
20x0

Fig. 5 Future scenario of single-wire driven computing

4.　自己評価

The current progress fully matches my initial proposal. All of three tiers including physics, mechanism, and architecture levels have been explored, and the performances of our developed platforms are superior in some specific features. The resource management and budget exactly meet the demands of this project. So far, there was no any computing architecture being developed and widely concluded specifically suitable for the approximate computing mechanism. Then, our proposals are ones of bridges to explore the society of “non-Boolean computing architecture”. People are seeing our analog-stochastic-spike hybrid, DiaNet, and flash computing technologies, but the impact to real-world applications are still under observation. From this project, we start collaborating with physicists in super conductor field and kick-off new scenario of “cool data center by combing CMOS and AQFP on the basis of single-wire computing”, which is expected to be one practical solution of almost-zero-carbon computer system.

5.　主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：10件

j-1. Man Wu, Yirong Kan, Tati Erlina, Renyuan Zhang, and Yasuhiko Nakashima: ″DiaNet: An Elastic Neural Network for Effectively Re-configurable Implementation″, Elsevier-Journal Neurocomputing, Sep. (2021), Vol. 464, p. 242-251

This work is a detailed description of our proposed reconfigurable approximate computing architecture, which was named “DiaNet” in its parents work. The original prototype of DiaNet is organized as a symmetrical bisection neural network, which is feasible to be partitioned into arbitrary pieces of neural networks (NNs) without redundancy. To prevent the depth explosion in implementing complex tasks (complicated pattern recognition for instance), the evolution of DiaNets is investigated in this work. Compared with the LeNet5 model as state-of-the-art, the evolved DiaNet topology achieves the parameter reduction of 90.86% for MNIST recognition with the negligible loss of accuracy. As an approximate computing technology, the sensitivity to the decline of computational precision is investigated to suggest the guideline for efficient hardware implementations such as analog or stochastic.

j-2. Y. Kan, M. Wu, R. Zhang and Y. Nakashima: ″MuGRA: A Scalable Multi-Grained Reconfigurable Accelerator Powered by Elastic Neural Network″, IEEE Transactions on Circuits and Systems I: Regular Papers, 2021, Vol. early access, p. 1 - 14

This work is the hardware implementation of “DiaNet” for approximating arbitrary arithmetic functions. The proposed DiaNet architecture is reconfigurable in fine-grained (arbitrary functions), mid-grained (flexible function feature, accuracy, and number of operands), and coarse grained (organization of cores). By implementing a large scale of novel bisection neural network (BNN) on hardware, the reconfiguration is conducted by partitioning entire BNN into any specific pieces without redundancy. Each piece of BNN retrieves the arbitrary function approximately. From the hardware implementation results, compared with other traditional function approximation methods, our method provides fewer parameter storage requirements. The comparison against related works proves that our accelerator effectively reduces the calculation latency with slight accuracy loss.

j-3. Renyuan Zhang, Noriyuki Uetake, Takashi Nakada and Yasuhiko Nakashima: ″Design of Programmable Analog Calculation Unit by Implementing Support Vector Regression for Approximate Computing″, IEEE MICRO, 2018 Vol. 38, p.73-82

In this work, we design a programmable analog calculation unit (ACU) for approximately computing arbitrary functions with two operands. By implementing an efficient scheme of support vector regression (SVR), the target functions are retrieved by VLSI circuits in one clock cycle with only 600 transistors. From the circuit simulation results, the proposed ACU calculates all the target functions with the average error less than 1.7%. The performances over energy, flexibility, and hardware efficiency of proposed ACU are superior to a basic four-bit digital Arithmetic Logic Unit (ALU) and look-up table (LUT) based architectures. To conveniently integrate the proposed ACUs in ordinary digital systems, we also design the compact memory circuits which offer dual-mode (analog and binary) data storage/access.

（2）特許出願

研究期間全出願件数:0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

c-1. Van Tinh NGUYEN, T. -K. Luong, E. Popovici, Q. -K. Trinh, Renyuan Zhang and Yasuhiko Nakashima: ″An Accurate and Compact Hyperbolic Tangent and Sigmoid Computation Based Stochastic Logic″, IEEE International Midwest Symposium on Circuits & Systems, pp.386-390, Aug. (2021)

c-2. Y. Kan, M. Wu, R. Zhang and Y. Nakashima: ″A Multi-Grained Reconfigurable Accelerator for Approximate Computing″, 2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Limassol, CYPRUS, Jul. (2020)

c-3. 【Best Paper Run-up Award】Renyuan Zhang, Tati Erlina, Tinh Van Nguyen, and Yasuhiko Nakashima: ″Hybrid Stochastic Computing Circuits in Continuous Statistics Domain″, IEEE Int. System-on-Chip Conf., pp.225-230, Sep. 8th-11th, (2020)

c-4. Tati Erlina, Yan Chen, Renyuan Zhang and Yasuhiko Nakashima: ″An Efficient Time-based Stochastic Computing Circuitry Employing Neuron-MOS″, GLSVLSI2019, pp.51-56, May. (2019)

c-5. [Invited Talk] Renyuan Zhang, ″Analog Approximate Computing″, The 6th International Symposium on Brainware LSI, RIEC, Tohoku Univ, Mar. 1-2, 2019