AI Chips on Things for Sustainable Society: A 28-nm CMOS, Fully Spin-to-spin Connected 512-Spin, Multi-Spin-Thread, Folded Halved-Interaction Circuits Method, Annealing Processing Chip

Satoshi Kitamura, Ryoma Iimura, Takayuki Kawahara

Department of Electrical Engineereing, Tokyo University of Science, Katsushika, Tokyo, Japan

***Abstract*— For sustainable society, next-generation Internet of Things (IoT) requires ultra-low-power information processing that is useful for both sensor area expansion and high-speed low-power feature extraction using a new signal processing artificial intelligence (AI) large-scale integration (LSI) chip developed for not only the cloud side but also the "things" side (edge) with an attached sensor. For this purpose, a fully spin-to-spin connected Ising model (annealing processing) AI LSI chip was successfully demonstrated for the first time. Its specifications are as follows: 521 spins, 262,144 interactions (with a halved-interaction circuit method), and 4-bit interaction accuracy. The chip was designed and fabricated using a 28-nm CMOS process. The new circuit technologies confirmed in actual operation of the chip are a block configuration realizing all spin-to-spin interactions, 8-spin-threads (core) method, and folded halved-interaction circuit method.**

Keywords—IoT, artificial intelligence, Ising model, annealing processing, CMOS

I. Introduction

The continuously growing Internet of Things (IoT) is expected to have a large market and be important in creating sustainable societies. However, increases in the amount of data and energy consumption are bottlenecks in the development of next-generation IoT.

For this reason, in next-generation IoT, as shown in Fig. 1, on not only the cloud side but also the "things" side (edge) with the sensor attached, a new signal processing and artificial intelligence (AI) processing technology has been developed to expand the sensor area, further reduce communication load, and improve data processing capacity and energy efficiency.

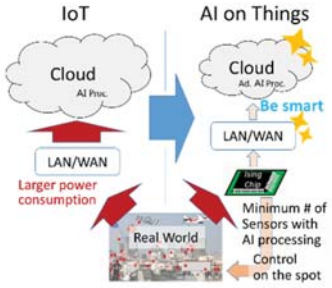

Figure 1 Next-generation IoT: AI on Things

Here, as shown in Fig. 2, features of signal processing on the things side are different from those on the cloud side. On the cloud side, processing can be performed in a suitably air-conditioned environment using highly integrated large-scale integrations (LSIs) using state-of-the-art devices. On the other hand, on things (edge), mounting size and power supply are limited. Also, for AI processing, the cloud side performs both learning and inference, but most tasks on things side are inference and control, and operation with low power consumption is required although hardware is limited. However, this AI processing on the things side is not easy. Only necessary information is searched for and extracted from complex raw data in the real world. The processing required to extract significant information from this complex raw data in the real world is specifically the same as the process for solving an optimization problem. Moreover, operation with low power consumption is required although hardware is limited.

Processor	GHz	kHz - MHz
Memory	TB, DRAM, Hybrid NVDRAM	GB, SRAM, NV-RAM
AI processing	Learning and inference	Inference and control
Power	MW, dedicated power supply	Tens of mW - μW, limited battery power, harvesting
Implementation	Air conditioning room, maintained rack	Factory, engine room, wind and rain environment

Figure 2 Comparison of computing on Cloud and Things (Edge)

To overcome these hurdles, in this paper, we develop a versatile, low-power, high-performance edge-side AI LSI chip on the basis of the Ising model that will be indispensable for realizing sustainable societies.

II. Ising Model and Fully Spin-to-spin Connection

A. Ising model and optimization problem

Fig. 3 shows a schematic of the two-dimensional fully connected Ising model [1]. Total energy E in a Ising model is defined as

$$E = -\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i \qquad (1)$$

where σ is the direction of spin. If a spin is upward, σ equals +1, and if the spin is downward, σ equals -1. J is the interaction between spins, and h is an external magnetic field acting on spin. It has a state-transition mechanism that gives a minimum-energy state. This can be treated as a solution to the general optimization problem [1] if "interactions between all spins" can be considered. This is because the optimization problem represented by the Traveling-Salesman Problem

978-1-7281-3149-8/20/$31.00 ©2020 IEEE

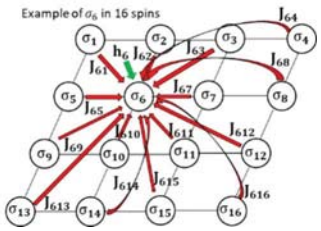

Figure 3 Ising Model considering interactions among all spins

(TSP) can be expressed as an evaluation function that is the difference between the optimal solution and the current state using the same mathematical formula. Therefore, reducing this difference corresponds to lowering the energy, which is the operation of the Ising machine, which is the same mathematical expression. Thus, the optimization problem can be solved by the Ising machine as shown in Fig. 4. At this time, the operation of the Ising model (Ising machine) changes depending on the total coupling of the spin and the interaction,

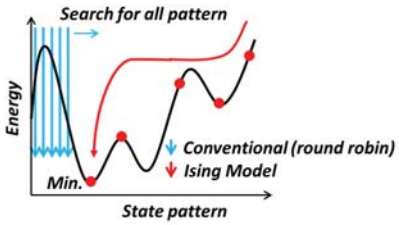

Figure 4 Energy transition in Ising model and conventional computing

so that the speed can be dramatically increased above that of conventional computing that searches all solution spaces. For example, if a PC with a 3.6-GHz processor core were used, it would take about 1.3 billion years to find the solution to the TSP with 25 cities. However, a property of the Ising model is its tendency to minimize *E*. Therefore, by utilizing this property, a 25-city TSP can be solved much more quickly (in less than one minute) than by using sequential computing.

B. Fully Spin-to-spin Connection

There are two development trends in the Ising model, which is a method of AI computing. One handles the coupling of adjacent spins, and the other handles the coupling between all spins.

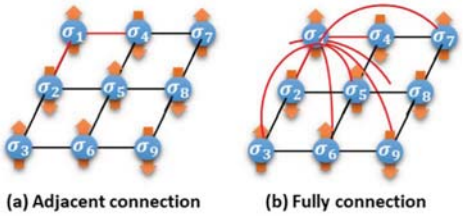

Figure 5 (a) Ising model in statistical mechanics, considering adjacent interactions. (b) Ising model to solve optimization problem, considering all interactions

The original Ising model is a statistical mechanics model, in which interactions exist only between adjacent spins as shown in Fig.5 (a). Due to its simple structure, the model is appropriate to implement on an LSI chip [2][3] and a chip with a quantum device [4].

This adjacent connected spin system can also solve the optimization problem. However, to do so, the form of full

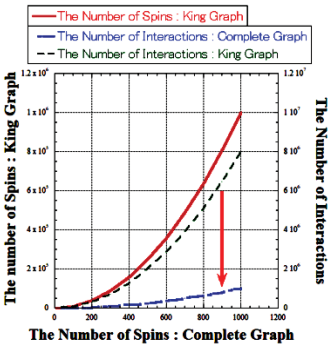

Figure 6 Comparison of spin and interaction numbers for full and adjacent pair of spins

connection that expresses the original optimization problem must be expanded to an "adjacent" form. This requires many spins and connections. As shown in Fig. 6, for example, if 512 fully connected spins are expanded to the adjacent connection, over 260,000 spins are required, and 1,024 fully connected spins result in one million adjacent ones. (The numbers of spins in the complete and King graphs correspond to full and adjacent connections, respectively (graph topology is not exactly the same as in Fig.5).) Moreover, even in an adjacent connection, the number of connections increases as well. In the King graph example, 8 times as many cells are required. In other words, the AI chip using the adjacent connection is huge, not suitable for processing on the things (edge) side, and not useful for a sustainable society.

Therefore, it is desirable to realize a fully connected spin system as shown in Figs. 3 and 5(b). The problem here is whether all the spins can be connected on the actual LSI chip. Although it can be implemented on software or field-programmable gate array (FPGA) on a processor is possible [5][6], this AI has been difficult to implement on an LSI chip with a limited number of wiring layers.

III. AI Chip Implementation

We aimed to implement the Ising model on the LSI by considering all interactions. One method for calculating this is a method in which σ_i is the upper layer spin and σ_j is the lower layer spin and arranged so as to intersect the same number as the other party and have an interaction J_{ij} corresponding to the connecting portion [1]. This configuration is similar to that in cellular automaton. All interactions are realized by arranging lower layer spins side by side, which is easy to implement with less wiring on LSI, so that each upper layer spin interacts with all spins. Each upper layer spin σ_i receives an interaction from a lower layer spin σ_j, adds all the interactions J_{ij} received by the same upper layer spin connected horizontally, determines the final state, and feeds back to the lower layer spin σ_j as one cycle in the update. This time, we have evolved this method into a simpler and more efficient form.

That is, from equation (1), equations (2), (3), and (4) can be used.

$$E = \sum_i E_i \tag{2}$$

$$E_i = -\sigma_i\left(\sum_j J_{ij}\sigma_j + h_i\right) = -\sigma_i \Delta E_l \tag{3}$$

$$\sigma_i^{new} = sign(\Delta E_l \pm T) \quad (4)$$

The elements necessary for these models are the upper layer spin σ_i, the lower layer spin σ_j, each interaction J_{ij}, and the calculation unit. Considering the energy E_i of a certain upper layer spin σ_i, the model can be updated by calculating ΔE and making the signs of σ_i and ΔE_l the same, so the requirements for the update are finally included in ΔE_l, the lower layer spin σ_j, each interaction J_{ij}, and the calculation unit only. Furthermore, the same spins in multiple calculation units and lower layer units can be integrated. The resultant circuit, as shown in Fig.7, comprises (1) the integrated

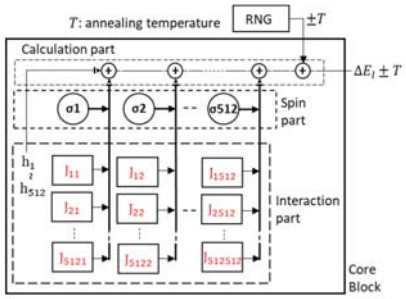

Figure 7 Core block configuration

calculation unit, (2) the lower layer spin that combines the same spins, and (3) each interaction. By adding an external field h_i , ΔE_l can be calculated. This proposed method considers all interactions and only the required number of spins.

By calculating ΔE using the Ising model, the method can be updated only in the direction of decreasing energy. Therefore, a simulated annealing (SA) technique with the temperature T is used [1][6]. We used a method to achieve SA without using a sigmoid function. In this method, T is used, and the spin is updated by the sign of $\Delta E \pm T$. The coefficient of T is random. When T is high, the probability P that the spin takes a state of -1 approaches random, and when T is low, the spin state is determined only by ΔE. Therefore, SA is achieved by updating while lowering the temperature. Unlike a method using a sigmoid function that needs to generate a random number with an accuracy that can be expressed in 100 gradations, the SA method only needs to generate a random number ±1 with an accuracy of only 1 gradation, simplifying the implementation. Fig. 8 shows our circuit incorporating the SA method, in which the Ising model with 512 spins was implemented. First, the spin σ_i to be updated in E_i is selected, ΔE is calculated for the selected σ_i, and the new state σ_i^{new} is determind and reflected.

The block diagram in Fig. 9 represents the overall configuration of our circuit. The interaction part stores the values of each interaction J_{ij}<3: 0> and external field h_i<7: 0>. The spin part stores 1 bit of each spin σ_i state. The

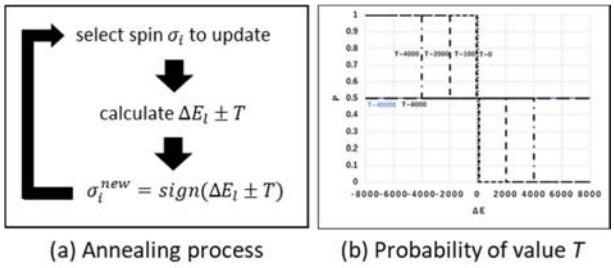

Figure 8 Annealing process

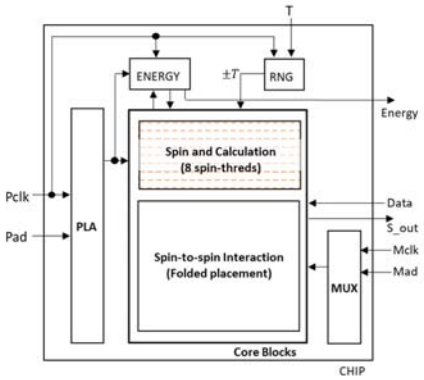

Figure 9 Chip block diagram

calculation part calculates $\Delta E \pm T$ for the selected σ_i. RNG generates a random number ± 1, combines the temperature T <12: 0> from the outside, and outputs ± T <13: 0> to the calculation part. PLA selects the spin σ to be updated in accordance with the value of Pad. The multiplexer (MUX) selects an interaction / external field that stores Data <3: 0>, which is an interaction / external field value, in accordance with the value of Mad. ENERGY calculates the Ising model energy E <22: 0>. The final output is the spin value Scout <8: 0> and the energy value Energy <7: 0>.

We implemented two additional concepts into our circuit. The first is the concept of spin-threads shown in Fig. 10. If we solve a large-scale optimization problem using an Ising model

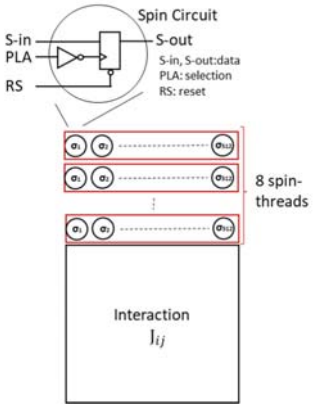

Figure 10 Spin circuit and 8-spin-thread structure

only once, it may converge to an approximate solution without obtaining an optimal one. To obtain an optimal solution, the same problem must be solved several times. We believed that the result of solving the same optimization problem multiple times could be obtained by adding a set of spins and a dedicated calculation unit that shared interactions and external fields and operating them in parallel. The spin circuit itself updates the calculation result as S-in by inputting the PLA. As shown in simulation results of the TSP in Fig. 11, when

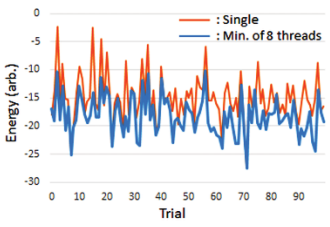

Figure 11 Energy transition in spin-thread

comparing the cases where the solution is obtained one and eight times, with the shortest route obtained as the solution, the eight-time solution case gives better results. In this study, eight sets of spin groups and dedicated calculation units were implemented.

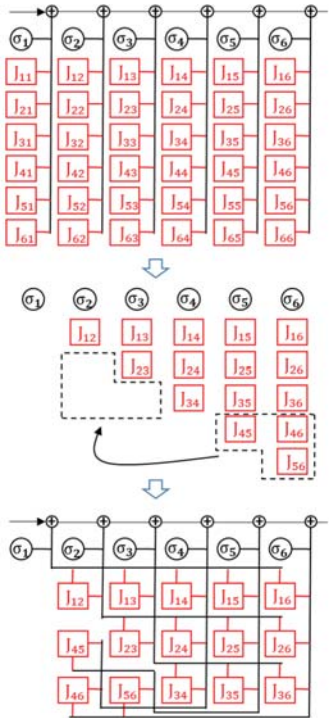

Figure 12 Folded halved-interaction placement

The second concept is a folded halved-interaction circuits placement as shown in Fig. 12. When we reviewed each interaction in the Ising model, we assumed that $J_{ij} = J_{ji}$ and that only J_{ij} was needed. Self-interaction J_{ii} was also omitted. Using only $J_{ij}(i < j)$ causes the whole interaction to become a triangle, so a part is cut out and moved to form a rectangle. In this circuit, the interaction arrangement is 256×511 as shown in the figure, and the area of the entire core was successfully reduced by about 38% after applying the method as shown in Fig. 13.

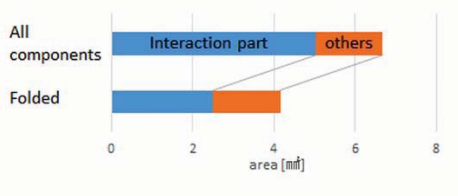

Figure 13 Area reduction effect

IV. Fabricated AI Chip and Evaluation

The AI LSI chip was fabricated with 28-nm CMOS technology. Fig. 14 shows the layout of the chip we developed and the actual chip. Table 1 shows the performance specifications. The number of spins is 512 per thread, and since there are 8 threads, 4096 spins are implemented. The number of interactions is 262,144, but half of them are installed by the proposed folded halved-interaction circuit method and the accuracy of the interaction is 4 bits. For new circuits described in the previous chapter, this chip has a circuit block configuration that realizes all spin-to-spin

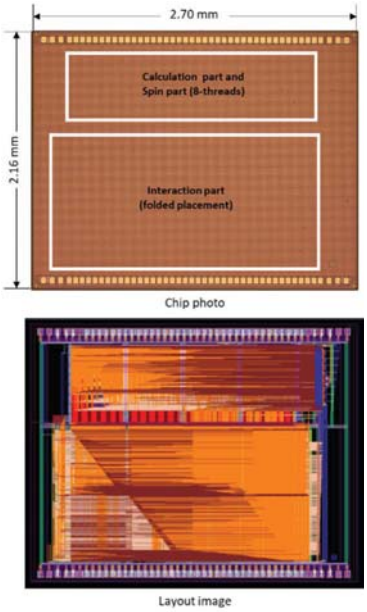

Figure 14 AI chip layout and result

Table 1 Al chip specifiations

Process		28-nm CMOS
Chip size		5.83 mm^2
Number of spins		512/thread
Spin Thread		8
Interaction		Fully connected between spins
		4-bit accuracy
		Folded placement
Supply Voltage	Circuits	0.9 V
	I/O	3.3 V
Current Dissipation	1MHz	19 mA

interactions or connections, and adopts an 8-spin-threads (core) structure and a folded halved-interaction circuit configuration. As far as we know, this chip is the world's first fully connected Ising model LSI, and is excellent in AI processing.

Fig. 15 shows an evaluation board and setup prepared for evaluating this AI LSI chip. The command and setting data from the PC are stored in the FPGA mounted on the evaluation board, and then the input signal is input from the FPGA to the fabricated chip. In terms of the circuit configuration, the operation of the chip can be divided into three phases: interaction input, calculation, and calculation result output.

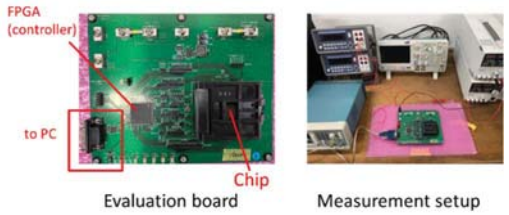

Figure 15 Experiment setup

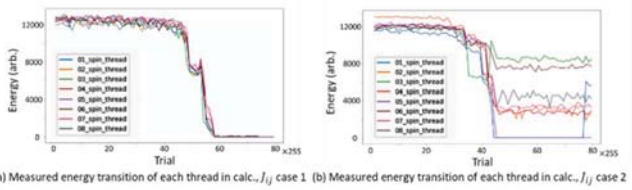

Figure 16 Experimental results

Fig. 16 shows the experimental results. After input of the interaction and external field data, a calculation was performed while changing the temperature, and finally output was obtained. The spin state Scout × 512 of the spin of 8 spin-threads, the spin of the lowest energy spin-thread, and the energy value ENERGY × 23 of each of the 8 spin-threads are obtained as outputs. Two J_{ij} examples shown here were confirmed with the fully spin-to-spin connected AI LSI chip to be in the desired spin state when the energy was lowered.

V. Application of AI chip

Using the LSI description of this chip, the simulation results of an eight-city TSP are shown in Fig. 17. In the TSP, the aim is to find the shortest route that goes through each city exactly once and returns to the first city. When the temperature is high, the spin state is determined by a random number. As the temperature gradually decreases, the spin state gradually stabilizes, and when the temperature reaches 0, a solution can be obtained. The chip is also compatible with Support Vector Machine (SVM), which is a basic function of classification at the things side or edge. SVM finds the recognition boundary that is the maximum distance from the nearest training data.

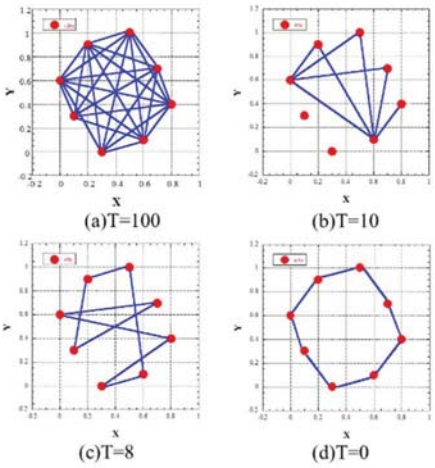

Figure 17 Example results of solving TSP

If the high-level description of this AI LSI chip is provided as an IP, it can be incorporated into a company-side system and a verification experiment can be performed as shown in Fig. 18. Since the IP has been confirmed to work with 28-nm CMOS technology, it will be easy for companies to adopt. Assumed applications are autonomous driving, high-efficiency agricultural and industrial production, infrastructure monitoring, and route search in the event of a disaster. For example, in autonomous vehicles, both physical information and cyber information move around the car body, but power is extremely limited. The proposed LSI promises to overcome this power limitation, making autonomous vehicles easier to produce and use.

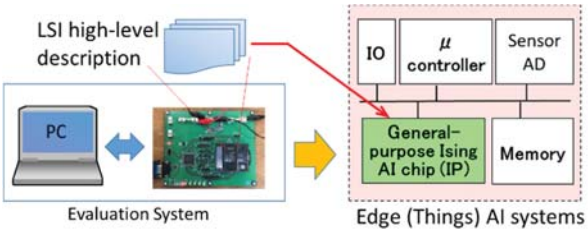

Figure 18 Deployment of chip results to AI system

VI. Conclusion

Next-generation Internet of Things (IoT) requires ultra-low-power information processing that is useful for both sensor area expansion and high-speed feature extraction using a new signal processing artificial intelligence large-scale integration chip (AI chip) developed for not only the cloud side but also the "things" side (edge).

For this purpose, the world's first fully spin-to-spin connected Ising model (annealing) AI chip was successfully demonstrated. Its specifications are as follows: 512 spins, 262,144 interactions (half are arranged with the proposal method), and 4-bit interaction accuracy. The chip was designed and fabricated using a 28-nm CMOS process.

The proposed methods in this chip are as follows. In the fully-connected Ising model, starting from the spin of a two-layer structure, the resultant circuit comprises (1) the integrated calculation unit, (2) the lower layer spin that combines the same spins, and (3) each interaction. In this proposed method, the necessary number of spins is the same as the number of original spins while considering all interactions. In the multi-spin-thread method, by utilizing the fact that the elements are decomposed in this structure, eight pairs of spins are calculated at one time by one interaction. That is, the throughput of the Ising machine, which is an approximate calculation of the solution candidates, was increased by eight times. In the folded halved-interaction circuit method by paying attention to the fact that the interaction between all spins is a symmetric matrix, the number of interaction is halved in the implementation.

Acknowledgments

Part of this paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), METI, Japan.

References

[1] K. Someya, R. Ono, and T. Kawahara, "Novel Ising Model Using Dimension-Control for High-Speed Solver for Ising Machines," Paper ID 4137, Session B2P-F, NEWCAS 2016, doi: 10.1109/NEWCAS.2016.7604797, (2016).

[2] Masanao Yamaoka, et al. "20k-spin Ising chip for combinational optimization problem with CMOS annealing," IEEE ISSCC Dig. Tech. Papers, pp.432–433, doi: 10.1109/ISSCC.2015.7063111, (2015).

[3] Takashi Takemoto, et al., "A 2×30k-Spin Multichip Scalable Annealing Processor Based on a Processing-In Memory Approach for Solving Large-Scale Combinatorial Optimization Problems," ISSCC, 2019, doi: 10.1109/ISSCC.2019.8662517, (2019).

[4] P. I. Bunyk et al., "Architectural Considerations in the Design of a Superconducting Quantum Annealing Processor," in IEEE Transactions on Applied Superconductivity, vol. 24, no. 4, pp. 1-10, Aug. 2014, Art no. 1700110, doi: 10.1109/TASC.2014.2318294.

[5] S. Tsukamoto, M. Takatsu, S. Matsubara, and H. Tamura, "An accelerator architecture for combinatorial optimization problems," Fujitsu Sci. Tech. J, 53(5), 8-13 2017.

[6] Akira Minamisawa, Ryoma Iimura, Takayuki Kawahara, "High-speed Sparse Ising Model on FPGA," MWSCAS, 2019, doi: 10.1109/MWSCAS.2019.8885105, (2019