2022 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2022, Virtual, December 12-15, 2022

Prospects of Energy-Efficient Edge-AI Accelerator Architecture Using Nonvolatile Logic

M. Natsui, D. Suzuki, Y. Takako, A. Tamakoshi, and T. Hanyu

Research Institute of Electrical Communication, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai, 980-8577 Japan
Email: masanori.natsui.a8@tohoku.ac.jp

Abstract– The realization of energy-efficient edge-AI hardware is an important issue for utilizing it as a fundamental technology for the next-generation IoT society. Nonvolatile logic-circuit technology based on MTJ devices is a key to solving this challenge. In this paper, we discuss design guidelines and prospects for edge-AI hardware that enables energy-saving operation by maximizing the use of nonvolatile memory functions.

1. Introduction

With the advancement of Internet-of-Things (IoT) sensor node applications, the number of devices connected to the Internet is increasing. Applications of IoT devices range from improving manufacturing and harvesting efficiency to smart homes, and are expected to be used in a variety of fields. The big data collected by IoT devices and the knowledge obtained from their analysis will open the door to a new era of intelligent computing paradigm.

In the present server-oriented IoT system, the cloud server performs overall control and data processing sent from the sensor nodes. In this system, however, a vast amount of data traffic will become a serious issue in the future. As a solution to this issue, the importance of so-called edge-AI hardware is attracting attention. By replacing some of the processing performed by the server with AI processing by the sensor node itself, and exchanging only the feature values of the sensing data with the server, data traffic can be reduced (Fig. 1).

For the implementation of new IoT systems based on edge AI, it is necessary to implement energy-efficient hardware that can perform AI processing with the extremely low energy consumption allowed by edge devices as its platform technology. For this purpose, we have been working on a new circuit technology that utilizes nonvolatile memory technology, i.e., nonvolatile logic [1-3]. In this paper, we provide a brief overview of recent developments in edge-AI hardware utilizing nonvolatile memory based on recent reports, and discuss issues to be considered for the realization of energy-efficient edge-AI hardware, including the recent works of our research group.

2. Basic Study on Energy-Efficient Edge-AI Hardware

The most common performance indicator for AI hardware is the number of executable operations per watt (OPS/W). If the supply voltage to the hardware equals to 1 V, OPS/W is equal to the reciprocal of the energy required for executing one multiply-and-accumulate (MAC) operation. This energy includes dynamic energy required for the operation and the energy required to transfer the corresponding data (inputs and weights), as well as static energy consumed at all times and energy required to control the computation unit per operation. In other words, OPS/W represents the energy efficiency of the edge-AI hardware.

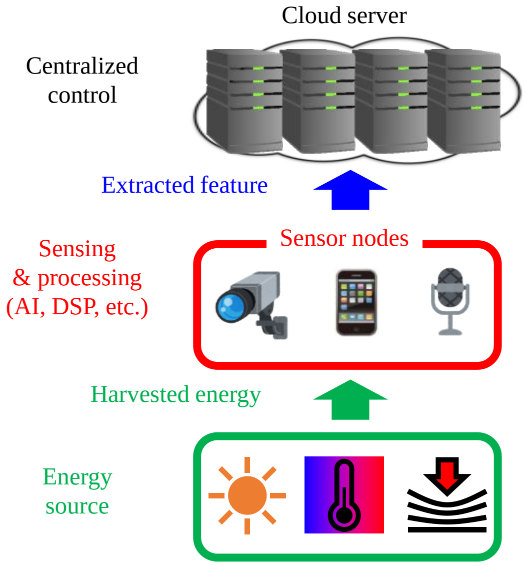


Fig. 1. Edge-AI-based IoT system.

To increase OPS/W, it is necessary to reduce the energy consumption per operation. For example, If the target performance is 10 TOPS/W, the energy per operation must be less than 1 / (10×10^{12}) = 100 fJ. Considering that the energy consumption for 32-bit fixed-point addition in a 45 nm CMOS process is approximately 100 fJ [4], this is a very challenging task.

Based on the energy breakdown shown above, the following approaches can be taken to reduce energy:

1. Develop energy-efficient MAC operation circuit,
2. Reduce the amount of data transfer as much as possible,
3. Reduce static energy,
4. Simplify control as much as possible.

Quantization [5], which is widely used in recent edge-AI hardware, is positioned as an approach for 1 and 2, while

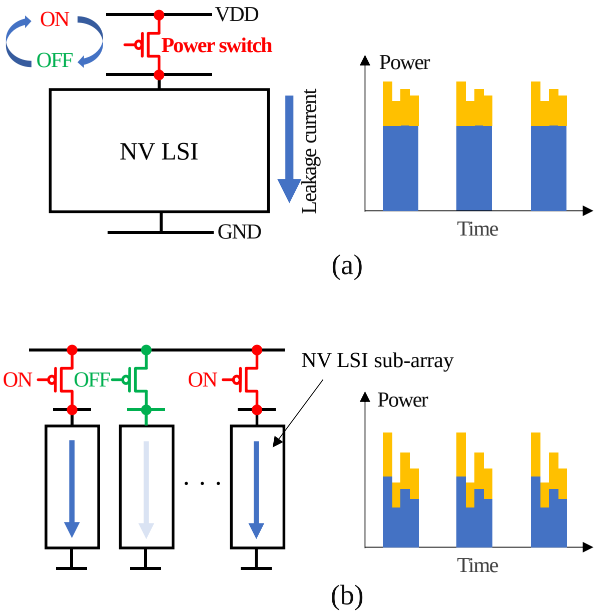


Fig. 2. Differences in power consumption trends for various nonvolatile LSI implementations: (a) with chip-level power gating, (b) with sub-array-level power gating.

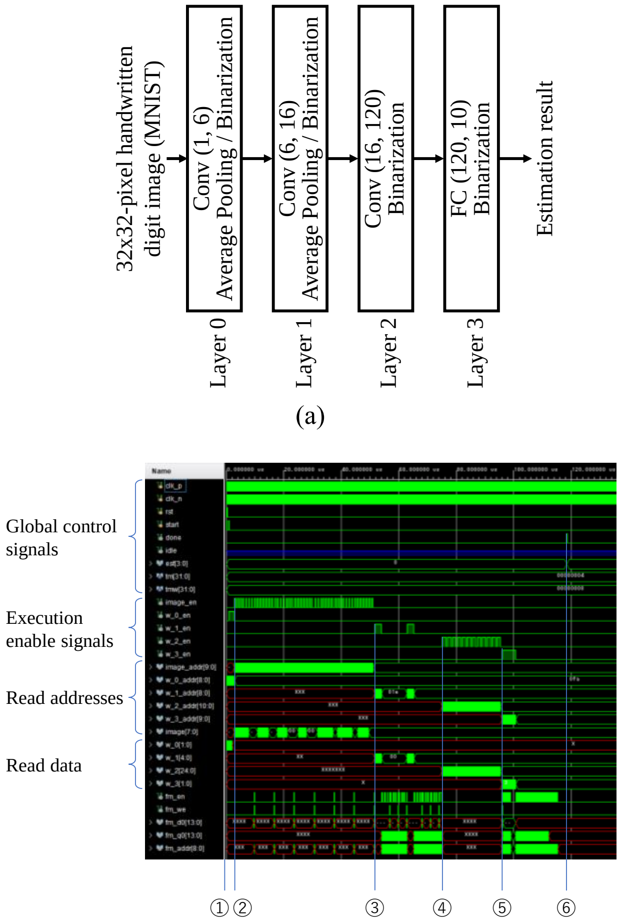
computing-in-memory (CIM) architecture [6, 7] is mainly aimed at 1, 2, and 4. For approach 3, it is expected to be reduced by downsizing the circuit using quantization and other methods. In contrast, nonvolatile logic-circuit technology, in which nonvolatile memory elements are placed near logic circuits, and the active application of power gating enabled by this technology are direct approaches to reduce static energy. Therefore, the use of this technology is expected to achieve further energy savings that cannot be achieved with the technologies currently used in AI hardware alone.

3. Recent Trend and Future Prospects of AI Hardware Utilizing Nonvolatile Devices

The objectives of utilizing nonvolatile memory in AI hardware can be roughly categorized as follows:

- Implementing CIM structures,
- Multi / analog value retention,
- 3D implementation,
- Nonvolatile memory for weight coefficients.

Most of the AI accelerator chips presented at ISSCC (~2021) that utilize nonvolatile memory are based on ReRAM. Considering the nature of the conference, ease of implementation may be the main reason for this. Several cases of using nonvolatile memory as an analog storage device will be presented around 2020. On the other hand, at technology-related conferences such as Symposium on VLSI Technology and IEDM, there are many presentations other than ReRAM. However, most of them are device-level studies, and there are still not many examples of chip-level implementations.


32x32-pixel handwritten digit image (MNIST)
Conv (1, 6) Average Pooling / Binarization
Conv (6, 16) Average Pooling / Binarization
Conv (16, 120) Binarization
FC (120, 10) Binarization
Estimation result
Layer 0
Layer 1
Layer 2
Layer 3
(a)
Global control signals
Execution enable signals
Read addresses
Read data
①②
③
④
⑤
⑥

① Store layer-0 weight to WBUF
② Store input data to IBUF → Execute MAC ops.
③ Store layer-0 weight to WBUF → Execute MAC ops.
④ Store layer-0 weight to WBUF → Execute MAC ops.
⑤ Store layer-0 weight to WBUF → Execute MAC ops.
⑥ Output estimation result and finish inference op.

(b)

Fig. 3. Design example of a BNN hardware: (a) network structure, (b) operation waveforms.

We are conducting research toward the realization of energy-efficient edge-AI hardware using MTJ devices, which are attracting attention as next-generation nonvolatile memory devices [8, 9]. The advantage of MTJ devices is that they can be implemented by distributing nonvolatile memory functions close to the logic circuitry. This structure makes it possible to apply power gating with low overhead without the need to save data to external memory, even in applications that require intermediate data retention. Therefore, power gating can be applied in a fine-tuned manner on a per-circuit-module basis [10], and the amount of wasted static energy can be reduced as much as possible, depending on the operating conditions of the circuit as shown in Fig. 2.

On the other hand, the drawback of MTJ devices is their high write energy. Therefore, it is desirable to choose a circuit configuration in which data stored in nonvolatile memory is rewritten as little as possible. An example of such a circuit architecture is one in which all layers are placed on a chip and all data required for one process (i.e., image data and weight coefficients for an inference process of convolutional neural network) are held in an input buffer with nonvolatile storage or nonvolatile registers near logic blocks. Although scalability issues remain, such a structure would be suitable for energy saving because data is exchanged with external memory only once, and buffers and parts of the logic circuit can be finely power-gated in the middle of processing.

As a design example, Figure 3 shows the configuration of the binarized neural network (BNN) hardware presented in [11] and its operating waveforms. Since the main focus of [11] is to improve the performance of the elemental circuits (nonvolatile look-up tables) of the BNN hardware, the overall architecture itself was constructed in a straight-forward manner. From the operation waveforms, it can be confirmed that there is room for further performance improvement by changing to a configuration that fully utilizes the nonvolatile memory function. Specifically, the following specifications can serve as a design guideline for energy-efficient edge-AI hardware which maximizes the advantages of the use of nonvolatile memory functions.

- Pipelined processing is applied to the computation blocks corresponding to each layer of the network.
- Weight data is not stored in buffers, but in nonvolatile registers located near the MAC operation units.
- Input data is held in a buffer with nonvolatile storage.
- The hardware configuration of each layer is designed to enable energy-saving technologies such as quantization, and to reuse data once read as much as possible.
- The degree of parallelism of operations at each layer is adjusted to equalize the processing time per stage as much as possible.
- Nonvolatile memory function is also implemented in the pipeline registers to hold the output of MAC operations.
- When an operation on a layer that is not on the critical path is completed, power gating is applied until the previous operation is completed.
- The input buffer is divided into multiple banks, and power gating is applied to the banks that hold data not subject to processing.

Further studies will be conducted to realize edge-AI hardware, which will be a fundamental technology for the next-generation IoT society.

4. Conclusion and Prospects

This paper discussed design guidelines for edge-AI hardware that maximizes nonvolatile memory capability and achieves energy-efficient operation, including recent trends of edge-AI hardware and how to utilize MTJ devices in consideration of their advantages and disadvantages. Nonvolatile logic-circuit technology enables flexible power gating, which is expected to expand the available design space and enable further energy savings and performance improvements that are not possible with conventional technologies.

In the future, demand for Edge-AI hardware that is not only energy efficient but also robust in the unstable operating environment expected in IoT applications is expected to increase. In addition to its effectiveness as a memory that holds inputs and weights, the programmability provided by nonvolatile logic circuit technology is also effective in making circuits highly reliable, and we have demonstrated the potential of this technology [3]. By comprehensively utilizing these technologies, high performance, energy-efficient, and highly-reliable edge-AI hardware based on nonvolatile logic-circuit technology is expected to be established as a platform technology for the next-generation IoT society.

Acknowledgments

This research was supported by JST CREST (JPMJCR19K3), JST OPERA, and JSPS Grant-in-Aid for Scientific Research (17KK0001, 21H03405, and 21H04868).

References

[1] M. Natsui, et al., "A 47.14μW 200MHz MOS/MTJ-Hybrid Nonvolatile Microcontroller Unit Embedding STT-MRAM and FPGA for IoT Applications," IEEE Journal of Solid-State Circuits, Vol.54, No.11, pp.2991-3004, 2019.

[2] T. Hanyu, et al., "Standby-Power-Free Integrated Circuits Using MTJ-Based VLSI Computing," Proceedings of the IEEE, Vol.104, No.10, pp.1844-1863, 2016.

[3] M. Natsui, T. Chiba and T. Hanyu, "Impact of MTJ-Based Nonvolatile Circuit Techniques for Energy-Efficient Binary Neural Network Hardware," Japanese Journal of Applied Physics, Vol.59, No.5, pp.050602-1-050602-7, 2020.

[4] M. Horowitz, “Computing's energy problem (and what we can do about it),” IEEE International Solid-State Circuits Conference Digest of Technical Papers, pp. 10-14, 2014.

[5] I. Hubara, et al., “Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations,” arXiv preprint, 1609.07061, 29 pages, 2016.

[6] W-H Chen, et al., "A 16Mb dual-mode ReRAM macro with sub-14ns computing-in-memory and memory functions enabled by self-write termination scheme," Proceedings of the 2017 IEEE International Conference on Electron Devices Meeting, pp. 657-660, 2017.

[7] S. Jain, et al., "Computing in Memory with Spin-Transfer Torque Magnetic RAM," IEEE Transactions on Very Large-Scale Integration Systems, no. 99, pp. 1-14, 2017.

[8] S. Ikeda et al., "A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction," Nature Mater., vol. 9, no. 9, pp. 721-724, Jul. 2010.

[9] H. Sato, et al., "MgO/CoFeB/Ta/CoFeB/MgO recording structure in magnetic tunnel junctions with perpendicular easy axis," IEEE Trans. Magn., vol. 49, no. 7, pp. 4437-4440, Jul. 2013.

[10] F. Zhong, M. Natsui, and T. Hanyu, "Dynamic activation of power-gating-switch configuration for highly reliable nonvolatile large-scale integrated circuits," Japanese Journal of Applied Physics, Vol.61, No.SC, pp.SC1035-1-SC1035-10, 2022.

[11] D. Suzuki, T. Oka, and T. Hanyu, "Design of an Active-Load-Localized Single-Ended Nonvolatile Lookup-Table Circuit for Energy-Efficient Binary-Convolutional-Neural-Network Accelerator" Japanese Journal of Applied Physics, vol.61, no.SC, pp.1083-1~1083-10, 2022.