Extended Abstracts of the 2019 International Conference on Solid State Devices and Materials, Nagoya, 2019, pp409-410

Opportunities and Challenges of Atom Switch for Next AI Hardware

Munehiro Tada, Toshitsugu Sakamoto and Makoto Miyamura

NEC Corporation
16-1 Onogawa, Tsukuba, Ibaraki 305-8569, Japan,
Phone: +81-29-893-5481 E-mail: m-tada@bl.jp.nec.com

Abstract

In this paper, an atom switch (AS) technology is reviewed in view point of realizing a next AI hardware. The nonvolatile AS enables us to route signals flexibly with lower power, providing an energy-efficient inference accelerator. The nonvolatile AS-memory provides on-chip memory with high bandwidth. The technology gives the new active function, riding next wave of mobile/edge AI computing in system on chip.

Introduction

In current artificial intelligence (AI) systems in cloud sever, GPUs serve as good accelerators of deep learning training [1]. Looking to the future, the evolution of AI expands from could to edge, and training to inference, in which the neural network (NN) inference needs high speed, energy efficiency and flexibility to support various applications in the edge. Therefore, beyond von Neumann (e.g. new computing-in-memory (CIM) architecture) is strongly desired. Recently, magneto-resistive RAM (MRAM) [2] and resistive RAM (ReRAM) [3] are of great interest and becoming research topics to realize the CIM.

On the other hand, a field programmable gate array (FPGA) is becoming the next possible solution for NN inference accelerator, having better speed and energy efficiency than GPUs [4], in which FPGA can implement high parallelism and enables developers to implement only the necessary logic in hardware according to the target algorithm. However, the FPGA's logic overhead for reconfigurability and the high static power consumption are challenges for the mobile/edge inference. Therefore, a nonvolatile FPGA with smaller foot-pint and instant-on is becoming a strong candidate for the application.

In this paper, we review an atom switch technology as non-volatile reprogrammable switch and embedded memory for realizing the AI inference accelerator.

Atom Switch

The atom switch (a.k.a. NanoBridge™) is one of the electrochemical resistive-change devices with the cation types [5-11] are called solid-electrolyte switch, programmable metallization cell (PMC), electrochemical metallization (ECM), or conductive-bridging (CB) cell, which controls metal cations extracted from an active electrode to form a metal bridge (metallic pathway, or filament) in the ionically conductive insulator. The switching mechanism of the atom switch is explained by electrolysis, and it is composed of a solid-electrolyte sandwiched between Cu and Ru electrodes (Fig.1). The very thin solid-electrolyte makes a high electric field at the interface between Cu and the electrolyte, initiating the ionization of Cu without mobile ions in the electrolyte. Once the Cu precipitation makes a conducting bridge between both electrodes, the resistive state turns into on-state. The polymer-solid electrolyte (PSE) enables the forming free programming due to the high free volume of the polymers [12]. The on-resistance of the atom switch can be tuned by a programming current [13].

Complementary Atom Switch (CAS)

An off-state lifetime under stress voltage is one of the reliability concerns of the atom switch because the longtime voltage stress causes an unintentional precipitation of Cu into the electrolyte. A Complementary Atom Switch (CAS) is proposed to improve the off-state lifetime by sharing the stress voltage with two atom switches [14]. The 1T1CAS structure can replace the conventional CMOS switch composed of SRAM and transmission gate (TMG) (Fig.2), giving non-volatility, smaller foot-print and smaller input capacitance. Atom Switch FPGA (AS-FPGA) with 51Mb atom switches is fabricated in a 40nm-node, 1P9M CMOS platform (Fig.3).

Atom Switch FPGA

The each logic element (LE) in the AS-FPGA composes of four pairs of 4-input LUTs and DFFs [15]. Differently from a SRAM-based FPGA, the LE includes AS memory for LUTs and AS multiplexer (AS-MUX) for connecting LE to and routing the inter-wires. The AS-MUX is characterized by a single-stage routing and small parasitic capacitance, resulting in 3.8x higher operation speed, and 3x higher power efficiency than those of the conventional SRAM-based FPGA (Fig.4).

Atom Switch SoC for Next AI Hardware

We propose an energy-efficient, atom switch system on chip (ASSoC) to realize a mobile/edge inference in AI hardware (Fig.5). The AS-FPGA-IP can work for the inference accelerator, giving us high performance computing with low power off/on overhead. The AS-memory gives us low power memory access with high bandwidth. Also, programmable I/O and bus will be designed by using atom switch, giving high flexibility to support various applications. Totally, the inference energy is expected to be 1/10 in the ASSoC.

Conclusions

The recent progress of the atom switch technology is discussed. The embedded nonvolatile switch/memory combination contributes to realize the energy-efficient AI hardware, in which software and hardware optimization will be is a key.

Acknowledgement A part of this work was supported by NEDO. A part of the device processing was operated by AIST, Japan. The authors thank to Drs. N. Banno, R. Nebashi, K. Okamoto, N. Iguchi, X. Bai, H. Numata, T. Sugibayashi and H. Hada for their support.

References

[1] J. Nickolls and W. J. Dally, "The GPU Computing Era", IEEE Micro, vol. 30, no. 2, 2010.

[2] N. Sakimura, et al., "A 90nm 20MHz Fully Nonvolatile Microcontroller for Standby-Power-Critical Application", IEEE ISSCC, pp.184-185, Feb. 2014.

[3] 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", IEDM, Dec. 2017.

[4] K. Guo, et al., "A Survey of FPGA-based Neural Network Inference Accelerators", ACM Trans. Reconfigurable Technol. Syst. 12, 1, Article 2 (March 2019), s26 pages.

[5] Y. Hirose and H. Hirose, "Polarity - dependent memory switching and behavior of Ag dendrite in Ag - photodoped amorphous As_2S_3 films", Journal of Applied Physics. vol. 47. no.6, pp.2767-2772, 1976.

[6] W. C. West, et al., "Equivalent Circuit Modeling of the Ag|$As_{0.24}S_{0.36}Ag_{0.40}$|Ag System Prepared by Photodissolution of Ag", J. Electrochem. Soc., vol. 145, no. 9, pp.2971-2974, 1998.

[7] K. Terabe, et al., "Formation and disappearance of a nanoscale silver cluster realized by solid electrochemical reaction," Journal of Applied physics, vol. 91, 10110, 2002.

[8] T. Sakamoto, et al., "Nanometer-scale switches using copper sulfide," Appl. Phys. Lett., 82, 3032, 2003.

[9] T. Sakamoto, et al., "Electronic transport in Ta_2O_5 resistive switch", Appl. Phys. Lett. 91 (2007) 092110.

[10] R. Waser and M. Aono, "Nanoionics-based resistive switching memories", Nature Material, vol. 6, pp. 833-839, 2007.

[11] E. Vianello, et al., "Sb-doped GeS_2 as performance and reliability booster in Conducting bridge RAM," IEEE International Electron Devices Meeting, pp. 741-744, 2012.

[12] M. Tada, et al., "Polymer Solid-Electrolyte (PSE) Switch Embedded on CMOS for Nonvolatile Crossbar Switch," IEEE Transactions on Electron Devices, vol. 58, no. 12, pp. 4398-4405, 2011.

[13] M. Tada, et al., "Set/Reset Switching Model of Cu Atom Switch based on electrolysis", IEEE Transactions on Electron Devices, vol. 64, no. 4, pp.1812-1817, 2017.

[14] M. Tada, et al., "Improved Off-state Reliability of Nonvolatile Resistive Switch with Low Programming Voltage," IEEE Transactions on Electron Devices, vol. 59, no. 9, pp. 2357-2362, 2012.

[15] M. Miyamura, et al., "0.5-V highly power-efficient programmable logic using nonvolatile configuration switch in BEOL", Proceedings of the 2015 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, Pages 236-239, 2015.

[16] X. Bai, et al., "A Low-Power Cu Atom Switch Programmable Logic Fabricated in a 40nm-node CMOS Technology", VLSI technology symposium, T28-29, 2017.

[17] R. Nebashi, et al., "High-Density and Fault-Tolerant Cu Atom Switch Technology Toward 28nm-node Nonvolatile Programmable Logic" VLSI technology symposium, pp.127-128, 2018.

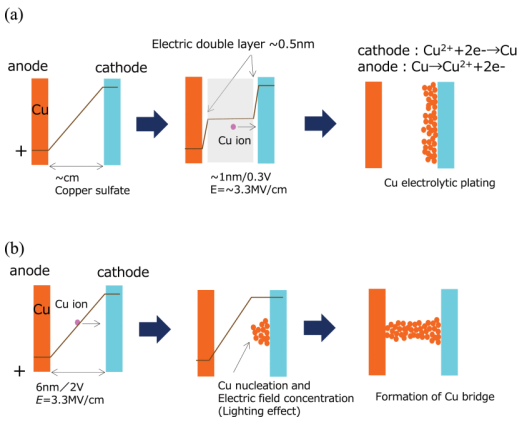
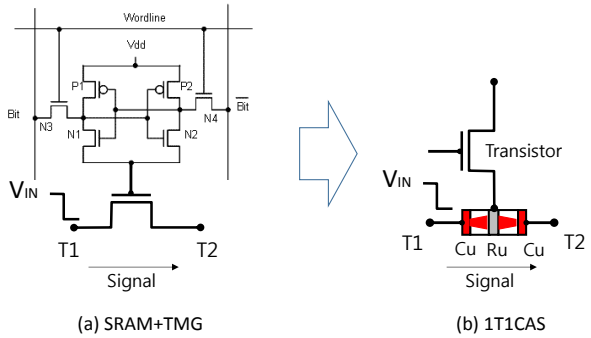


Fig.1 Schematic images of electrochemical reactions of (a) Cu electrolytic refining (liquid-electrolyte) and (b) Cu atom switch (solid-electrolyte). Copyright 2017 IEEE. Reprinted, with permission, from [13].



(a) SRAM+TMG (b) 1T1CAS

Fig.2 Schematic images of switch components in FPGAs, (a) SRAM + TMG, and (b) 1T1CAS [14].

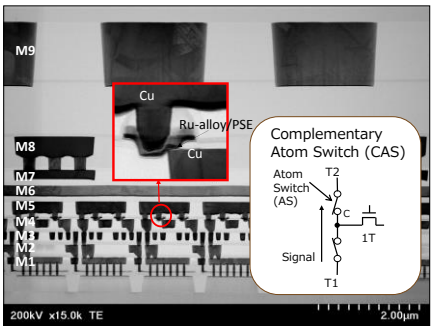
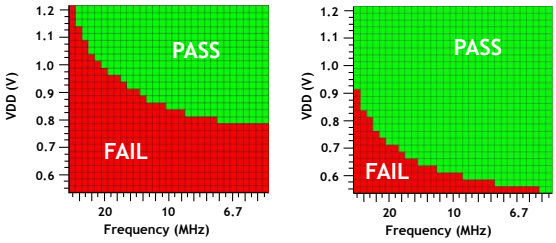
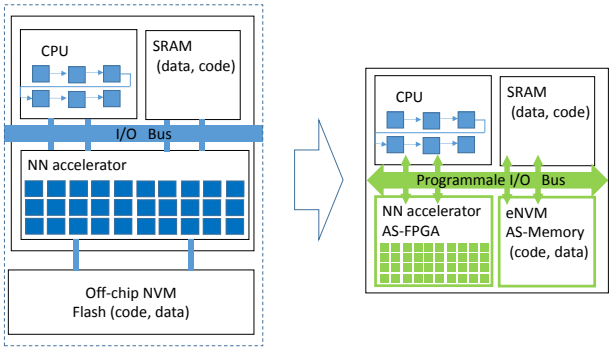


Fig.3 Cross sectional TEM image of a 40nm-ndoe AS-FPGA. Copyright 2018 IEEE. Reprinted, with permission, from [17].



(a) 40nm SRAM-based FPGA (b) 40nm Atom-Switch FPGA

Fig.4 Shmoo pots of application (16b-ALU) on (a) SRAM-based FPGA and (b) AS-FPGA. Copyright 2017 IEEE. Reprinted, with permission, from [16].



(a) Conventional SoC (b) Atom switch SoC (ASSoC)

Fig.5 Schematic images of AI edge devices, (a) conventional SoC, (b) ASSoC.