ON-state retention of Atom Switch eNVM for IoT/AI Inference Solution

Koichiro Okamoto, Ryusuke Nebashi, Naoki Banno, Xu Bai, Hideaki Numata, Noriyuki Iguchi, Makoto Miyamura, Hiromitsu Hada, Kazunori Funahashi, Tadahiko Sugibayashi, Toshitsugu Sakamoto, and Munehiro Tada
System Platform Research Laboratories
NEC Corporation
Tsukuba, Japan
Phone: +81-29-893-5481, e-mail address: k-okamoto@nec.com

***Abstract*— An ON-state retention of a 40nm-node atom switch embedded nonvolatile memory (eNVM) has been carefully investigated for IoT/AI inference solution. Based on ON-conductance (G_{on}) tuning model of atom switch, one order of magnitude lower programming power is achieved while keeping the same G_{on}. Smaller ON-state retention dependences on temperature (E_a = 0.2eV) and time (n = 0.11) are experimentally clarified and the lifetime is predicted to be more than 10 years at 150°C under +20% shift criteria of G_{on}.**

***Index Terms*—Atom switch, CBRAM, nonvolatile memory, ON-state retention**

I. INTRODUCTION

In current artificial intelligence (AI) systems in cloud server, GPUs serve as good accelerators of deep learning training [1]. Looking into the future, evolution of the AI system expands from cloud to edge, and training to inference, which needs high energy efficiency, flexibility and environmental suitability to support various applications in the IoT/edge. Especially for in-situ AI training/inference, nonvolatility is essential. An atom switch is a kind type of electrochemical resistive-change device using migration of cations (e.g. CBRAM) [2]-[6]. Fig. 1(a) shows a schematic image of the atom switch stack. The atom switch is composed of a polymer solid-electrolyte (PSE) [7]-[9] with an adjacent buffer layer sandwiched between Cu and Ru-alloy [10], [11] electrodes. Atom switches have one and only advantage of their wider tuning range of the on/off conductance ratio which can be controlled by programming condition. Fig. 1(b) shows a new concept of atom switch system-on-chip (SoC) for AI hardware used in mobile/edge, which attains the multiple-use of atom switches as not only routing switches and lookup table (LUT) memory bits in nonvolatile field programmable gate array (FPGA) [12], but also memory bits in embedded nonvolatile memory (eNVM) [13]. For the nonvolatile memory application, the retention should be carefully investigated.

In this paper, we discuss (1) ON-state conductance (G_{on}) tuning by changing the programming current and pulse width and (2) ON-state retention after endurance at high temperature of 150°C of the atom switches fabricated in a 40nm-node 1P9M CMOS platform.

A part of this work was supported by NEDO.

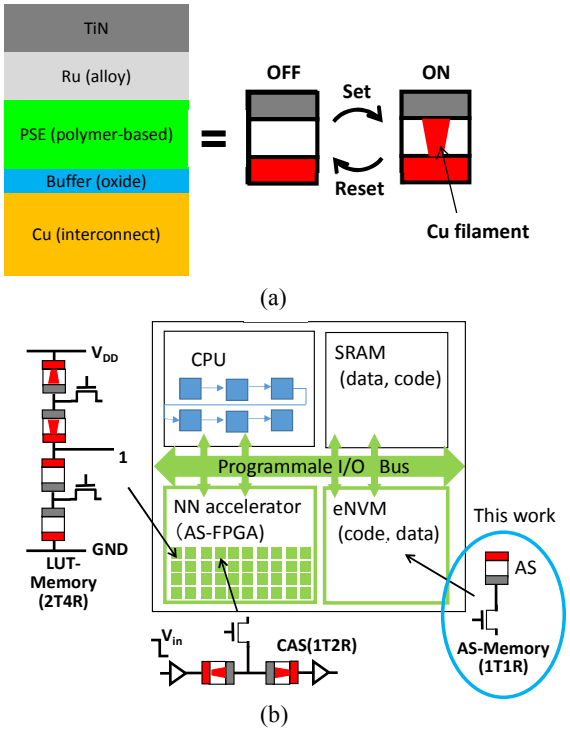

Figure 1. (a) Simplified schematics of the atom switch stack. (b) Concept of atom switch SoC (ASSoC) for realizing the AI hardware used in the edge/mobile. Atom switches can be simultaneously used as rooting switches and lookup table (LUT) memory bits in eFPGA, and memory bits in eNVM.

II. ATOM SWITCH CHARACTERISTICS

A. *Fabrication of Atom swtich*

Fig. 2 shows a cross-sectional TEM image of the fabricated atom switches. Atom switches were fabricated on the 4th Cu interconnects (metal 4) using only two additional photomasks. Buffer metal oxide, PSE, and Ru-alloy top electrode layers were deposited on the Cu interconnects through contact holes. The Cu interconnects are used as bottom electrodes of the atom switches. The buffer oxide layer effectively prevents oxidation of the underlying Cu interconnects, and serves as a part of the solid electrolyte at the same time [14]-[16]. The PSE enables the forming-less programming and shows high breakdown voltage [7], [8]. The deposited stack of the atom switch was

978-1-7281-3199-3/20/$31.00 ©2020 IEEE

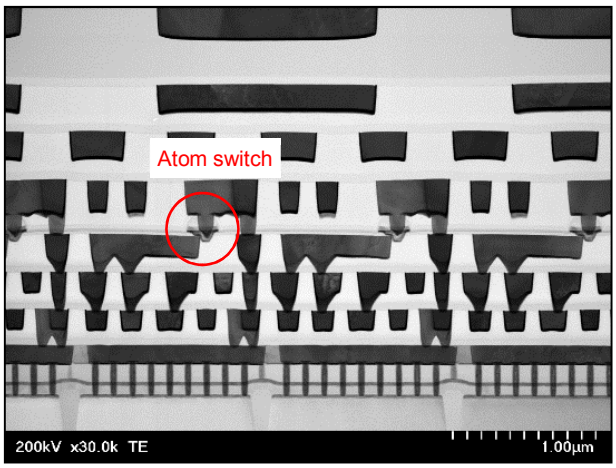

Figure 2. Cross-sectional TEM image of the atom switch integrated between metal 4 and 5 interconnects on the 40nm-node CMOS platform.

subsequently etched by using a dry process after lithography patterning, and was then covered with a SiN layer to suppress Cu diffusion into the nearby insulator layers. The atom switches were successfully integrated between metal 4 and 5 interconnects on the 40nm-node CMOS.

B. Tuning of ON-conductance

Fig. 3 shows dependence of set programming energy on the obtained G_{on}. For the routing switch application in FPGAs, a high on/off conductance ratio is necessary to transfer/isolate logic signals. However, such higher conductance ratio requires higher set programming power. For the eNVM application, it is essential to reduce the programming power compared to switch application for FPGAs. In order to clarify the effect of set programming current (I) and pulse width (t) on G_{on}, we introduce a modified Faraday's law [17] considering set model of an atom switch shown in Fig. 4;

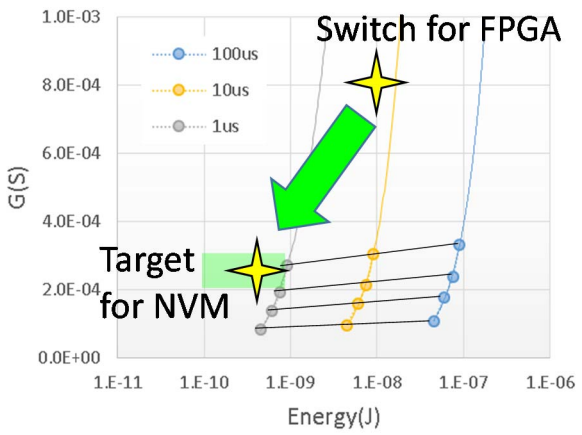

Figure 3. Dependence of set programming energy on the obtained G_{on}. Target G_{on} range of atom switches for eNVM application is indicated by a rectangle. It is essential to reduce the programming power for the eNVM compared to switch application in FPGAs.

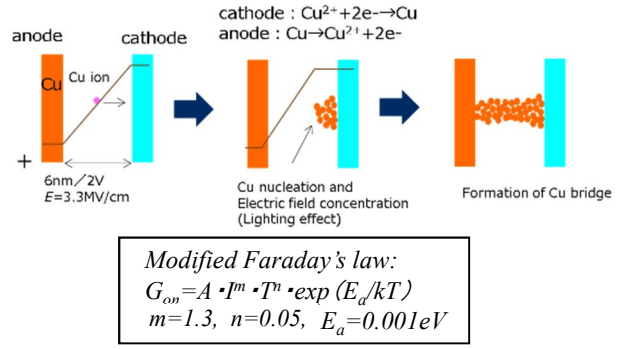

Figure 4. Set model of an atom switch. The migration behavior of Cu atoms in the electrolyte layer during set programming is explained by a modified Faraday's law [17].

$$G_{on} = A \cdot I^m \cdot t^n \cdot exp(E_a/kT), \quad (1)$$

where A is a proportional constant, k is a Boltzmann constant, and T is programming temperature. The exponents of the set programming current m = 1.3, pulse width n = 0.05, and activation energy E_a = 0.001eV are obtained from the equation (1). For reducing the programming power with keeping the same G_{on} for eNVM application, it is revealed that the higher current with smaller pulse width is the most effective way.

Fig. 5 shows G_{on} after set programming under various tuned programming conditions based on the model. The programming power required to obtain target G_{on} of about 2.5×10^{-4} S for the eNVM application is significantly reduced by one order of magnitude according to the equation (1). In addition to that, each of the G_{on} levels decreases to approximately 2×10^{-4} S after subsequent 260°C baking for 1 hour. This indicates the G_{on} also shows almost the same thermal stability as each other.

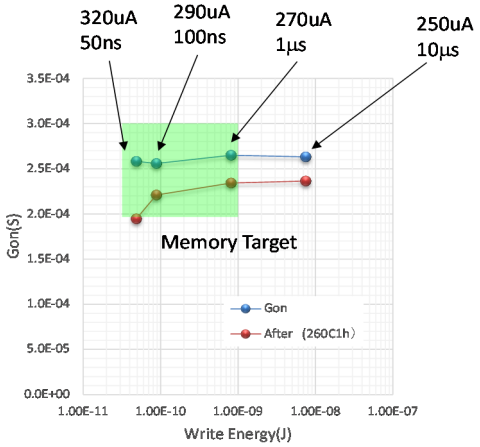

Figure 5. G_{on} after set programming and after subsequent 260°C baking for 1 hour under various tuned programming conditions based on the model. Set programming power is drastically reduced as the set pulse width decreases while keeping the same G_{on}. In addition, the thermal stability of each G_{on} is also almost the same.

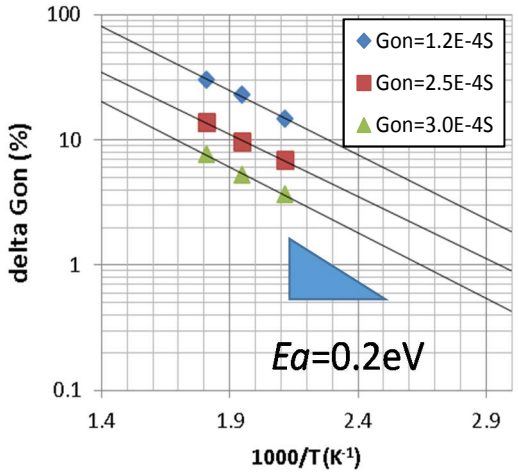

Figure 6. Temperature dependence of delta-G_{on} (G_{on} shift) of the atom switch after one hour baking at various temperatures. E_a=0.2eV is obtained irrespective of G_{on}.

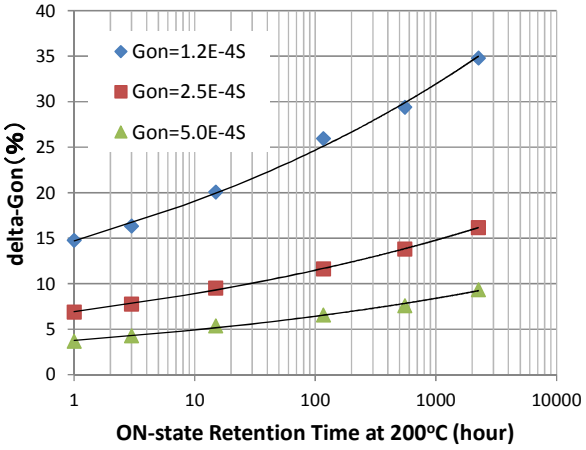

Figure 7. ON-state retention time dependence of delta-G_{on} of the atom switch with different G_{on} levels during high temperature retention test of the ON-state bits at 200°C.

C. *Retention of ON-state atom switch*

Reduction of G_{on} during the retention test indicates that the formed Cu bridge is thermally degraded and collapsed, resulting in smaller memory window. In order to estimate the G_{on} shift after the retention test, dependences of G_{on} on temperature and time are evaluated using 16-kb array. Fig. 6 shows temperature dependence of G_{on} shift (delta-G_{on}) of the atom switch after baking at various temperatures for 1 hour. All the bits had been switched by 10^3 on/off cycles before the retention test. The median delta-G_{on} is used as the typical value. The delta-G_{on} increases with increase in the baking temperature, and E_a = 0.2eV is estimated from the linear relationship between logarithmic delta-G_{on} and the reciprocal of the temperature. The estimated E_a of 0.2eV is irrespective of the prepared G_{on} ranging at least from $1.2\text{x}10^{-4}$ to $3.0\text{x}10^{-4}$ S. Fig, 7 shows ON-state retention time dependence of delta-G_{on} of the atom switch with different G_{on} levels during high temperature retention test of the ON-state bits at 200°C.The delta-G_{on} is also increased with time, indicating the higher G_{on} is more stable. From these results, the lifetime of ON-state at a field operation is estimated by following an empirical fitting model;

$$delta\text{-}G_{on} = A' \cdot G_{on}^{m'} \cdot t^{n'} \cdot exp(E_a/kT), \qquad (2)$$

where A' is a proportional constant. The exponents of G_{on} m' = -0.95, pulse width n' = 0.11, and E_a = 0.2eV are obtained at the temperature ranging from 25 to 280°C and the retention test time ranging from 1 to 2000 hours. For an automotive grade, e.g. 20 years at 150°C, required G_{on} and delta-G_{on} criteria are derived from the equation (2). Fig. 8 shows lifetime estimation as a function of G_{on} with the automotive grade requirements. It is shown that G_{on} = $2.5\text{x}10^{-4}$S with delta-G_{on} = criteria of 20% is essential for securing the lifetime of 20 years at 150°C. This stability is comparable to and/or better than the previous reports [18], [19].

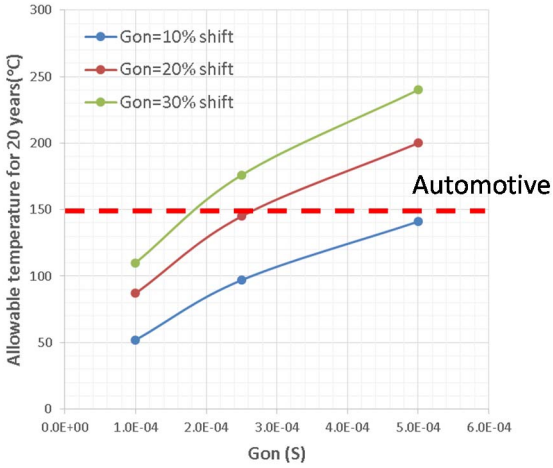

Figure 8. Lifetime estimation as a function of G_{on}. For G_{on}=$2.5\text{x}10^{-4}$S, +20% shift should be allowed for the lifetime of 20 years at 150°C.

III. CONCLUSIONS

Based on a new G_{on} tuning and ON-state retention models of atom switches, the programming power is successfully reduced while keeping the same G_{on} levels. The stable retention characteristic of G_{on} on temperature and lifetime is experimentally estimated to be more than 10 years at 150°C.

ACKNOWLEDGMENT

A part of the device processing was operated by the National Institute of Advanced Industrial Science and Technology (AIST), Japan.

REFERENCES

[1] J. Nickolls and W. J. Dally, "The GPU Computing Era," *IEEE Micro*, vol. 30, no. 2, pp. 56-69, Mar. 2010.

[2] Y. Zhao, P. Huang, Z. Zhou, C. Liu, S. Qin, L. Liu, X. Liu, H.-S.P. Wong, and J. Kang, "A Physics-Based Compact Model for CBRAM Retention Behaviors Based on Atom Transport Dynamics and Percolation Theory," *IEEE Electron Device Letters*, vol. 40, no. 4, pp. 647-650, Apl. 2019.

[3] J. Guy, G. Molas, P. Blaise, M. Bernard, A. Roule, G. Le Carval, V. Delaye, A. Toffoli, G. Ghibaudo, F. Clermidy, B. De Salvo, and L. Perniola, "Investigation of forming, SET, and data retention of conductive-bridge random-access memory for stack optimization," *IEEE Trans. Electron Devices*, vol. 62, no. 11, pp. 3482–3489, Nov. 2015.

[4] J. Guy. G. Molas, E. Vianello, F. Longnos, S. Blanc, C. Carabasse, M. Bernard, J. F. Nodin, A. Toffoli, J. Cluzel, P. Blaise, P. Dorion, O. Cueto, H. Grampeix, E. Souchier, T. Cabout, P. Brianceau, V. Balan, A. Roule, S. Maitrejean, L. Perniola, and B. De Salvo, "Investigation of the physical mechanisms governing data-retention in down to 10nm nano-trench Al2O3/CuTeGe conductive bridge RAM (CBRAM)," *2013 IEEE International Electron Device Meeting (IEDM)*, pp. 30.2.1–30.2.4.

[5] K. Ota, A. Belmonte, Z. Chen, A. Redolfi, L. Goux,and G.S. Kar, "Impact of the filament morphology on the retention characteristics of Cu/Al2O3-based CBRAM devices," *2016 IEEE International Electron Device Meeting (IEDM)*, pp. 21.2.1-21.2.4.

[6] R. Ichihara, S. Fujii, M. Yamaguchi, Y. Yoshimura, Y. Mitani, and M. Saitoh, "Investigation of Switching-Induced Local Defects in Oxide-Based CBRAM Using Expanded Analytical Model of TDDB," *IEEE Trans. Electron Devices*, vol. 66, no. 5, pp. 2165-2171, May 2019.

[7] M. Tada, T. Sakamoto, K. Okamoto, M. Miyamura, N. Banno, Y. Katoh, S. Ishida, N. Iguchi, N. Sakimura, and H. Hada, "Polymer Solid-Electrolyte (PSE) Switch Embedded in 90nm CMOS with Forming-free and 10nsec Programming for Low Power, Nonvolatile Programmable Logic (NPL)," *2010 IEEE International Electron Device Meeting (IEDM)*, pp. 16.5.1-16.5.4.

[8] M. Tada, K. Okamoto, T. Sakamoto, M. Miyamura, N. Banno, and H. Hada, "Polymer Solid-Electrolyte Switch Embedded on CMOS for Nonvolatile Crossbar Switch," *IEEE Trans. Electron Devices*, vol. 58, no. 12, pp. 4398-4406, Dec. 2011.

[9] K. Okamoto, M. Tada, N. Banno, N. Iguchi, H. Hada, T. Sakamoto, M. Miyamura, Y. Tsuji, R. Nebashi, A. Morioka, X. Bai, and T. Sugibayashi, "Robust Cu atom switch with over-400°C thermally tolerant polymer-solid electrolyte (TT-PSE) for nonvolatile programmable logic," *2016 Symposium on VLSI Technology,* pp. 124-125.

[10] M. Tada, T. Sakamoto, N. Banno, K. Okamoto, M. Miyamura, N. Iguchi, and H. Hada, "Improved Reliability and Switching Performance of Atom Switch by Using Ternary Cu-alloy and RuTa Electrodes," *2012 IEEE International Electron Device Meeting (IEDM)*, pp. 29.8.1-29.8.4.

[11] M. Tada, T. Sakamoto, N. Banno, K. Okamoto, N. Iguchi, H. Hada, and M. Miyamura, "Improved ON-State Reliability of Atom Switch Using Alloy Electrodes," *IEEE Trans. Electron Devices*, vol. 60, no. 10, pp. 3534-3540, Oct. 2013.

[12] X. Bai, T. Sakamoto, M. Tada, M. Miyamura, Y. Tsuji, A. Morioka, R. Nebashi, N. Banno, K. Okamoto, N. Iguchi, H. Hada, and T. Sugibayashi, "A low-power Cu atom switch programmable logic fabricated in a 40nm-node CMOS technology," *2017 Symposium on VLSI Technology,* pp. T28-T29.

[13] T. Sakamoto, Y. Tsuji, X. Bai, M. Miyamura, A. Morioka, R. Nebashi, N. Banno, K. Okamoto, N. Iguchi, H. Hada, T. Sugibayashi, and M. Tada, "Atom Switch with Improved Cycle Endurance using Field Enhancement for Nonvolatile SoC," *2018 IEEE International Memory Workshop (IMW)*, pp. 165-168.

[14] N. Banno, M. Tada, T. Sakamoto, K. Okamoto, M. Miyamura, N. Iguchi, T. Nohisa, and H. Hada, "Nonvolatile 32x32 Crossbar Atom Switch Block Integrated on a 65-nm CMOS Platform," *2012 Symposium on VLSI Technology*, pp. 39-40.

[15] N. Banno, M. Tada, T. Sakamoto, K. Okamoto, M. Miyamura, N. Iguchi, and H. Hada, "Improved Switching Voltage Variation of Cu Atom Switch for Nonvolatile Programmable Logic," *IEEE Trans. Electron Devices*, vol. 61, no. 11, pp. 3827-3832, Nov. 2014.

[16] N. Banno, M. Tada, T. Sakamoto, M. Miyamura, K. Okamoto, N. Iguchi, T. Nohisa, and H. Hada, "Cu Atom Switch With Steep Time-to-ON-State Versus Switching Voltage Using Cu Ionization Control," *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 2966-2971, Sep. 2015.

[17] M. Tada and T. Sakamoto, "Set/Reset Switching Model of Cu Atom Switch Based on Electrolysis," *IEEE Trans. Electron Devices*, vol. 64, no. 4, pp. 1812-1817, Apr. 2017.

[18] J.R. Jameson, J. Dinh, N. Gonzales, S. Hollmer, S. Hsu, D. Kim, F. Koushan, D. Lewis, E. Runnion, J. Shields, A. Tysdal, D. Wang, and V. Gopinath, "Towards Automotive Grade Embedded RRAM," *2018 48th IEEE European Solid-State Device Research Conference (ESSDERC)*, pp. 58-61.

[19] M. Ueki, Y. Hayashi, N. Furutake, K. Masuzaki, A. Tanabe, M. Narihiro, H. Sunamura, K. Uejima, A. Mitsuiki, K. Takeda, and T. Hase, "Stabilizing Schemes for the Minority Failure Bits in Ta_2O_5-Based ReRAM Macro," *IEEE Trans. Electron Devices*, vol. 64, no. 2, pp. 419-426, Feb. 2017.