Stochastic Computing Based on Volatile Ovonic Threshold Switching Devices

Z. Chai[1,2], W. Zhang*[1], J. F. Zhang[1]
[1]School of Engineering, Liverpool John Moores University, UK, [2] State Key Laboratory for Mechanical Behavior of Materials, and School of Materials Science and Engineering, Xi'an Jiaotong University, Xi'an, China
*Corresponding author, email: w.zhang@ljmu.ac.uk

Abstract

Ovonic Threshold Switching (OTS) selector is essential in the **1S** (selector) **-1R** (resistive switching device) crossbar memory array to suppress the sneak current paths. OTS exhibits inherent stochastic characteristics in its switching process and can be used for implementing true random number generators (TRNGs). Stochastic computing (SC) can be further designed and realized by exploiting the probabilistic switching behavior in the OTS. The stochastic bit streams generated by OTS are demonstrated with good computation accuracy in both multiplication operation and edge detection circuit for image processing. Moreover, the distribution of random bit in the stochastic streams generated by OTS has been statistically studied and linked to the defect de/localization behavior in the chalcogenide material. Weibull distribution of the delay time supports the origin of such probabilistic switching, facilitates further optimization of the operation condition, and lays the foundation for device modelling and circuit design. Considering its other advantages such as simple structure, fast speed, and volatile nature, OTS is a promising material for implementing SC in a wide range of novel applications, such as image processors, neural networks, control systems and reliability analysis.

1. Introduction

Stochastic computing (SC) is a special type of digital compute strategy where numbers are represented by the probability of 1 and 0 in stochastic bit streams. It is an approximate computing paradigm that enables low-cost implementations of arithmetic operations using standard logic elements, leading to superior hardware simplicity, and also provides high tolerance to soft errors [1-2]. Therefore, SC is particularly suitable for applications requiring parallel processing techniques, such as image processing [3], neural networks [4] and control systems [5]. However, the performance of SC is limited by the quality of bit streams: correlation in the bit streams could dramatically degrade the computing accuracy. As demonstrated in Fig. 1a, a multiplication operation can be simply realized with a single AND gate with two random input operand streams, but correlated input streams could cause errors and should be avoided (Fig.1b).

Emerging nanotechnologies such as memristors including resistive switching devices (ReRAM) [6], phase change devices (PCRAM) [7] and magnetic-tunnel junction devices (MTJ) [8] are suitable for SC due to their natural probabilistic behavior, by utilizing their probabilistic mechanisms. However, the non-volatile memristors and MTJ devices need an erase (reset) operation and a separate read-out operation in each programming cycle, which increases operational complexity and energy consumption, and limits the bit generation frequency.

Ovonic threshold switching chalcogenide materials, such as GeSe, GeAsTe and SiGeAsSe, have been used as selector devices to suppress the sneak current in crossbar emerging 1S1R memory arrays due to their favorable characteristics such as CMOS-compatibility, volatile switching, fast speed and excellent endurance [9-10]. OTS switching is an electronic process which involves defect localization/delocalization in a volatile conductive filament formed during the first-fire operation [9]. OTS selectors have high on-state drive current (> 10 MA/cm^2), good half-bias non-linearity, fast switching speed, and excellent endurance when compared with other selectors [10]. The endurance of Ge_xSe_{1-x} based OTS device is more than 10^{10} cycles by using a simple recovery scheme. After composition optimization, Se-Ge-As-Te OTS has achieved an excellent endurance of more than 10^{11} cycles without recovery operation. The statistical analysis of the OTS switching voltage and switching time and their correlation is critically important for its SC applications.

True random number generation (TRNG) is essential in SC, and in many other applications such as communication systems, statistical sampling, computer simulation and cryptography systems. Unlike the software-based pseudo-RNGs, hardware-based true RNGs use local physical phenomena to produce truly random outputs, which cannot be replicated or predicted externally, and are particularly critical in hardware security applications. OTS's probabilistic switching has been exploited to implement the true random number generators (TRNGs) with good randomness and stability.

The time-to-switch-on/-off (t_{on}/t_{off}) in OTS at a constant bias is found following the Weibull distribution and is

dependent on both the pulse bias and time. This stochastic nature of OTS switching can be used to implement TRNGs [11]. The volatile nature of OTS makes reset operations unnecessary, simplifying the operation conditions and improving the generation frequency. The stochastic bit streams generated by OTS-based TRNG have shown good computation accuracy in both multiplication operation and edge detection circuit for image processing. OTS has simple structure, fast speed, and volatile nature is promising for SC implementation in a wide range of novel applications, such as image processors, neural networks, and control systems.

2. Devices and Experiments

Amorphous Ge_xSe_{1-x} films are prepared by room temperature physical vapor deposition (PVD). TiN/GeSe/TiN selector devices were integrated in a 300 mm process flow, using a pillar (TiN) bottom electrode which defines the device size down to 50 nm. A Ge_xSe_{1-x} chalcogenide films control from 20 nm down to 5 nm thickness was achieved and passivated with a low-temperature BEOL process scheme (Fig.1c). The OTS device has a balanced probability to be switched on, which can become a good source of random "0" and "1" generation. Fig.1d shows the schematic of the bit stream generation waveform for the TRNG. Current is measured at the end of each pulse. Since device might be switched on immediately or not switched or switched on after some time during the pulse, such stochasticity is used as the source of the randomness generation to convert it to a bit stream.

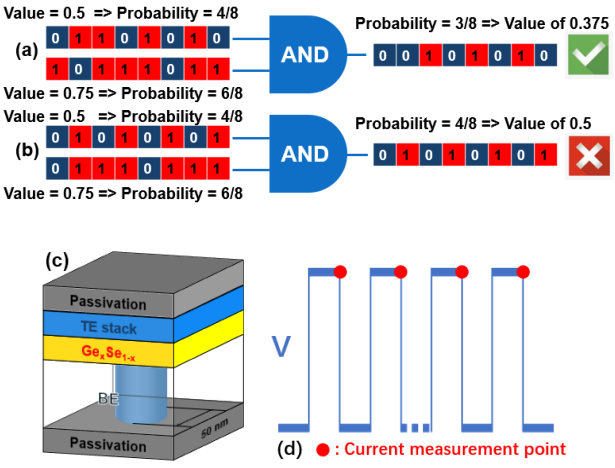

Fig.1. (a) A multiplication operation is implemented with a single AND gate, (b) but the error is unacceptable with correlated input bit streams. (c) Schematic of the OTS structure. GeSe is the switching layer and device size is confined by the pillar bottom electrode (BE). (d) Schematic of the bit stream generation waveform.

3. Stochastic Switching and TRNG

OTS device is switched on and off by a square pulse for 100 cycles and the I-t waveforms are measured and plotted together. It is observed that the time-to-switch-on (t_{on}) recorded at the constant pulse top bias spreads over a wide time range and follows the Weibull distribution, as shown in Fig. 2a. t_{on} at different biases shifts towards shorter t_{on} at higher V_{OTS}. The dependence of t_{on} on V_{OTS} is clearly shown in Fig. 3b. A 99.7% switching-on probability, equivalent to the 3σ of the normal distribution (t_{on}, 3σ), can be obtained via a linear extrapolation. An increase of bias by 0.2 V will lead to a $t_{on,99.7\%}$ decrease by nearly two orders. V_{OTS} of 3.5 V is needed in order to achieve a t_{on} of 10 ns with 99.7% of probability by a linear extrapolation. This supports that the t_{on} at constant biases can be controlled by the bias (Fig.3a), which provides a reliable method to generate the random bit streams for SC.

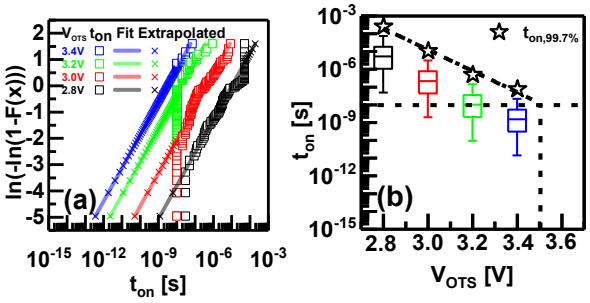

Fig. 2. (a) Weibull plot of measured data (□), linear fitting (——) and extrapolation beyond measurement resolution and pulse duration (×) of t_{on} at different biases. (b) Boxplot of t_{on} at different biases and the extracted t_{on} to ensure the probability of switching-on reaches 99.7% (☆) as obtained from the Weibull distribution. V_{OTS} of 3.5 V is needed to achieve a t_{on} of 10 ns with 99.7% of probability by a linear extrapolation.

4. Stochastic Computing

SC using OTS is demonstrated first in a multiplication operation. With a fixed pulse width, OTS's switching probability is dependent on the pulse amplitude (Fig. 3a). The bit streams generated by the 1,000 pulses at 2.7 V and 2.8 V are sent to an AND gate for multiplication (Fig. 3b). After SC, the output bit stream represents a value of 0.463 (Fig. 3c), very close to the arithmetic product of 0.460 and supporting the good stochasticity and uncorrelatedness of bit streams generated by OTS.

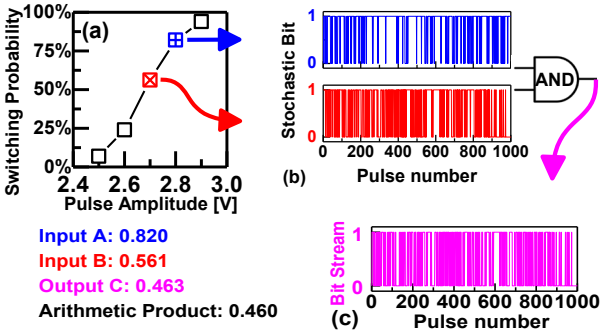

Fig.3. (a) Switching probability at different pulse amplitudes (tpulse = 1μs). (b) 1000-bit stochastic streams generated at 2.7 V and 2.8 V representing the values of 0.561 and 0.820 respectively sent to an AND gate for multiplication. (c) The output bit stream after SC represents a value of 0.463, close to the arithmetic product of 0.460.

The segment length, i.e. the number of consecutively appearing "0" or "1" as demonstrated in Fig. 4a, follows exponential distribution (Fig. 4b), which indicates that the stochastic generation process can be considered as a memoryless discrete-time Markov chain. The mean value of the segment lengths can be named as segment length constant τ (Fig. 4b). τ_0 and τ_1 are oppositely dependent on the pulse amplitude (Fig. 4c) and pulse width (Fig. 4d).

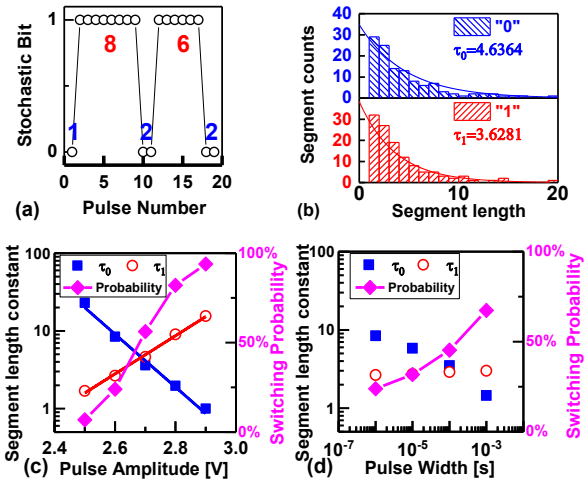

Fig.4. (a) Demonstration of "0" and "1" segments in a 20-bit stream. (b) Exponential distribution of segment length, from the 1,000-bit stream generated with Vpulse = 2.7 V and tpulse = 1 μs. (c) Dependence of segment length constants, τ0 and τ1, on Vpulse with fixed tpulse = 1 μs. The corresponding τ0 and τ1 of any analog probability value can be obtained via interpolation (straight lines). (d) Dependence of segment length constants on tpulse with fixed Vpulse = 2.6 V.

This facilitates the modelling of bit streams of any probability values and the simulation of an OTS-based edge detection circuit in an image processing system. Such system uses an OTS array to convert pixel values into stochastic bit streams in parallel (Fig. 5a), which are further processed by the Robert cross algorithm (Fig. 5b) to highlight significant gradients in the diagonal direction across the array. The array and circuit have been reported in [6]. In this way, image edge can be detected. Fig. 4(c-d) compares the edge detection result with streams of 100 and 1,000 bits respectively. Edge detection is successful and increasing the stream length can significantly improve detection quality.

The origin of such excellent stochasticity of OTS-based SC can be attributed to the stochastic delay time during switch-on, as demonstrated in Fig. 6a. It is further found that the delay time within the measurement range follows Weibull distribution at different pulse amplitudes (Fig. 6b). It is well known that the time-dependent dielectric breakdown (TDDB) follows the Weibull distribution and is triggered by the formation of a filamentary conductive percolation path [21][23]. The Weibull distribution of t_{delay} in OTS can be attributed to a volatile filamentary formation process, therefore, induced by a different mechanism such as electronic defect de/localization [23-25].

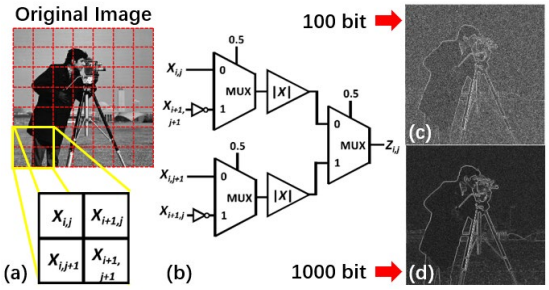

Fig.5. (a) A 256x256 image for edge detection with schematic pixel array. (b) Edge detection circuit based on Robert cross algorithm. Xi,j is the stochastic bit stream generated by OTS, representing the pixel value at (i,j) in the array.(c-d) Edge detection results using (c) 100-bit and (d) 1000-bit stochastic streams.

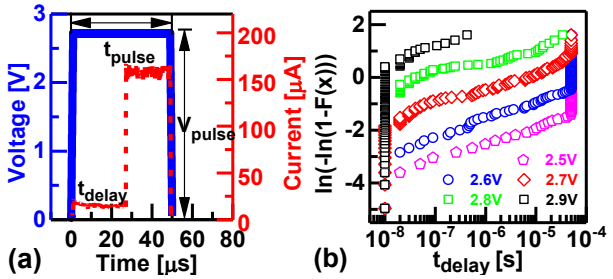

Fig.6. (a) Current of OTS during a constant voltage pulse (50 μs, 2.7V), with probabilistic the delay time (tdelay) before switch-on. (b) Weibull plot of t_{delay} under various pulse amplitudes. The t_{delay} is limited by the measurement resolution (10 ns) and pulse width of 50 μs

The dependence of OTS switching probability on both pulse amplitude and width is further investigated. In Fig. 6b, t_{delay} is reduced at higher pulse amplitude, as the Weibull distribution parallel shifts to left, in agreement with Fig.2a. This demonstrates that a pulse with either higher amplitude or longer width can increase the switching probability measured at the end of the pulse. The details of the Weibull plot can be found in [9] [12]. The Weibull parameters, α and β, in Fig. 7a, are extracted from Fig. 6b and can be fitted well linearly with pulse amplitude. Based on this observation, the switching probability at a wide range of pulse conditions, i.e. amplitude and/or width, can be simulated, as shown in the heat map of Fig. 7b, where the scattered coloured dots are experimental switching probability measured by applying 1,000 pulses at the corresponding conditions. The good agreement supports that whilst the switching of OTS is stochastic, the switching probability can be precisely

controlled by either tuning the pulse amplitude or width in a wide range, which provides further flexibility for its SC application.

A practical OTS-based SC system will be challenged by a range of reliability issues, such as switching probability drift induced by cycling and device-to-device (D2D) variability. The endurance of GeSe OTS has been significantly improved by introducing a recovery scheme which prolongs the endurance to $>10^{11}$ [9], which is limited by the measurement instrument. Whilst SC has demonstrated good robustness against reliability issues thanks to its error-tolerate nature [2], these issues can be further migrated by solutions at the peripheral circuitry level, such as a real-time switching probability monitor circuit utilizing counters/comparators to adjust and map the input pulse conditions accordingly. Additionally, the switch-off process of OTS is also a probabilistic process but at a much faster speed, which could be further exploited in future work to enhance the performance of OTS based SC.

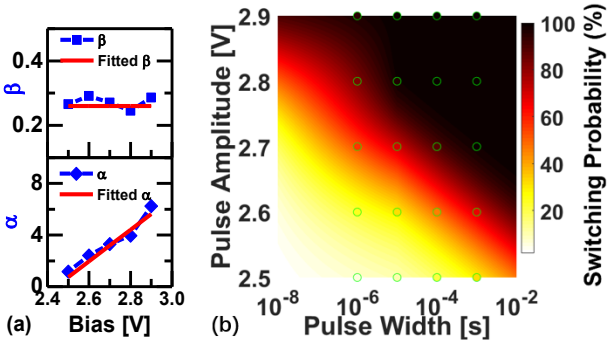

Fig.7. (a) Extracted Weibull parameters at different pulse amplitudes. (b) Simulated dependence of switching probabilities on pulse amplitude and width. Experimental data (scatters "o" with filled colours) indicates the measured probability.

5. Conclusions

In this paper, stochastic computing using OTS selector device is demonstrated by exploiting its probabilistic switching behavior. The high-quality stochastic bit streams generated by OTS lead to good computation accuracy in both multiplication operation and edge detection circuit for image processing. The bit distribution in the stochastic streams has been statistically studied and linked to the defect localization/ delocalization behavior in the chalcogenide material. Weibull distribution of the delay time supports the filamentary origin of such probabilistic switching, facilitates optimization of operation conditions, and lays the foundation for device modelling and circuit design. The simple structure, fast speed, and volatile nature of OTS make it promising for stochastic computing in a wide range of novel applications, such as image processors and neural networks.

Acknowledgments

This work was supported by the Engineering and Physical Science Research Council of UK under the grant no. EP/M006727/1 and EP/S000259/1. The authors would like to thank colleagues at IMEC, Belgium, for supply of test samples used in this work and fruitful discussions.

References

[1] B.R. Gaines, Stochastic Computing Systems, Advances in Information Systems Science, Boston, MA, Springer, pp. 37-172, 1969, DOI: 10.1007/978-1-4899-5841-9_2.

[2] A. Alaghi, The Logic of Random Pulses: Stochastic Computing, University of Michigan, Ph.D Dissertation, 2015.

[3] P. Li, et al, IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 22, no. 3, pp. 449 - 462, Apr. 2014 DOI: 10.1109/TVLSI.2013.2247429

[4] J. Yu, et al, IEEE 35th International Conference on Computer Design, Boston, MA, USA, Nov. 2017, DOI: DOI: 10.1109/ICCD.2017.24.

[5] D. Zhang et al, IEEE Trans. Ind. Electron., vol. 55, no. 2, pp. 551-561, Feb. 2008, DOI: 10.1109/TIE.2007.911946

[6] Y. Zhao, et al, in IEDM 2019, DOI: 10.1109/IEDM19573.2019.8993559.

[7] S. Gaba, et al, IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne VIC, Australia, July 2014, DOI: 10.1109/ISCAS.2014.6865703.

[8] Y. Lv et al, IEDM 2016, DOI: 10.1109/IEDM.2017.8268504.

[9] Z. Chai, et al, VLSI Symp. Tech. Dig., 2019.

[10] F. Hatem, et al, IEDM 2019, DOI: 10.1109/IEDM19573.2019.8993448

[11] Z. Chai , et al, IEEE Electron Device Lett., vol. 41, no. 2, DOI: 10.1109/LED.2019.2960947

[12] Z. Chai, et al, IEEE Electron Device Lett.,vol. 40, no. 8, 2019, DOI: 10.1109/LED.2019.2924270