

# A neuromorphic core based on threshold switching memristor with asynchronous address event representation circuits

Jinsong WEI<sup>1,2,4†</sup>, Jilin ZHANG<sup>3†</sup>, Xumeng ZHANG<sup>2</sup>, Zuheng WU<sup>2</sup>, Rui WANG<sup>2</sup>, Jian LU<sup>1,2,4</sup>, Tuo SHI<sup>2,4</sup>, Mansun CHAN<sup>5</sup>, Qi LIU<sup>2,4\*</sup> & Hong CHEN<sup>3\*</sup>

<sup>1</sup>*School of Microelectronics, University of Science and Technology of China, Hefei 230027, China;*

<sup>2</sup>*Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China;*

<sup>3</sup>*Department of Microelectronics and Nanoelectronics, Tsinghua University, Beijing 100084, China;*

<sup>4</sup>*Zhejiang Lab, Hangzhou 200240, China;*

<sup>5</sup>*Department of Electronic & Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China*

Received 18 November 2020/Revised 31 December 2020/Accepted 9 March 2021/Published online 16 December 2021

**Abstract** The full memristive network hardware features high density and excellent scalability. However, recent researches on the full memristive network have been limited to a single-layer network, due to the lack of effective and flexible communication between neurons. In this design, we demonstrate a neuromorphic core based on Ag/SiO<sub>2</sub>/Au threshold switching memristor, which has built-in asynchronous address event representation (AER) circuits to provide flexible communication between neurons. Since temporally sparse spikes are the medium of communication between neurons, the AER circuits are designed to transmit spikes serially which have been encoded with neurons' addresses before transmission. With the asynchronous circuits design, the AER circuits will detect neurons' output in real-time. To test the performance of the neuromorphic core, we designed a behavioral simulator for the neuromorphic core to simulate the liquid state machine (LSM) network, which achieves a 100% recognition rate in the free spoken digital dataset. The simulation results show that the neuromorphic core obtains 35 times higher performance than the CPU and 111 times higher energy efficiency than the GPU.

**Keywords** leaky-integration-and-fire (LIF), memristor, threshold switching, artificial neuron, AER circuits, asynchronous circuits, on-chip communication

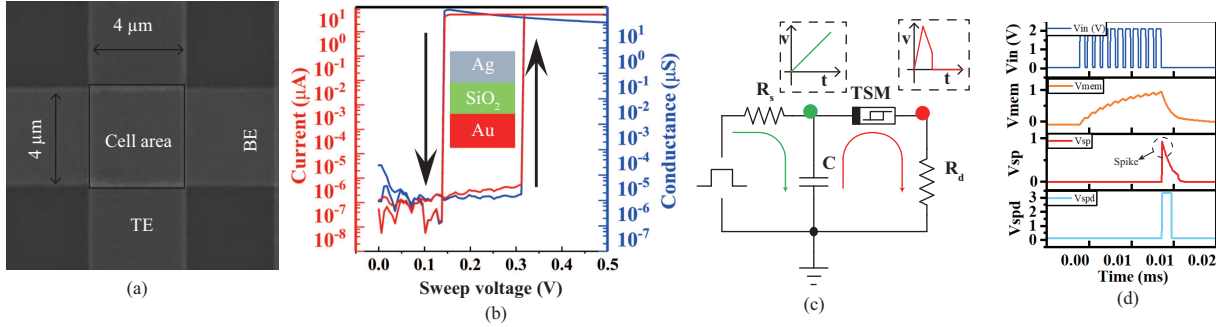
**Citation** Wei J S, Zhang J L, Zhang X M, et al. A neuromorphic core based on threshold switching memristor with asynchronous address event representation circuits. *Sci China Inf Sci*, 2022, 65(2): 122408, <https://doi.org/10.1007/s11432-020-3203-0>

## 1 Introduction

Spiking neural networks (SNNs) draw inspiration from the brain to improve the capabilities and energy efficiency of machine learning. Some researchers adopt digital or analog circuits based on complementary metal-oxide-semiconductor (CMOS) technology to realize neurons and use synapses based on static random-access memory (SRAM) [1–4]. With network-on-chip, multi-cores are integrated into a single chip to handle a huge network. However, as the area of CMOS neuron and SRAM synapse is difficult to be reduced, the total number of physical neurons in a chip is hard to be increased. To improve the density and energy-efficiency of synapses, the memristor is used to design neural networks, such as face classification network [5], reinforcement learning network [6], and long short-term memory network [7], in which the multi-level conductance of memristor is used as the weight of the synapse, and the neurons are used to design analog-digital converts (ADCs) which consume a lot of areas. Furthermore, the leaky-integration-and-fire (LIF) neuron based on the threshold switching memristor (TSM) proposed in [8] is

\* Corresponding author (email: liuqi@ime.ac.cn, hongchen@tsinghua.edu.cn)

† Wei J S and Zhang J L have the same contribution to this work.



**Figure 1** (Color online) (a) The scanning electron microscope image of a TSM device. (b) IV characteristics of TSM and the corresponding switching of the resistance. According to the  $I-V$  curve, TSM will turn on when voltage is larger than 0.31 V and turn off when lower than 0.13 V. (c) The schematic of TSM neuron circuits, consisting of a charging loop and a discharging loop. (d) The membrane voltage and spike of a neuron when a 1 kHz spike train is applied to the neuron.

used in the fully memristive network to reduce power consumption and area [9,10]. As we know, the TSM based neuron has a simple structure and features high density and excellent scalability. However, because the TSM neuron is composed of passive devices, its driving ability is too limited to drive the next neuron. This problem limits the flexibility of interconnection between neurons, so it is impossible to build a multi-layer fully memristive network. To implement a multi-layer network, a good solution is to digitally shape spikes and design neuron interconnection circuits in the digital domain. The address event representation (AER) circuits are adopted to transmit spikes from an array of neurons in one chip to other chips [11], which are also widely used in spike-based-image sensors [12]. The AER circuits convert parallel and discrete spikes to a serial address sequence and transmit the addresses to other chips. The connection of neurons can be flexibly configured by address mapping, and the scalability of the network is greatly improved. Moreover, the asynchronous AER circuits are clock-free circuits that bring two advantages: no clock power consumption and no quantization error in time.

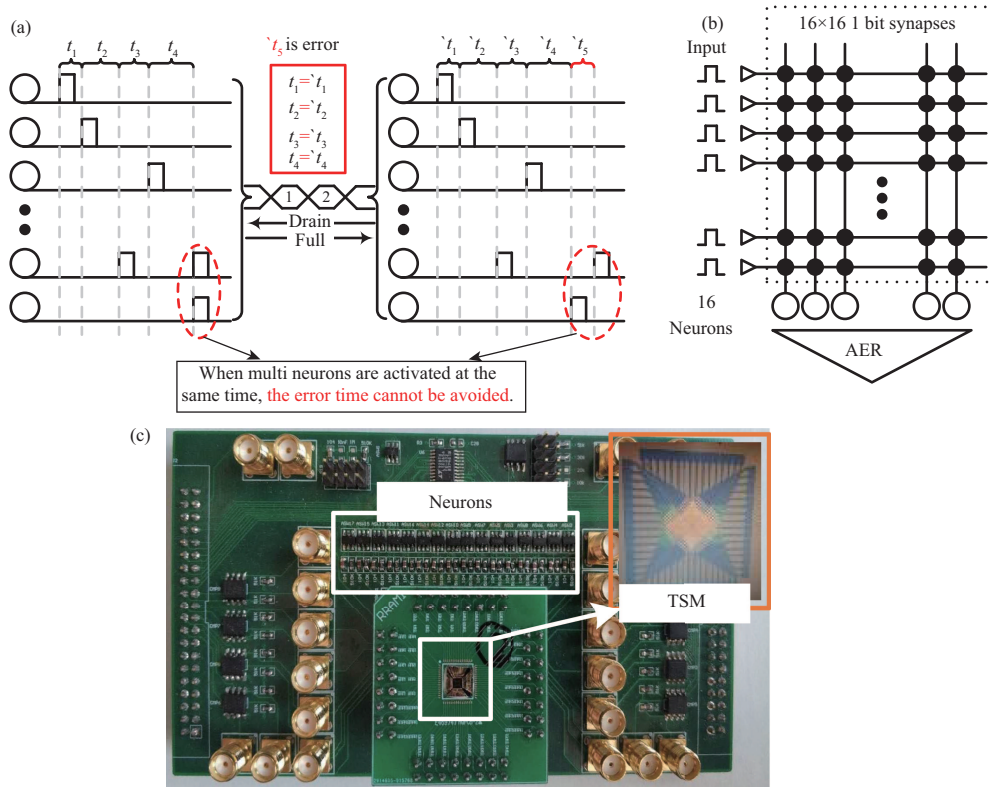
In this work, we design a neuromorphic core with 16 LIF neurons based on Ag/SiO<sub>2</sub>/Au TSM and asynchronous AER circuits. The neuromorphic core is composed of TSM neuron circuits, voltage level shifter circuits, and a field-programmable gate array (FPGA). In particular, we propose the asynchronous AER circuits, in which an OR gate and a D-flip-flop are used to receive spikes accurately, which greatly reduces the demand of the receiving circuits for the neuron's driving ability.

## 2 TSM and neuron

The threshold switch memristive devices in this work consist of three layers: an Au/Ti (40 nm/10 nm) bottom electrode, a SiO<sub>2</sub>:Ag (10 nm) as function layers, an Ag (40 nm) top electrodes. The area of the device is 4 μm × 4 μm, as shown in Figure 1(a). The detailed fabrication process of the devices can be found in our previous work [8]. The electrical characteristics of a single TSM are tested by Agilent B1500A. Figure 1(b) is the  $I-V$  curve and  $I-R$  curve of the TSM, from which we can find that, when the voltage between top and bottom electrodes of TSM is beyond the threshold voltage (about 310 mV), the TSM will switch from high resistance state (HRS) to low resistance state (LRS); when the voltage is lower than the threshold voltage (about 130 mV), the TSM will switch from LRS to HRS, as shown in Figure 1(b).

The neuron is composed of one input resistor  $R_s$ , one capacitor  $C$  and one load resistor  $R_d$ , as shown in Figure 1(c). When a series of spikes (1 kHz, 2 V) are fed into the neuron through  $R_s$ , the capacitor will accumulate the charges and lift the voltage potential. During charging, the TSM remains in the HRS. After the voltage is beyond the threshold voltage of TSM, the TSM switches from HRS to LRS, and the capacitor will discharge through  $R_d$  rapidly. A spike will be generated by the neuron, as shown in Figure 1(d).

Because the output spike voltage of a TSM neuron is a short pulse, which is not able to directly drive the next layer of the network. In this design, we first convert the spike voltage to a digital signal. The network cascading is completed in the digital domain, and an AER circuit is employed to convert multi-channel spike signals into serial signals, as shown in Figure 2(a). In this design, 16 neurons based on TSM are bounded to a printed circuits board (PCB), and the 16 × 16 1 bit synapses are designed with the on-chip memory of FPGA, as shown in Figure 2(b). The AER circuits are implemented in FPGA.



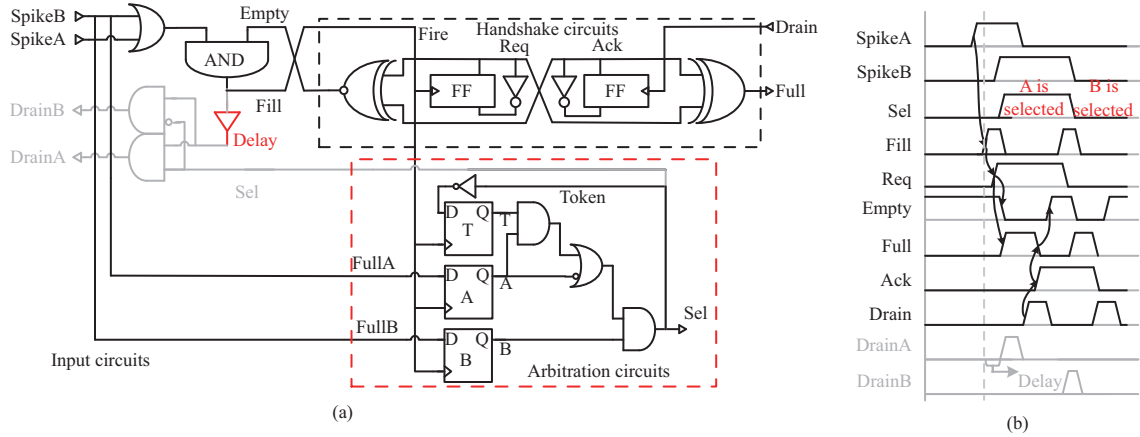
**Figure 2** (Color online) (a) The working principle of AER. The parallel and discrete spikes are converted to serial address sequences and then transmitted to the next layer. When multi neurons are activated at the same time, the time error cannot be avoided. (b) The schematic of TSM based neuromorphic core with 16 neurons, 16 × 16 1 bit synapses, and AER circuits. (c) This is the PCB of the neuromorphic core, the TSMs are bounded to a sub PCB, and the FPGA is on the back.

The input spikes are first generated by control circuits in FPGA, analog switches, and a 12 bit DAC with eight channels. An oscilloscope (MSO9404A) collects all the input spikes, output spikes, and membrane potentials. The signals in the AER circuits such as address and handshake signals are collected by JTAG circuits in FPGA because of the limited channels of the oscilloscope. The PCB is shown in Figure 2(c).

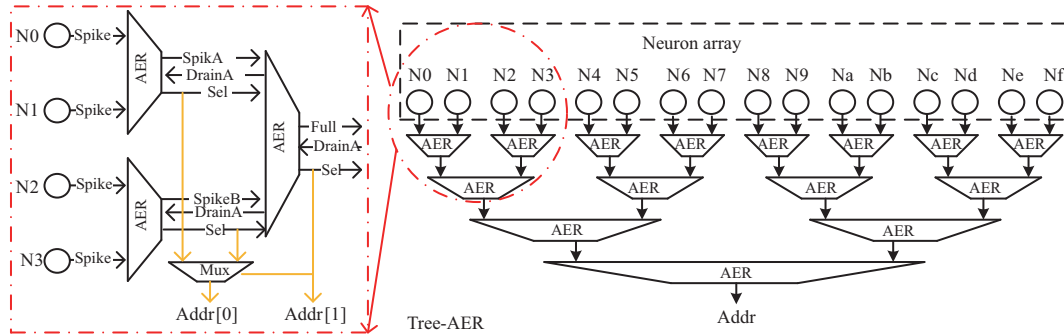
### 3 AER circuits and neuromorphic core

The AER circuits convert multiple spikes into serial signals and transmit them to the next layer. However, when multi neurons are activated at the same time, the time errors in the process of spike transmission cannot be avoided. To solve the problem, in this design, an adaptive priority arbitration scheme is put forward to distribute the time errors evenly among different neurons in Figure 2(a). As a result, the adaptive priority arbitration will reduce the influence on the neural networks. Besides, the asynchronous circuits design method is employed which brings two advantages: no clock-power computation and no quantization error in time.

Figure 3(a) illustrates the two-port AER circuits, that consist of input port circuits, arbitration circuits, and handshake circuits. The input port circuits are composed of the XOR gate, AND gate, and NAND gate. A rising edge of either of the two signals (SpikeA, SpikeB) will trigger the Fill signal, which will then trigger the handshake circuits and arbitration circuits. DrainA and DrainB are the ACK signals of SpikeA and SpikeB, respectively. According to the arbitration result, DrainA or DrainB is pulled up to respond to the input port. If previous circuits do not need the ACK, the gray circuits in Figure 3(a) can be removed. The arbitration circuits generate an arbitration priority, which is composed of three D-Flip-Flops, two AND gates, and one OR gate. Once Fill is triggered, the status of SpikeA and spikeB will be latched immediately until the next Fill signal arrives, and the priority will be updated. The arbitration logic is  $Sel = (A || A \& \& T) \& \& B$ , in which Sel is the result and Token is the priority. The adaptive priority is adopted in the arbitration process to distribute the time errors evenly among different neurons. In this work, click-based link-joint handshake circuits are adopted, which handle multi-port arbitration results.



**Figure 3** (Color online) (a) The schematic of asynchronous two-port AER circuits. (b) The timing information of the two-port AER circuits.



**Figure 4** (Color online) Tree-AER circuits.

Linke-Joint is an event-driven 4-phase handshake protocol [13], in which Fill/Empty and Drain/Full are backward and forward handshake signals respectively. When Empty is 1'b1, the Link circuits will receive Fill. When Fill is 1'b1, Empty is pulled down, and the previous circuits are blocked to ensure that the information will not be lost. At the same time, Req is reversed to make Full signal pulled up. After Full being pulled up, the subsequent circuits can receive the arbitration result (Sel signal). The timing constraints are adopted to ensure that the Sel signal is stable before the Full is high. The Drain is generated by subsequent circuits to reset Link to idle status when the Sel has been read. The waveforms in Figure 3(b) show the timing of the connection to form multi-port AER circuits.

Multiple two-port AER circuits can be connected to form multi-port AER circuits. In this work, we design a 4-stage Tree-AER with 16 input ports and 15 two-port AER circuits, as shown in Figure 4. The output of each neuron is connected to the input port (SpikeA or SpikeB) of the first-stage AER circuits. The Full signals of the first-stage of AER circuits are connected to the SpikeA or SpikeB signals of the second-stage AER circuits. The Drain signals of the first-stage of AER circuits are connected to the DrainA or DrainB signals of the second-stage AER circuits. The connection of the AER circuits of other stages is the same as that of the first-stage AER circuits. The Sel signals of the selected AER circuits at each stage constitute a 4-bit ADDR signal, which is the address of the activated neuron and is transmitted to the next layer of the network.

Furthermore, we design refractory circuits to achieve controllable refractory period time for neurons and protect the TSM device. According to the model of LIF neurons, when a neuron is activated, it enters the refractory period. In the refractory period, the input spikes to the neuron will not affect the membrane potential, and the neuron will not generate spikes in the refractory period. Another function of refractory circuits is that it protects the TSM device to be break down from continuous input when the TSM switches to LRS. When the TSM switches to LRS, the continuous input spikes will cause a large current, which may break down the TSM. The refractory circuits are designed as shown in Figure 5(a), in which a Mask signal is used to shield the input spike to the neuron, when the neuron is activated. Figure 5(b) is the finite-state-machine (FSM) of the refractory circuits. The key signal is generated by

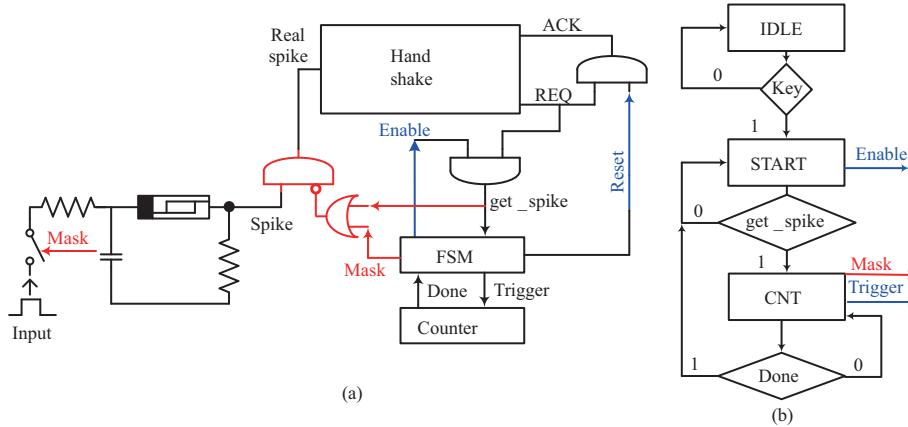


Figure 5 (Color online) (a) Refractory circuits; (b) FSM of refractory circuits.

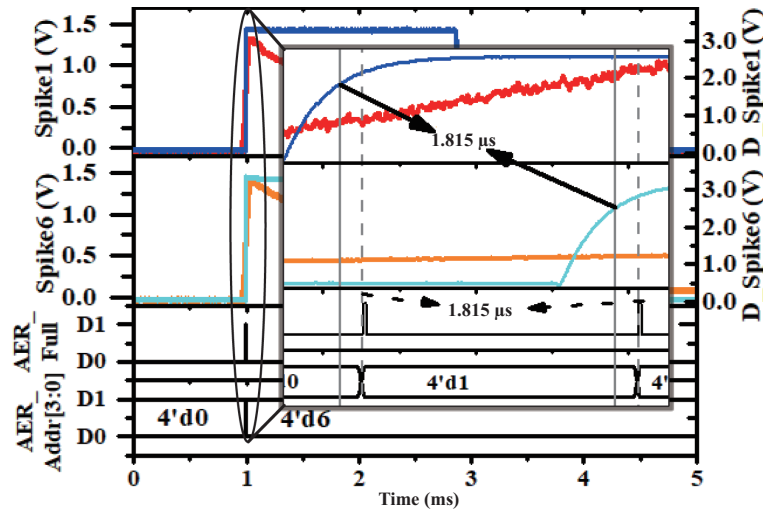


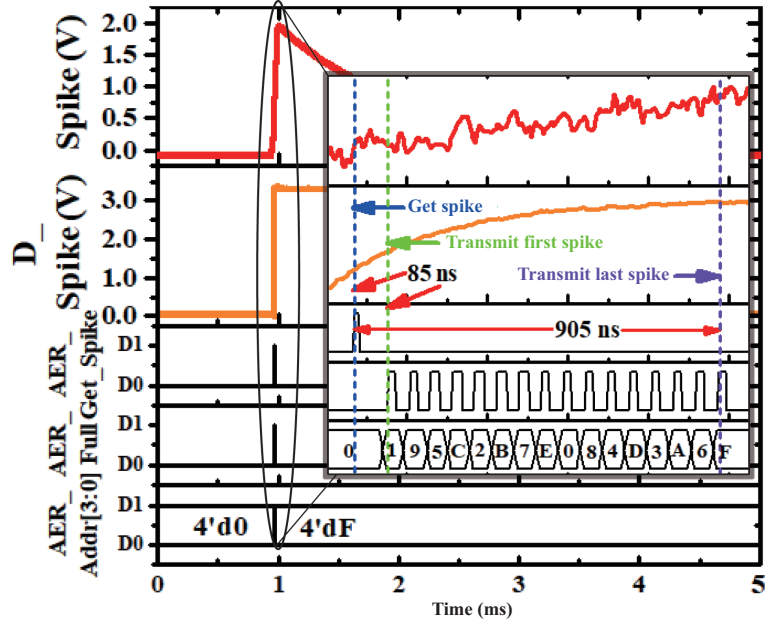
Figure 6 (Color online) The spike waveforms of neuron1 and neuron6. The red line and orange line are spikes of neuron1 and neuron6. The blue line is digital spike of neuron1; the cerulean line is the digital spike of neuron6. AER\_FULL is the spike that is transmitted to the next layer, and AER\_Addr is the address of spikes.

the reset button in the PCB. When the button is pressed, the system will be activated. If the neuron is activated (get\_spike is 1'b1), the FSM switches to “CNT” state and the Mask signal is 1'b1. After a pre-set period, the system switches the “start” state. This period is the refractory time of neurons, which is defined by users.

## 4 Results and discussion

With the neuromorphic core, two experiments are carried out: each neuron spikes at different times, and all neurons spike at the same time. When AER circuits serially transmit a spike of a neuron, it is necessary to ensure that the time information carried by the spike is not lost [14,15]. We adopt an oscilloscope to accurately measure the time when neurons generate spikes and the time taken by AER for transmitting spikes. The measurement results of neuron1 and neuron6 in the neuromorphic core are shown in Figure 6, in which the time when neuron1 generated a spike was 1.815  $\mu\text{s}$  earlier than the time when neuron6 generated a spike ( $\Delta t_1 = 1.815 \mu\text{s}$ ). When the handshake signal of AER (Full) is high, it indicates that the AER circuits have finished encoding the spike. The neurons in the post layer obtain Addr and restore spikes according to the Addr when Full is high. The rising edge of the Full signal means that the spike transmission is finished. In Figure 6, the time interval between neuron1 and neuron6 reaching the next layer is  $\Delta t_2 (= 1.815 \mu\text{s})$  equal to  $\Delta t_1$ , which proves that the timing information is not lost.

When multi neurons are activated at the same time, spikes congestion will occur. As a result, time errors will appear when blocked spikes are transmitted. To detect the errors in an extreme case, we

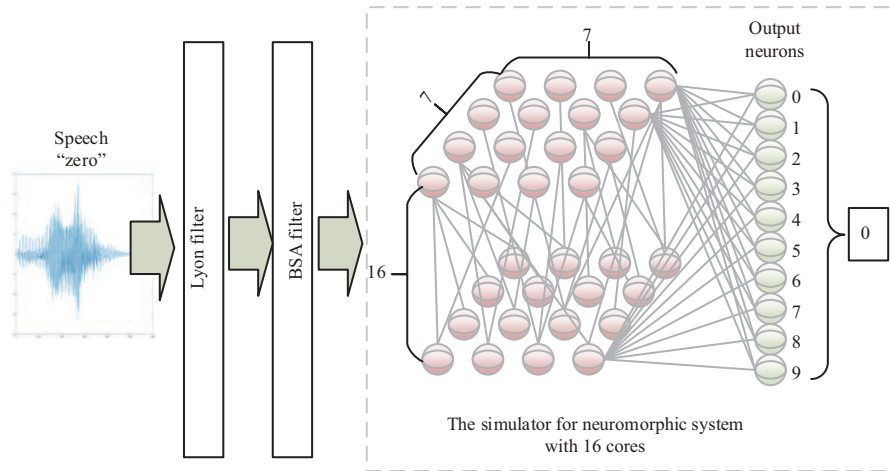


**Figure 7** (Color online) The spike waveform of neurons when 16 neurons generate spikes at the same time.

connect one neuron to all 16 channels of Tree-AER to simulate the situation when 16 neurons generate spikes at the same time. From the waveform shown in Figure 7, we can find that the delay of the first transmitted spike (neuron0) is 85 ns and the delay of the last transmitted neuron (neuronF) is 905 ns. Therefore, the last transmitted spike causes the greatest time error of 825 ns. However, the probability of time errors is very small because the TSM neuron generates spikes at a speed of several thousands of spikes per second [8], but the speed of AER circuits transmitting spikes is 11.76 Mspike/s. Another method to deal with spike congestion is adoptive priority arbitration. In this design, every two-port AER circuits have a Token signal. When the Token signal is 1'b0, SpikeA channel has the highest priority. When the spikes reach the SpikeA channel and SpikeB channel at the same time, the spike of the SpikeA channel is transmitted to the next stage first. When the Token signal is 1'b1, SpikeB channel has the highest priority. When AER circuits finished the transmission of a spike, the Token signal will be updated to !Sel. This mechanism makes the time errors be distributed evenly among different neurons.

To benchmark the neuromorphic core, we design a behavior level simulator of the neuromorphic system with 16 neuromorphic cores, as shown in Figure 8. In the system, the cores are connected by the AER circuits and routers. To accurately simulate the system behavior and performance, the event-based driven mechanism and timing synchronization mechanism are adopted by the simulator. With these two mechanisms, the transmission delay and time errors of the AER circuits will be accurately simulated. A sound recognition liquid state machine (LSM) is mapped on the system. The sound signal is first transformed to spike trans by Lyon passive ear model and Ben coding algorithm (BSA) filters [16]. The output spikes of every neuron in LSM are collected and classified by a linear layer. The LSM consists of 16 layers. The neurons in each layer are arranged in 7 rows and 7 columns. The connection between each neuron is random.

Such a neuromorphic system performs 100% accuracy on the free spoken digital dataset (FSDD) speech corpus. The running time of the neuromorphic system is 5.6 ms which is 1/35 of the running time (195.2 ms) of CPU (i7 9700). We also map the LSM in NVIDIA V100. In order to make the most effective use of GPU, we set batchsize to 128. The result shows that the average time to run each case is 5.3 ms and the power consumption is 10.34 mW. The result of our neuromorphic simulator shows that the average time to run each case is 5.6 ms, which is very close to GPU. And the power consumption is 93  $\mu$ J. The simulation shows that the neuromorphic core achieves 1/1 performance and 111/1 energy efficiency to GPU.



**Figure 8** (Color online) The LSM network for digital speech recognition. The speech signal is first transformed to spike trains by Lyon and BSA filters. The LSM consists of 16 layers with 49 neurons, each layer is mapped in a neuromorphic core. The neurons are connected by random 1 bit synapses. The spikes of LSM are collected and classified by a linear layer.

## 5 Conclusion

In this paper, a neuromorphic core with the AER circuits is built for the SNN chip. The TSM neuron successfully achieves the LIF neuron model and the AER circuits successfully achieve serial transmission of multi neurons' spikes without loss of time information when there is no spike congestion. The simulation results show that the neuromorphic core obtains 35 times higher performance than the CPU, and 111 times higher energy efficiency than GPU. It is ready for the SNN chip applications in future work.

**Acknowledgements** This work was supported by National key R&D Program of China (Grant No. 2018AAA0103300), National Natural Science Foundation of China (Grant Nos. 61751401, 61804171, 61825404, 61732020, 61674090), Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDB44000000), Major Scientific Research Project of Zhejiang Lab (Grant No. 2019KC0AD02), National Science and Technology Major Project from Minister of Science and Technology, China (Grant No. 2018AAA0103100), and CAS-Croucher Funding (Grant Nos. CAS18EG01, 172511KYSB20180135).

## References

- 1 Davies M, Srinivasa N, Lin T H, et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro*, 2018, 38: 82–99
- 2 Akopyan F, Sawada J, Cassidy A, et al. TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. *IEEE Trans Comput-Aided Des Integr Circ Syst*, 2015, 34: 1537–1557
- 3 Qiao N, Mostafa H, Corradi F, et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Front Neurosci*, 2015, 9: 141
- 4 Benjamin B V, Gao P, McQuinn E, et al. Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations. *Proc IEEE*, 2014, 102: 699–716
- 5 Yao P, Wu H, Gao B, et al. Face classification using electronic synapses. *Nat Commun*, 2017, 8: 15199
- 6 Wang Z, Li C, Song W, et al. Reinforcement learning with analogue memristor arrays. *Nat Electron*, 2019, 2: 115–124
- 7 Li C, Wang Z, Rao M, et al. Long short-term memory networks in memristor crossbar arrays. *Nat Mach Intell*, 2019, 1: 49–57
- 8 Zhang X, Wang W, Liu Q, et al. An artificial neuron based on a threshold switching memristor. *IEEE Electron Device Lett*, 2018, 39: 308–311
- 9 Wang Z, Joshi S, Savel'ev S, et al. Fully memristive neural networks for pattern classification with unsupervised learning. *Nat Electron*, 2018, 1: 137–145
- 10 Midya R, Wang Z, Asapu S, et al. Artificial neural network (ANN) to spiking neural network (SNN) converters based on diffusive memristors. *Adv Electron Mater*, 2019, 5: 1900060
- 11 Chan V, Liu S C, van Schaik A. AER EAR: a matched silicon cochlea pair with address event representation interface. *IEEE Trans Circ Syst I*, 2007, 54: 48–59
- 12 Shoushun C, Bermak A. Arbitrated time-to-first spike CMOS image sensor with on-chip histogram equalization. *IEEE Trans VLSI Syst*, 2007, 15: 346–357
- 13 Roncken M, Gilla S M, Park H, et al. Naturalized communication and testing. In: *Proceedings of the 21st IEEE International Symposium on Asynchronous Circuits and Systems*, 2015. 77–84
- 14 Maass W, Bishop C M. *Pulsed Neural Networks*. Cambridge: MIT Press, 1999
- 15 Kheradpisheh S R, Ganjtabesh M, Thorpe S J, et al. STDP-based spiking deep convolutional neural networks for object recognition. *Neural Netw*, 2018, 99: 56–67
- 16 Zhang Y, Li P, Jin Y, et al. A digital liquid state machine with biologically inspired learning and its application to speech recognition. *IEEE Trans Neural Netw Learn Syst*, 2015, 26: 2635–2649