Realizing Memristive Three-dimensional Neuromorphic Computing System

Hongyu An

Email: Hongyu51@vt.edu

Committee members: Dong Ha (Chair), Yang Yi (Advisor), Mariusz Orlowski

1 Introduction

In this report, I would introduce my recent research on the topic of realizing a Memristive Three-dimensional Neuromorphic Computing System. As we known, scaling down the transistor to gain more computational power will eventually reach the unsurmountable physical limitation [1]. In order to achieve more computing capability with low power budget, novel computing architectures and methodologies need to be developed. Neuromorphic Computing (NC), which is based on non-von Neumann architecture that mimics bio-neurological signal processing methodology with a non-binary data encoding scheme, holds great promise for the next generation of low power computing machine. This report introduces the fundamental background of the neuromorphic computing from the concept to the state-of-the-art hardware implementations. Moreover, I would discuss the proposed novel three-dimensional (3D) neuromorphic computing architecture combining the nanoscale device so-called memristor and monolithic 3D integration technology, which could reduce system power consumption, provide high connectivity, resolve routing congestion issues, and offer massively parallel data processing capability.

2 Backgrounds

2.1 Neuromorphic Computing

Neuromorphic computing is a novel computing methodology that is fundamentally different from the traditional digital computer. A modern digital computer is designed intentionally to perform Boolean algebra and arithmetic under von Neumann architecture that separates the computing units (CPUs) and memory physically connecting with a data communication bus. Because of incompatibilities involving the fabrication process and the size mismatch between the CPU and the memory, the von Neumann architecture is constrained by a physical separation between the CPUs and the memory, as illustrated in Figure 1(b). These separations require an inexorable back and forth transfer of data within the parallel bus; high data rates along the bus ultimately become too costly in terms of power consumption and hence become economically infeasible.

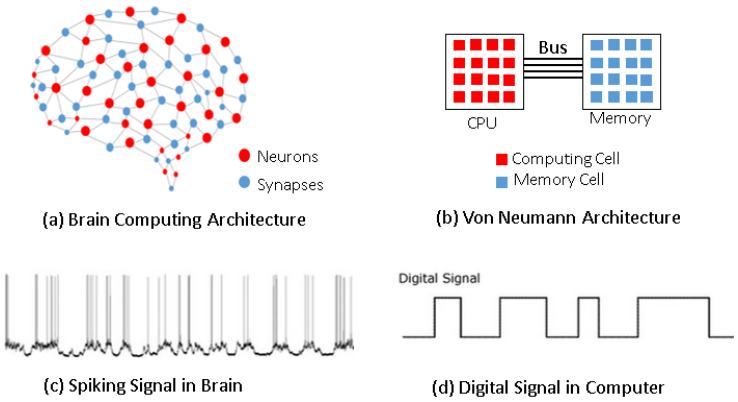

Figure 1: Comparison between brain computing architecture with von Neumann computing architecture.

On the contrary, the human brain employs the fundamental different computing schemes from the computing/memory unit level to the architecture level. Additionally, the information captured from the surrounding world is encoded in a spiking signals format in the brain [2], which is illustrated in Figure 1(c). At the architecture and device levels, the computing unit (neuron) and the memory unit (synapses) in a mammalian nervous system are placed adjacent to each other distributedly forming a network configuration, which is illustrated in Figure 1(a). Figure 1(c) depicts a spiking sequence that is an information representative form in the mammalian nervous system. Moreover, the main frequency of these spiking signals is usually low (~kHz). The low signal frequency and the network-based architecture with a small distance between the computing unit and the memory unit make the mammalian nervous system to be the most power efficient and fast computing machine in the world. Additionally, the power consumption of a human brain averages only about 20 Watts, which is significantly lower than the power consumption of any current computing system.

In order to enable the conventional computers to have similar computing efficiency and power consumption. Neuromorphic computing was proposed by Dr. Carver Mead [3] with the concept of physically rebuilding the human brain nervous system through very-large-scale integration systems (VLSI). Recently, this concept has been extended to use other emerging technologies like memristor, Magnetic tunnel junction, carbon nanotube FETs, etc.

Although, the fundamental functions of a mammalian nervous system are still under investigation, the two main elements: neuron and synapse are well studied at the cellular level.

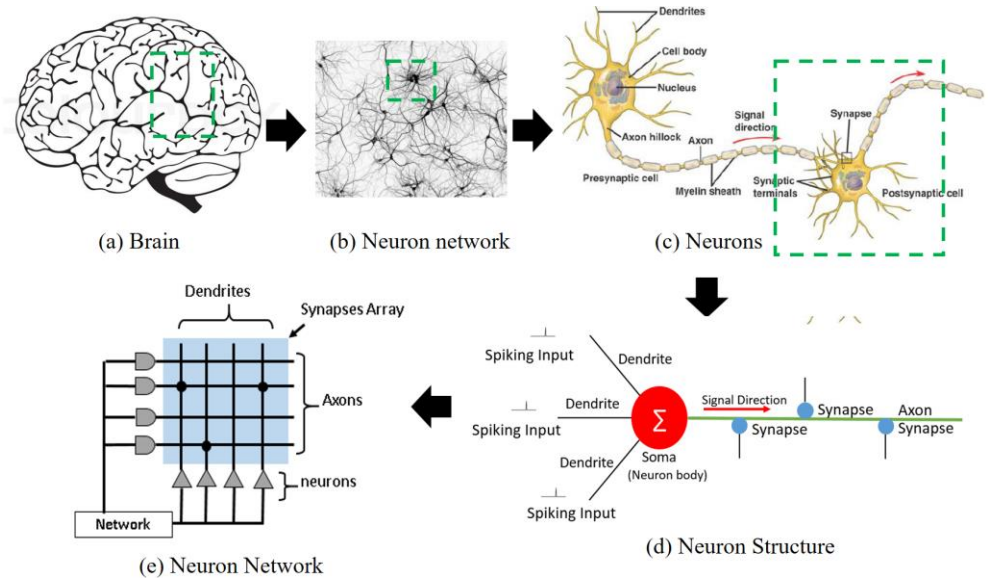

Figure 2: Neuron networks: (a) brain; (b) neural network (c) neuron connecting structure (d) neuron structure (e) neuron network architecture

The structure of a neuron is shown in Figure 2 (c) and (d). There are four main parts in each neuron:

- Dendrite: receives spiking signals from other neurons;
- Soma (neuron body): generates/sends spiking signals to the axon under the condition of the integration of received spiking signal levels from the dendrites exceeds the specific threshold voltage;
- Axon: propagates spiking signals generated by soma to other neurons, and connects to other neurons through synapses;
- Synapse: acts as a memory organ in the brain. It connects axon of the last neuron to the dendrite of next neuron. The connectivity strength can be modified by spiking signal stimulus.

As illustrated in Figure 2 (d), a neuron has a capability of receiving a multitude of input signals through dendrites simultaneously and integrating them into a membrane potential. Once the membrane potential exceeds a specific liminal value (i.e. the threshold voltage), the neuron generates a spiking signal sequence that propagates along the axon. Through the axon, firing signals are sent to the thousands of subsequent neurons. This unique connecting structure fold and signal processing method permits each neuron to receive and process thousands of signals simultaneously. In a mammalian brain, the degree of connectivity for a single neuron is at the ten-thousand-fold level; any discrete neuron can concurrently receive/send signals from/to more than 10,000 other neurons. Based on the deep understanding of neuron functions, several neuron models listed in Table 1 have been developed in the past century.

Table 1: State-of-the-art Neuron Models

Neuronal Mode	Year	Reference
Integrate and Fire	1907	[4]
Hodgkin-Huxley	1952	[5]
Leaky integrate-and-fire	1965	[6]

A synapse acting as a memory organ connects neurons in a mammalian nervous system. The connecting strength can be modified in two directions (strengthen or weaken) by both excitatory and inhibitory stimuli. This phenomenon is called as the plasticity of a synapse. The relationship between the plasticity of synapse and the memory is revealed by Dr. Kandel's research on Aplysia. In the experiments, the stimulus was repeatedly applied to the Aplysia's sensory neurons as shown in Figure 3. When the constant stimulus was repeatedly applied to the sensory neuron (1, 2, 5, 10, 15), the response signal magnitude of the motor neuron reduces gradually, indicating that the strength of the connection between sensory neuron and response neuron (motor neuron) becomes weak. This phenomenon is identified as the memory mechanism at the cellular level.

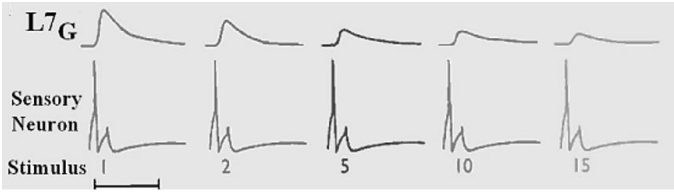

Figure 3: A sample of five identical action potentials number 1, 2, 5, 10, and 15 along with the corresponding motor response signals of diminishing strength recorded at the motor neuron (identified by L7G) (top)[7].

In order to build a neuromorphic computing system, the scientists initially attempted to simulate the brain's network structure with traditional computers or servers [8]. However, these projects suffer the inherent limitations of the von Neumann architecture-based computer; i.e. the high power consumption and the slow response time due to the inordinately high data transfer rate between CPUs and memories, especially whenever attempting to simulate any extremely large neural network that reaches the human brain level.

After a realization of the malefic traits posed by traditional digital computers, technologists sought out a novel, but natural hardware implementation and architecture: a network formed by analog neurons and synapses. These novel hardware platforms which intend neuromorphic

computing are called neuromorphic chips. Table 2 summarizes the contemporary state-of-the-art fabricated “analog” and “digital” neuromorphic chips with the comparison to human brain on critical parameters.

Table 2: The state-of-the-art neuromorphic chips

	TrueNorth [9, 10]	ROLLs[10, 11]	Neurogrid[10, 12]	HICANN Chip/ BrainScaleS[10]	SpiNNaker[10, 13, 14]	Human Brain[10, 15]
Neurons	1,048,576	256	65,535	511	20,833	20 Billions
Synapses	256 millions	128,000	N/A	113,636	20,833,333	200 Trillions
Area/volume	430 mm^2	51.4 mm^2	168 mm^2	50 mm^2	102 mm^2	1130 cm^3
Neuron Density	2438.55 per mm^2	5 mm^2	390 per mm^2	10 per mm^2	204 per mm^2	17,699 per mm^3
Synapses Density	0.595 million per mm^2	2490	N/A	2272 per mm^2	204,248 per mm^2	177 million per mm^3
Ratio of synapses to neurons	244	500	N/A	222	1,000	10,000
Power density	0.15 mW/ mm^2	0.078 mW/ mm^2	18 mW/ mm^2	57 mW/ mm^2	0.012 mW/ mm^2	0.0177 mW/ mm^3

2.2 Memristor as Synapse

The plasticity of a synapse can be implemented as a nanoscale non-volatile device “Memristor”, also called as resistive RAM (RRAM), since voltage/current pulses can alter its resistance/conductivity reversibly [16, 17]. A typical memristor is formed as a metal-insulator-metal (MIM) structure, as illustrated in Figure 4(a). The metals form the top and bottom contacts and the insulator is usually a resistive switching material [18, 19]. When the voltage/current stimulus is applied on its two terminals, the resistance of switching material would gradually change between its low resistance state (LRS) and high resistance state (HRS). The decrease in resistance of the switching material is due to the formation of the conductive filament (CF) marked as red lines in Figure 4 (a). The real TEM (Transmission electron microscopy) images are illustrated in Figure 4(b). The phenomenon is called a soft breakdown in material science. This breakdown of the switching material can be recovered by applying a reversed stimulus at the terminals, which consequently resets the memristor to its high resistance state as shown in Figure 5.

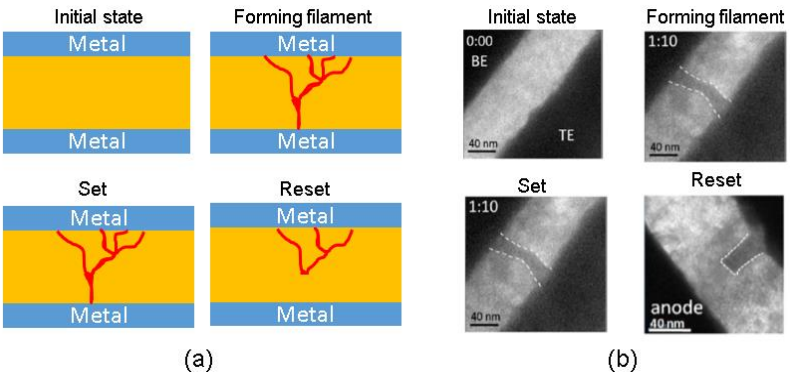

Figure 4: Illustration of the switching mechanism of a memristor: (a) switching process (b) the TEM images of the dynamic evolution of conductive filaments [10]

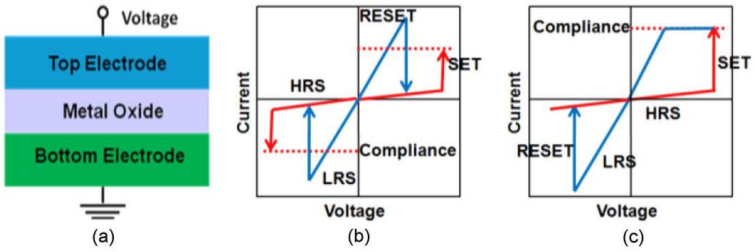

Figure 5: Schematic of MIM structure for metal–oxide RRAM, and schematic of metal–oxide memory's I–V curves, showing two modes of operation: (b) unipolar and (c) bipolar [20]

3 Memristive Three-dimensional Neuromorphic Computing System

To our best knowledge, the recent fabricated neuromorphic chips implement neurons and synapses using traditional two-dimensional CMOS and memory technology. In 2D placement methodology, a longer signal delivery distance is generally expected due to the routing density increases linearly with the number of the connections, which inevitably increases die area, power consumption [9, 21], and consequently diminishes the benefits offered by neuromorphic computing [22, 23]. Moreover, the conventional SRAM memory technology, which is used to implement synapse [9], needs extra static power for holding the stored value, which further augments consumption.

For the purpose of overcoming the limitations imposed by the state-of-the-art neuromorphic chip designs, I propose a novel 3D neuromorphic architecture utilizing three-dimensional integrated circuit (3D-IC) technology [24] to extend neuromorphic chips into the third dimension which is illustrated in Figure 6. Applying 3D integration technology to neuromorphic chips permits vertical routing paths of reduced nanoscale dimension, subsequently diminishing critical path lengths. It also decreases power consumption, and shrinks die areas with high-complexity, high-connectivity, and massively parallel signal processing capability.

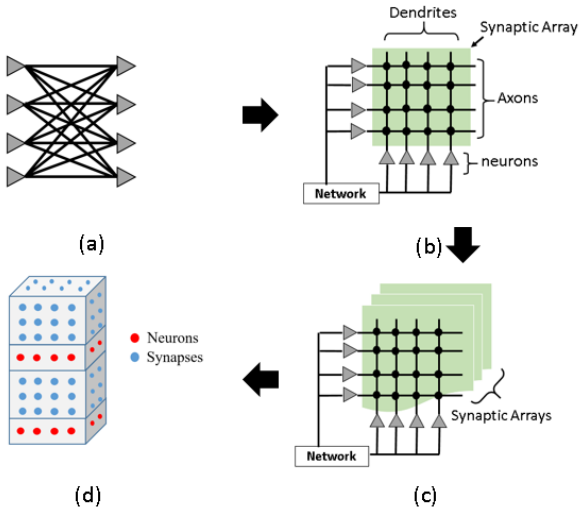

Figure 6: Evolution of neuromorphic IC architecture: from planar 2D to vertical 3D. (a) Neuron Network; (b) 2D Neuromorphic IC Architecture; (c) 3D Neuromorphic IC Architecture (d) 3D Neuromorphic IC Architecture Diagram: Stacking synaptic arrays and neuronal arrays vertically.

The benefits of applying 3D integration technology to neuromorphic chips design can be summarized as:

- Address the 2D neuron routing congestion problem, thereby increasing interconnectivity and scalability of the NC network and reducing the critical-path lengths;
- Allow numerous 3D interconnections between hardware layers that offer high device interconnection density, low power density, and broad channel bandwidth using fast and energy-efficient links;
- Provide a high-complexity, high-connectivity, and massively parallel-processing circuital system that can accommodate highly demanding computational tasks.

First, I am planning to extend the memristor-based synapse to the third dimension. In general, there are two types of three-dimensional (3D) RRAM (memristor) structures that can be used as 3D synaptic arrays: Horizontal RRAM (H-RRAM) and Vertical RRAM (V-RRAM), as shown in Figure 7.

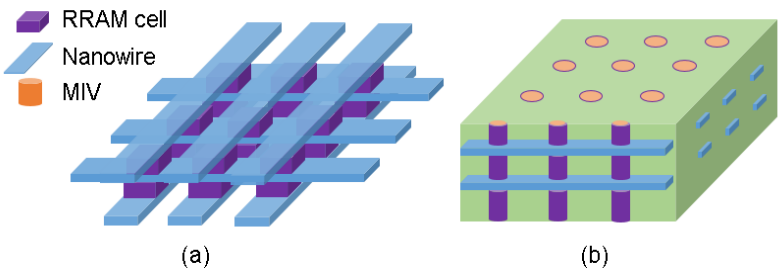

Figure 7: 3D RRAM Integration Structure: (a) Horizontal Structure (b) Vertical Structure

In both structures, the device sizes can be $4F^2/n$, where n is the number of the stacked layers and F is the minimal lithographic feather size dictated by technology node. The number of critical lithography masks for H-RRAM structure increases linearly with increasing the number of the stacked layers while the number of masks for V-RRAM is relatively independent of the stacking number. With increasing the number of the stacked layers, V-RRAM becomes, even more, cost effective[25] as shown in Figure 8.

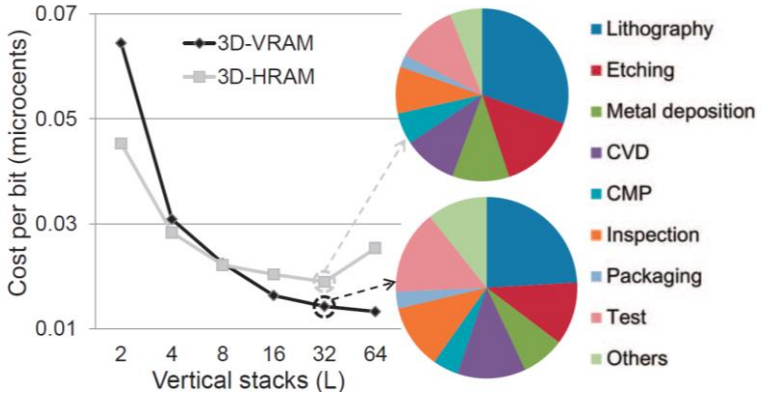

Figure 8: Cost breakdown for per bit projections

Next, there are two 3D integration technologies that can be used for integrating other transistor-based neurons and supportive circuitry to the 3D memristor-based synapse array. One is more traditional, mature and more close to commercialization level: TSV based 3D integration technology[24, 26]. The second one is a more emerging monolithic 3D technology that sequentially fabricates the transistors in a single wafer at the low process temperature. Figure 9 depicts the diagram of these two 3D integration technologies.

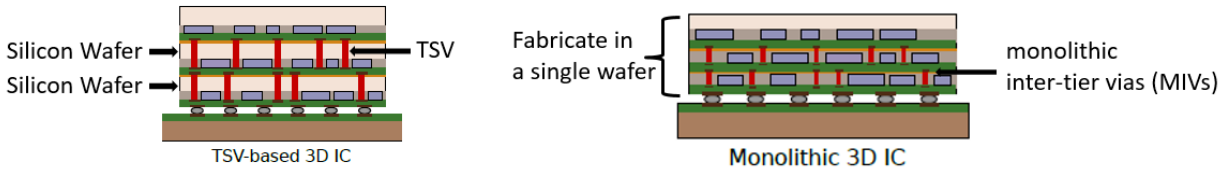

Figure 9: The TSV-based 3D-IC and Monolithic 3D-IC[27]

The main difference between two three-dimensional technologies is the bonding method. For TSV-based 3D technology, the circuitry is fabricated separately at different wafers individually with traditional CMOS technology. After that, the fabricated wafers would be stacked and bonded together through TSVs serving as vertical electrical connections. For the TSV-based 3D

integration technology, power delivery is one of the challenges. As multiple dies are stacked together with small footprints, delivering current to all circuitry located at different vertical layers while meeting the power noise and thermal constraints become more and more challenging. This is mainly caused by the number of TSVs available for power distribution networks is limited.

Table 3: The emerging transistors with low fabrication temperature[28]

3D Device	FinFET	Epi-like Si NWFET	Epi-like Si UTB	SOI-Si UTB	Poly-Si/Ge FinFET	IGZO OSFET
Thermal budget ℃	< 400	< 400	< 400	< 650	< 400	< 500
I_on/I_off	$>10^7$	$>5 \times 10^5$	$>5 \times 10^5$	$>10^7$	$>10^7$	$>10^{21}$

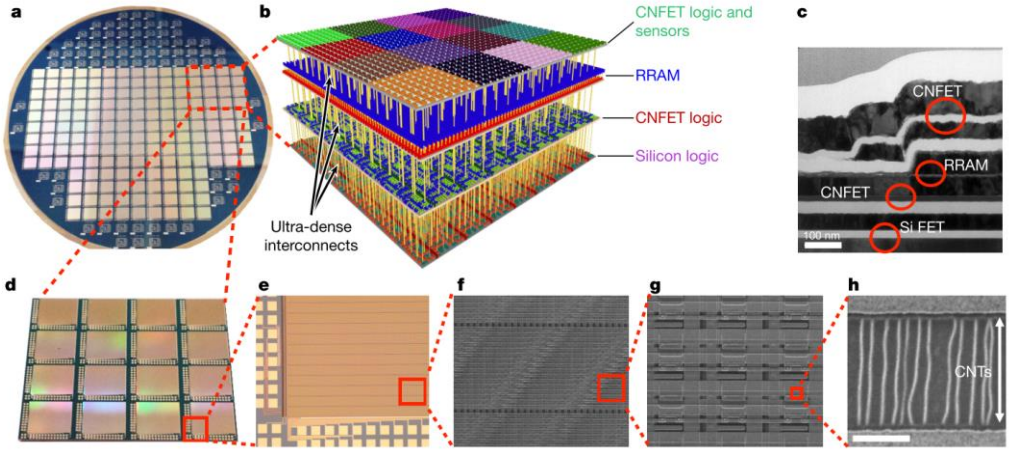

Figure 10: a 3D chip with RRAM, CNFET logics fabricated by Stanford [29].

Another emerging 3D integration technology so-called monolithic 3D integration technology. Unlike the TSV-based 3D technology uses separately fabricate processes, the monolithic 3D technology integrates different layers of devices at a single wafer sequentially with nanoscale inter-tier vias serving as vertical connections. The main challenge for the monolithic 3D integration technology is the low temperature fabrication constraint for the upper layers. In order to protect the former fabricated devices in lower layers, the higher layers need to be fabricated at a lower temperature. This low temperature requirement restricts the traditional CMOS transistor (fabricates at more than 1000℃) does not fit the requirements for the upper layer circuitry implementation. Fortunately, several low temperature transistors are potential candidates to fit this requirement, such as FinFETs [28], Carbon nanotube FETs [30, 31], etc. Table 3 summarizes the state-of-the-art transistors that are fabricated at low temperature and potentially can be

employed in the monolithic 3D integration technology[28]. Furthermore, the functional chip combing monolithic 3D integration technology, memristor, and CNTFETs has been fabricated recently by Stanford, which is demonstrated in Figure 10 [29].

In future, I would mainly focus on integrating the memristor array with the monolithic 3D integration technology with low temperature transistors (CNTFETs). Figure 11 demonstrates the modeling and design framework. I will first use the SPICE (Simulation Program with Integrated Circuit Emphasis) model of the memristor and CNTFETs SPICE model to build an integrate and fire neuron model in a three-dimensional structure. Through these SPICE models, the analysis on noise, computing efficiency, and the device variations would be performed. At last, the modeling and simulations on system levels for real neuromorphic computing applications would be demonstrated.

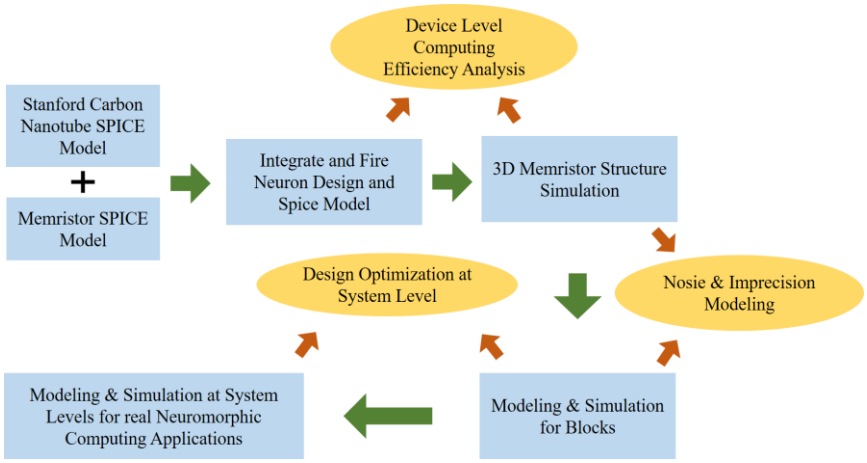

Figure 11: Modeling and Simulation Framework

4 The List of the Publications

In my Ph. D. period, my research mainly focuses on the three-dimensional memristive neuromorphic computing and the emerging biological learning schemes. So far, there are three journal articles and nine conference papers have been published [32-43].

Ref [33, 38] introduce the state-of-the-art three-dimensional memristive neuromorphic computing hardware research trends and challenges. Ref [32, 34, 35, 44, 45] mainly discuss the three-dimensional memristive neuromorphic architecture electrical characteristics at large-scale. The simulations results are obtained by the SPICE models of memristor and the neurons. The signal integrity and signal attenuation issues for a large-scale memristive electronic synaptic array is analyzed and discussed. Ref [36] mainly introduces a novel neuromorphic computing application on the cloud robotics. Furthermore, Ref [37] analyzes a biologically learning schemes

named as associative memory at the cellular level. Ref [40, 41] explore the TSV-based 3D neuromorphic computing architecture, which implements the capacitance form by the TSVs as the capacitor in a neuron model. Ref [42, 43] mainly investigate several hardware implementations and designs on the temporal encoding based neurons. Ref [39] discusses a 3D memristive-based adjustable deep recurrent neural network topology with a programmable attention mechanism.

5 References

[1] R. Courtland, "Transistors could stop shrinking in 2021," *IEEE Spectrum,* vol. 53, no. 9, pp. 9-11, 2016.

[2] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. Hudspeth, *Principles of neural science*. McGraw-hill New York, 2000.

[3] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE,* vol. 78, no. 10, pp. 1629-1636, 1990.

[4] L. F. Abbott, "Lapicque's introduction of the integrate-and-fire model neuron (1907)," *Brain research bulletin,* vol. 50, no. 5, pp. 303-304, 1999.

[5] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology,* vol. 117, no. 4, p. 500, 1952.

[6] R. B. Stein, "A theoretical analysis of neuronal variability," *Biophysical Journal,* vol. 5, no. 2, p. 173, 1965.

[7] L. Chua, "Memristor, Hodgkin-Huxley, and edge of chaos," in *Memristor Networks*: Springer, 2014, pp. 67-94.

[8] H. De Garis, C. Shuo, B. Goertzel, and L. Ruiting, "A world survey of artificial brain projects, Part I: Large-scale brain simulations," *Neurocomputing,* vol. 74, no. 1, pp. 3-29, 2010.

[9] F. Akopyan *et al.*, "True North: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip," (in English), *Ieee Transactions on Computer-Aided Design of Integrated Circuits and Systems,* Article vol. 34, no. 10, pp. 1537-1557, Oct 2015.

[10] F. Walter, F. Röhrbein, and A. Knoll, "Neuromorphic implementations of neurobiological learning algorithms for spiking neural networks," *Neural Networks,* vol. 72, pp. 152-167, 2015.

[11] N. Qiao *et al.*, "A Re-configurable On-line Learning Spiking Neuromorphic Processor comprising 256 neurons and 128K synapses," (in English), *Frontiers in Neuroscience,* Original Research vol. 9, 2015-April-29 2015.

[12] B. Benjamin *et al.*, "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," (in English), *Proceedings of the Ieee,* vol. 102, no. 5, pp. 699-716, May 2014.

[13] S. B. Furber *et al.*, "Overview of the SpiNNaker system architecture," *Computers, IEEE Transactions on,* vol. 62, no. 12, pp. 2454-2467, 2013.

[14] E. Painkras *et al.*, "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE Journal of Solid-State Circuits,* vol. 48, no. 8, pp. 1943-1953, 2013.

[15] D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, "Cognitive computing," *Communications of the ACM,* vol. 54, no. 8, pp. 62-71, 2011.

[16] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *nature,* vol. 453, no. 7191, pp. 80-83, 2008.

[17] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano letters,* vol. 10, no. 4, pp. 1297-1301, 2010.

[18] G. Stefanovich, A. Pergament, and D. Stefanovich, "Electrical switching and Mott transition in VO2," *Journal of Physics: Condensed Matter,* vol. 12, no. 41, p. 8837, 2000.

[19] J. Honig and T. Reed, "Electrical properties of Ti 2 O 3 single crystals," *Physical Review,* vol. 174, no. 3, p. 1020, 1968.

[20] H. S. P. Wong *et al.*, "Metal-oxide RRAM," *Proceedings of the IEEE,* vol. 100, pp. 1951-1970, 2012.

[21] Y. Yi *et al.*, "FPGA based spike-time dependent encoder and reservoir design in neuromorphic computing processors," *Microprocessors and Microsystems,* vol. 46, pp. 175-183, 2016.

[22] H. An, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical Modeling and Analysis of 3D Neuromorphic IC with Monolithic Inter-tier Vias."

[23] Y. Yi, P. Li, V. Sarin, and W. Shi, "Impedance extraction for 3-D structures with multiple dielectrics using preconditioned boundary element method," in *2007 IEEE/ACM International Conference on Computer-Aided Design*, 2007, pp. 7-10: IEEE.

[24] Y. Yi, P. Li, V. Sarin, and W. Shi, "A preconditioned hierarchical algorithm for impedance extraction of three-dimensional structures with multiple dielectrics," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 27, no. 11, pp. 1918-1927, 2008.

[25] C. Xu, D. Niu, S. Yu, and Y. Xie, "Modeling and design analysis of 3D vertical resistive memory—A low cost cross-point architecture," in *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2014, pp. 825-830: IEEE.

[26] Y. Yi, Y. Zhou, X. Fu, and F. Shen, "Modeling differential through-silicon-vias (TSVs) with voltage dependent and nonlinear capacitance," *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Microelectronics (JSAM),* vol. 3, no. 6, pp. 234-241, 2013.

[27] J. Knechtel and J. Lienig, "Physical Design Automation for 3D Chip Stacks," pp. 3-10, 2016.

[28] C.-C. Yang *et al.*, "Footprint-efficient and power-saving monolithic IoT 3D+ IC constructed by BEOL-compatible sub-10nm high aspect ratio (AR>7) single-grained Si FinFETs with record high Ion of 0.38 mA/μm and steep-swing of 65 mV/dec. and I<inf>on</inf>/I<inf>off</inf> ratio of 8," pp. 9.1.1-9.1.4, 2016.

[29] M. M. Shulaker *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature,* vol. 547, no. 7661, pp. 74-78, Jul 5 2017.

[30] M. M. Shulaker *et al.*, "Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs," in *Electron Devices Meeting (IEDM), 2014 IEEE International*, 2014, pp. 27.4. 1-27.4. 4: IEEE.

[31] M. M. Shulaker *et al.*, "Three-dimensional integration of nanotechnologies for computing and data storage on a single chip," *Nature,* vol. 547, no. 7661, pp. 74-78, 2017.

[32] H. An, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D Neuromorphic IC with Monolithic Inter-tier Vias," in *Electrical Performance Of Electronic Packaging And Systems (EPEPS), 2016 IEEE 25th Conference on*, 2016, pp. 87-90: IEEE.

[33] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi, "Design challenges and methodologies in 3D integration for neuromorphic computing systems," in *Quality Electronic Design (ISQED), 2016 17th International Symposium on*, 2016, pp. 24-28: IEEE.

[34] H. An, M. A. Ehsan, Z. Zhou, F. Shen, and Y. Yi, "Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons," *Integration, the VLSI Journal,* 2017.

[35] H. An, M. A. Ehsan, Z. Zhou, and Y. Yi, "Electrical modeling and analysis of 3D synaptic array using vertical RRAM structure," in *Quality Electronic Design (ISQED), 2017 18th International Symposium on*, 2017, pp. 1-6: IEEE.

[36] H. An, J. Li, Y. Li, X. Fu, and Y. Yi, "Three dimensional memristor-based neuromorphic computing system and its application to cloud robotics," *Computers & Electrical Engineering,* vol. 63, pp. 99-113, 2017.

[37] H. An, Z. Zhou, and Y. Yi, "Memristor-based 3D neuromorphic computing system and its application to associative memory learning," in *2017 IEEE 17th International Conference on Nanotechnology (IEEE-NANO)*, 2017, pp. 555-560.

[38] H. An, Z. Zhou, and Y. Yi, "Opportunities and challenges on nanoscale 3D neuromorphic computing system," in *Electromagnetic Compatibility & Signal/Power Integrity (EMCSI), 2017 IEEE International Symposium on*, 2017, pp. 416-421: IEEE.

[39] H. An, Z. Zhou, and Y. Yi, "3D memristor-based adjustable deep recurrent neural network with programmable attention mechanism," *Proceedings of Neuromorphic Computing Symposium,* pp. 1-6, July 17–19, 2017 2017.

[40] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi, "Adaptation of Enhanced TSV Capacitance as Membrane Property in 3D Brain-inspired Computing System," in *Proceedings of the 54th Annual Design Automation Conference 2017*, 2017, p. 86: ACM.

[41] M. A. Ehsan, H. An, Z. Zhou, and Y. Yi, "A Novel Approach for using TSVs as Membrane Capacitance in Neuromorphic 3D IC," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* 2017.

[42] C. Zhao, J. Li, H. An, and Y. Yi, "Energy efficient analog spiking temporal encoder with verification and recovery scheme for neuromorphic computing systems," in *Quality Electronic Design (ISQED), 2017 18th International Symposium on*, 2017, pp. 138-143: IEEE.

[43] C. Zhao, J. li, H. An, and Y. Yi, "When Energy Efficient Spike-Based Temporal Encoding Meets Resistive Crossbar: From Circuit Design to Application," *Proceedings of Neuromorphic Computing Symposium,* pp. 1-6, July 17-19, 2017.

[44] H. An, M. S. Al-Mamun, M. K. Orlowski, and Y. Yi, "Learning Accuracy Analysis of Memristor-based Nonlinear Computing Module on Long Short-term Memory," in *Proceedings of the International Conference on Neuromorphic Systems*, 2018, p. 5: ACM.

[45] H. An, K. Bai, and Y. Yi, "The Roadmap to Realize Memristive Three-Dimensional Neuromorphic Computing System," in *Advances in Memristor Neural Networks-Modeling and Applications*: IntechOpen, 2018.