Realizing Memristive Three-dimensional Neuromorphic Computing System

Hongyu An

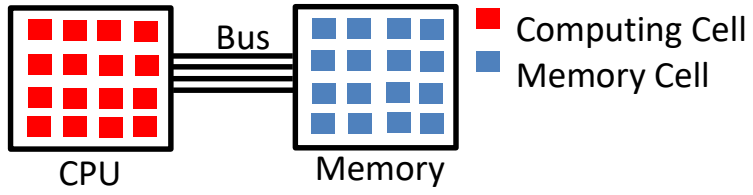
Dept. of Electrical and Computer Eng.
Virginia Tech

April 26, 2018

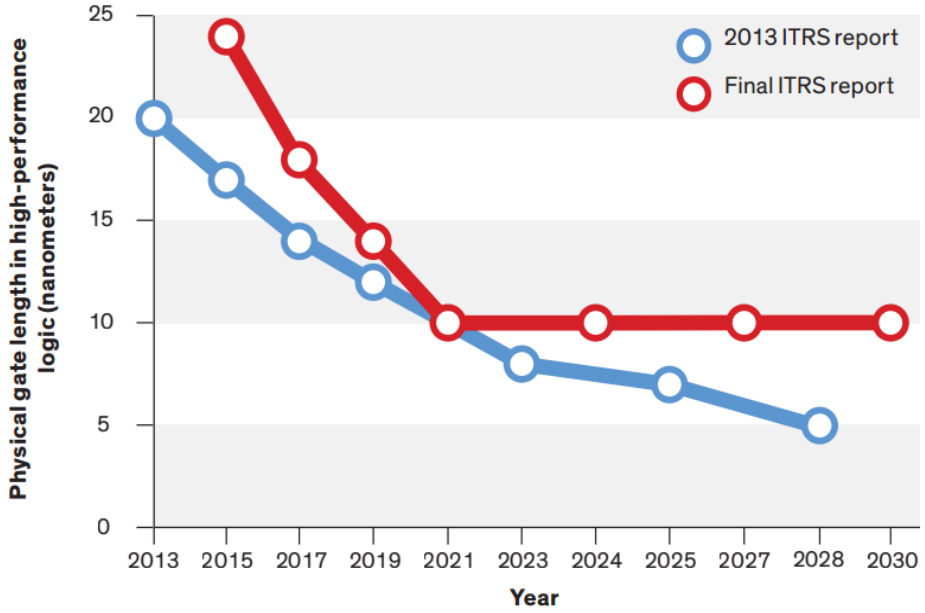
Outline

- Motivation
- Neuromorphic Computing
  - Neuromorphic Chips
- Memristive Three-dimensional Neuromorphic Computing System
  - Three-dimensional Integration technology;
  - Memristor as synapse;
  - Memristive Three-dimensional Neuromorphic Computing System
    - Objective
    - Methodology
    - Expected outcomes
- My Publications
- References

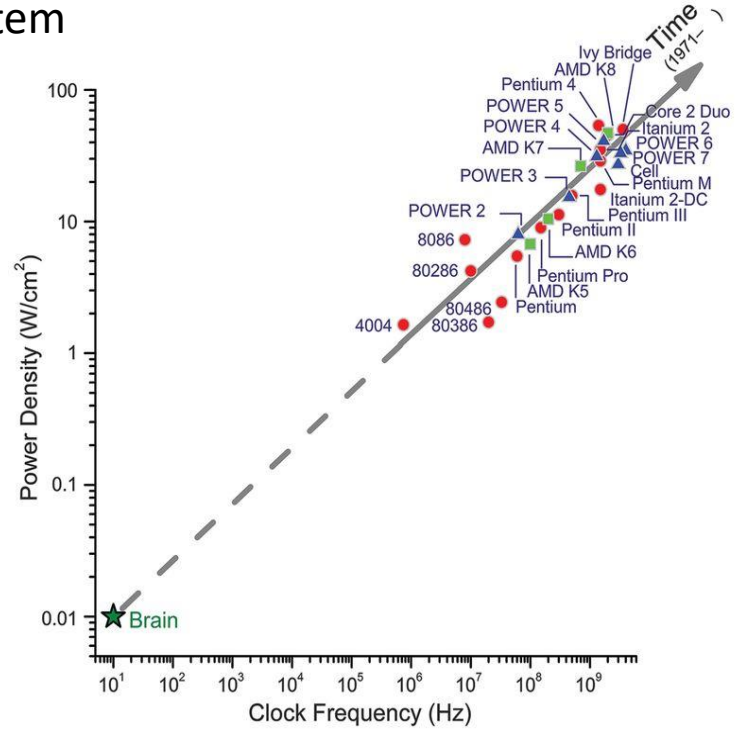
Physical Challenges of Traditional Von Neumann Computer



Von Neumann Computing System



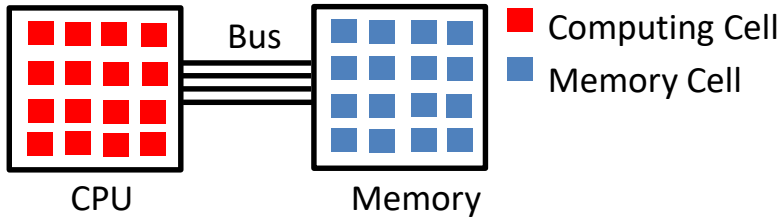
- **Transistor size shrinking challenges at 10 nm scale in following several years (ITRS)**



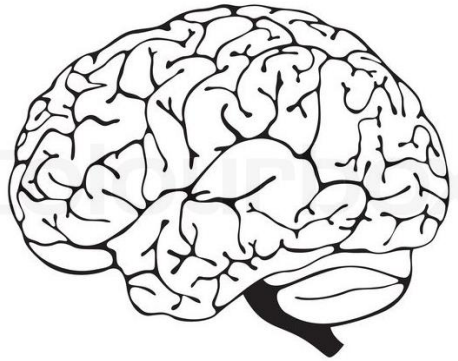
- **High Power and frequency Challenges**

P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, *et al.*, *"A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, pp. 668-673, 2014.*

Intelligent Challenges of Traditional Von Neumann Computer



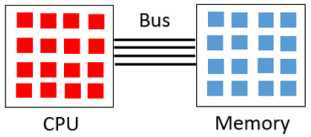
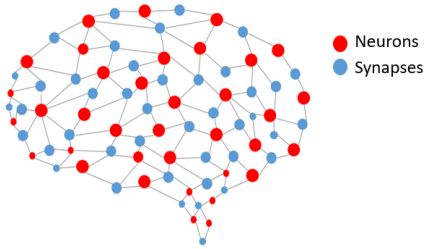
Von Neumann Computing System



- High Adaptivity with dynamic surrounding environment
- Spontaneous and independent Learning
- Perception
- Cognition

- High Adaptivity with dynamic surrounding environment
- Spontaneous and independent Learning
- Perception
- Cognition
- Low power consumption (~ 20W)
- High computing efficiency;

Brain VS. Computer

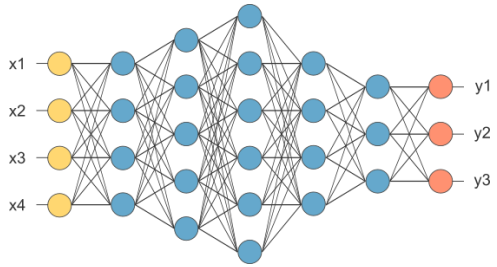
Computer		Human Brian
Binary Signals	Encoding Scheme	Spiking Signals
Von Neumann Architecture 	Architecture	 Brain Computing System
CPUs(Logic Gates, etc.), Memory(SRAM, etc.)	Devices	Biological Neurons & Synapses

Neuromorphic Computing

Neuromorphic Computing

Human Brian

Spiking Signals



Artificial Neurons and Synapses

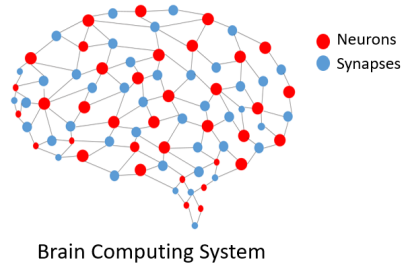
Reverse Engineering

Encoding Scheme

Architecture

Devices

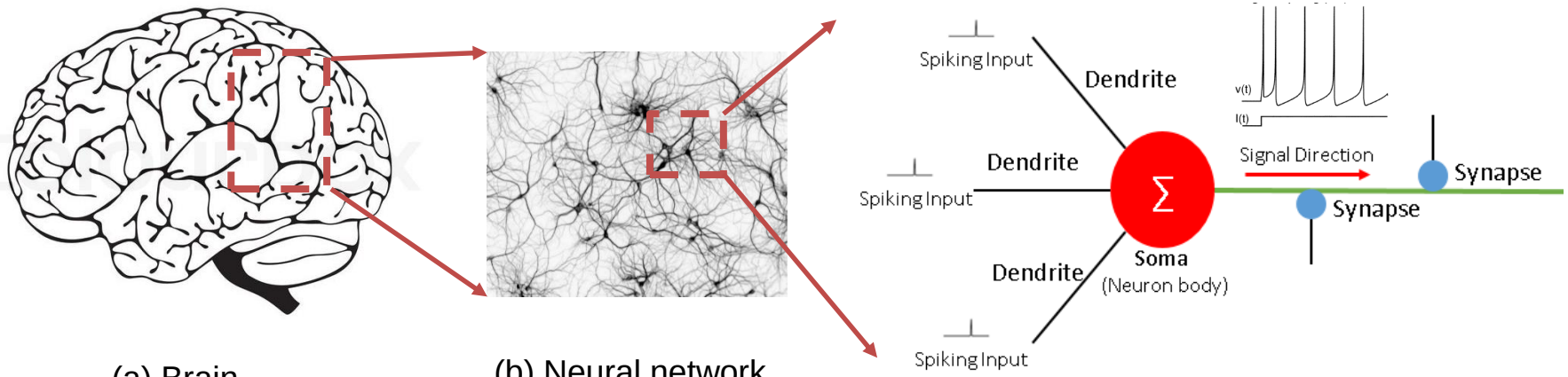
Spiking Signals



Brain Computing System

Biological Neurons & Synapses

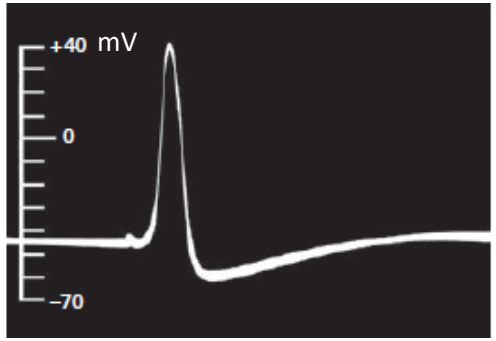
Functionalities of Neurons and Synapses



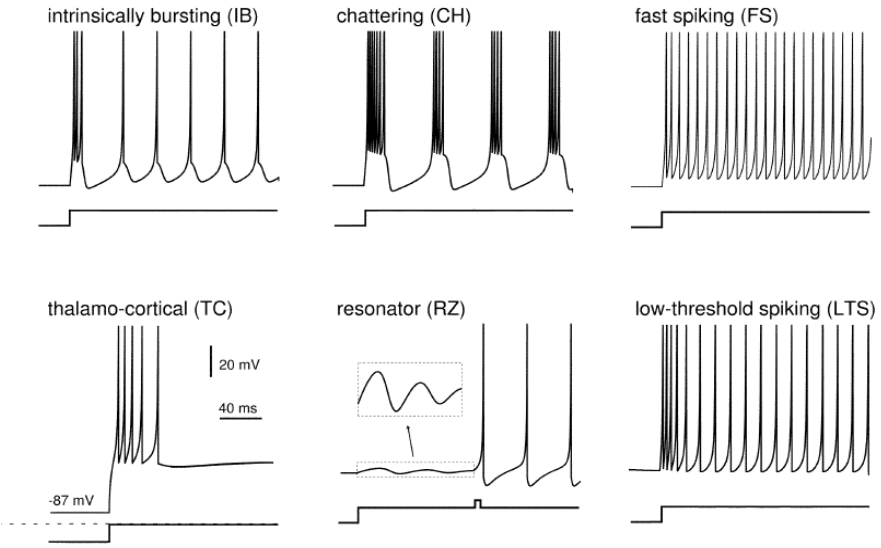
(a) Brain (b) Neural network (c) Neuron structure

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

Spiking Signals



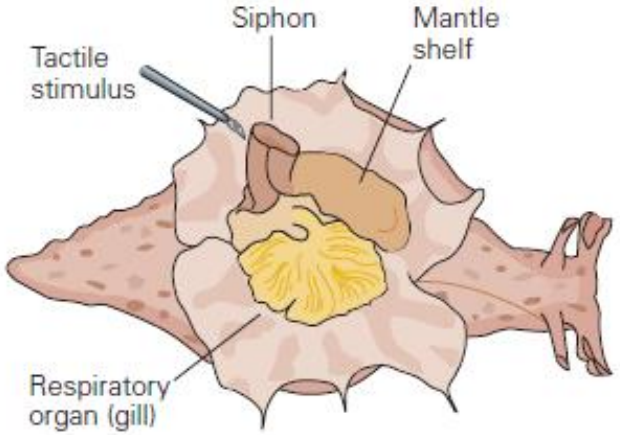
This historic tracing is the first published intracellular recording of an action potential. It was recorded in by Hodgkin and Huxley (captured through a probe attached on the axon)



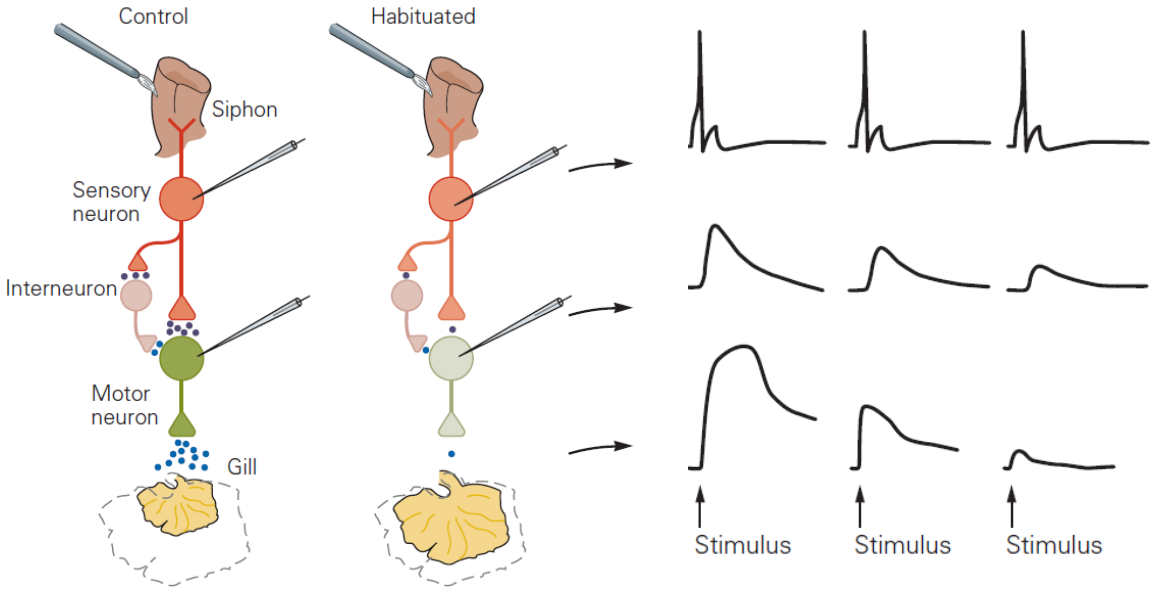
http://www.izhikevich.org/

- **Rate coding**
  - The rate coding model of neuronal firing communication states that as the intensity of a stimulus increases, the frequency or rate of action potentials, or "spike firing", increases;
- **Spike-count coding**
  - The information is encoded by the number of spikes that appear during a time period

Synapse & Memory

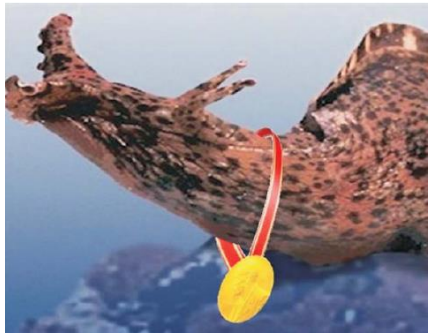


Sea Slug



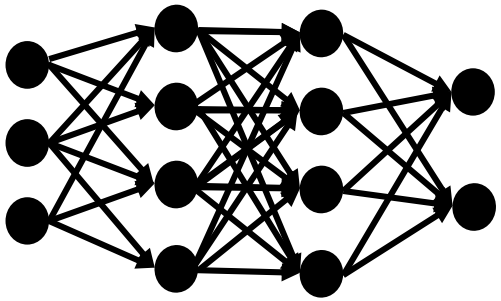
Synapse: acts as a memory organ in brain. It connects axon of last neuron to dendrite of next neuron. The connectivity strength can be modified by spiking signal stimulus. The connectivity strength of synapse is defined as weight of synapse.

E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. Hudspeth, *Principles of neural science vol. 4: McGraw-hill New York, 2000.*

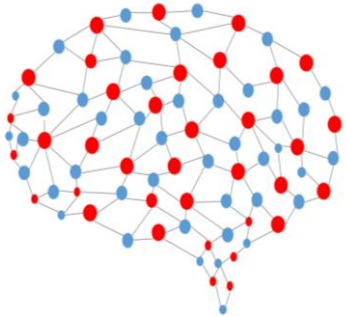


Dr. Eric Kandel was awarded the Nobel Prize in medicine and physiology in 2000 for uncovering the *molecular basis of memory*.

Neural Network Topologies

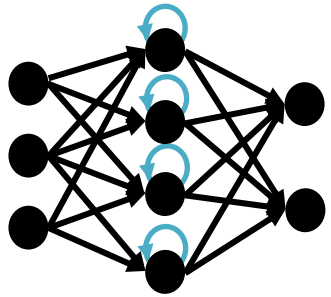


Deep Neural Networks

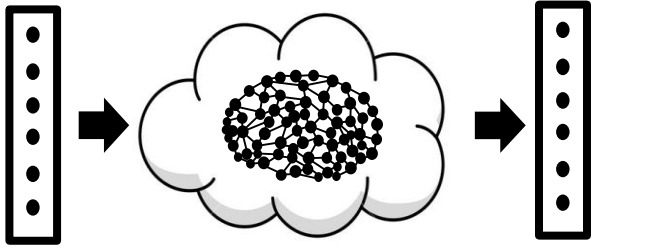


Neurons
Synapses

Time Lag τ



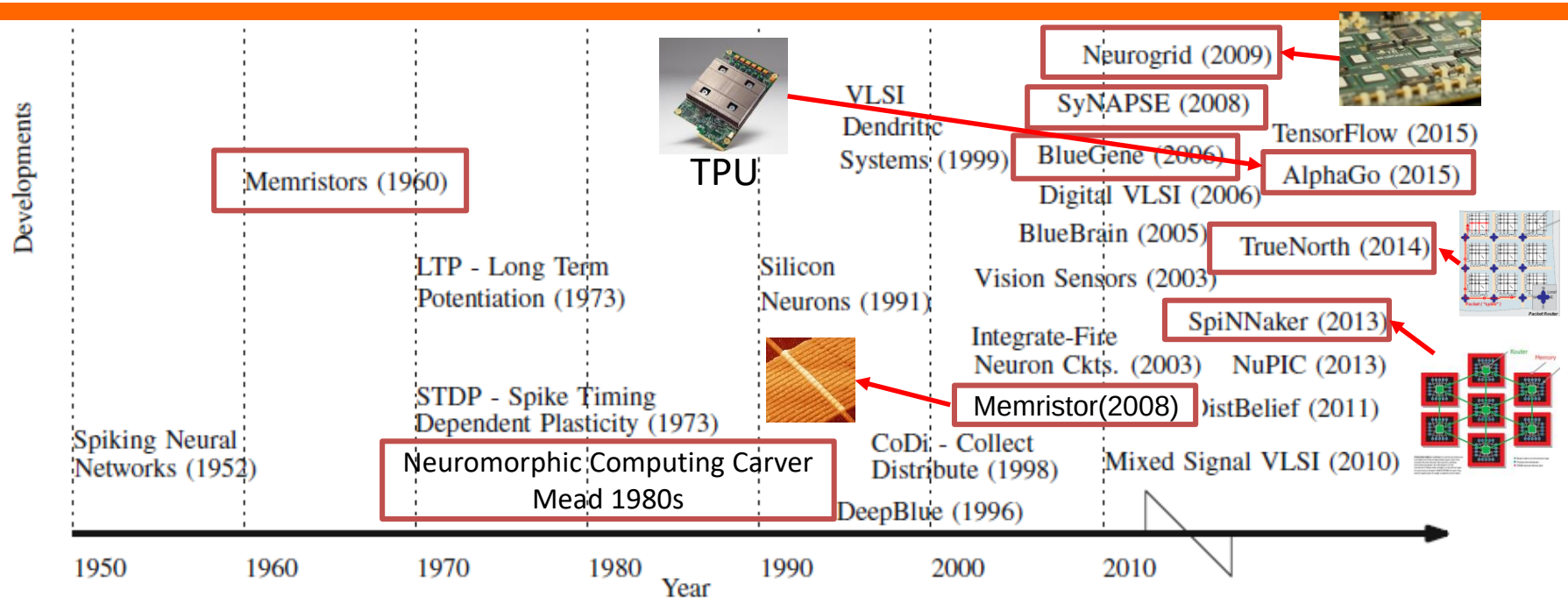
Recurrent Neural Networks



Input Layer
Reservoir
Output Layer

Reservoir Networks

Neuromorphic Platforms and Chips

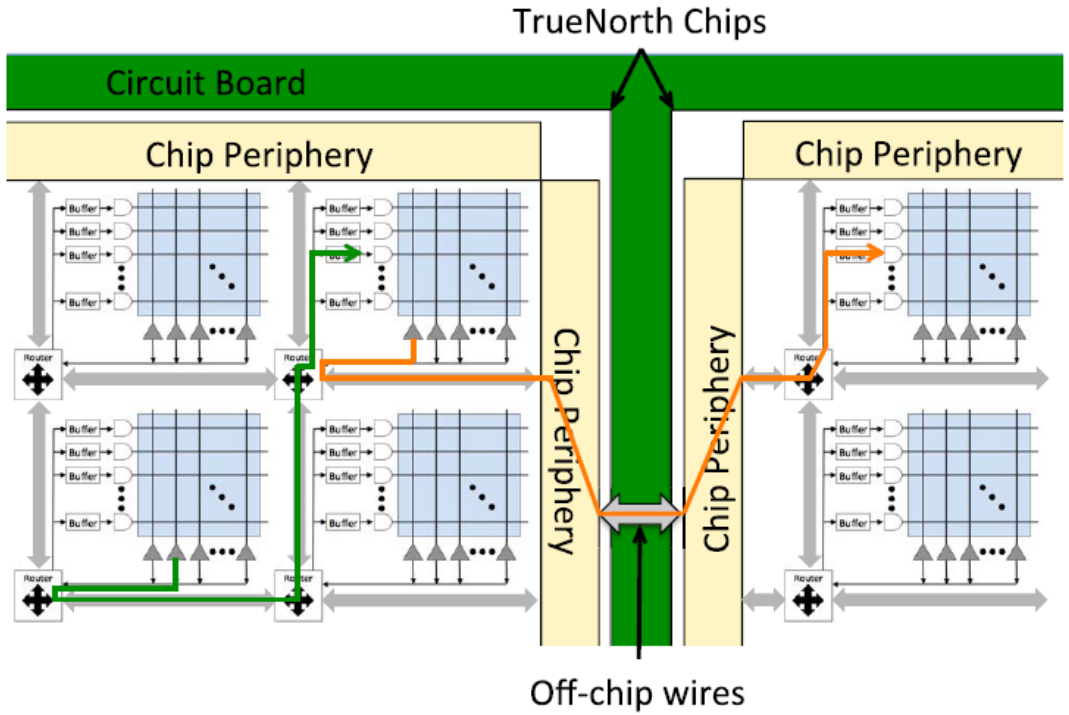


Soman, Sumit, and Manan Suri. "Recent trends in neuromorphic engineering." *Big Data Analytics* 1, no. 1 (2016): 15.

	TrueNorth [1]	Neurogrid [2]	SpiNNaker [3]	Human Brain
Neurons	1,048,576	65,535	20,833	20 Billions
Synapses	256 millions	N/A	20,833,333	200 Trillions
Area/Volume	430 mm^2	168 mm^2	102 mm^2	1130 cm^3
Power Density	0.15 mW/mm^2	18 mW/mm^2	0.012 mW/mm^2	0.0177 mW/mm^3

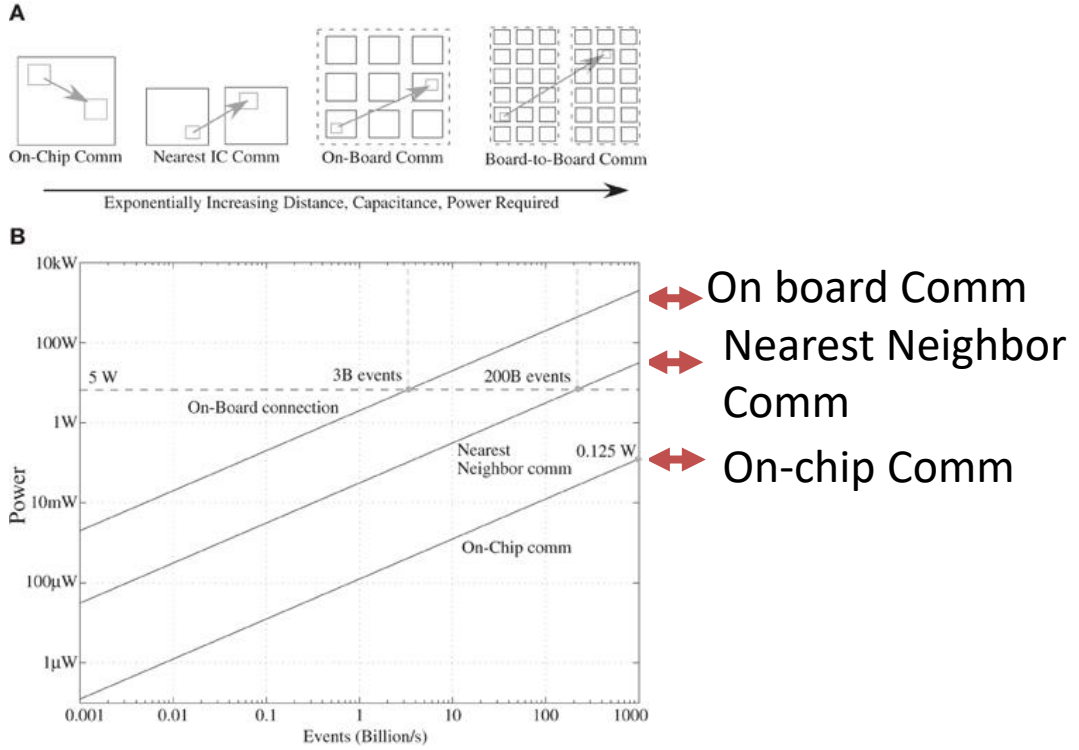
Limits of State-of-art Brain-inspired Chips

- 2D flat routing/placement method
- Large size CMOS based synapse design (RAM)



TrueNorth Spiking Signal Propagation Scheme [1]

- Long signal propagation distance leads large power consumption on signal delivery
- Large chip design area

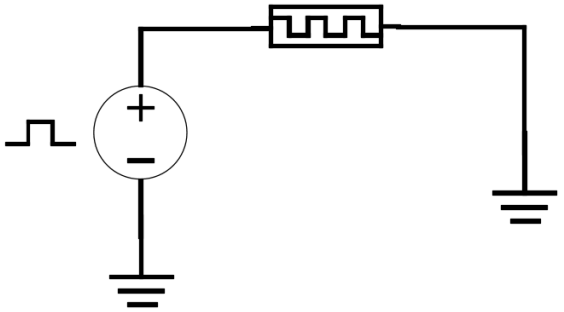


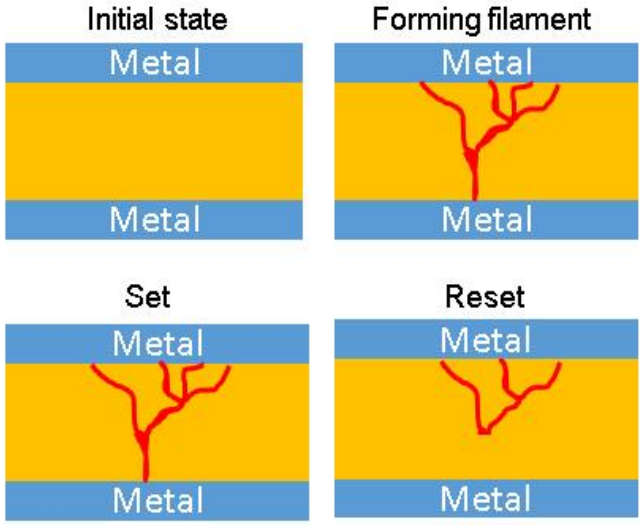
Power consumption increases with event number and signal propagation distance [4]

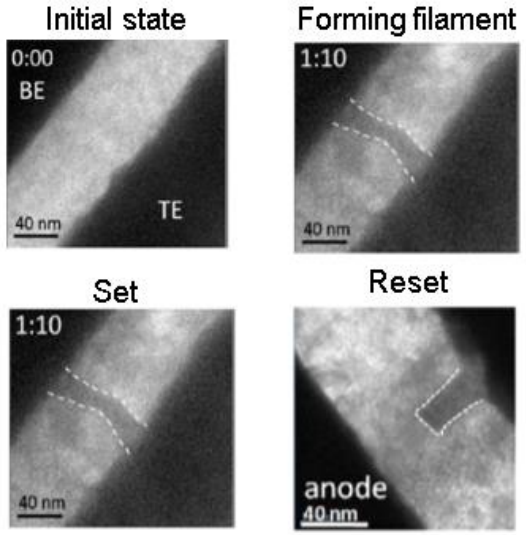
Memristive Three-dimensional Neuromorphic Computing System

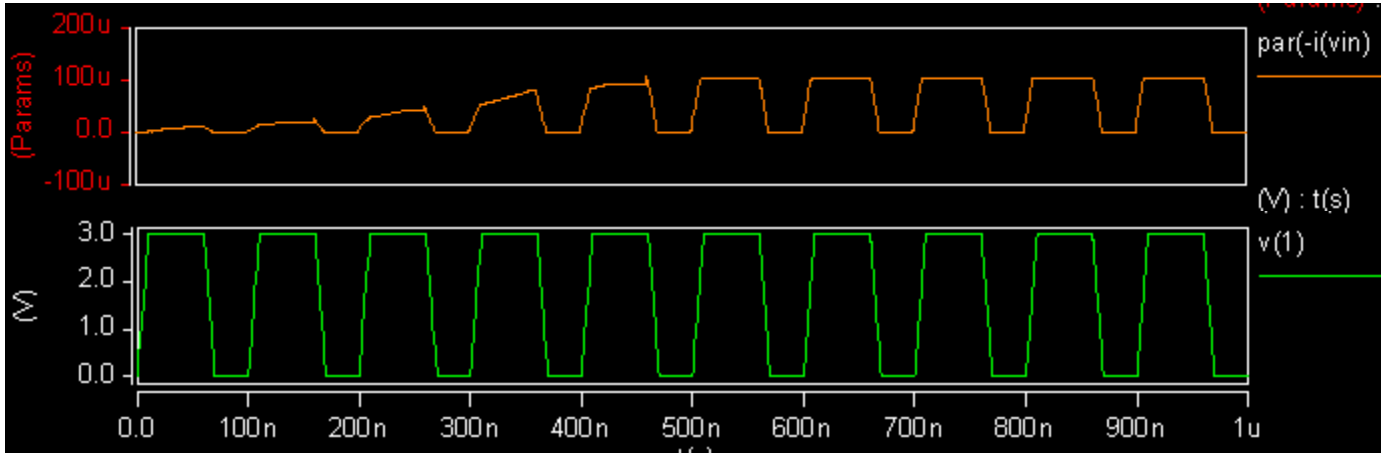
- Objectives
  - Reduce the signal transfer distance, consequently decrease the power consumption;
  - Reduce the die area by stacking the neuron and synapse circuitries vertically;
  - Memristors can reduce the size of synapse to nanoscale;
  - Reduce the wire length;
  - Increase the interconnection density;

Memristor as Synapse



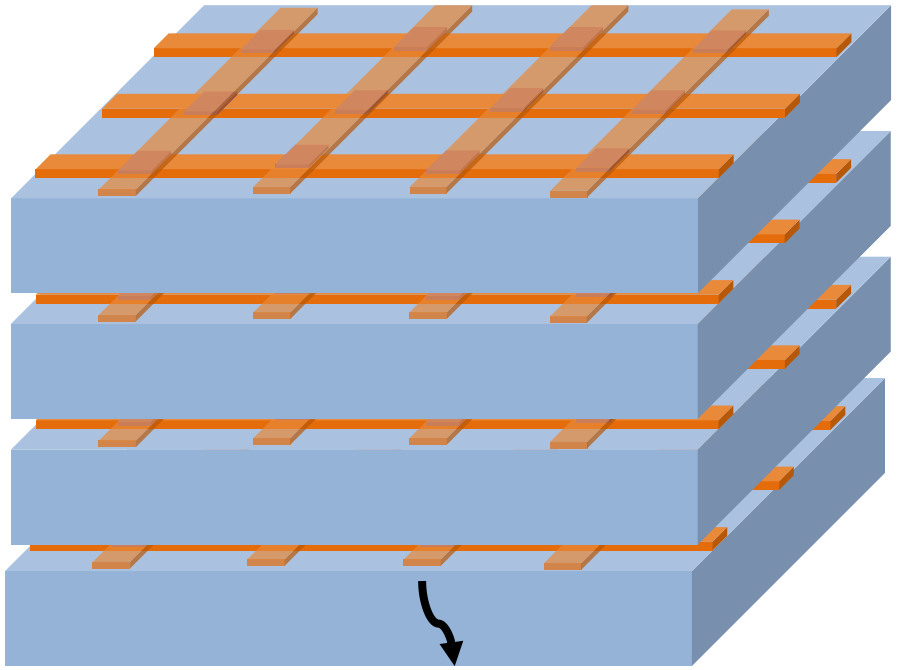






TEM image: J.-Y. Chen, C.-L. Hsin, C.-W. Huang, C.-H. Chiu, Y.-T. Huang, S.-J. Lin*, et al.*, "Dynamic evolution of conducting nanofilament in resistive switching memories," *Nano letters,* vol. 13, pp. 3671-3677, 2013.

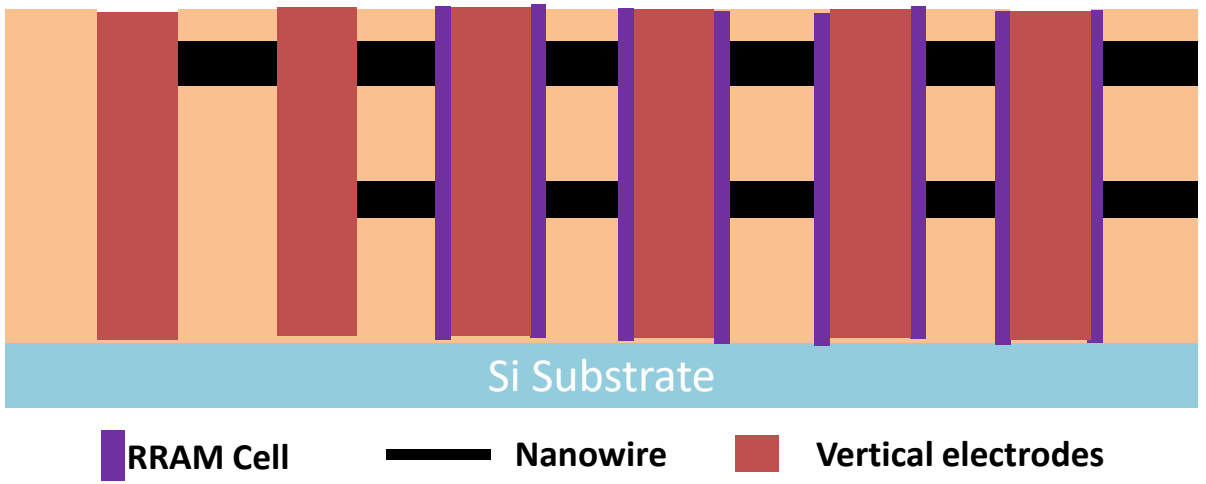
3D Memristor-based Synapse



Memristor Crossbar Structure

Horizontal RRAM (Resistive RAM) Structure

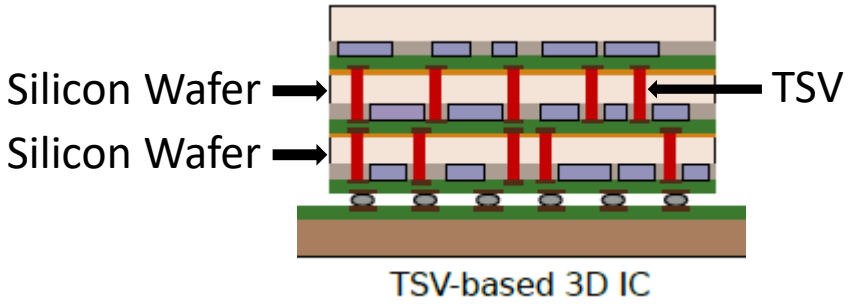
Vertical Horizontal RRAM (Resistive RAM) Structure (Side View)



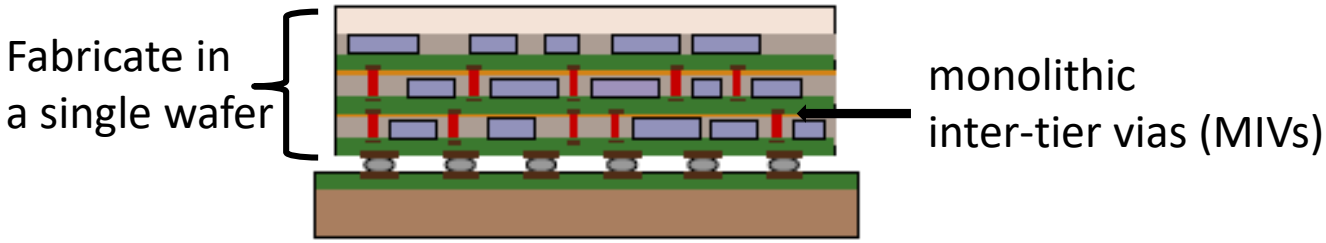
	SRAM	DRAM	NOR	NAND	RRAM
Cell area	>100F^2	6F^2	10F^2	<4F^2 (3D)	4F^2
Voltage	<1V	<1V	>10V	>10V	<3V
Write Energy(J/bit)	~Fj	~10fJ	~100 pJ	~ 10 fJ	~0.1 pJ

Comparison between RRAM with other Memory Technology [5]

3D Integration Technologies



TSV-based 3D IC



Monolithic 3D IC

TSV based 3D-IC Integration:

- Dies fabricated separately;
- Wafer thinned;
- Wafer aligned and bonded;

Monolithic 3D integration technology:

- fabricates two or more tiers of devices sequentially;
- No aligning and bonding procedure;
- No wafer thinning procedure;
- Monolithic inter-tier vias (MIVs) are at nanoscale level;
- Fabrication compatible with RRAM (Resistive RAM) array;

Challenges for Monolithic 3D Integration

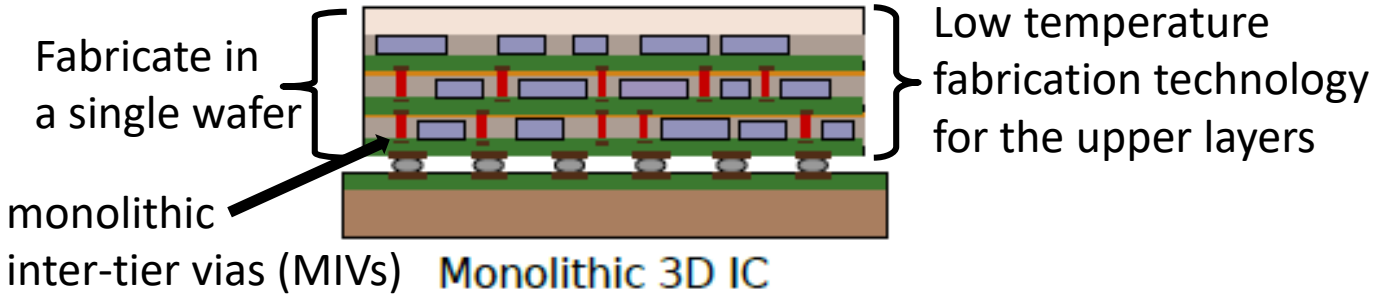


Table 3: The emerging transistors with low fabrication temperature [8]

Devices	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	$>10^7$	$>5 \times 10^5$	$>5 \times 10^5$	$>10^7$	$>10^7$	$>10^{21}$

IGZO: In-GA-Zn-O;
OSFET: Oxide semiconductor FET;

NWFET: Gate first nanowire FET;
UTB: ultra thin body;
SOI: Silicon on insulator;

CNFETs + Monolithic 3D Integration

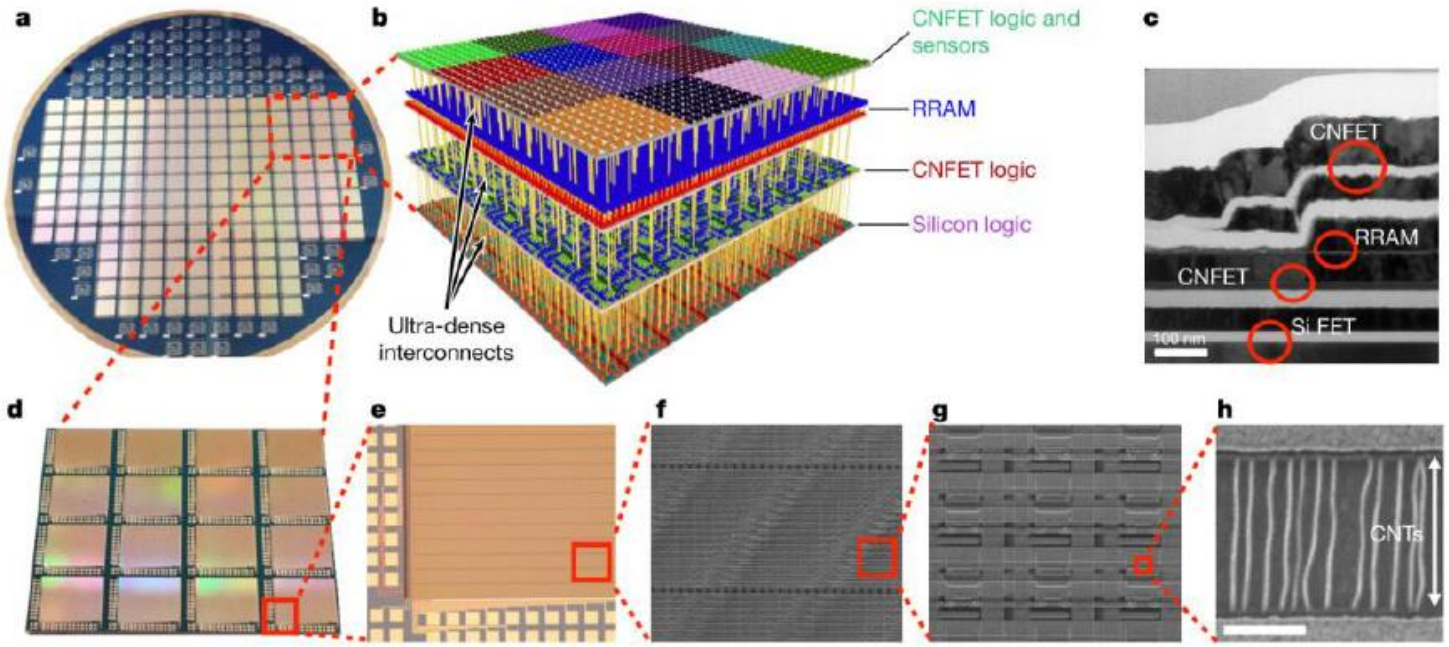
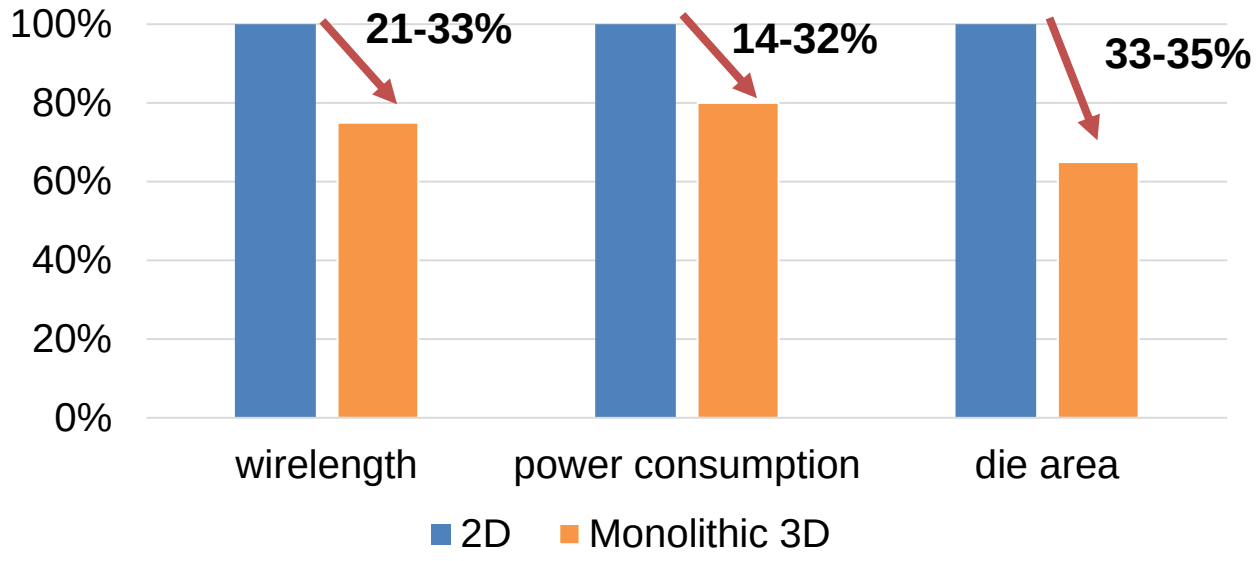


Figure 10: 3D chip with RRAM, CNFET logics fabricated by Stanford [28].

Stanford Carbon Nanotube FETs SPICE Model

Memristor SPICE Model

Comparison between 2D to 3D Monolithic Integration



Comparison between 2D to 3D Monolithic Integration on wirelength, power consumption and die area [6,7]. (45nm technology; Benchmarks: FPU;AES;DES; LDPC; M256)

FPU: a double precision floating point unit.
AES & DES: encryption engines.
LDPC: a low-density parity-check engine for the IEEE 802.3 standard.
M256: a simple partial-sum-add-based 256bit integer multiplier.

Methodologies

Stanford Carbon Nanotube FETs SPICE Model

Memristor SPICE Model

Neuron Circuit

Supportive Circuitry

Synapse

2D Memristor Structure Model

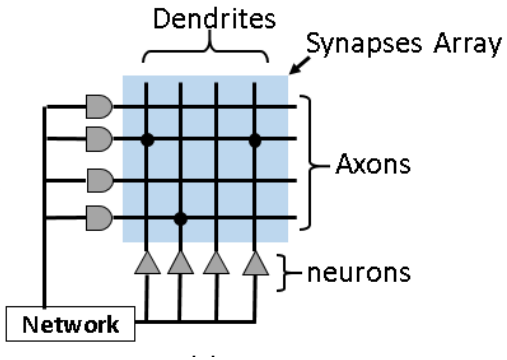
2D Neuron Circuit

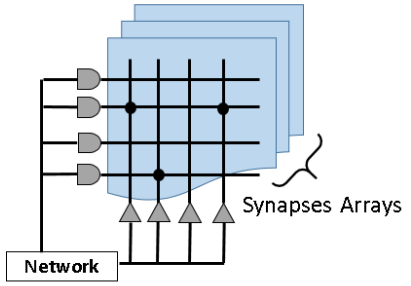
2D Supportive Circuitry

3D Memristor Structure Model

3D Neuron Circuit

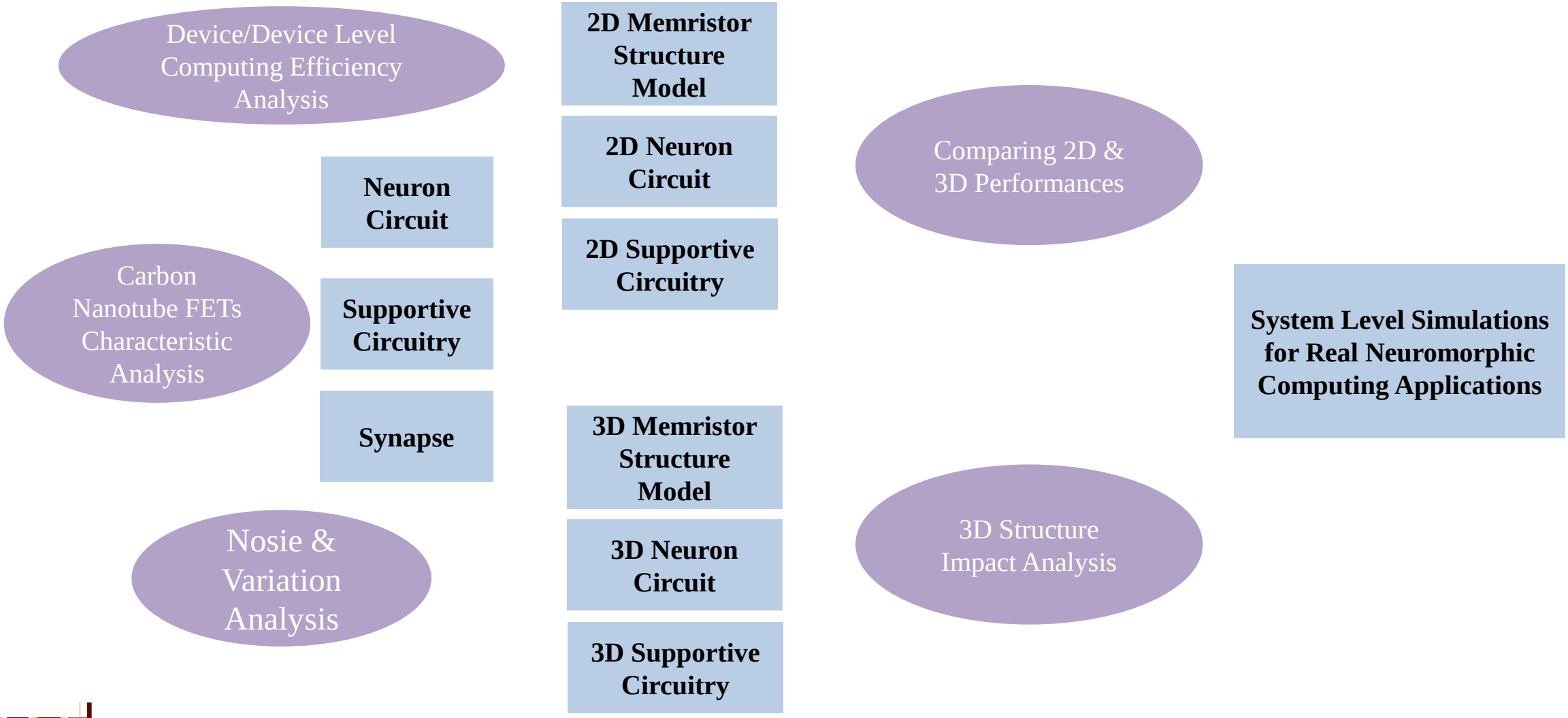
3D Supportive Circuitry





System Level Simulations for Real Neuromorphic Computing Applications

Analysis


Device/Device Level Computing Efficiency Analysis
Neuron Circuit
Carbon Nanotube FETs Characteristic Analysis
Supportive Circuitry
Synapse
Nosie & Variation Analysis
2D Memristor Structure Model
2D Neuron Circuit
2D Supportive Circuitry
3D Memristor Structure Model
3D Neuron Circuit
3D Supportive Circuitry
Comparing 2D & 3D Performances
3D Structure Impact Analysis
System Level Simulations for Real Neuromorphic Computing Applications

Expected Outputs

- 3D circuit level SPICE model of memristor-based synapse (V-RRAM);
- The neuron circuit design by using traditional CMOS technology and carbon Nanotube FETs;
- The supportive circuit design by using traditional CMOS technology and emerging carbon Nanotube FETs;
- The system level simulation and analysis on the power consumption, computing efficiency, and design area reduction, etc.

My Publications

1. 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 (TCAD), 2017
2. **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.
3. **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.
4. 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 (DAC) 2017, 2017, p. 86.
5. **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. (Best Paper Nomination)
6. 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.
7. **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.
8. **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.
9. **H. An**, Z. Zhou, and Y. Yi, “3D Memristor-based Adjustable Deep Recurrent Neural Network with Programmable Attention Mechanism,” in Proceedings of Neuromorphic Computing Symposium (NCS), 2017.
10. C. Zhao, J. Li, **H. An**, and Y. Yi, “When Energy Efficient Spike-Based Temporal Encoding Meets Resistive Crossbar: From Circuit Design to Application,” *in Proceedings of Neuromorphic Computing Symposium, 2017.*
11. **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.*
12. 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.*

References

[1] Akopyan, Filipp, et al. "TrueNorth: Design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 34.10 (2015): 1537-1557.

[2] B. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A. R. Chandrasekaran, J. M. Bussat, et al., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," Proceedings of the IEEE, vol. 102, pp. 699-716, May 2014.

[3] E. Painkras, L. A. Plana, J. Garside, S. Temple, F. Galluppi, C. Patterson, 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, pp. 1943-1953, 2013.

[4] Hasler, Jennifer, and Harry Bo Marr. "Finding a roadmap to achieve large neuromorphic hardware systems."

[5] Yu, Shimeng, and Pai-Yu Chen. "Emerging memory technologies: recent trends and prospects."

[6] Liu, Chang, and Sung Kyu Lim. "A design tradeoff study with monolithic 3D integration."

[7] Y.-J. Lee, D. Limbrick, and S. K. Lim, "Power benefit study for ultrahigh density transistor-level monolithic 3D ICs,"

[8] C.-C. Yang, J.-M. Shieh, T.-Y. Hsieh, W.-H. Huang, H.-H. Wang, C.-H. Shen, 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" pp. 9.1.1-9.1.4, 2016.


Thank you