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

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

- **High Power and frequency Challenges**

Intelligent Challenges of Traditional Von Neumann Computer

- 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**

- **Binary Signals**
- **Von Neumann Architecture**
  - CPUs (Logic Gates, etc.), Memory (SRAM, etc.)

**Human Brain**

- **Encoding Scheme**
- **Architecture**
- **Devices**
- **Spiking Signals**
- **Biological Neurons & Synapses**
Neuromorphic Computing

Artificial Neurons and Synapses

Spiking Signals

Reverse Engineering

Encoding Scheme

Architecture

Devices

Biological Neurons & Synapses
Functionalities of Neurons and Synapses

- **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)

- **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: 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.


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

Recurrent Neural Networks

Reservoir Networks

- Neurons
- Synapses

Input Layer → Reservoir → Output Layer

Time Lag $\tau$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurons</td>
<td>1,048,576</td>
<td>65,535</td>
<td>20,833</td>
<td>20 Billions</td>
</tr>
<tr>
<td>Synapses</td>
<td>256 millions</td>
<td>N/A</td>
<td>20,833,333</td>
<td>200 Trillions</td>
</tr>
<tr>
<td>Area/Volume</td>
<td>430 mm$^2$</td>
<td>168 mm$^2$</td>
<td>102 mm$^2$</td>
<td>1130 cm$^3$</td>
</tr>
<tr>
<td>Power Density</td>
<td>0.15 mW/mm$^2$</td>
<td>18 mW/mm$^2$</td>
<td>0.012 mW/mm$^2$</td>
<td>0.0177 mW/mm$^3$</td>
</tr>
</tbody>
</table>
Limits of State-of-art Brain-inspired Chips

- 2D flat routing/placement method
- Large size CMOS based synapse design (RAM)
- Long signal propagation distance leads large power consumption on signal delivery
- Large chip design area

TrueNorth Chips

On board Comm
Nearest Neighbor Comm
On-chip Comm

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

3D Memristor-based Synapse

Memristor Crossbar Structure

Horizontal RRAM (Resistive RAM) Structure

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

Comparison between RRAM with other Memory Technology [5]
3D Integration Technologies

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

Fabricate in a single wafer

don-layer inter-tier vias (MIVs)

Low temperature fabrication technology for the upper layers

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

<table>
<thead>
<tr>
<th>Devices</th>
<th>FinFET</th>
<th>Epi-like Si NWFET</th>
<th>Epi-like Si UTB</th>
<th>SOI-Si UTB</th>
<th>Poly-Si/Ge FinFET</th>
<th>IGZO OSFET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal budget (°C)</td>
<td>&lt; 400</td>
<td>&lt; 400</td>
<td>&lt; 400</td>
<td>&lt; 650</td>
<td>&lt; 400</td>
<td>&lt; 500</td>
</tr>
<tr>
<td>I_on/I_off</td>
<td>&gt;10^7</td>
<td>&gt;5 × 10^5</td>
<td>&gt;5 × 10^5</td>
<td>&gt;10^7</td>
<td>&gt;10^7</td>
<td>&gt;10^{21}</td>
</tr>
</tbody>
</table>

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].
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
Supportive Circuitry
Synapse

2D Memristor Structure Model
2D Neuron Circuit
2D Supportive Circuitry

Comparing 2D & 3D Performances

System Level Simulations for Real Neuromorphic Computing Applications

Carbon Nanotube FETs Characteristic Analysis

Noise & Variation Analysis

3D Memristor Structure Model
3D Neuron Circuit
3D Supportive Circuitry

3D Structure Impact Analysis
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


Thank you