Neuronic Chips: Building Blocks and System

Byung-Gook Park
Inter-university Semiconductor Research Center &
Department of Electrical and Computer Engineering
Seoul National University

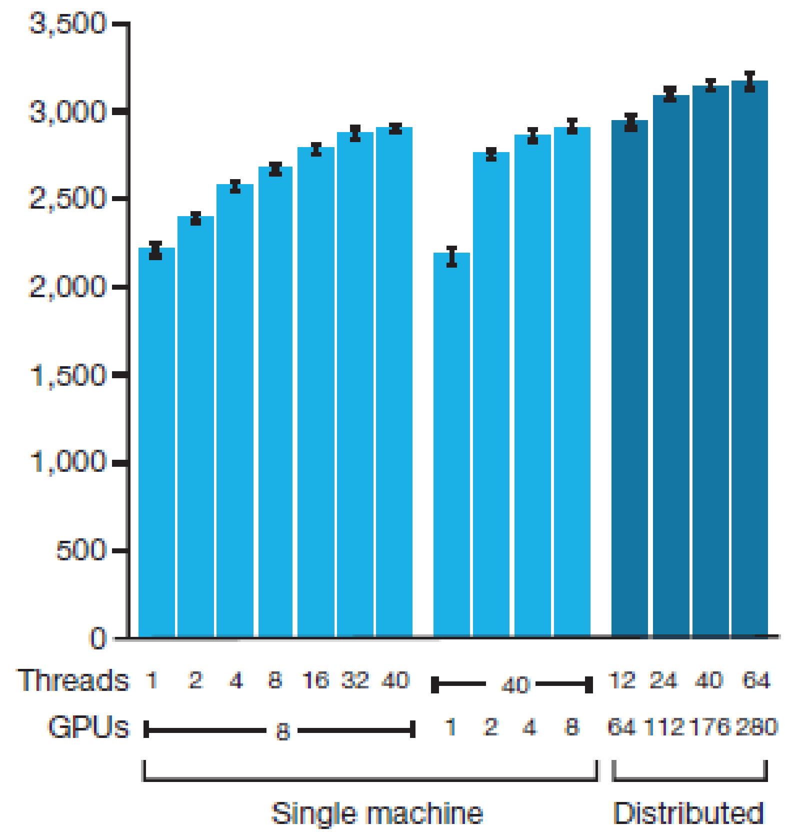
Human vs. AlphaGo

AlphaGo - Hardware

- Supercomputer

1920 CPU + **280 GPU**

- Performance



Comparison

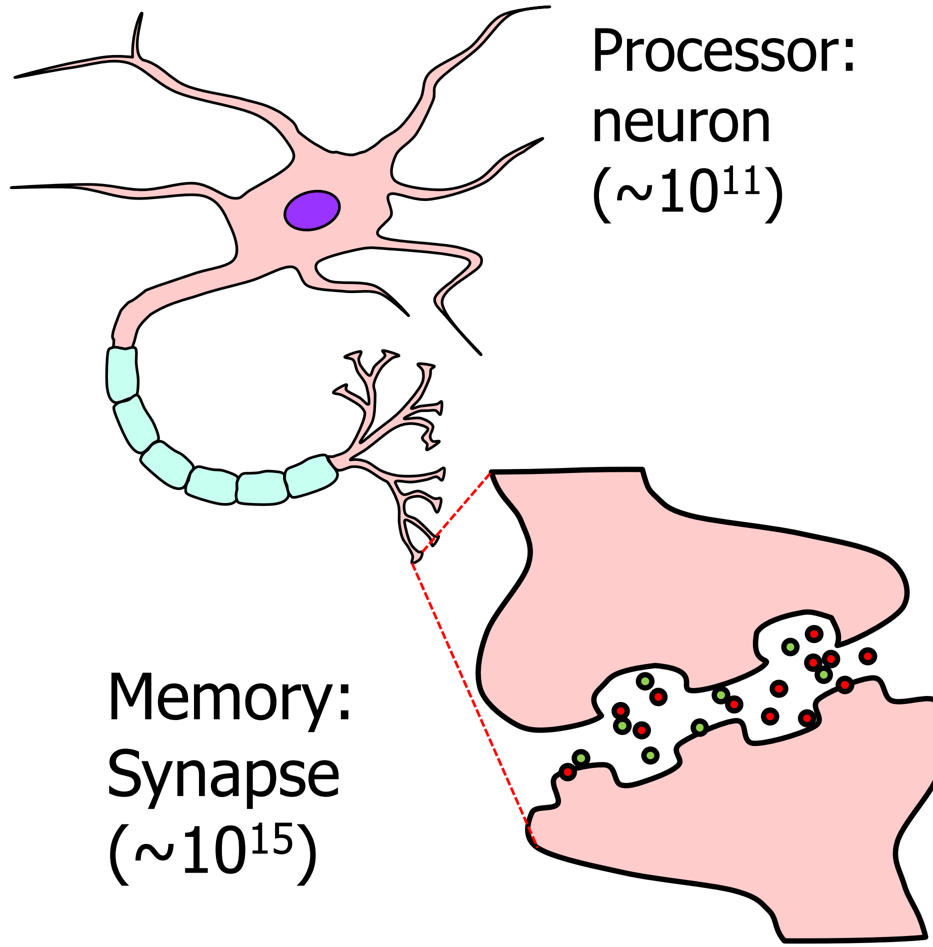
- Human Brain

- neuron + synapse
- massively parallel
- ~ ms speed
- low power (~20 W)
- recognition/reasoning

- Digital Computer

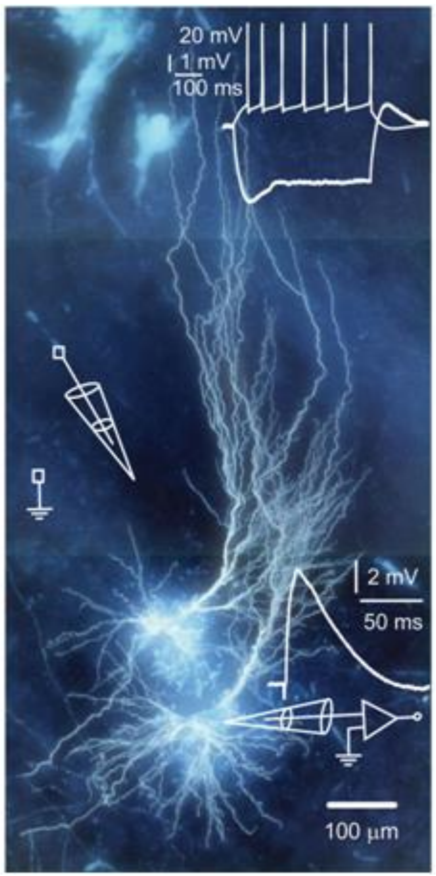
- CPU + memory
- serial
- ~ ns speed
- high power (~20 MW)
- computation

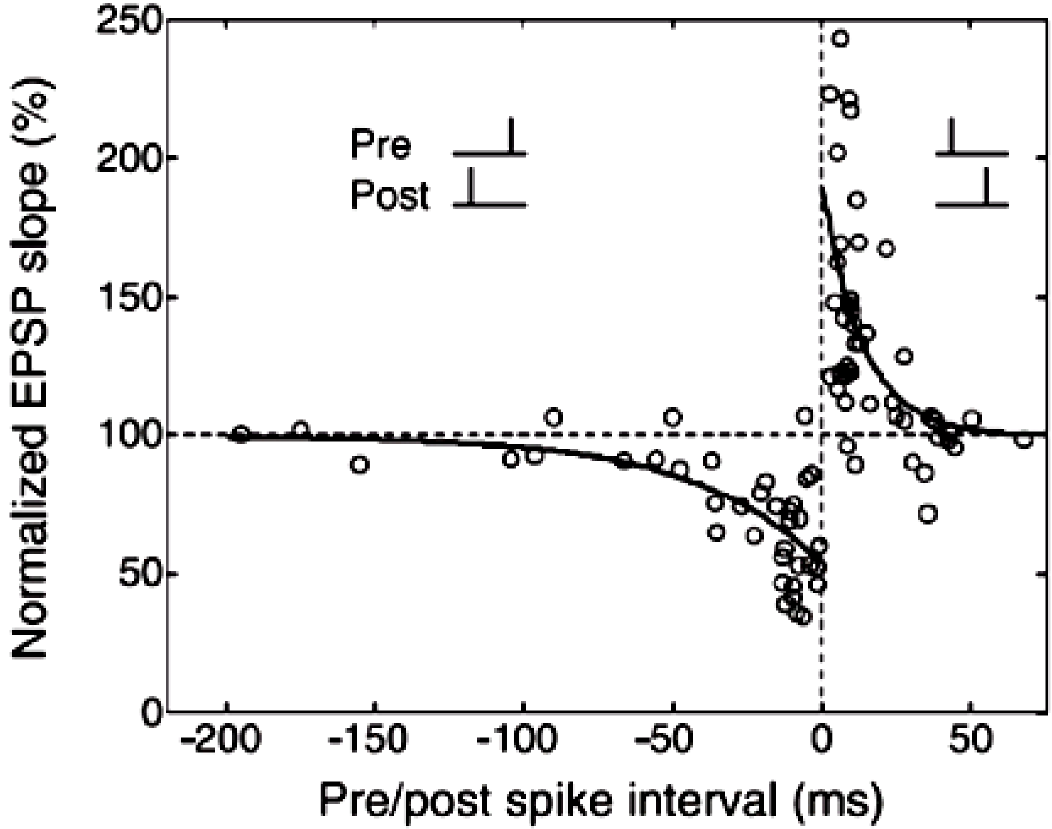
Human Brain and Its Building Blocks



Spike-Timing-Dependent Plasticity

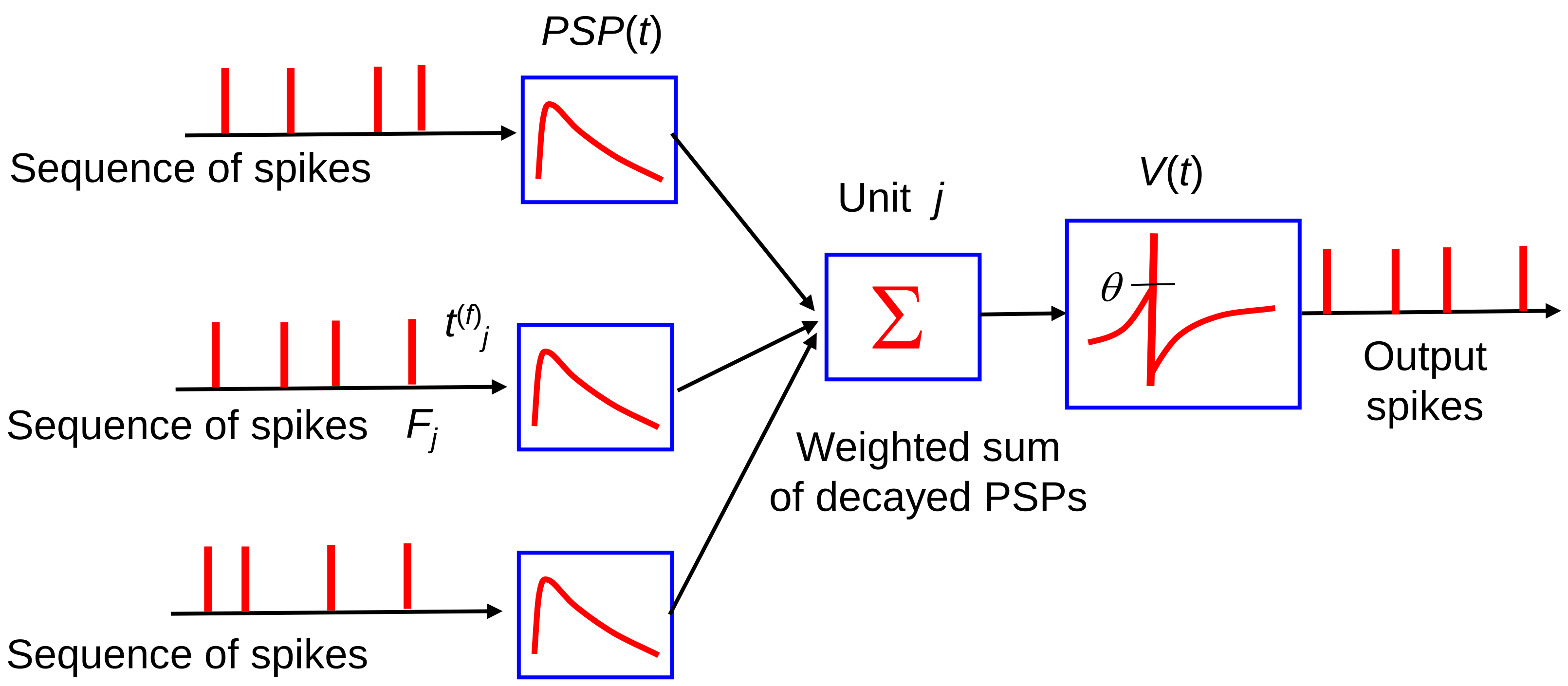
- Spike-timing-dependent plasticity – learning mechanism





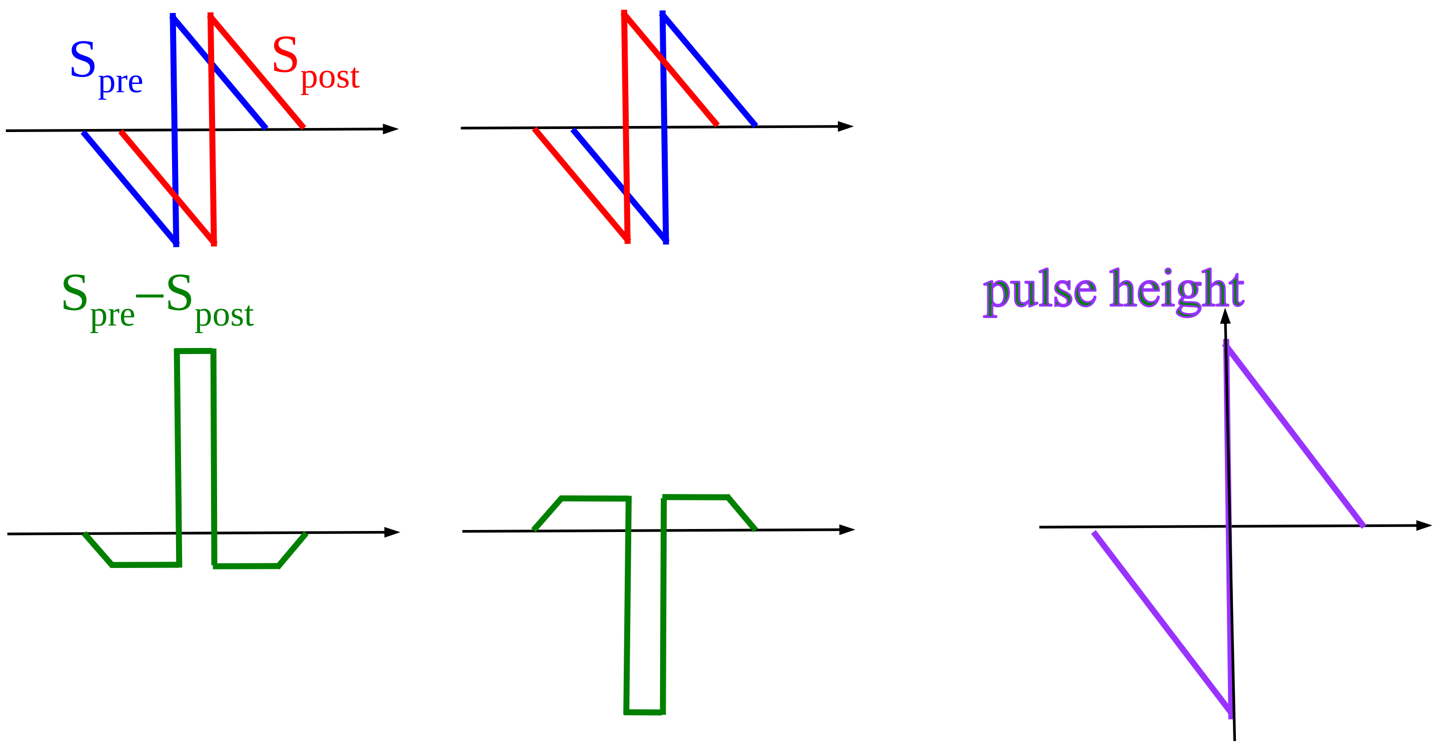
Spiking Neural Network (SNN) (1)

- 3^{rd} generation neural network model
  - input/output: spikes
  - signal intensity: firing rates



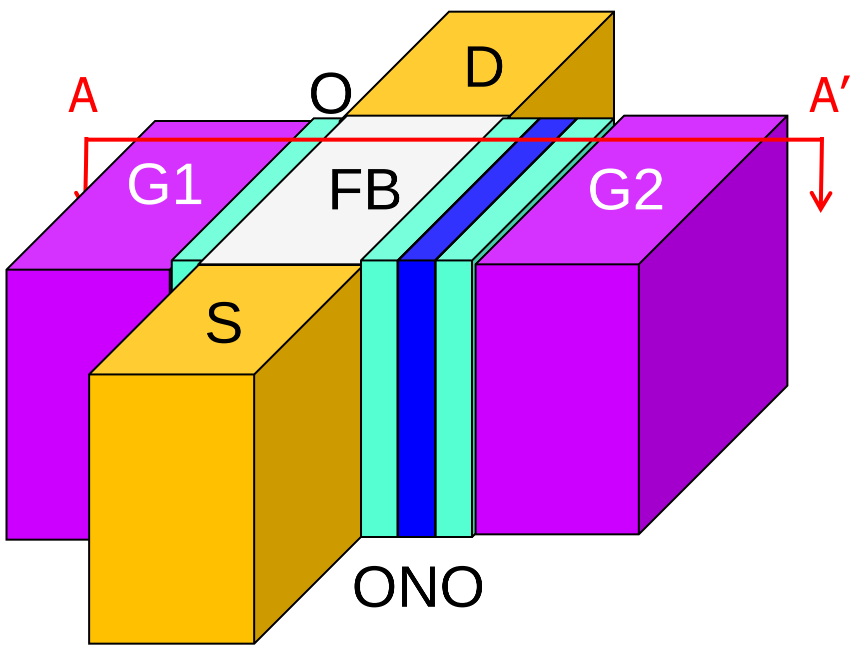
Spiking Neural Network (SNN) (2)

- Learning mechanism
  - error back-propagation with time coding
  - spike-timing-dependent plasticity (STDP)

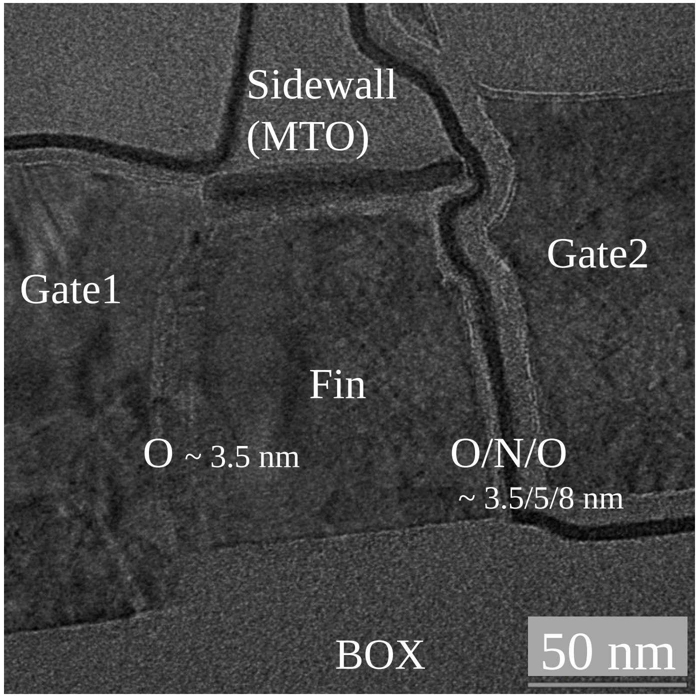


Floating-body Synaptic Transistor (1)

- Structure



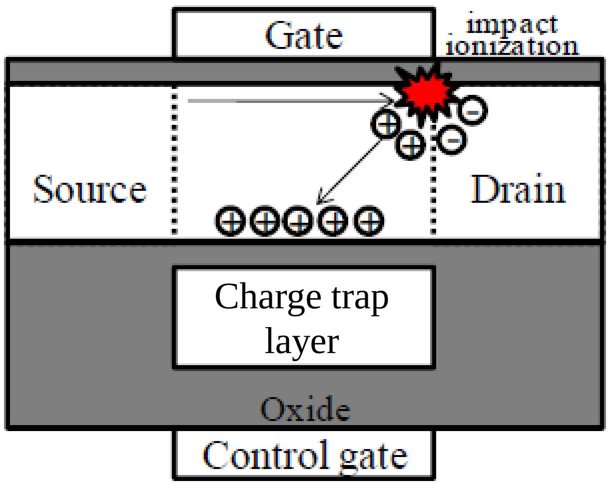
- TEM image



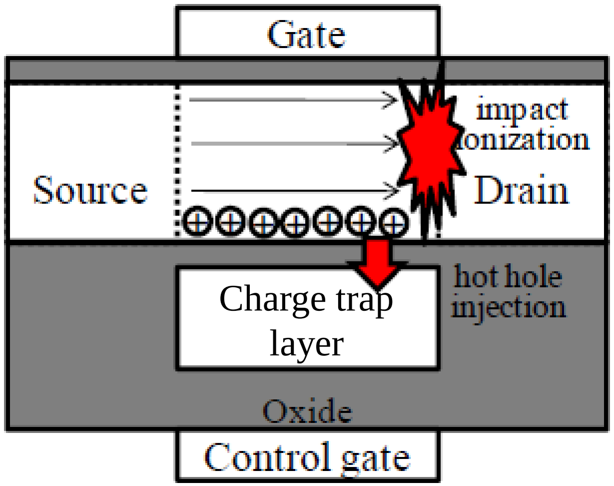
- Cross-section in A – A' direction

Floating-body Synaptic Transistor (2)

- Short-term memorization



- Long-term memorization

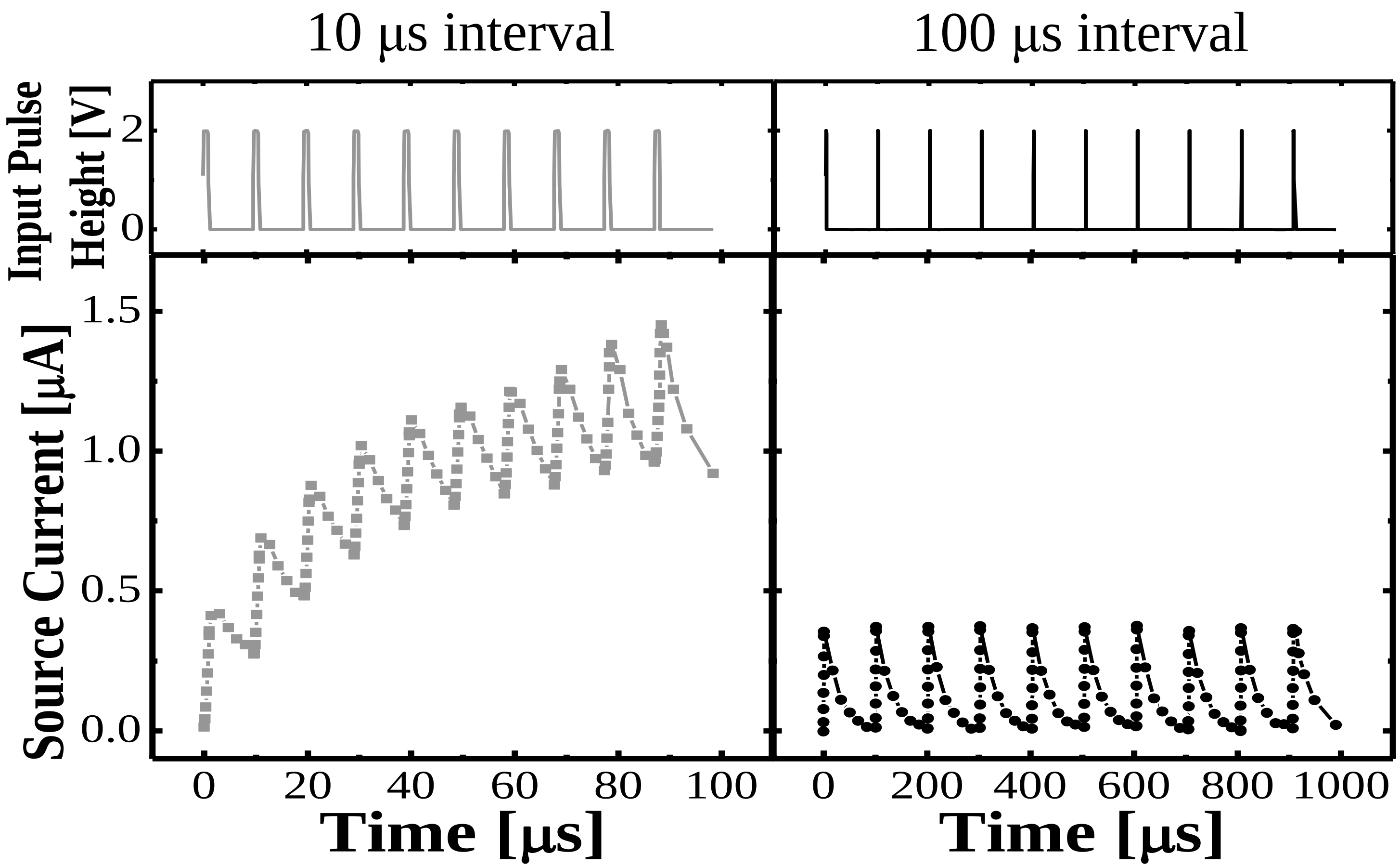


- Impact-generated holes are temporarily stored in the body.
- Without further inputs, these holes gradually disappear through recombination process.

- Hot holes are programmed to the floating gate.
- Even without further inputs, these charges do not disappear without special erasing actions.

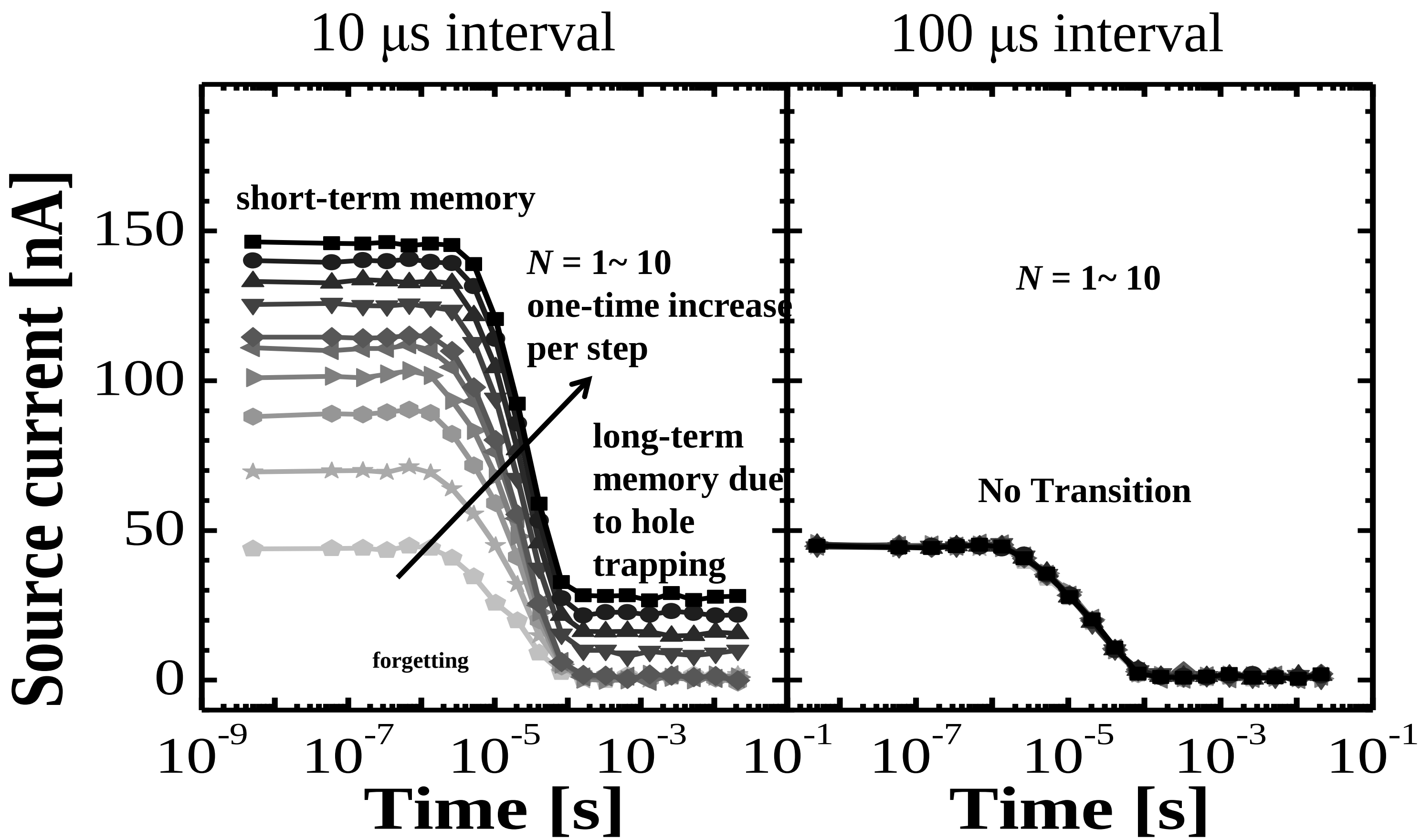
Floating-body Synaptic Transistor (3)

- Transient response of FST to spikes



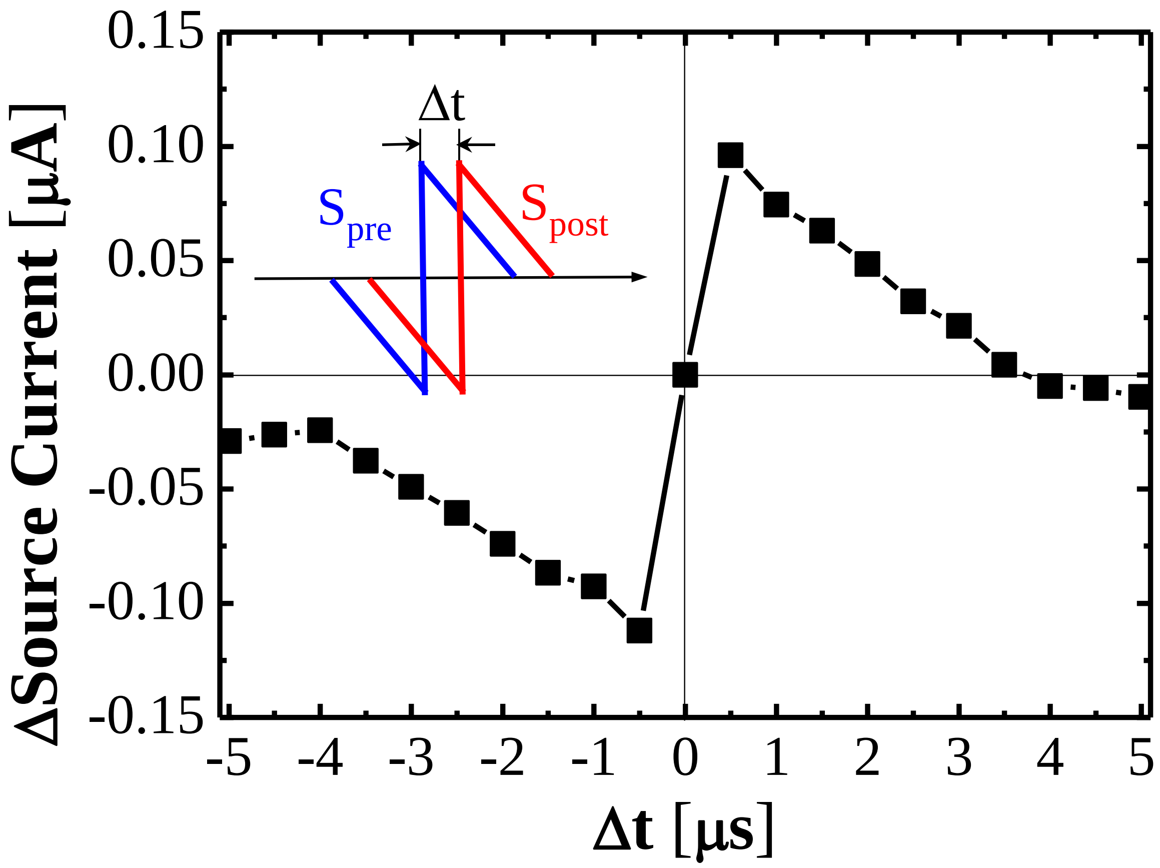
Floating-body Synaptic Transistor (4)

- Short-term to long-term memory transition



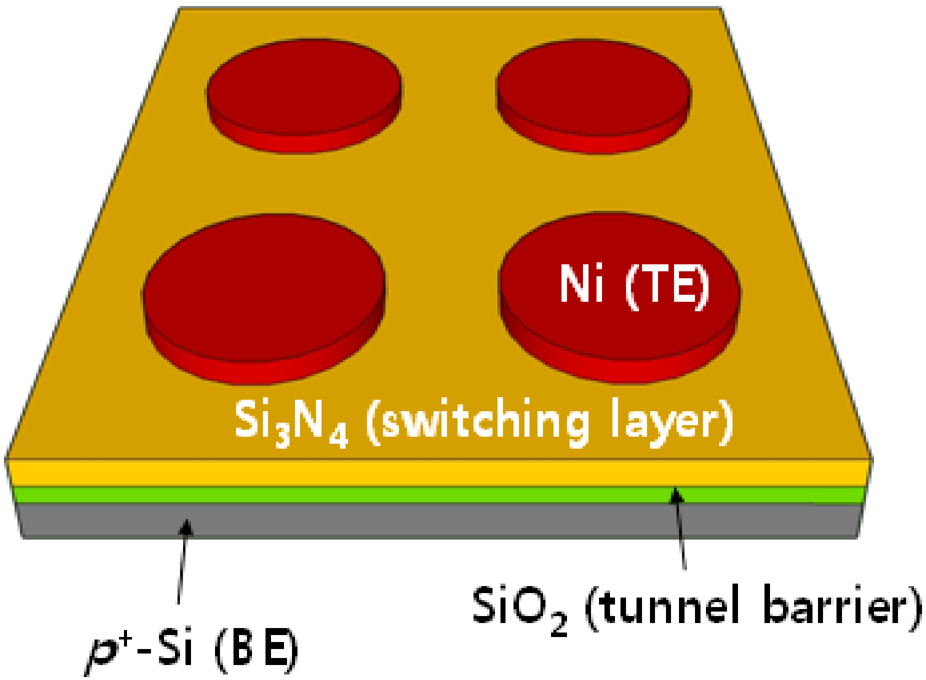
Floating-body Synaptic Transistor (5)

- STDP characteristic

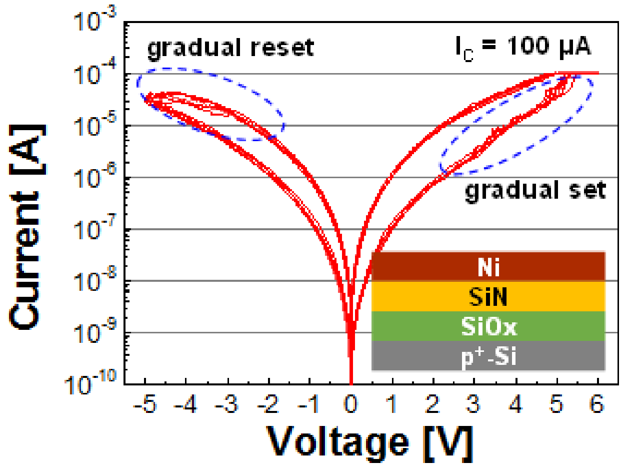


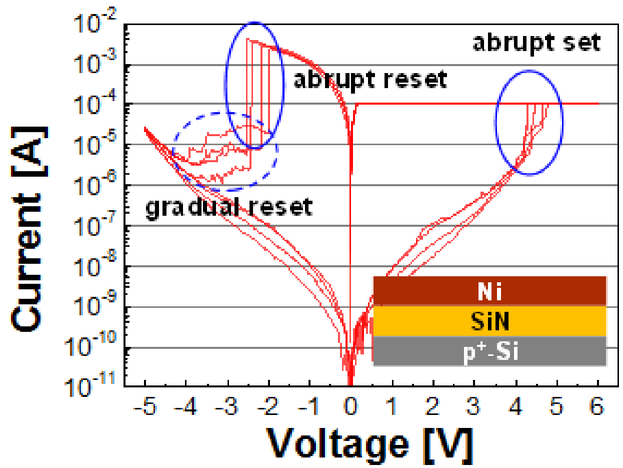
Resistive Memory Synapse (1)

- Structure



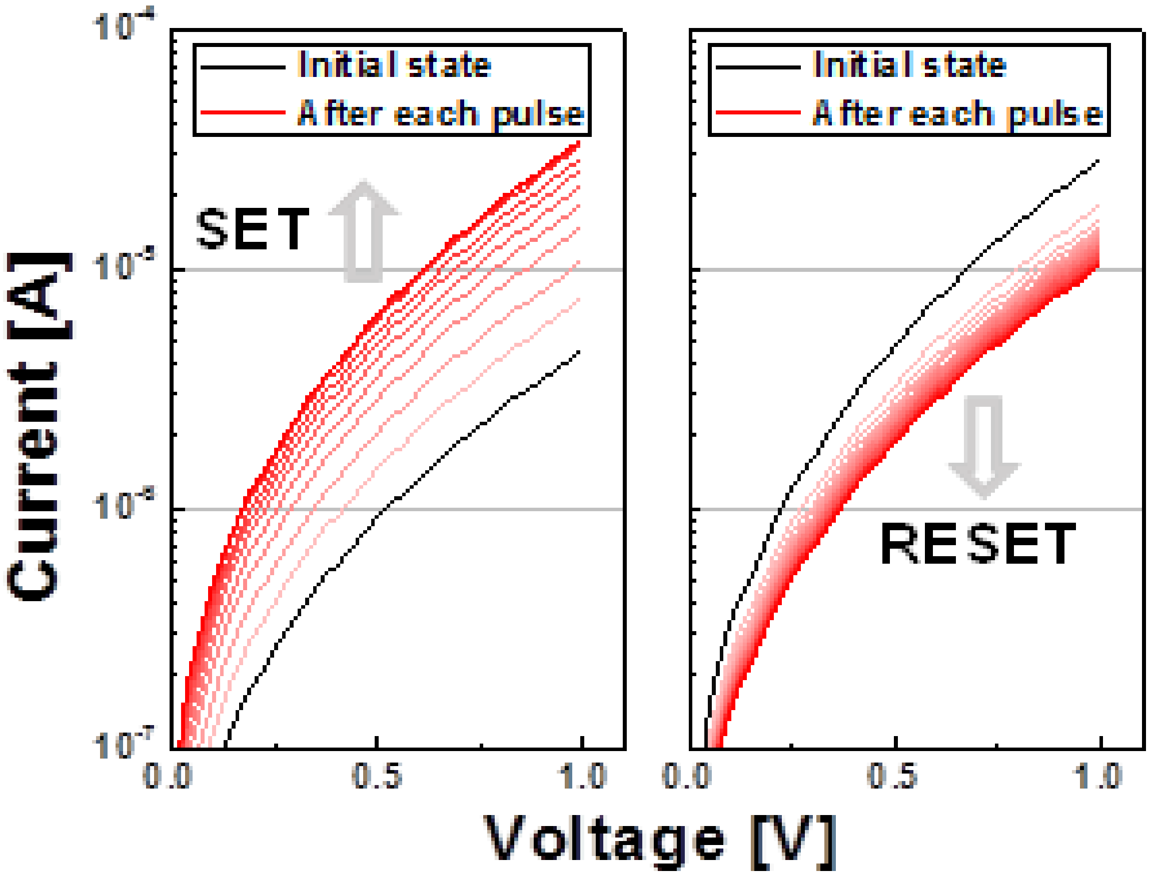
- Switching characteristics





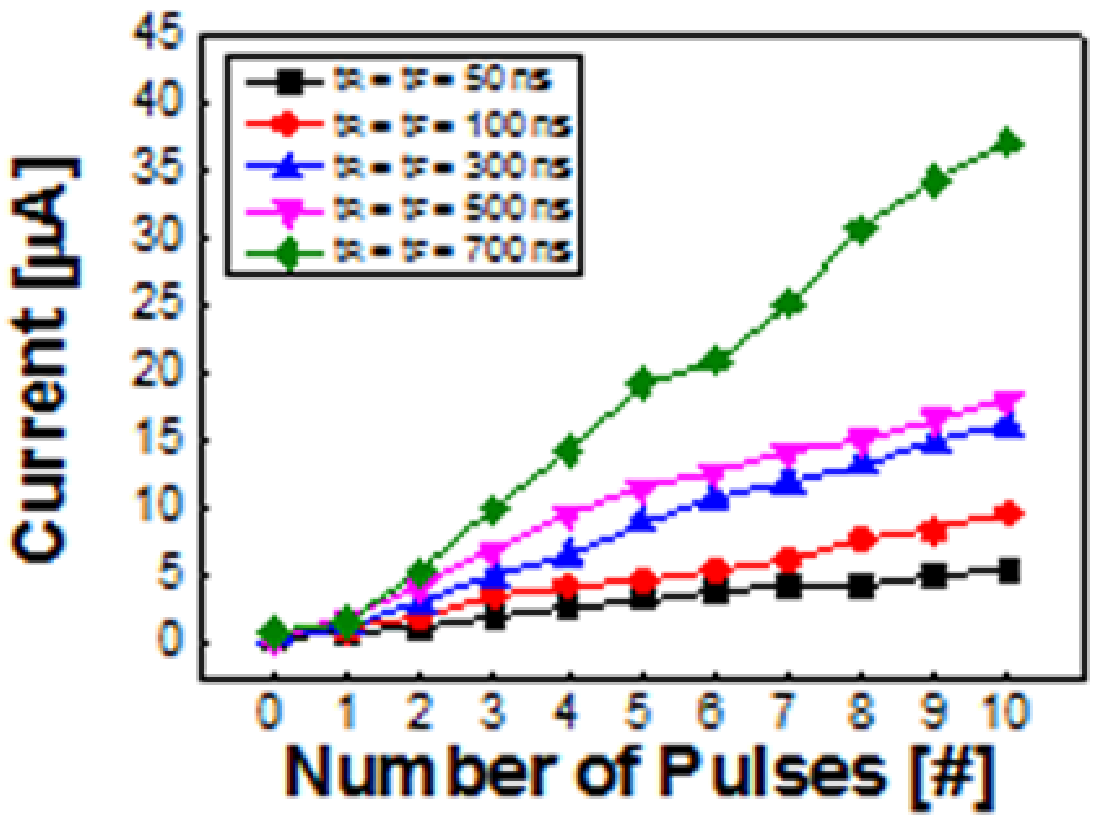
Resistive Memory Synapse (2)

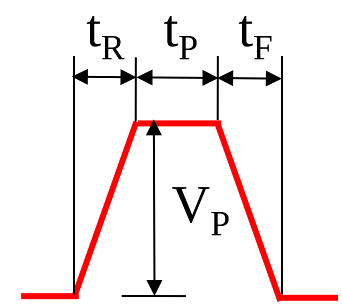
- Gradual switching characteristics



Resistive Memory Synapse (3)

- Read current as a function of the number of spikes



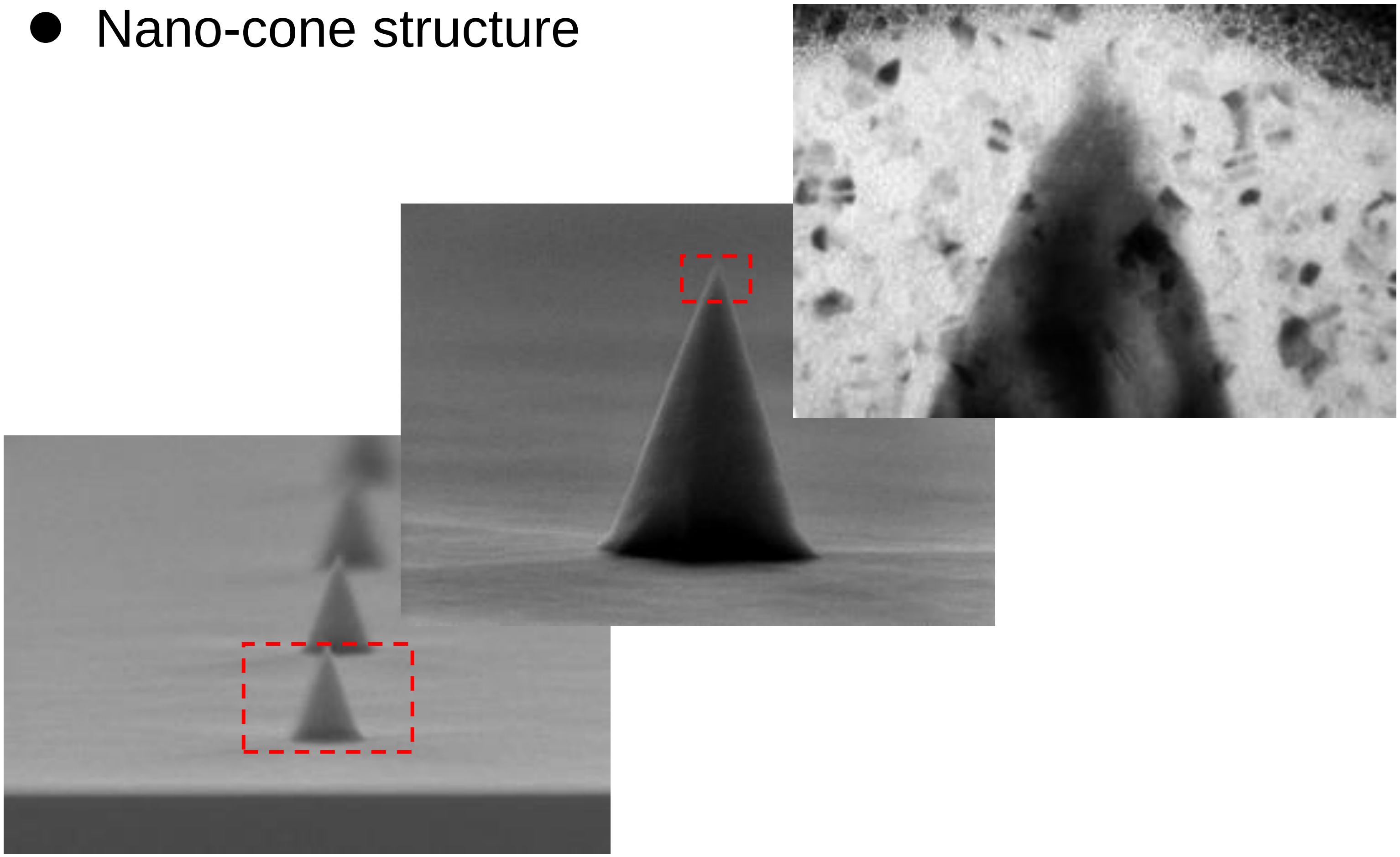


$t_P = 100$ ns

$V_P = 10$ V

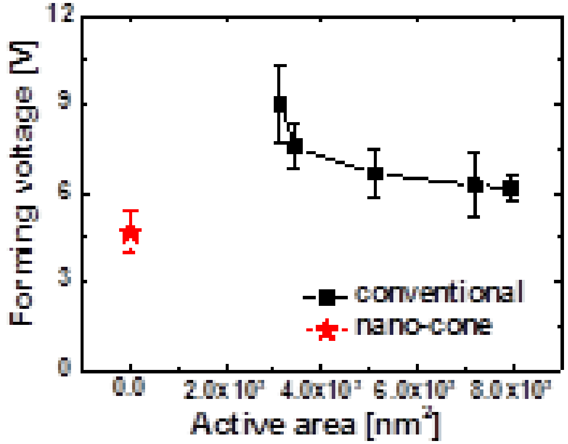
Resistive Memory Synapse (4)

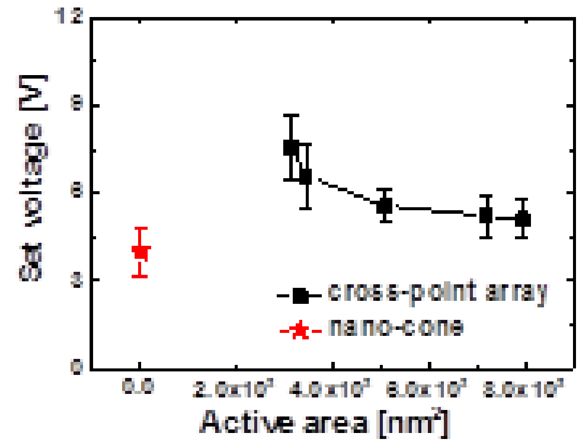
- Nano-cone structure

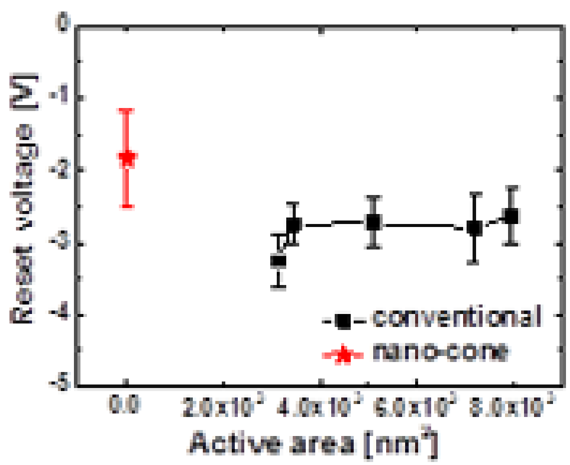


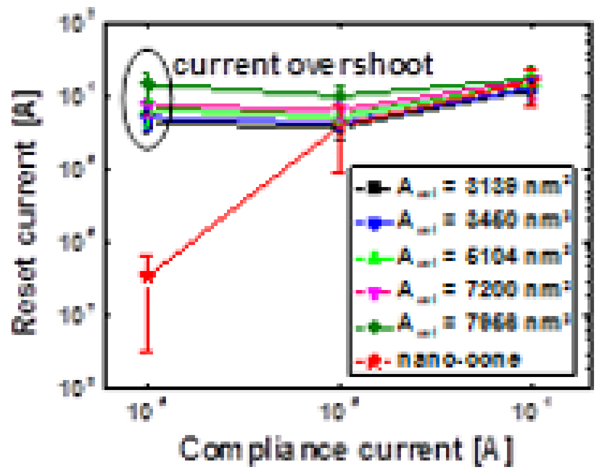
Resistive Memory Synapse (5)

- Reduction of operating voltage and current



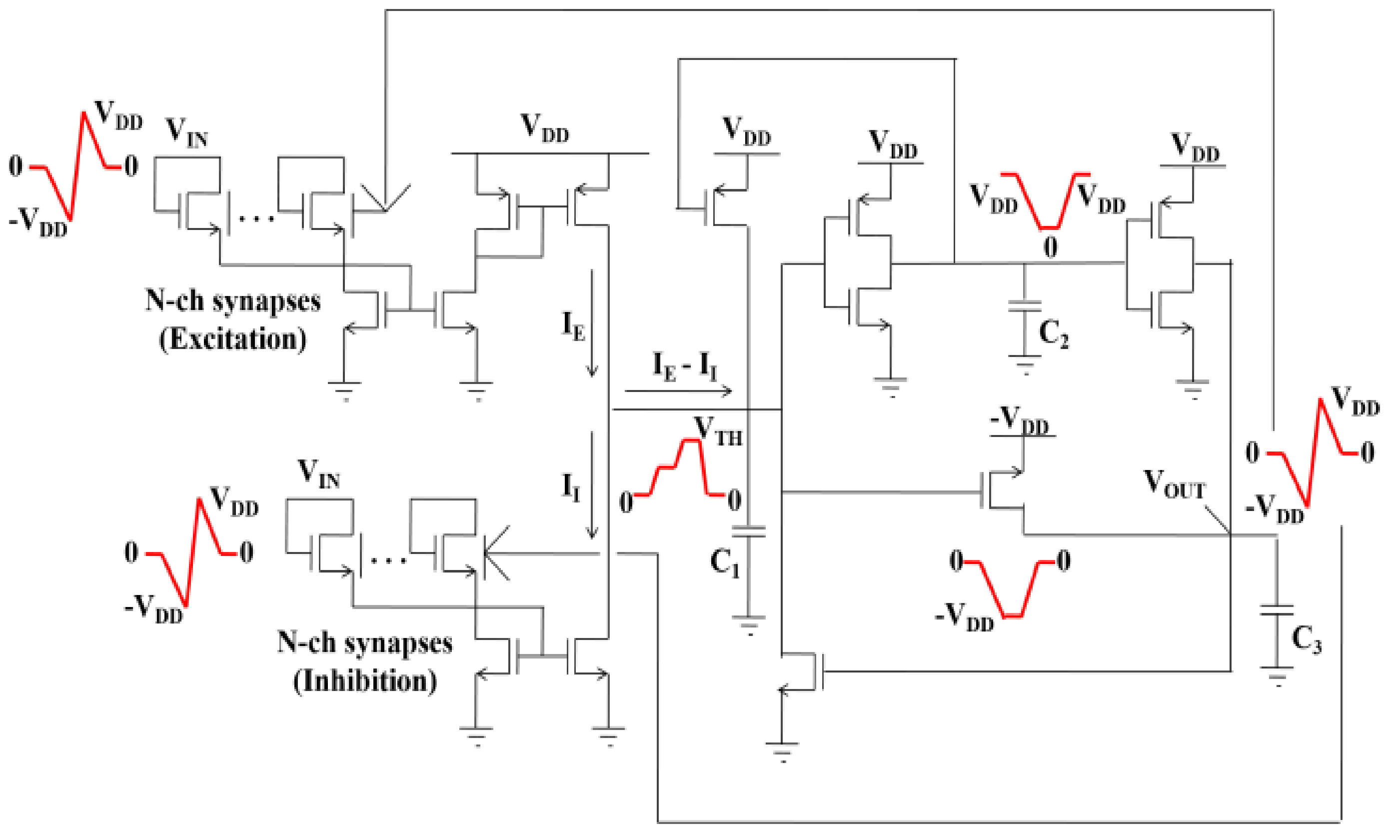






Neuron Circuit with Capacitors (1)

- Integrate-and-fire neuron circuit with capacitor integration

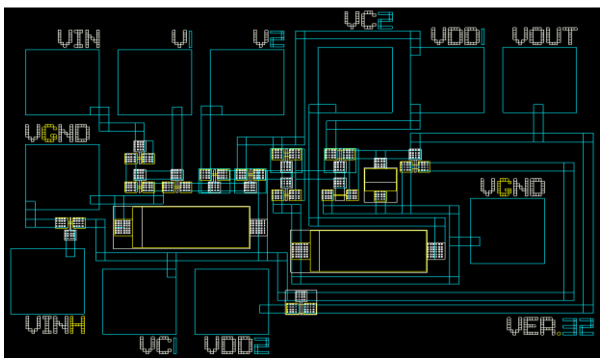


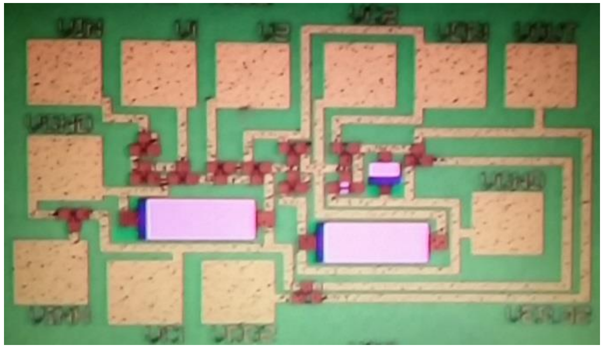
<synaptic integration part>   <spike generation part>

Neuron Circuit with Capacitors (2)

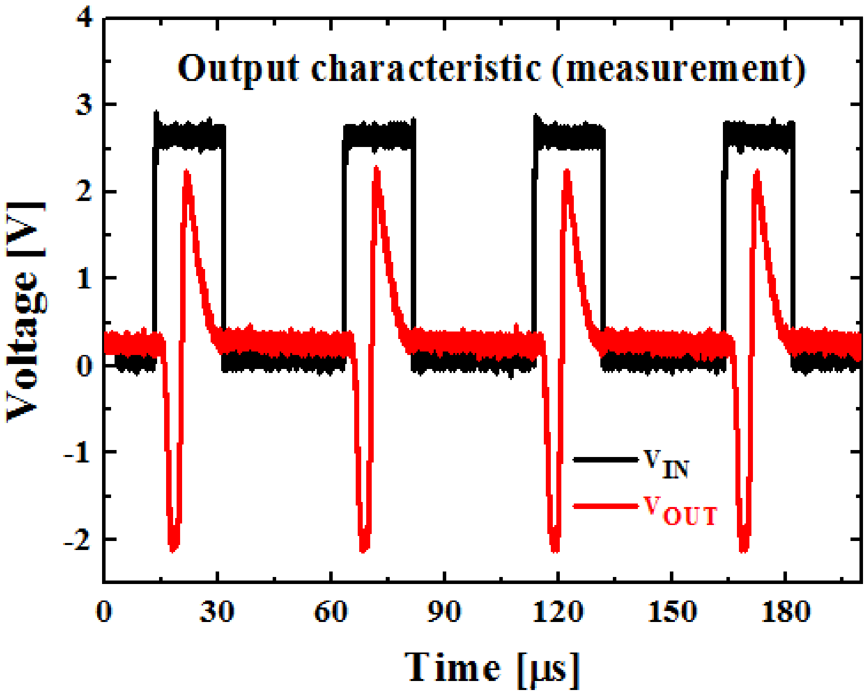
- Integrated circuit implementation

<Layout and chip image>



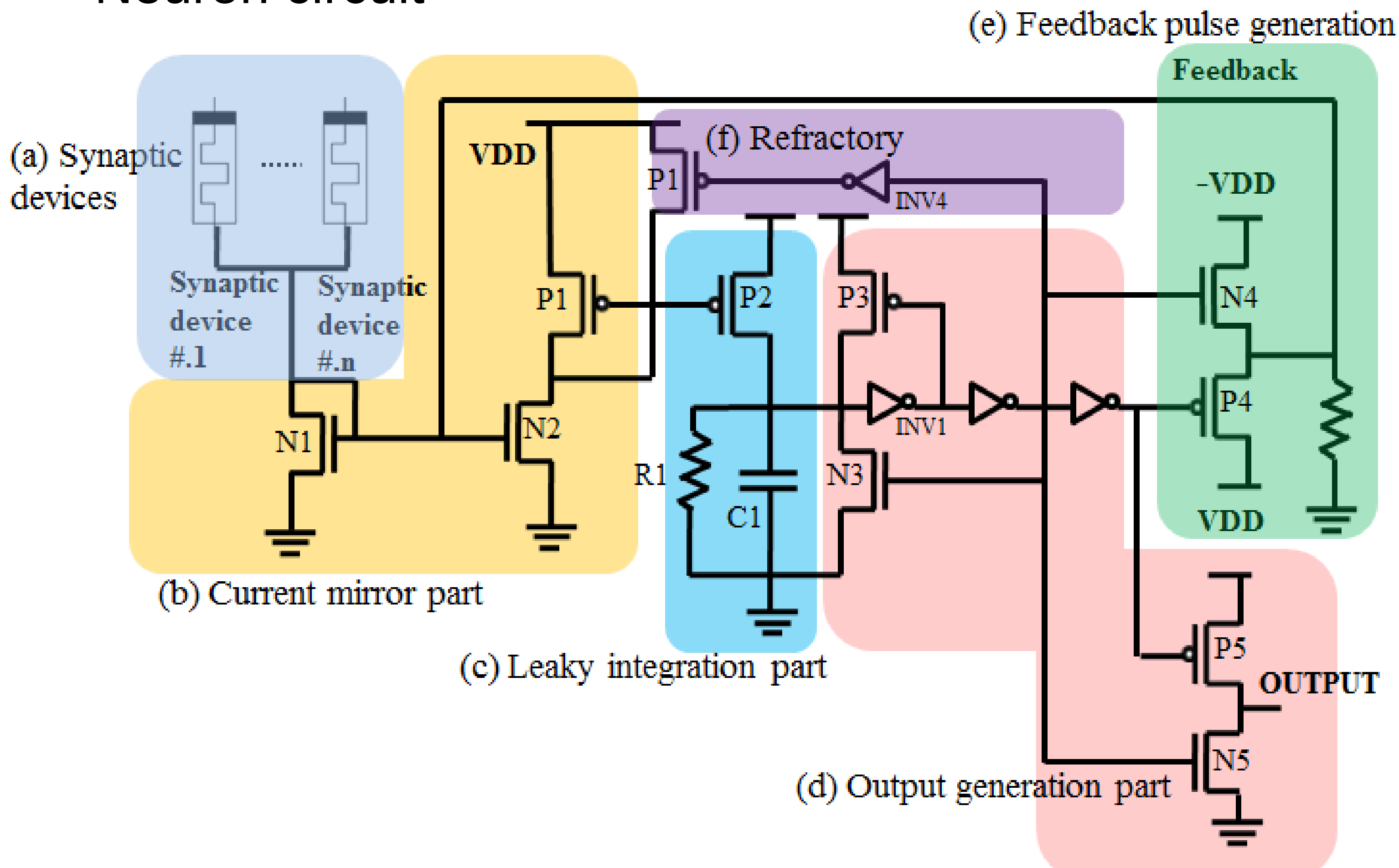


<Output of neuron>



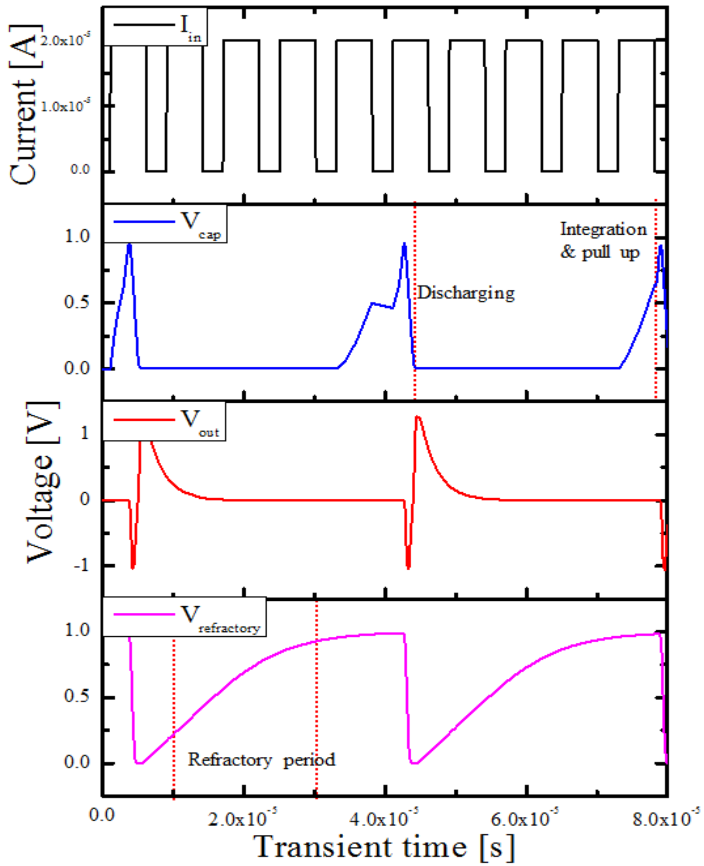
Neuron Circuit for Resistive Memory (1)

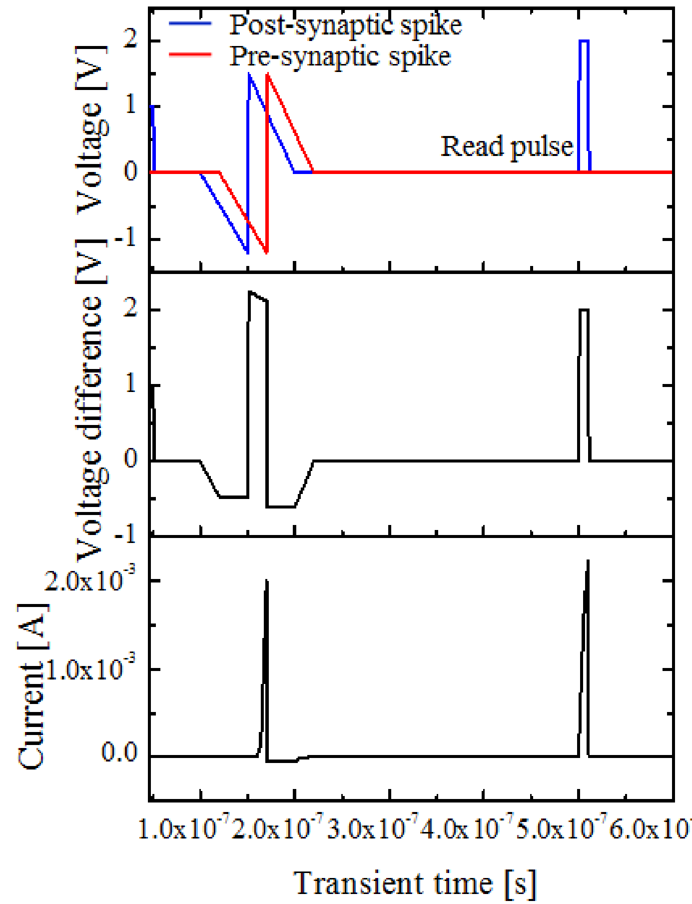
- Neuron circuit



Neuron Circuit for Resistive Memory (2)

- Characteristics of neuron

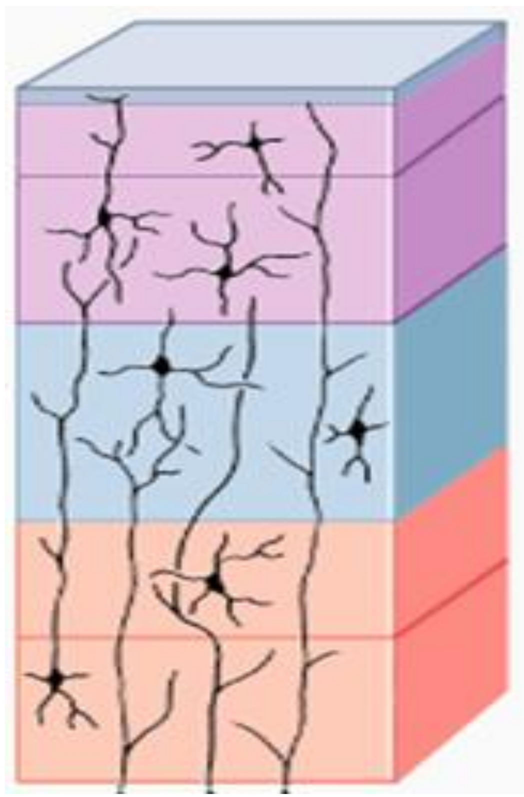




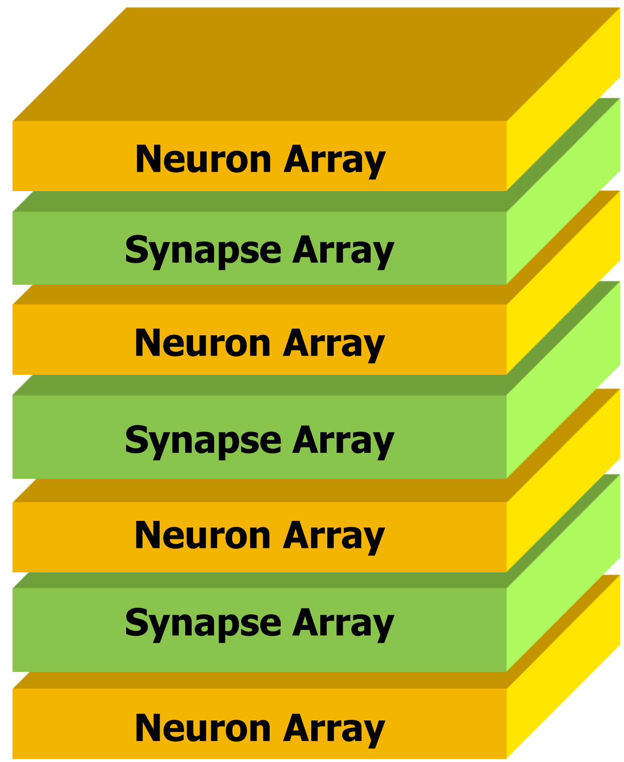
Integration of Neurons and Synapses

- Stacking of neuron and synapse arrays

<primary sensory cortex>

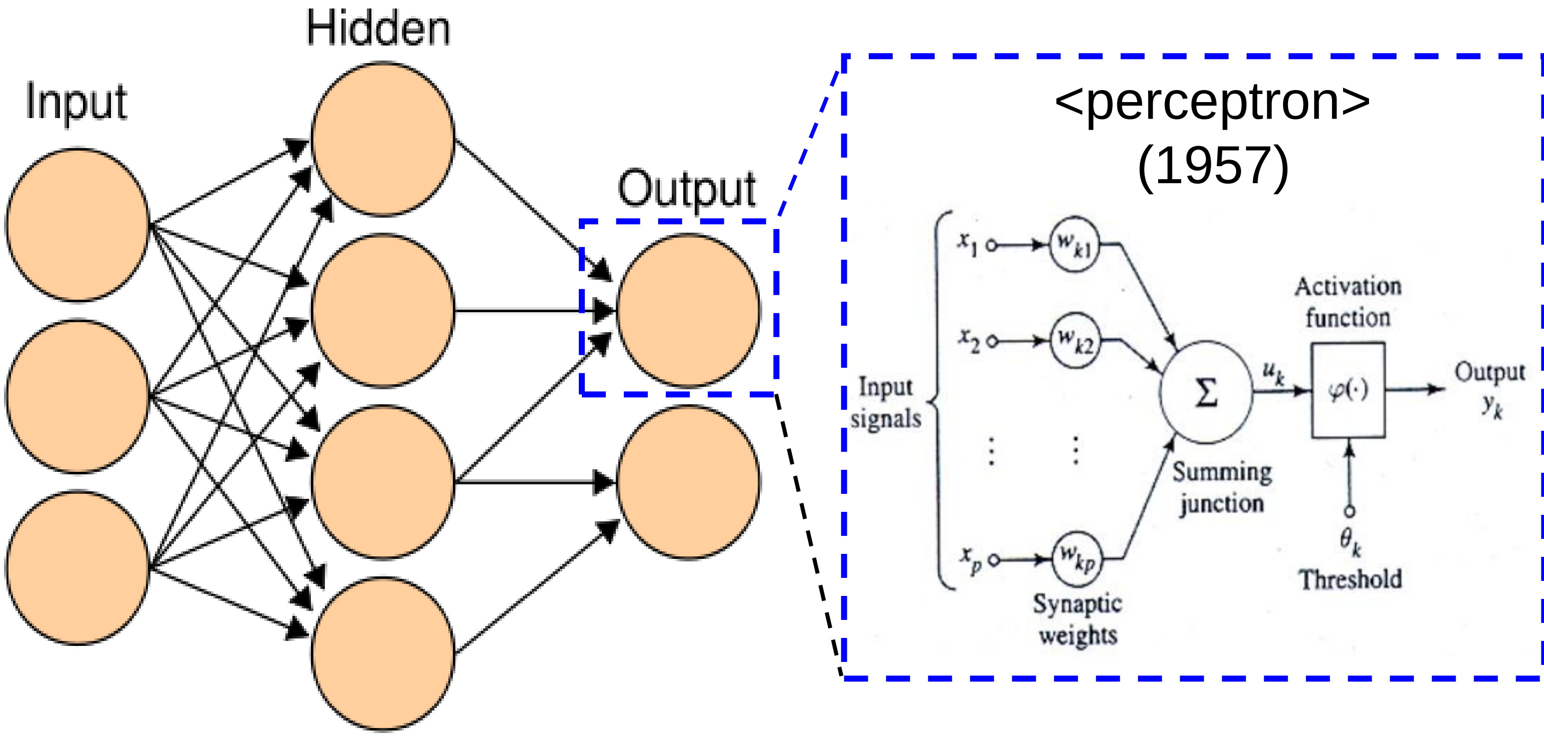


<neuronic system>



Artificial Neural Network (ANN)

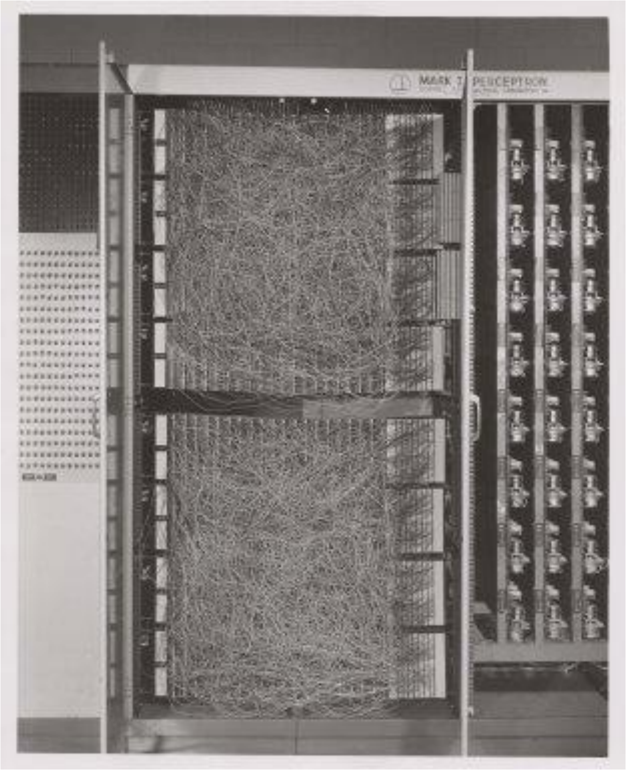
- Concept of neural network



** Various weight calculation methods were proposed, but a learning algorithm for general networks was unavailable.

Perceptron

- Invention of perceptron



Mark I Perceptron Machine

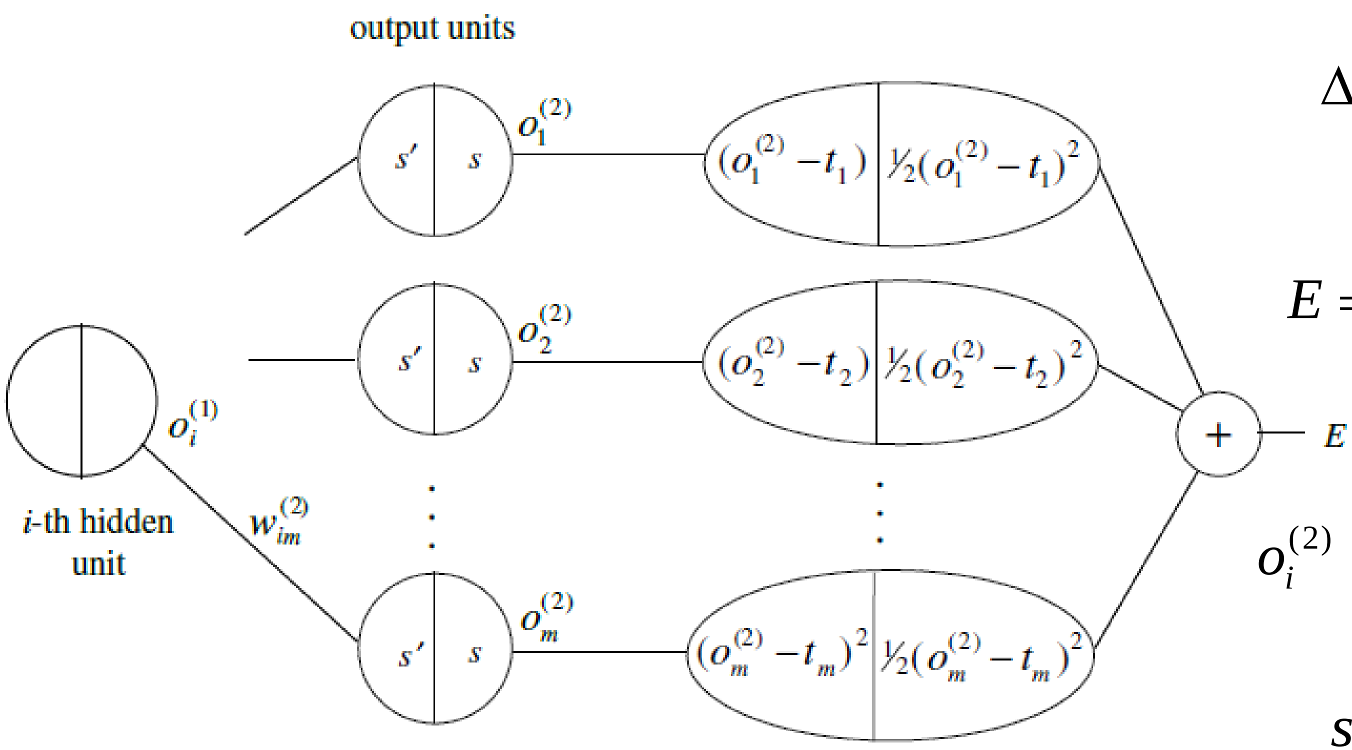
Frank Rosenblatt

Cornell Aeronautical Laboratory

"Perceptron is the embryo of an electronic computer that [the Navy] expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence"
- New York Times, 1958

Breakthrough (1986)

- Back propagation



$$\Delta w_{ij} = -\gamma \frac{\partial E}{\partial w_{ij}}$$

$$E = \sum_i \frac{1}{2}\left(o_i^{(2)} - t_i\right)^2$$

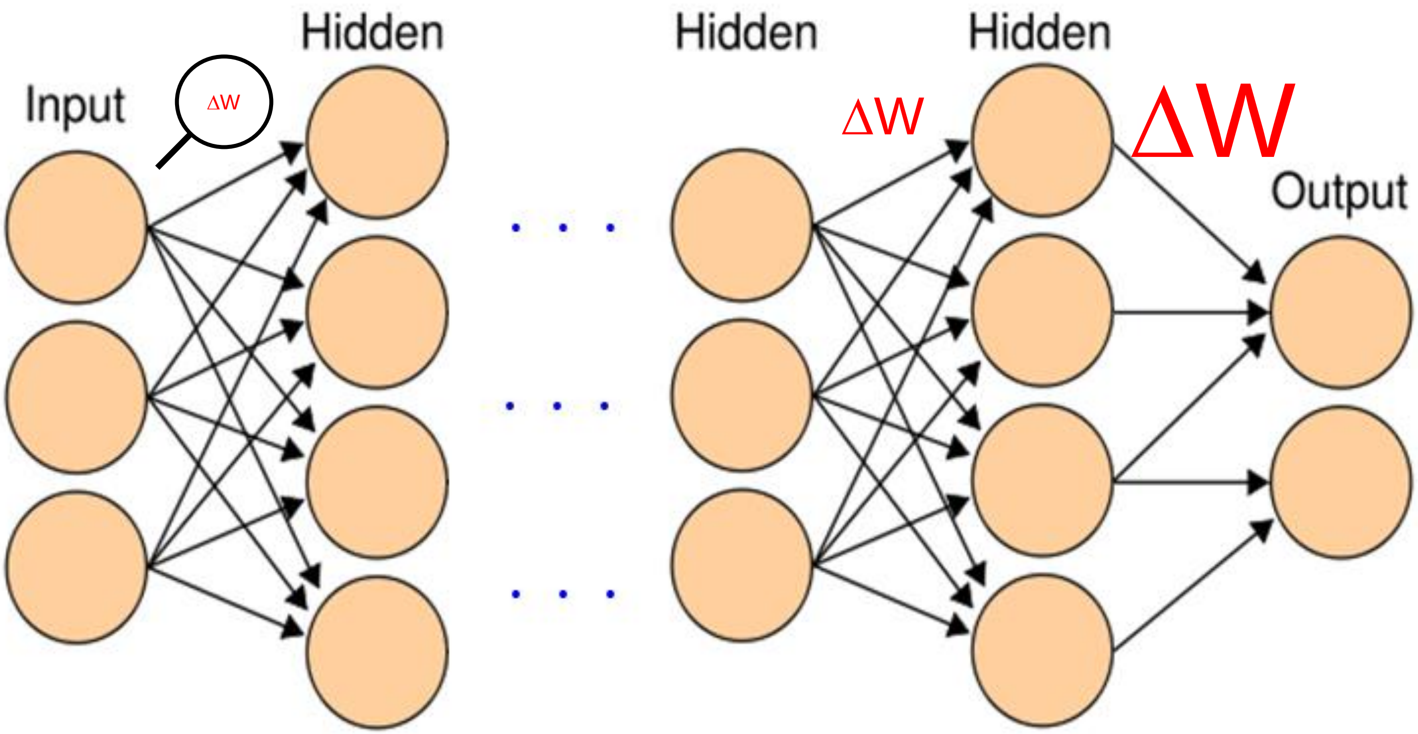
$$o_i^{(2)} = s\left(\sum_j w_{ij} o_i^{(1)} + b\right)$$

$$s' = o_i^{(2)}\left(1 - o_i^{(2)}\right)$$

** Weights are calculated by the gradient descent (chain rule) method .

Deep Neural Network (DNN)

- Multiple hidden layers



** Vanishing gradient problem (VGP) → new activation function (ReLU)

Breakthrough (2010)

- Rectified Linear Unit (ReLU)

** ReLU solves the vanishing gradient problem!!
(+ Concept of signal intensity included)

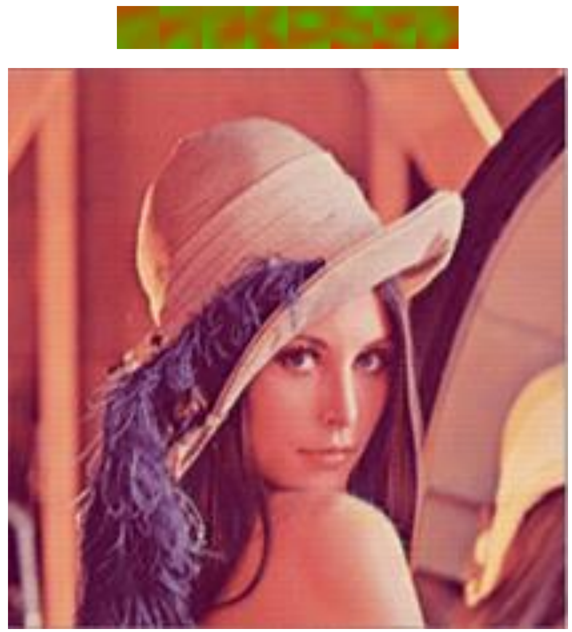
Comparison: ReLU vs. Sigmoid

- Speed of Learning: 8:1 Compression

<Original>

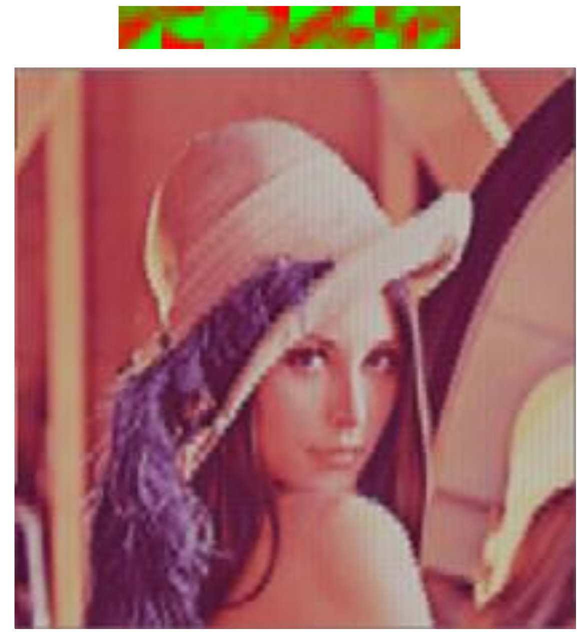
512x512
Image

<ReLU>



Epoch = 800
MSE = 0.00093

<Sigmoid>



Epoch = 800
MSE = 0.00142

** MSE (mean square error)

STDP and Error Back-propagation

- STDP

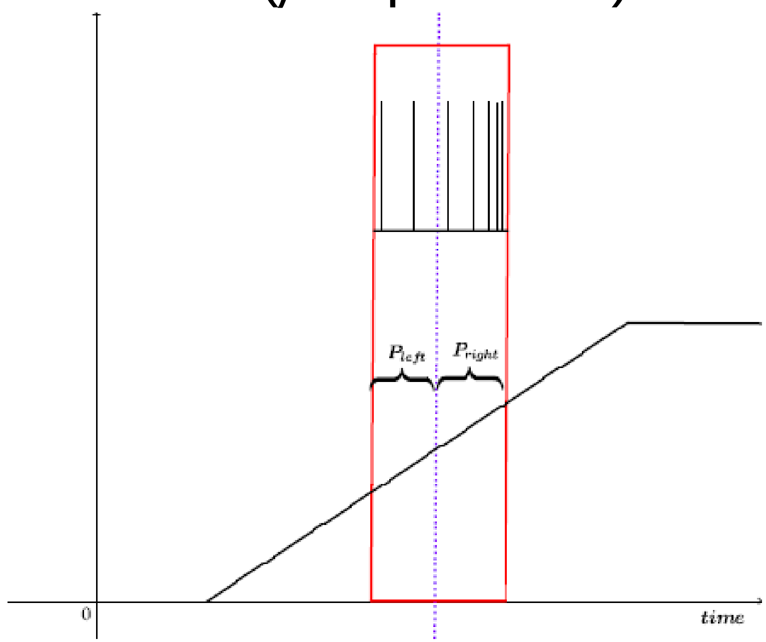
$$\Delta W_{ij} = \alpha \rho_i \dot{\rho}_j$$

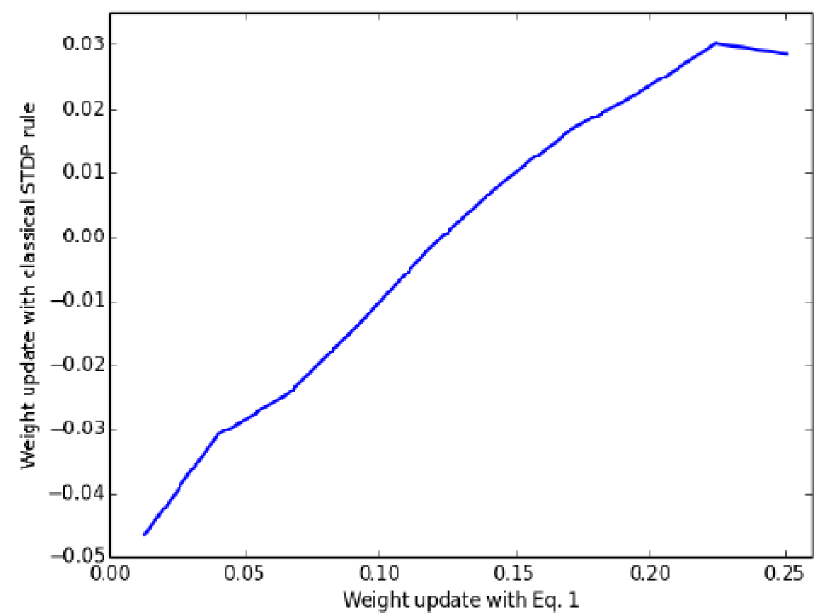
(ρ : spike rate)

- BP

$$\Delta W_{ij} = \alpha' x_i \dot{x}_j$$

(x : neuron output)

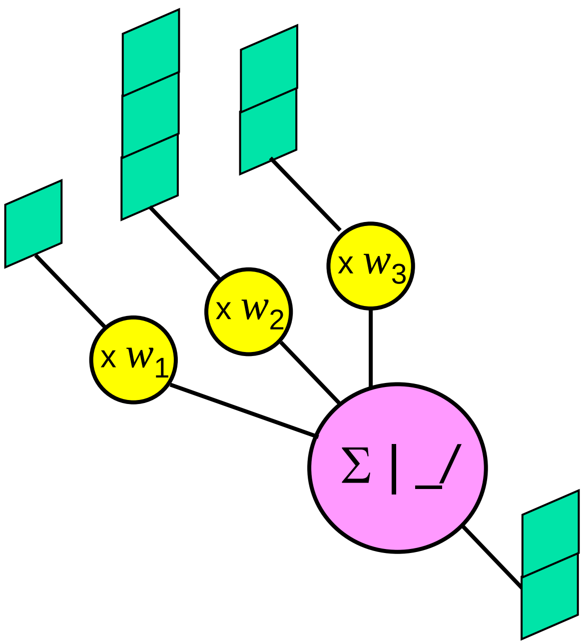


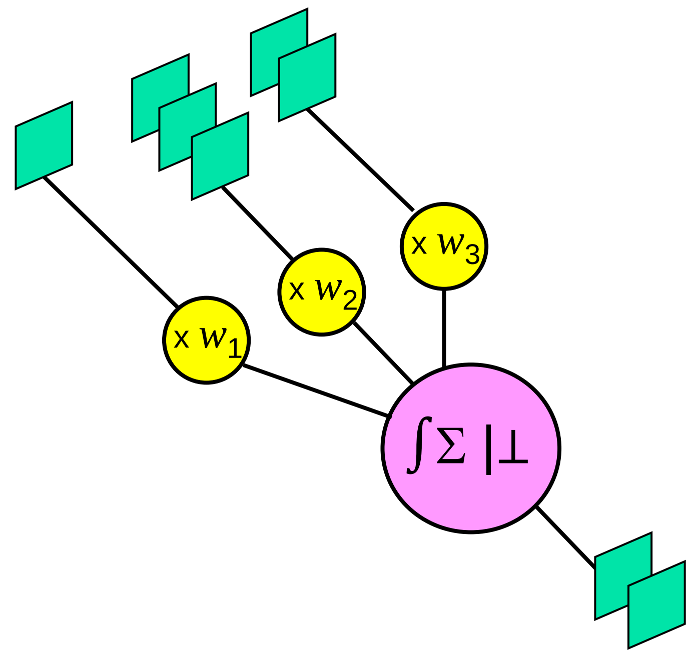


<Bengio, arXiv.org, (2016)>

ReLU Perceptron and Spiking Neuron

- ReLU Perceptron
- Spiking Neuron





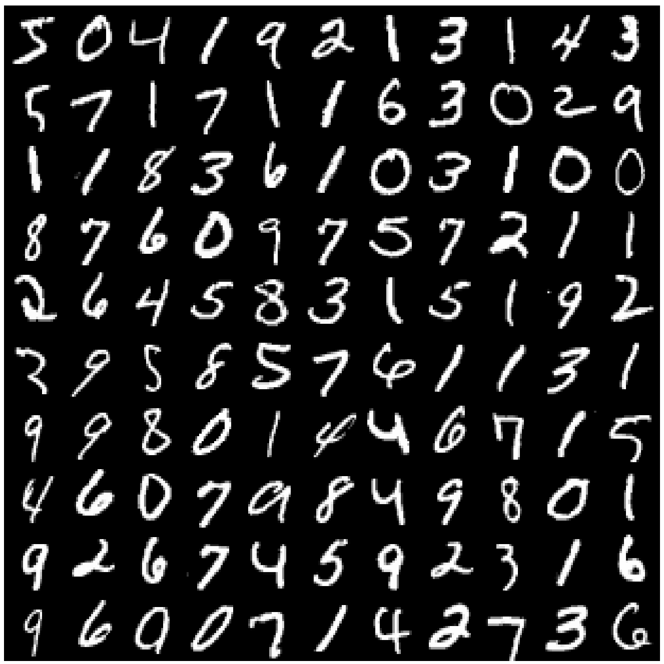
Equivalent in terms of inference!!!

<O'Connor, arXiv.org, (2016)>

High-level SCNN Simulation (1)

- MNIST Handwritten Digits

<train set>



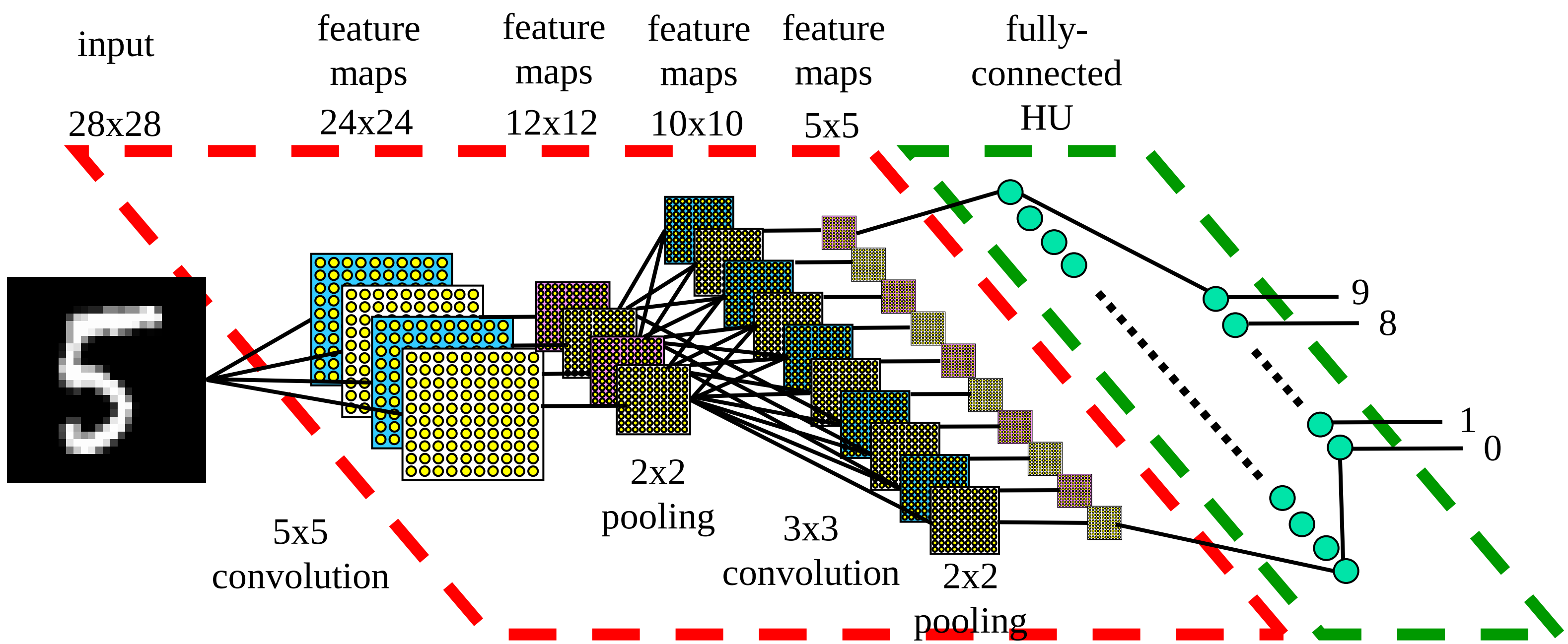
60,000 samples

<test set>

10,000 samples

High-level SCNN Simulation (2)

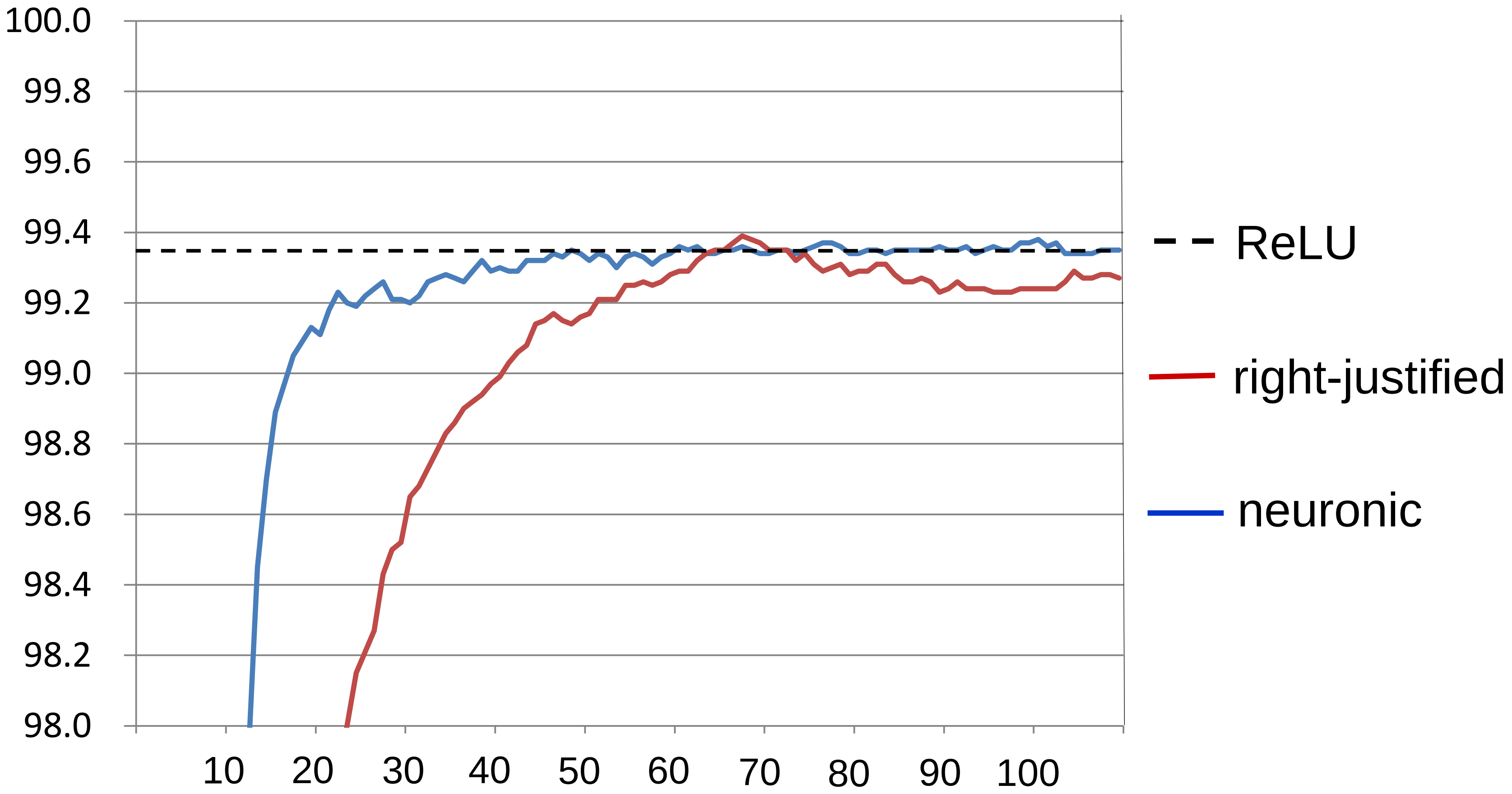
- Structure: Convolutional Neural Network (CNN)



1) Convolution + pooling (subsampling): feature extraction
2) Fully-connected layer: classification

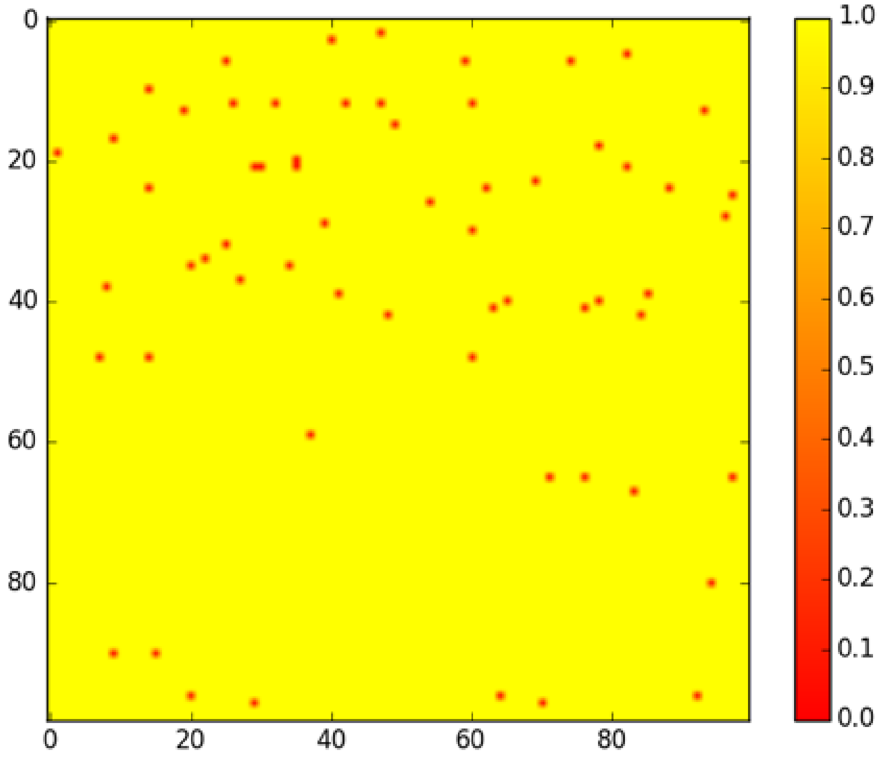
High-level SCNN Simulation (3)

- MNIST Handwritten Digits – SCNN Inference Accuracy



High-level SCNN Simulation (3)

- MNIST Handwritten Digits – SCNN Error Map

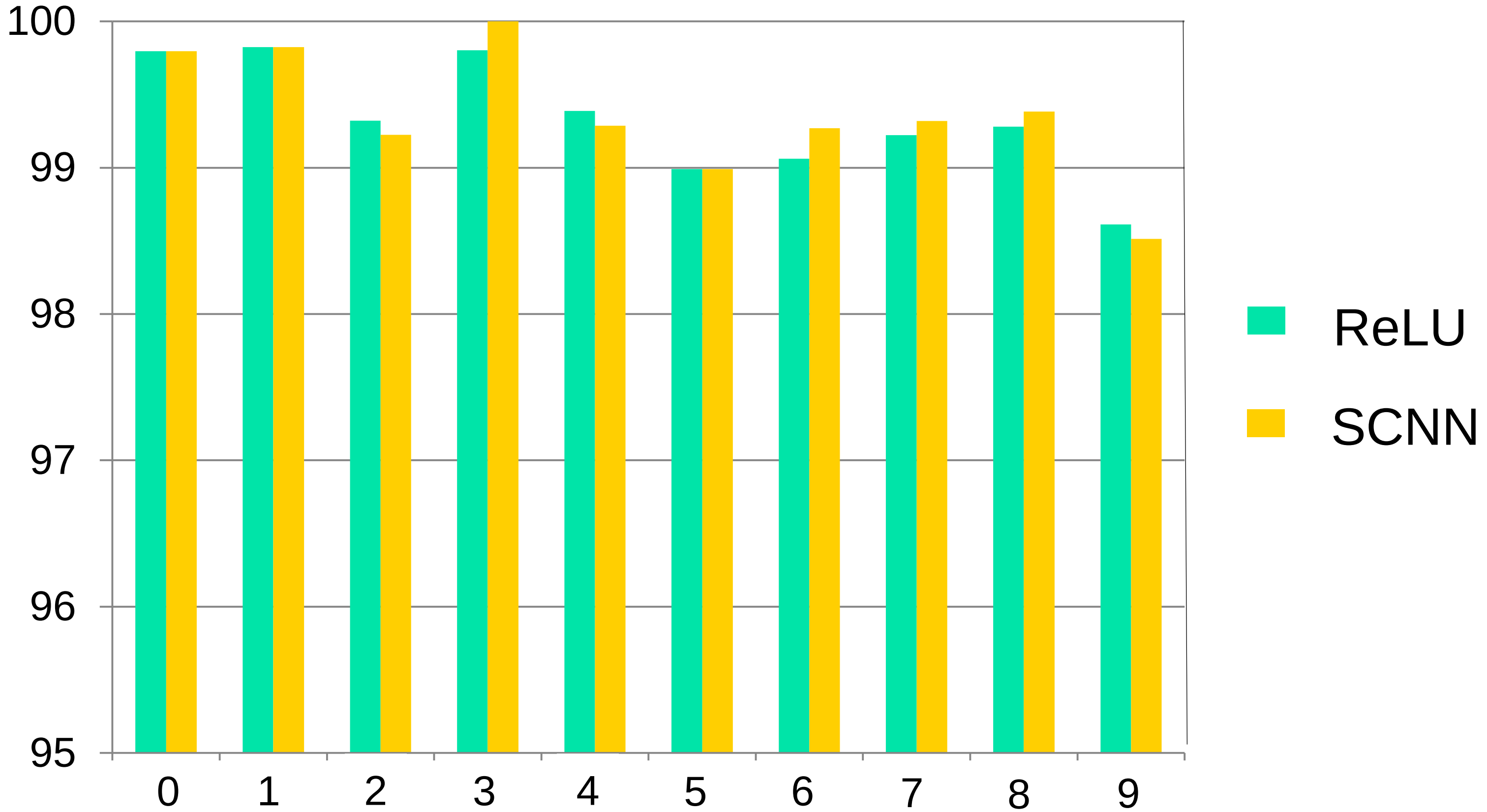


Time = 100
Error = 0.0062

High-level SCNN Simulation (4)

- MNIST Handwritten Digits – SCNN

<recognition rate for each digit>



High-level SCNN Simulation (5)

- MNIST Handwritten Digits – SCNN

<recognition rate vs. weight variation >

variation / trial	5%	10%	20%	30%
1	99.34	99.22	99.13	98.22
2	99.30	99.28	99.09	98.45
3	99.32	99.31	99.13	98.62
4	99.34	99.25	99.11	98.83
5	99.33	99.34	99.13	98.55
6	99.34	99.16	99.24	98.59
7	99.31	99.24	99.06	98.89
8	99.28	99.18	99.15	98.54
Average	99.32	99.25	99.13	98.59

Summary (1)

- The recent advancement of ANNs has been achieved by imitating the biological neural networks (BNNs) more closely. Spiking neural networks with STDP weight adjustment is the closest to the BNN.
- Combining the capacitor-less DRAM and SONOS flash memory, we have developed floating-body synaptic transistors (FSTs), which show short- and long-term memory and STDP.
- Resistive memory synapses are also investigated and nano-cone structures are proposed and fabricated for ultra-low power synapses.

Summary (2)

- Integrate-and-fire neuron circuits for FSTs are designed and fabricated.
- Various neuron circuits that can work with resistive memory synapses are discussed.
- System implementation scheme is designed and high-level simulation methods are developed for spiking neural networks with STDP capability.