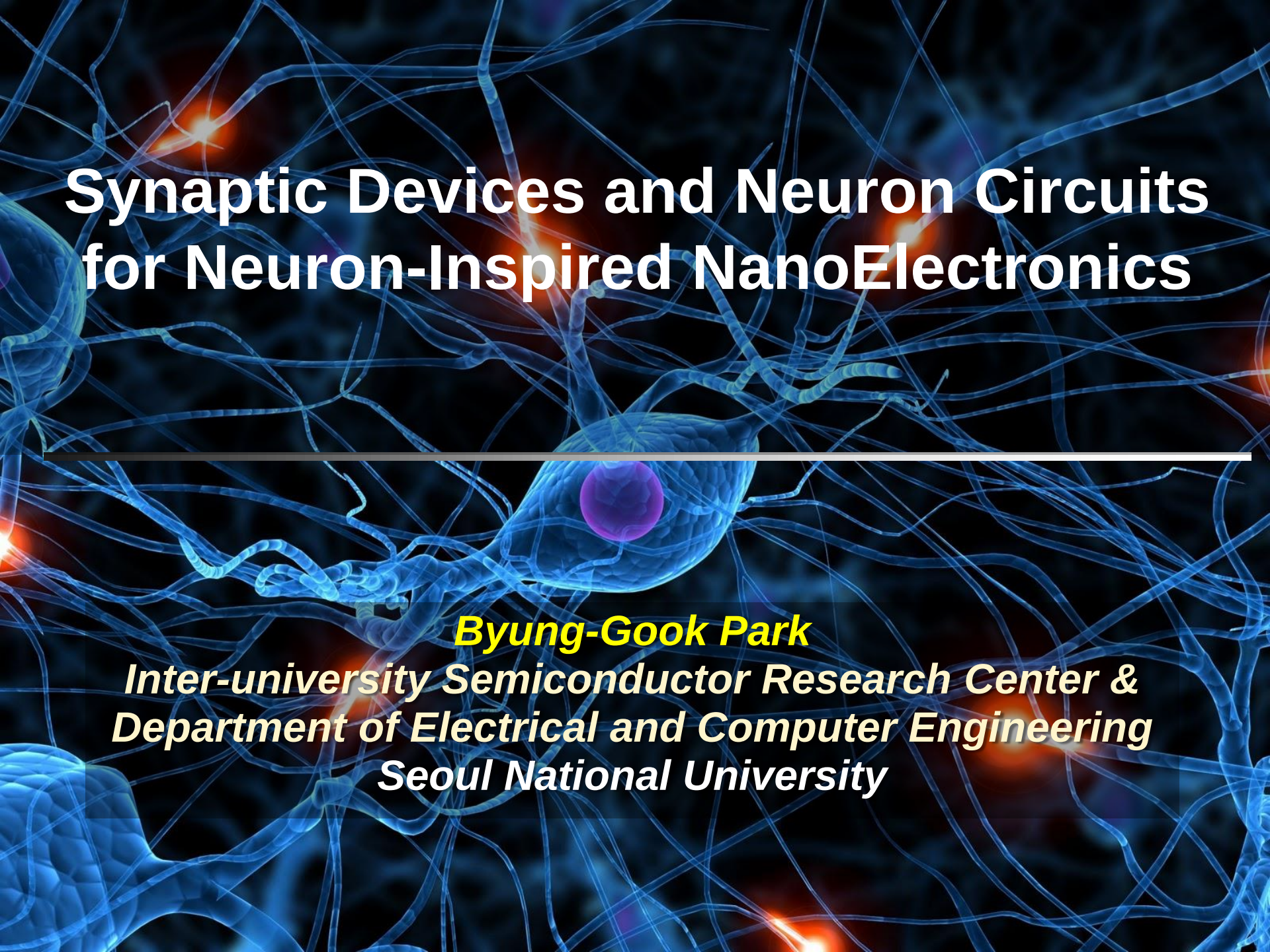


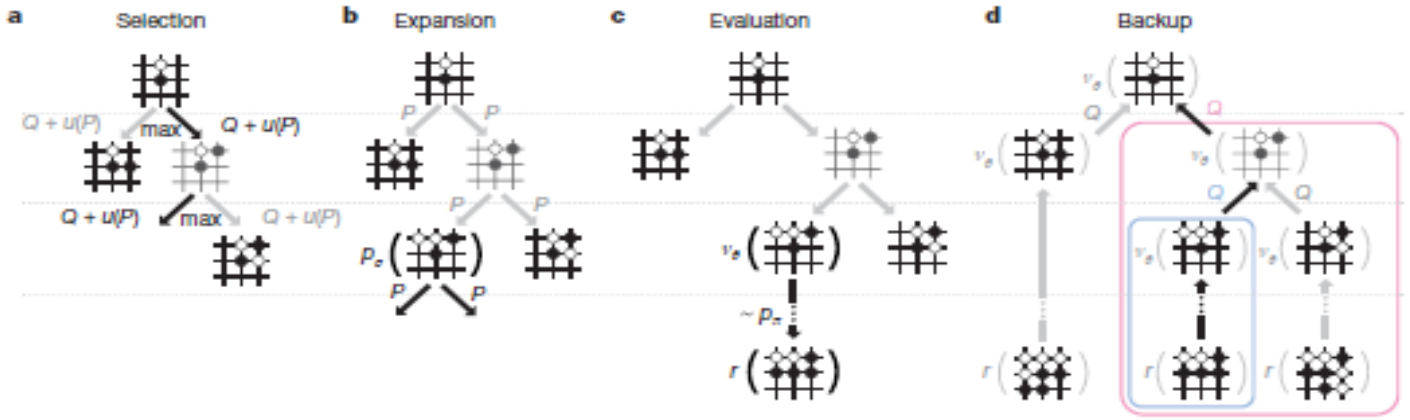
Synaptic Devices and Neuron Circuits for Neuron-Inspired NanoElectronics

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

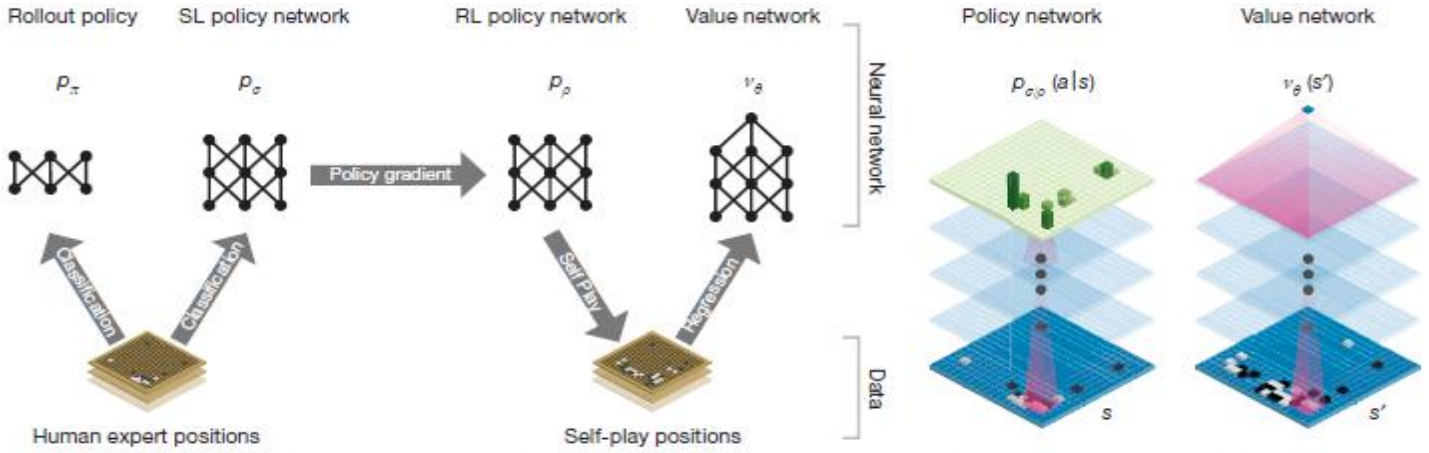
Human vs. AlphaGo

AlphaGo - Software

- Monte Carlo Tree Search (MCTS) <Silver, Nature (2016)>



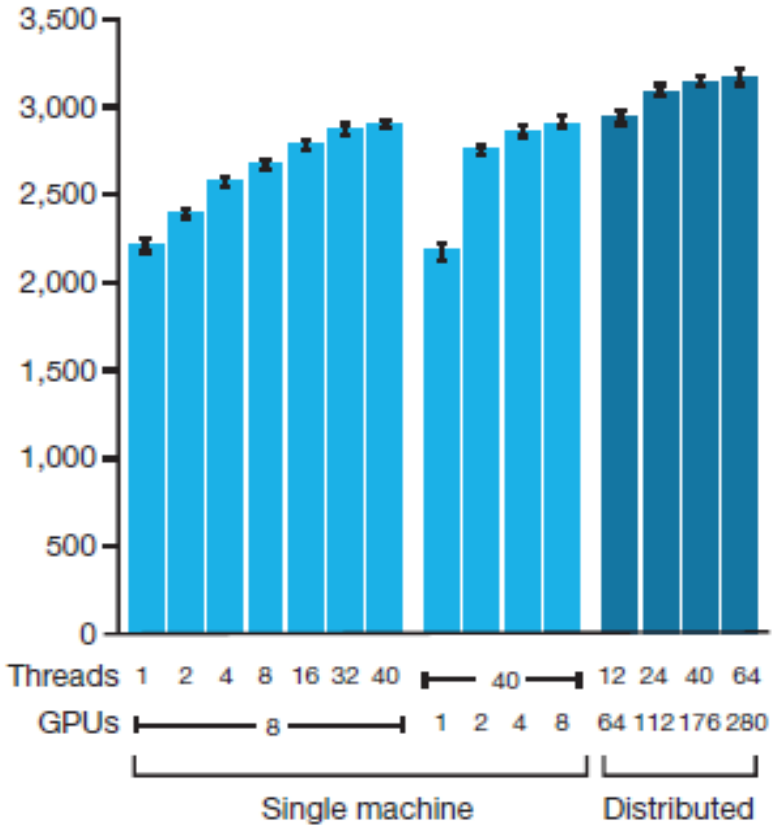
- Neural Networks



AlphaGo - Hardware

- Supercomputer

- Performance

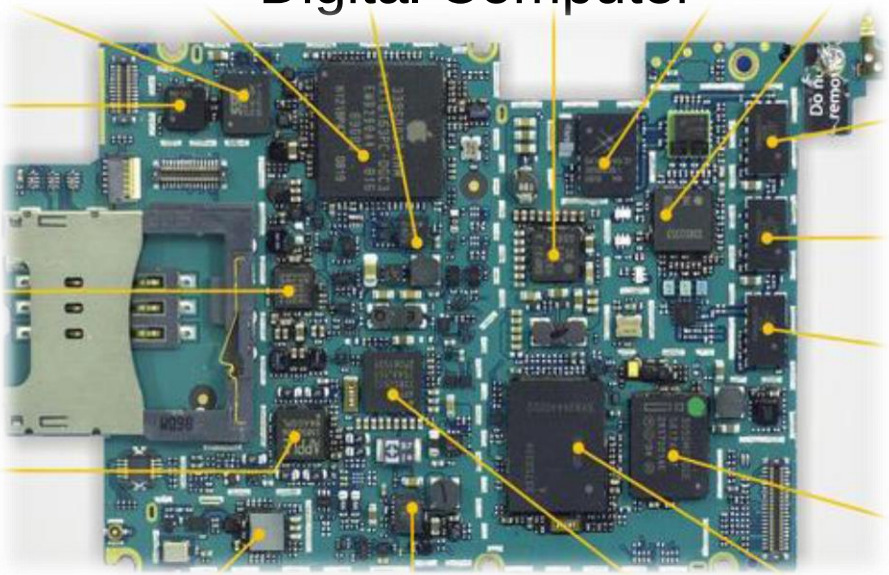


Comparison

- Human Brain

- neuron + synapse
- massively parallel
- ~ ms speed
- low power
- recognition/reasoning
- self-organized

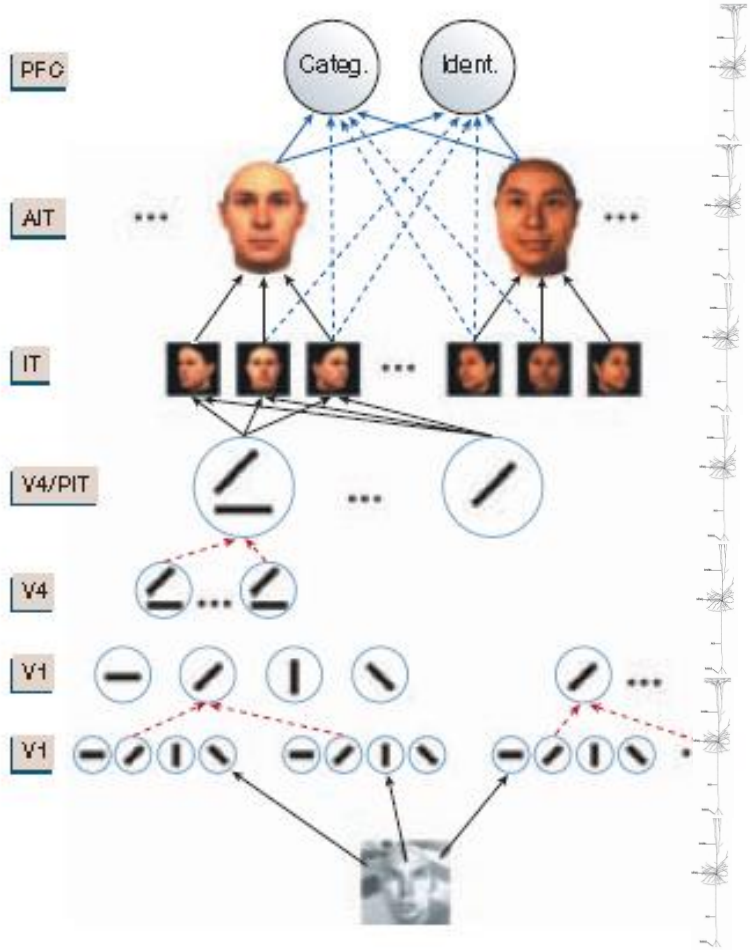
- Digital Computer



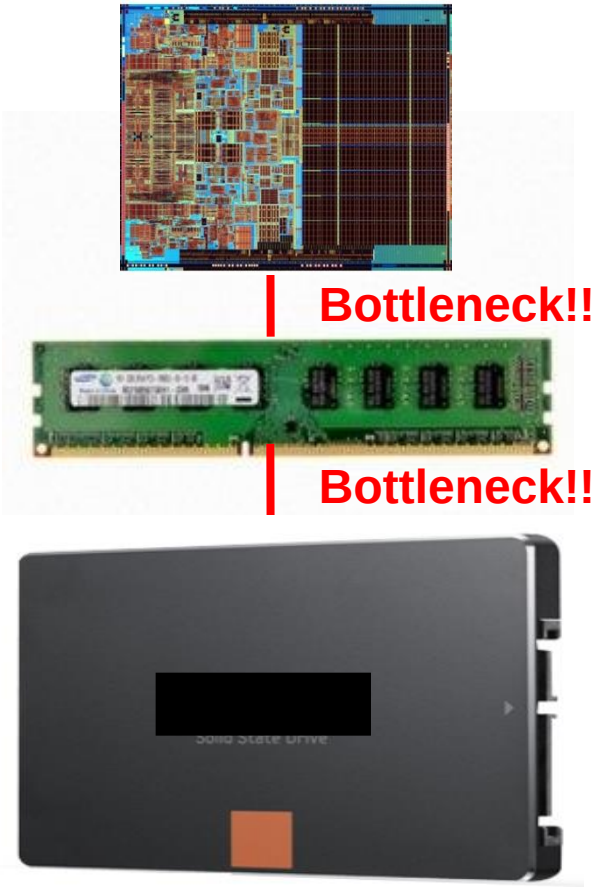
- CPU + memory
- serial
- ~ ns speed
- high power
- computation
- manufacturing

Interconnection Bottleneck

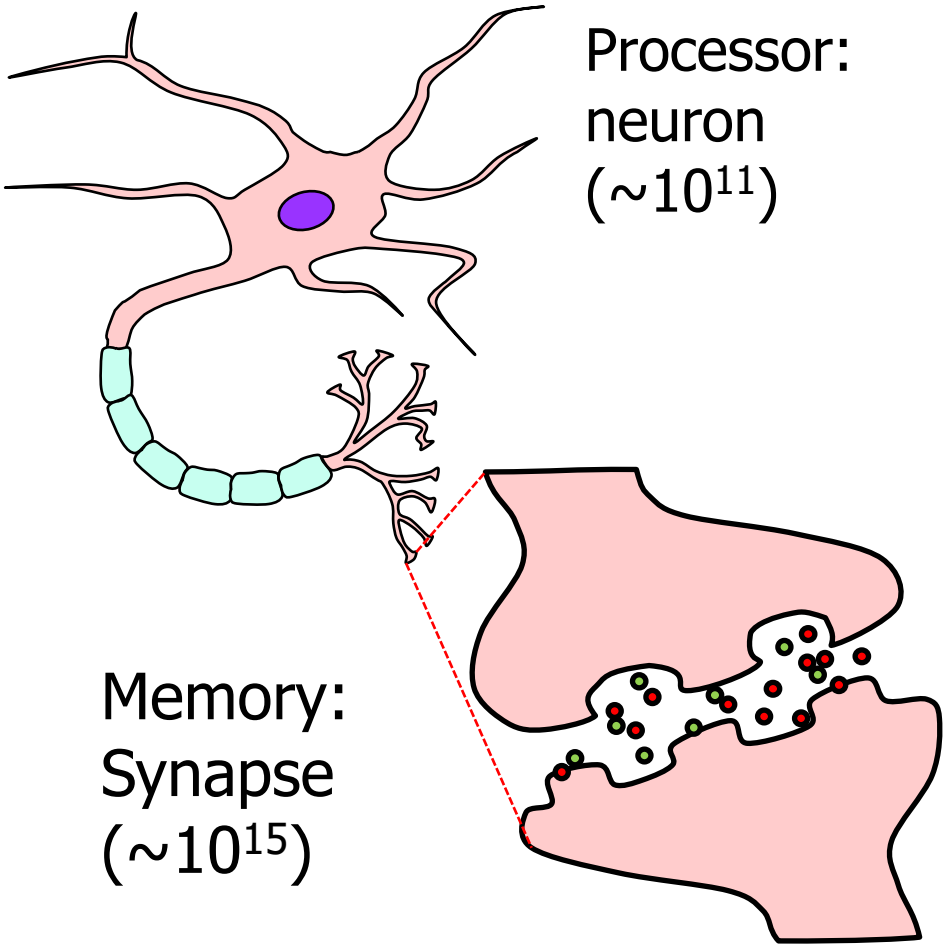
- Human Brain (Face Recognition)



- Computer System

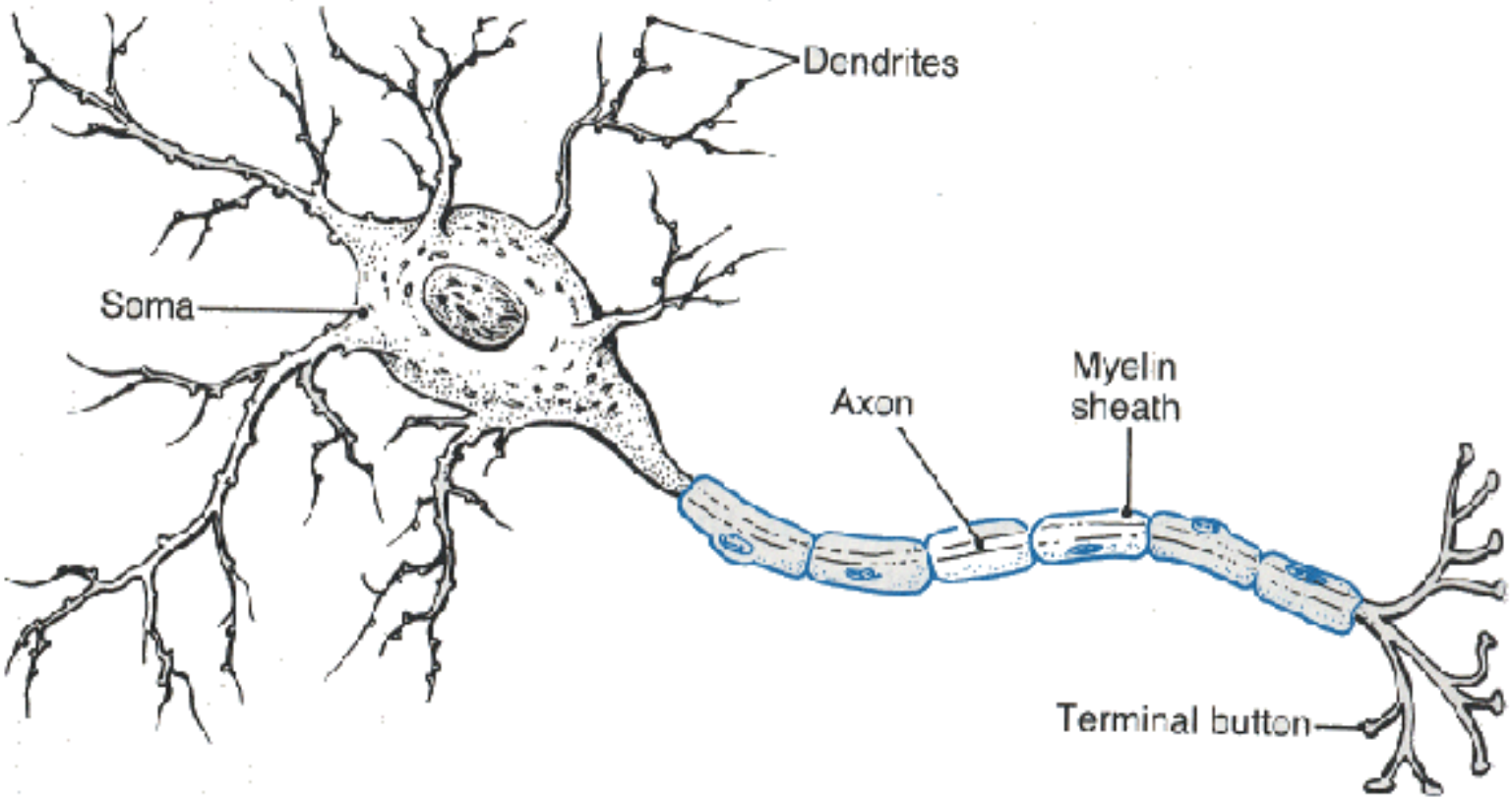


Human Brain and Its Building Blocks



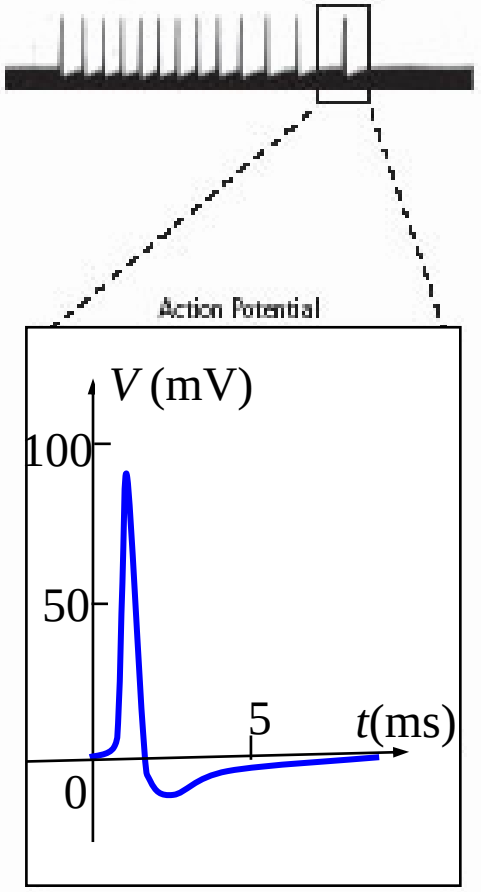
Neuron (1)

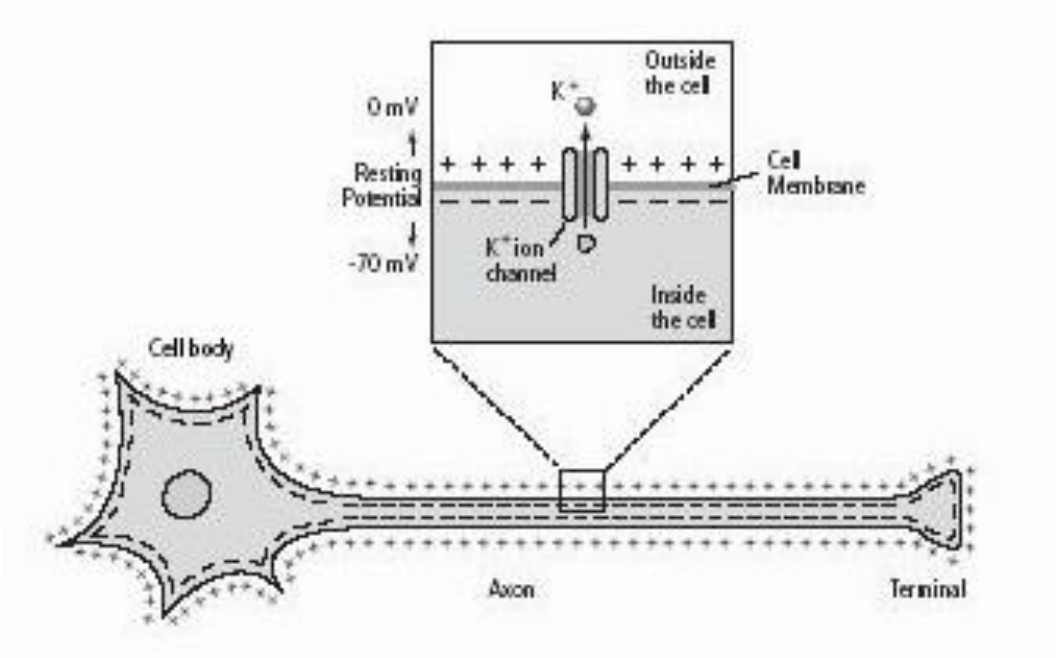
- Structure of a neuron



Neuron (2)

- Structure of a neuron

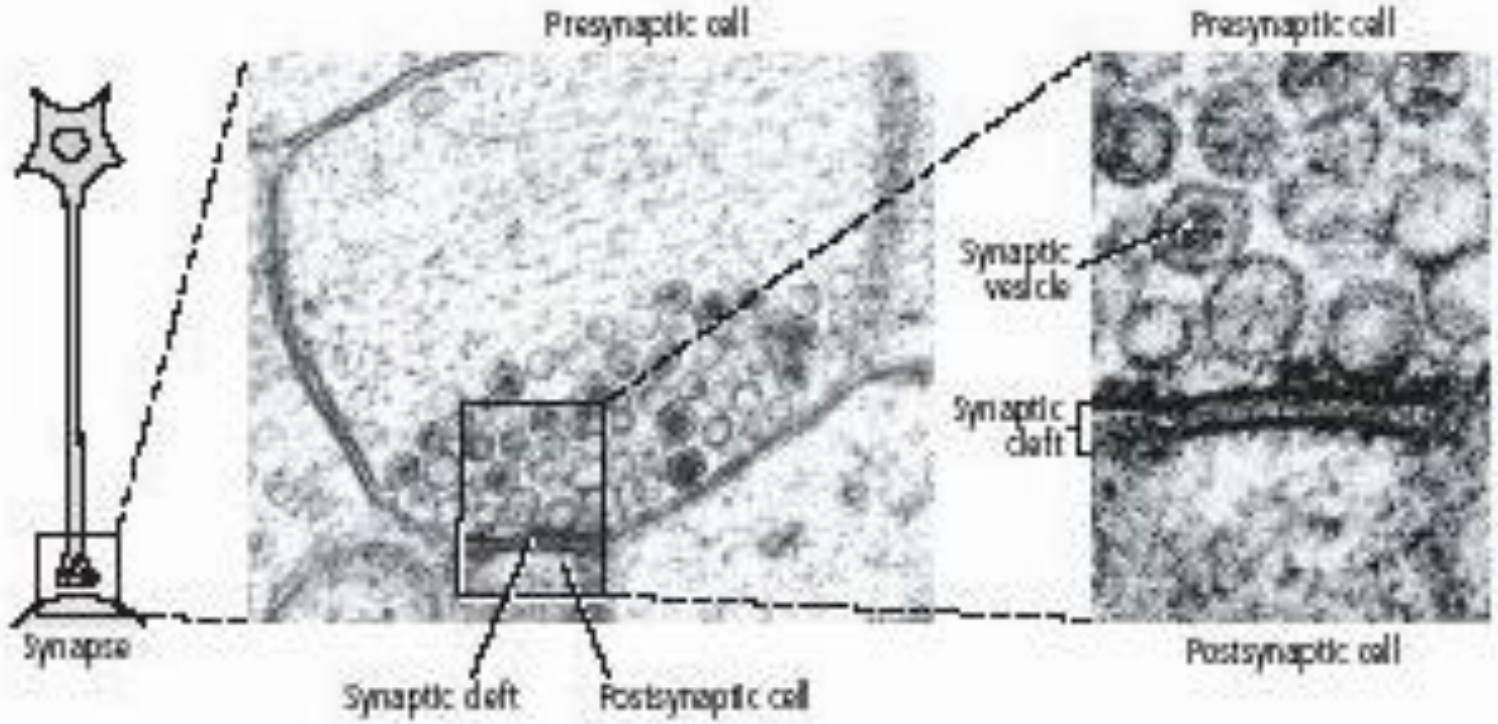




Potential change due to K^+ ion channel

Synapse (1)

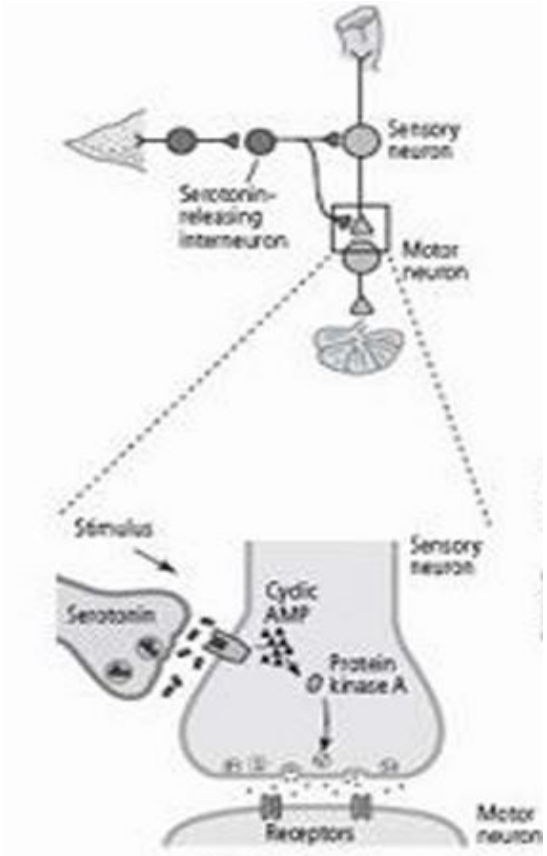
- Electrochemical signal transmission in a synapse

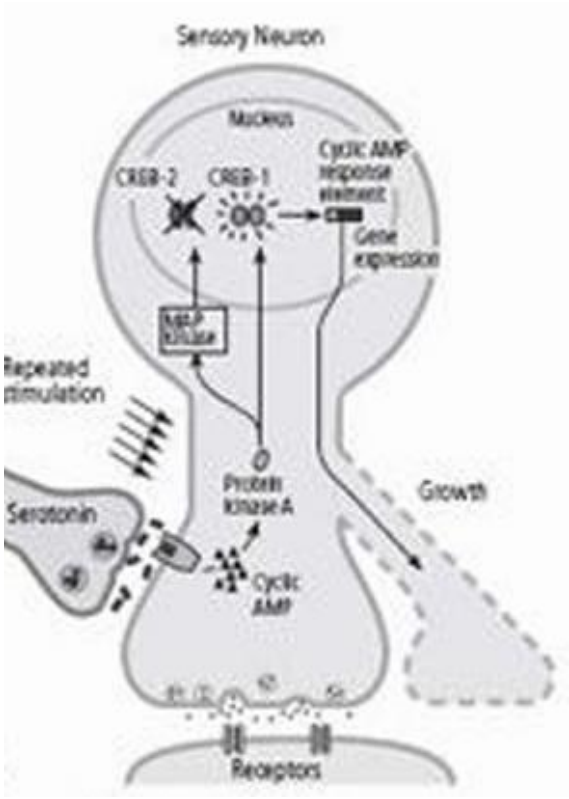


Neurotransmitter is emitted in the unit of a vesicle → conveyed to the post-synaptic cell through the cleft

Synapse (2)

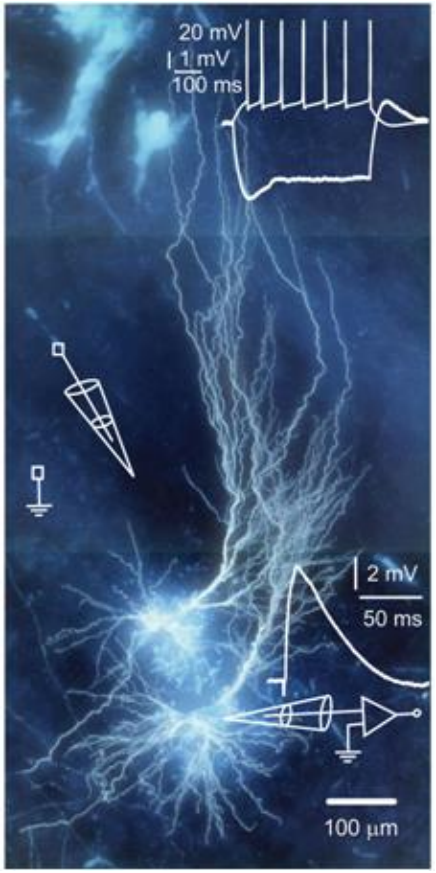
- Short term memory and long term memory

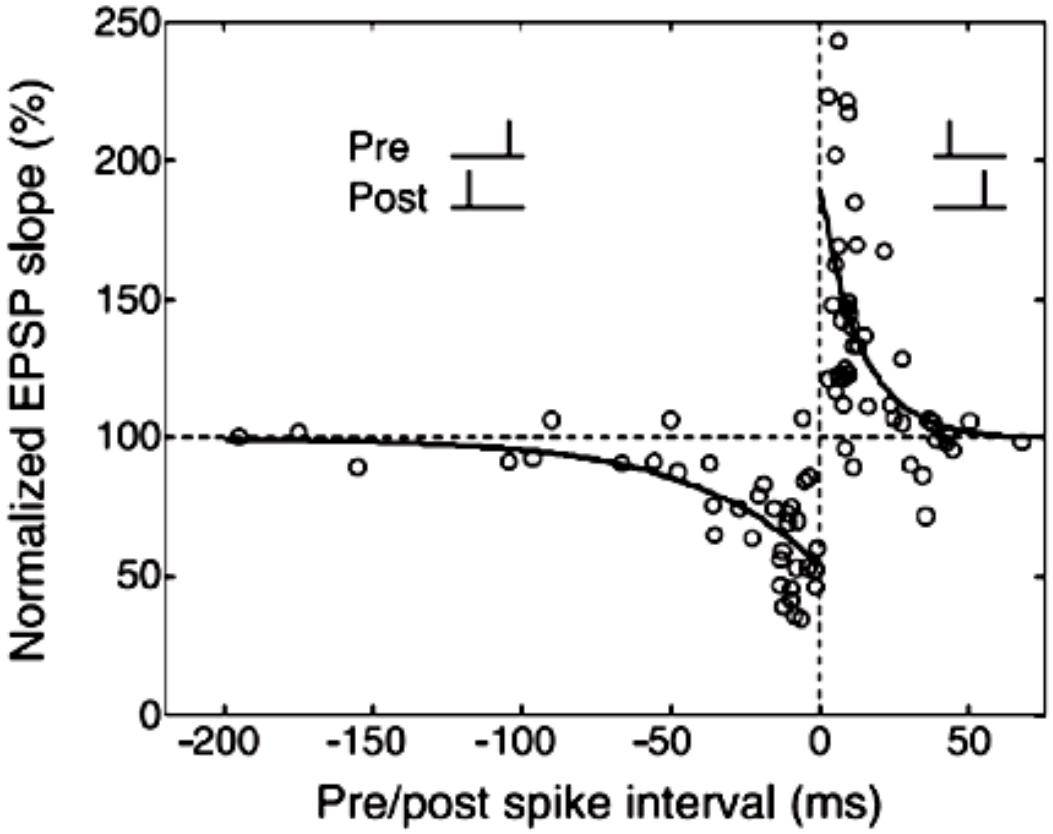




Spike-Timing-Dependent Plasticity

- Spike-timing-dependent plasticity – learning mechanism

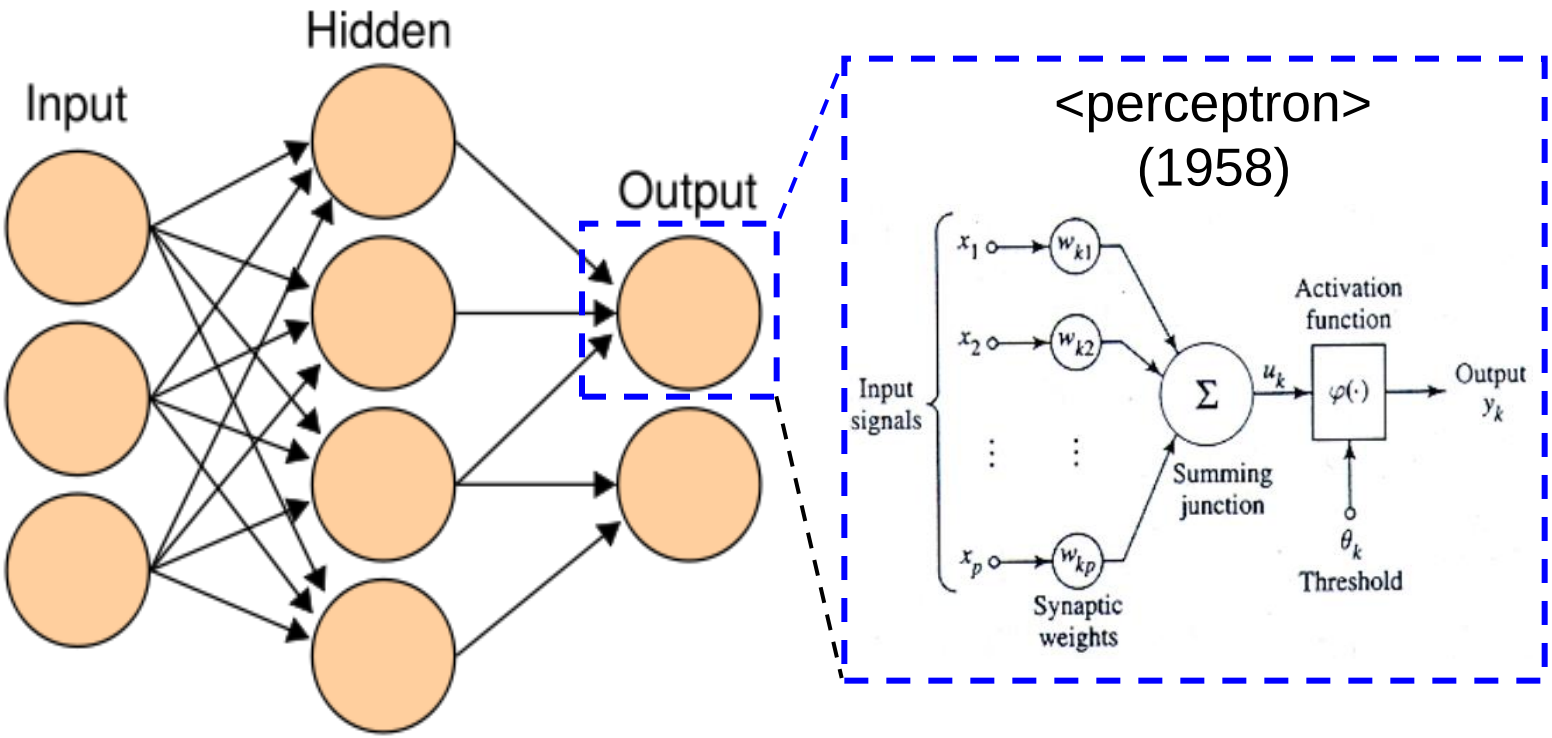




Building Blocks: Biological vs. Electronic

Artificial Neural Network (ANN)

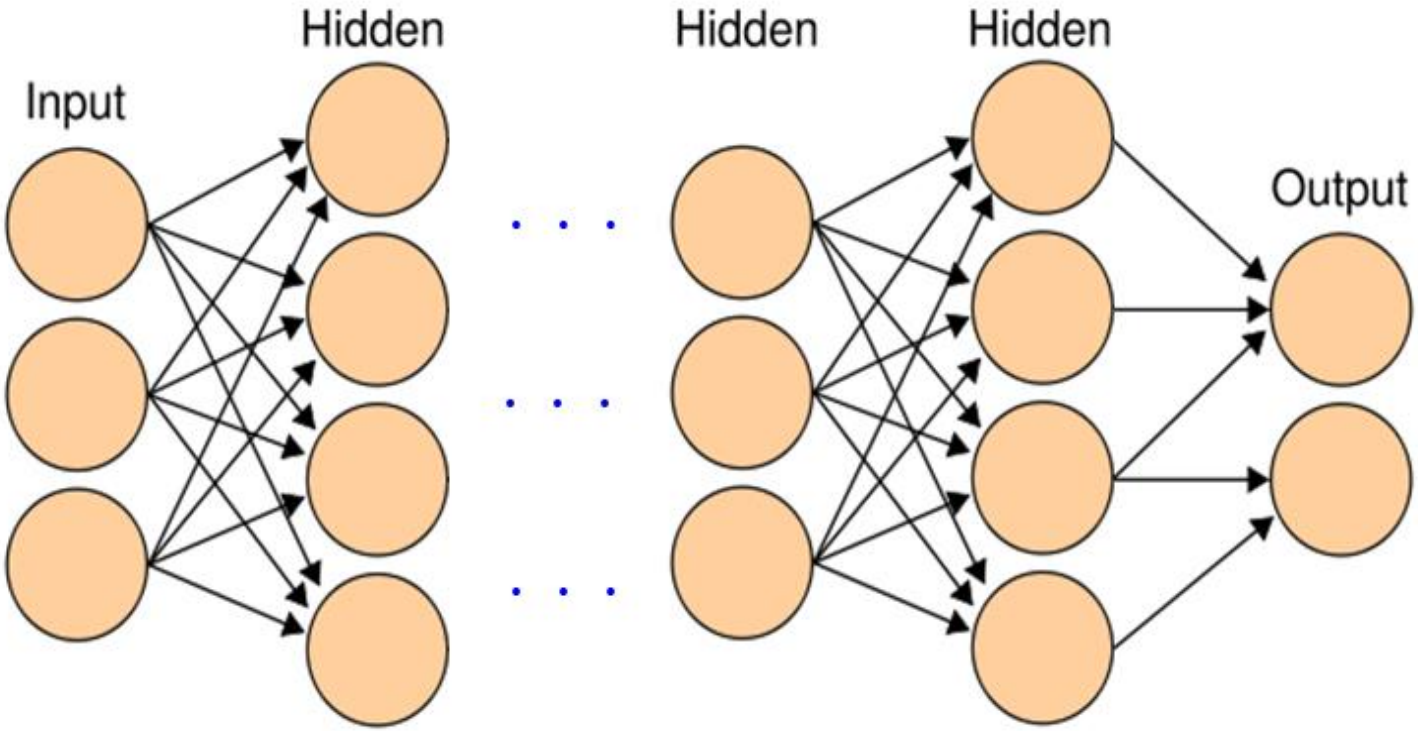
- Concept of neural network



** Weights are calculated by the back-propagation (BP) method (1986).

Deep Neural Network (DNN)

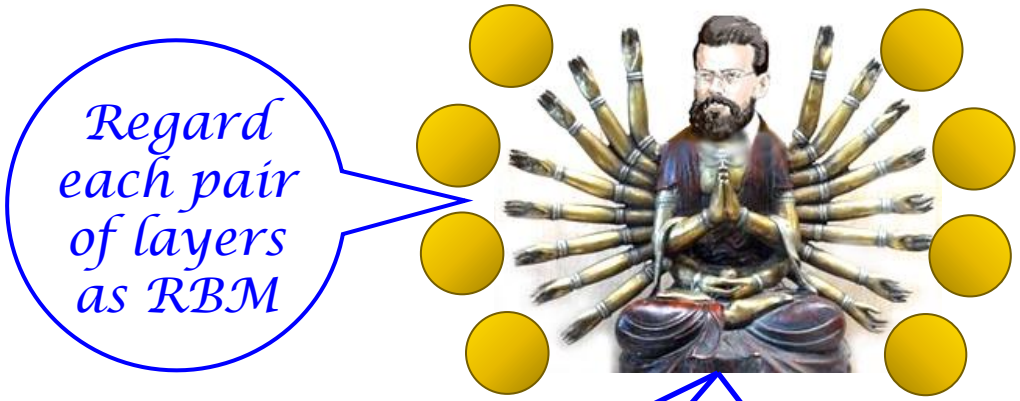
- Multiple hidden layers



** Vanishing gradient problem (VGP) → unsupervised pre-training (RBM) + modified activation function (ReLU)

1st Breakthrough (2006)

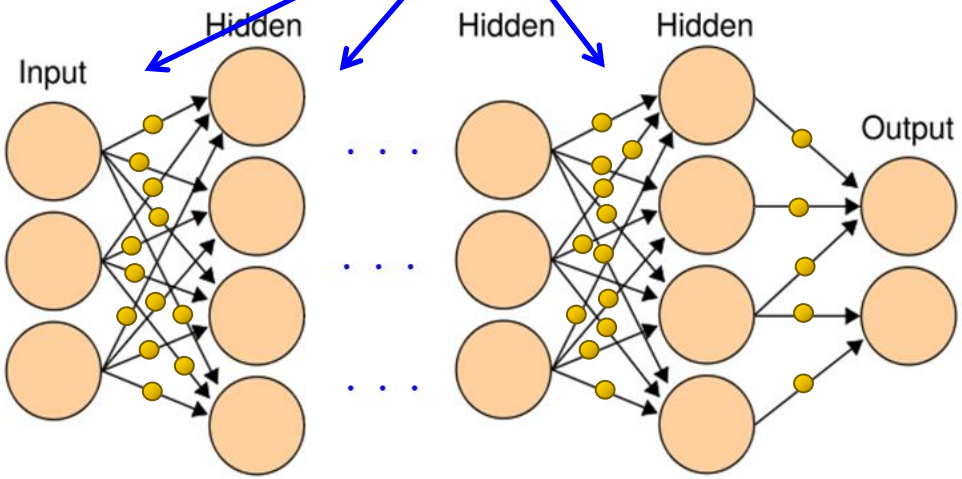
- Pre-training with restricted Boltzmann machine (RBM)



$$E_\theta(\mathbf{x}) = -\mathbf{b}^T\mathbf{x} - \frac{1}{2}\mathbf{x}^T\mathbf{W}\mathbf{x}$$

$$P_\theta(\mathbf{x}) = \frac{\exp[-E_\theta(\mathbf{x})/\tau]}{\sum_{\mathbf{x}'}\exp[-E_\theta(\mathbf{x}')/\tau]}$$

$$W_{ij} \leftarrow W_{ij} + \alpha\frac{\partial(\sum \ln P_\theta)}{\partial W_{ij}}$$



2nd 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> <ReLU> <Sigmoid>

512x512
Image

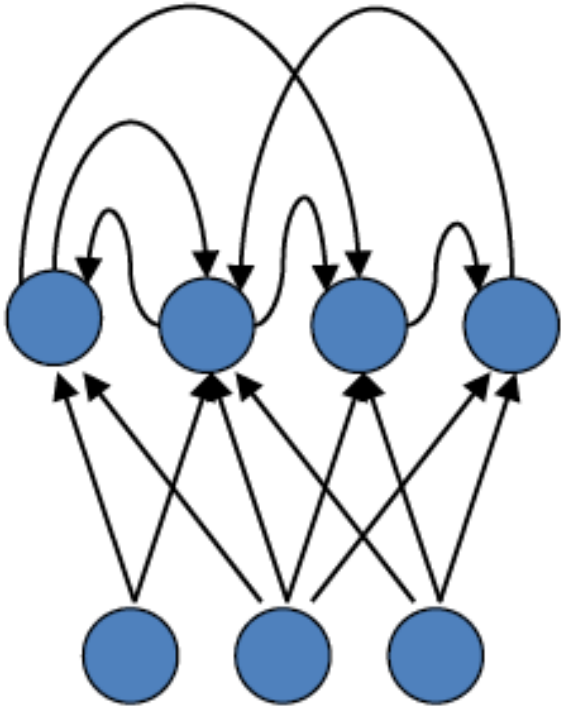
Epoch = 800
MSE = 0.00093

Epoch = 800
MSE = 0.00142

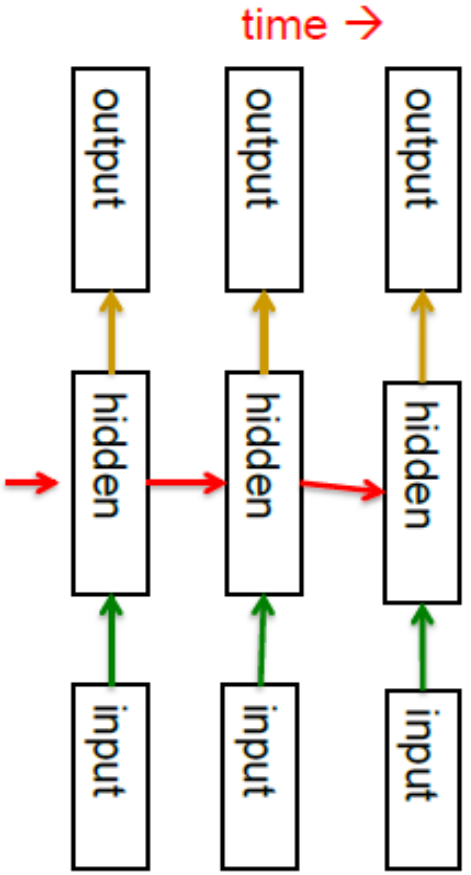
** MSE (mean square error)

Recurrent Neural Network (RNN)

- Directed cycles in connections
- Sequential data modeling

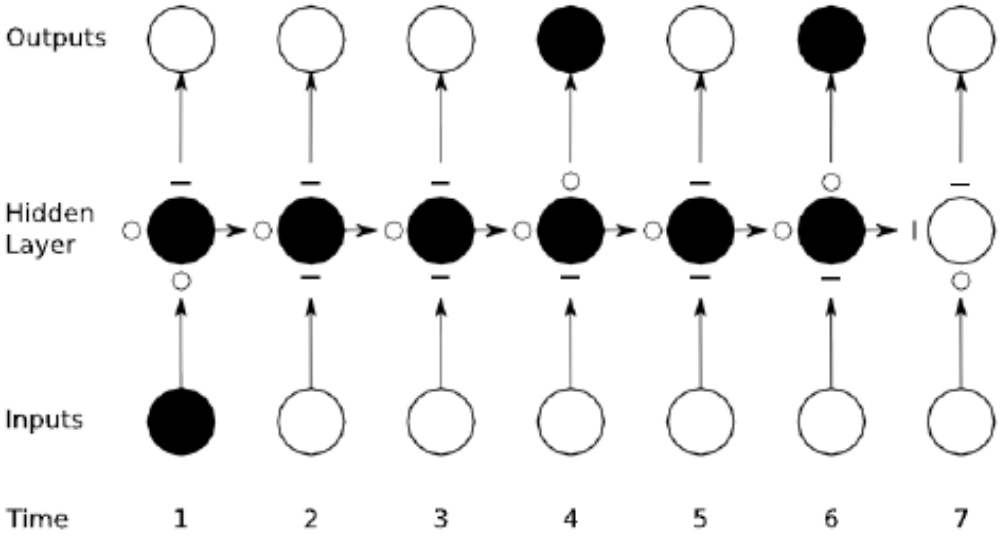


** Complicated dynamics and difficulty in training

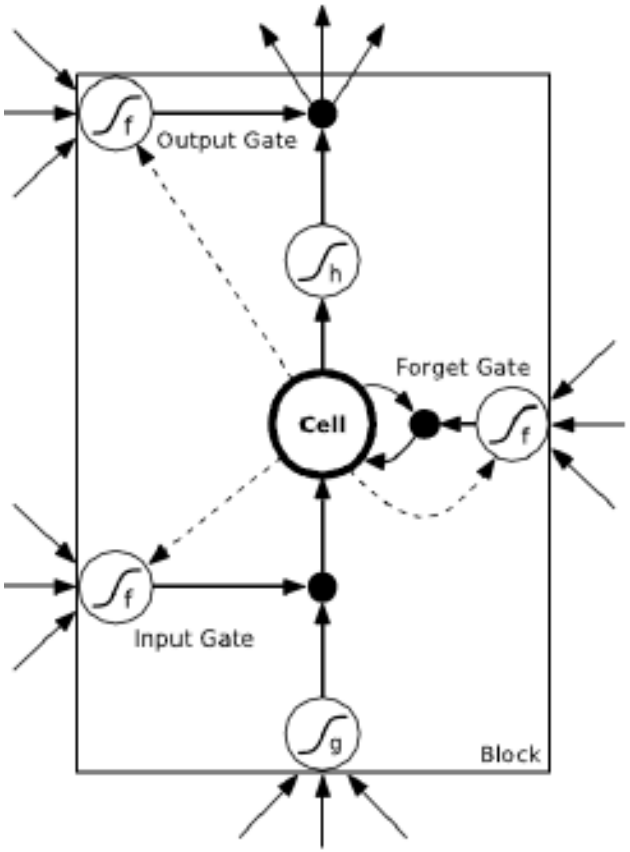


Long Short-Term Memory (LSTM)

- RNN with LSTM



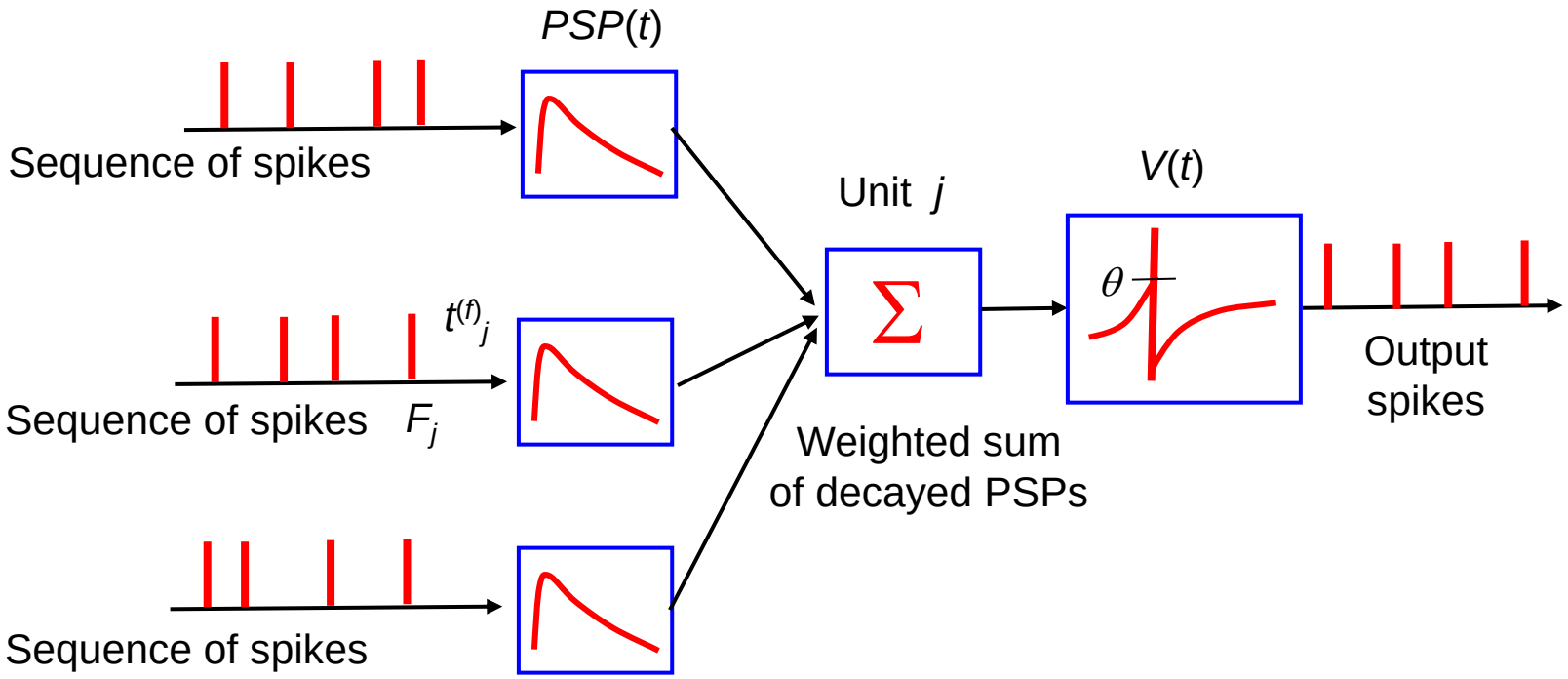
- LSTM block



** Hidden layer units with LSTM can store and access information over a long period of time.

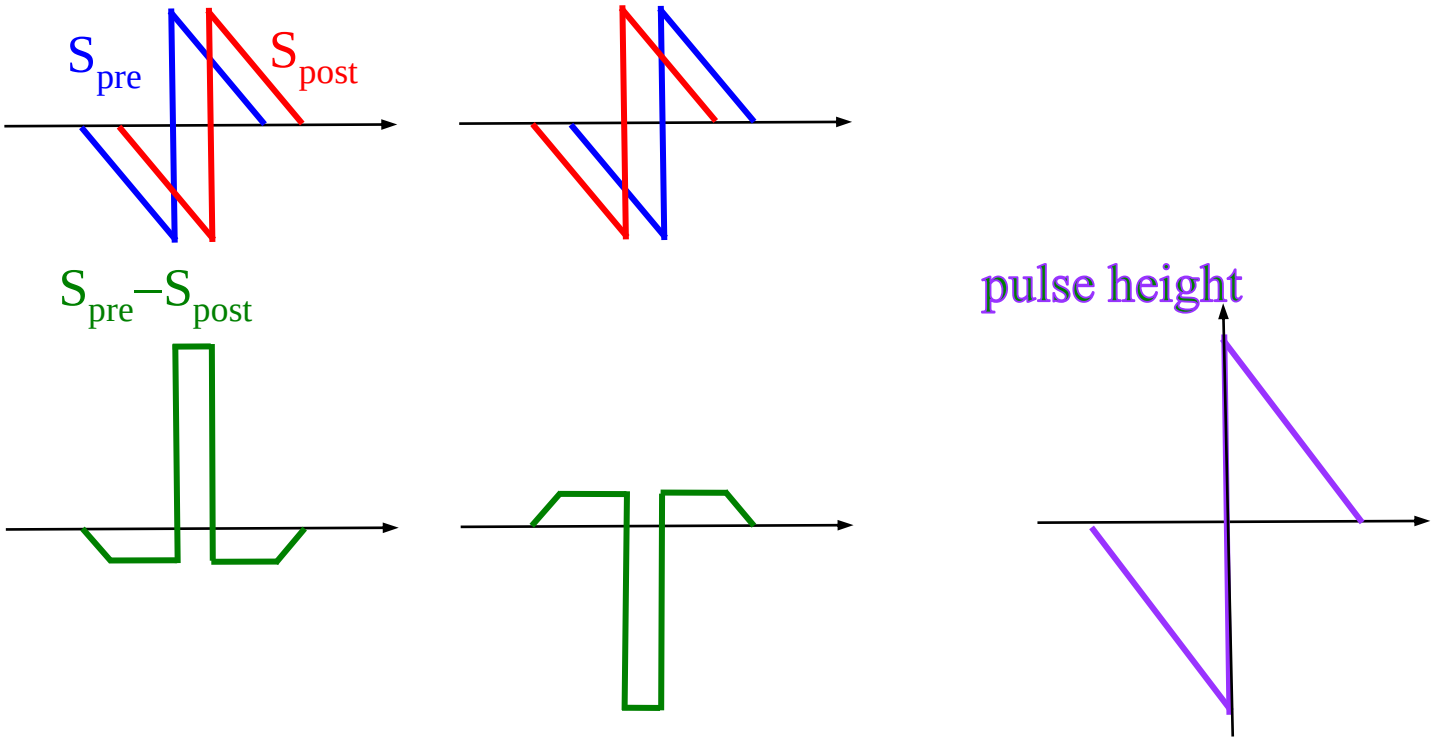
Spiking Neural Network (SNN) (1)

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



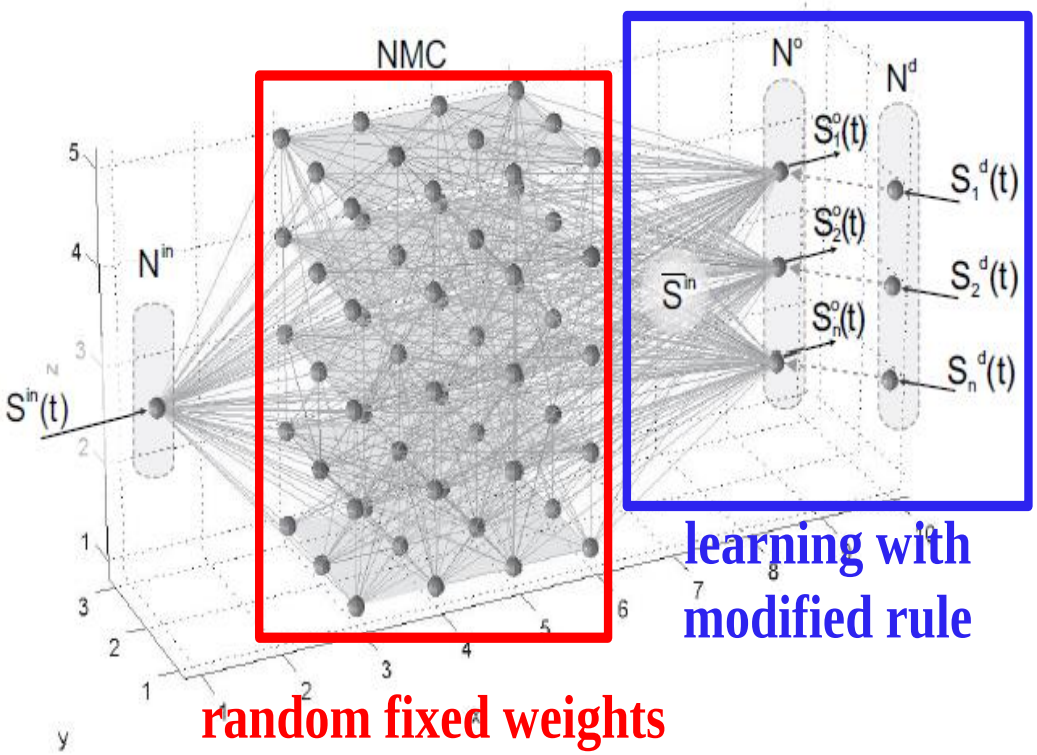
Spiking Neural Network (SNN) (2)

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

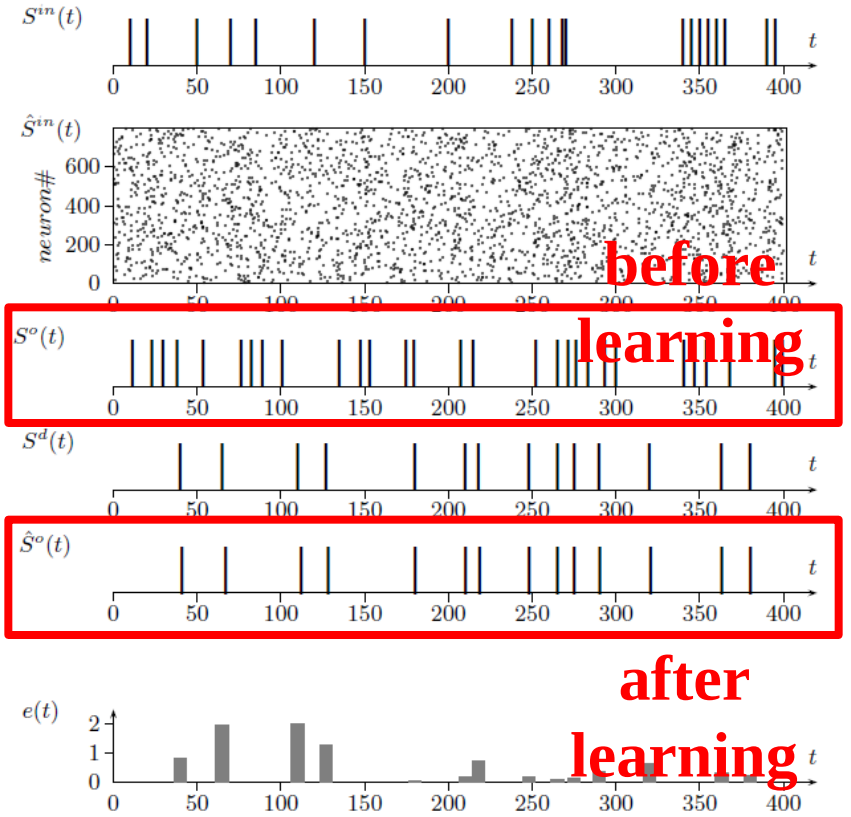


Spiking Neural Network (SNN) (3)

- Structure



- Supervised learning

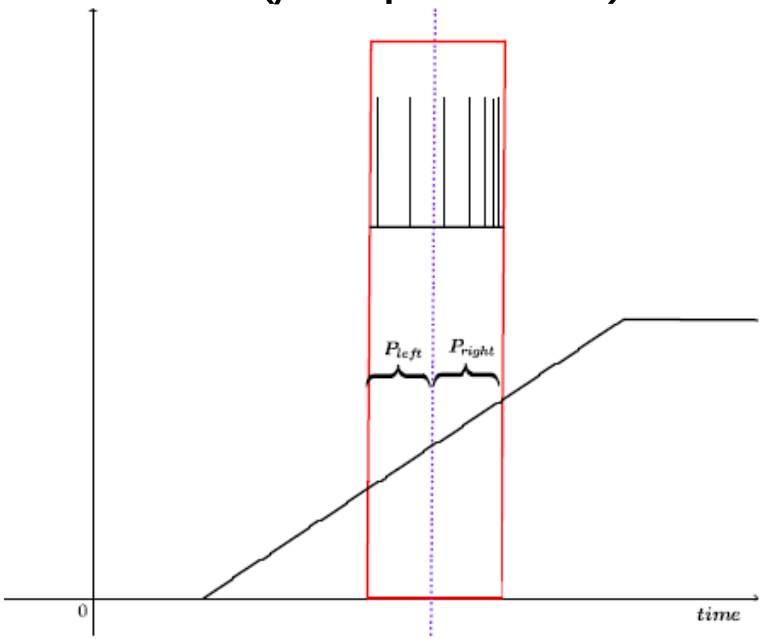


STDP and Error Back-propagation

- STDP

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

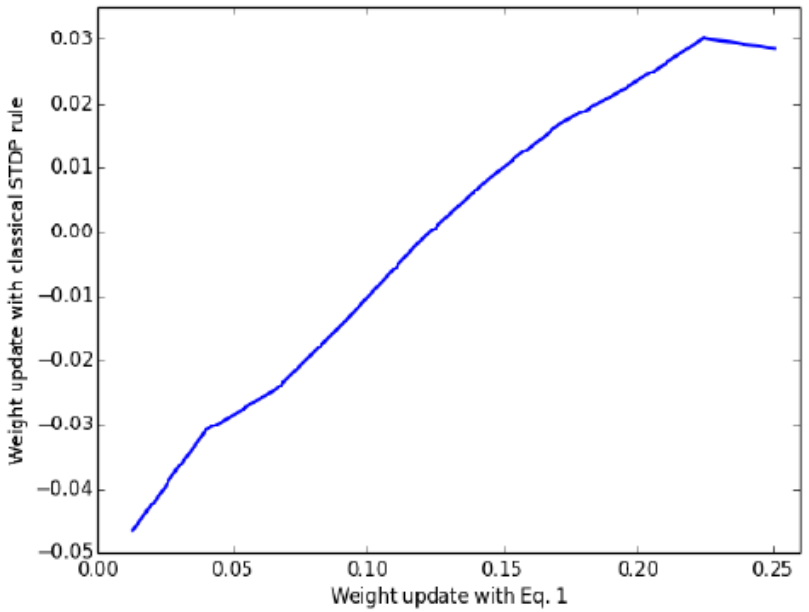
(ρ : spike rate)



- BP

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

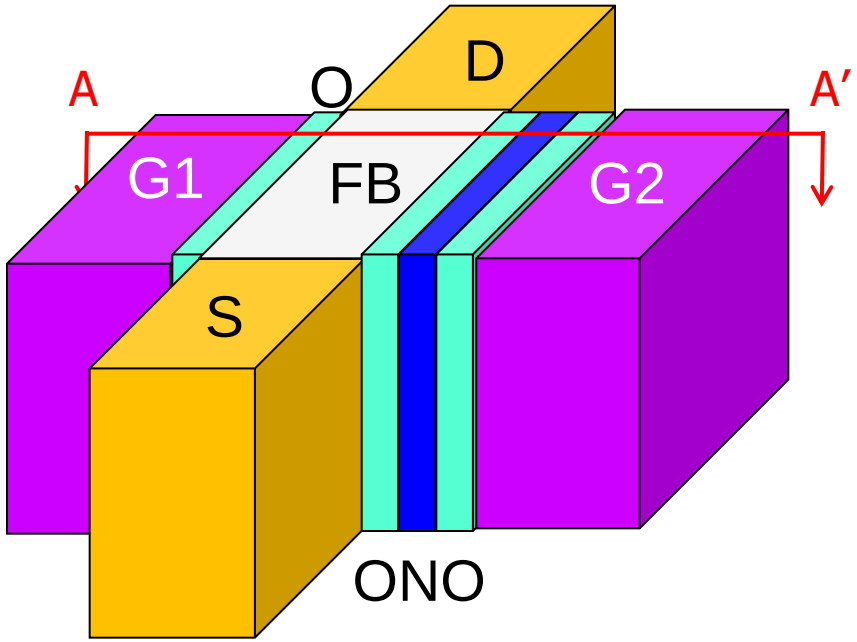
(x : neuron output)



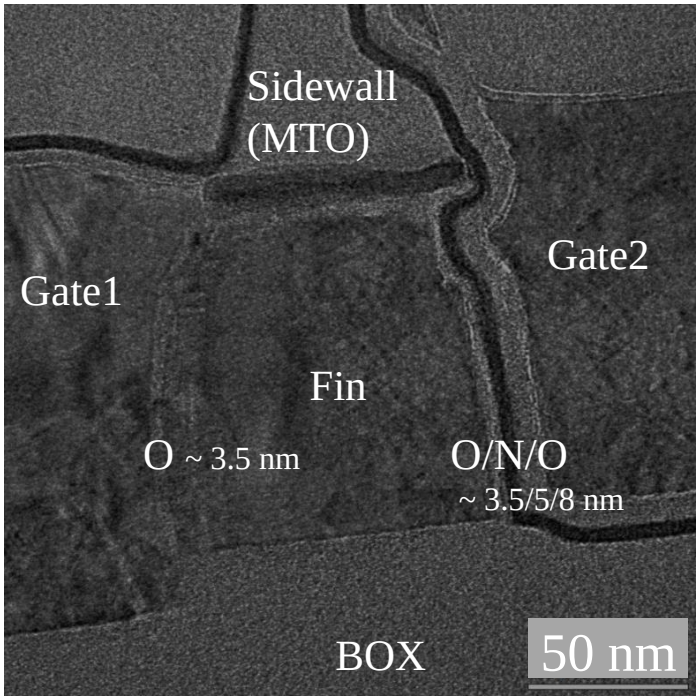
<Bengio, arXiv.org, (2016)>

Floating-body Synaptic Transistor (1)

- Structure



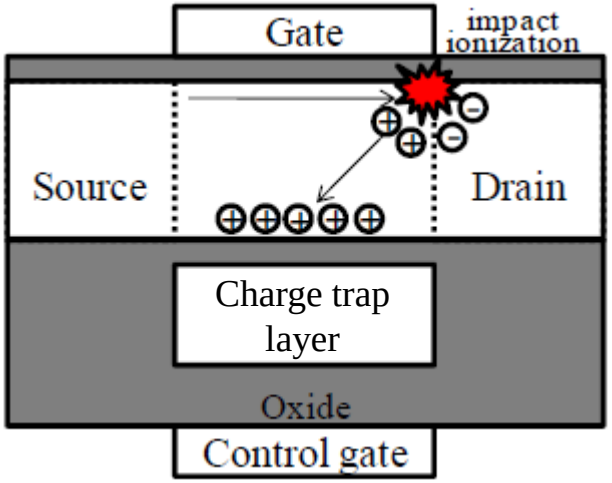
- TEM image



  - Cross-section in A – A' direction

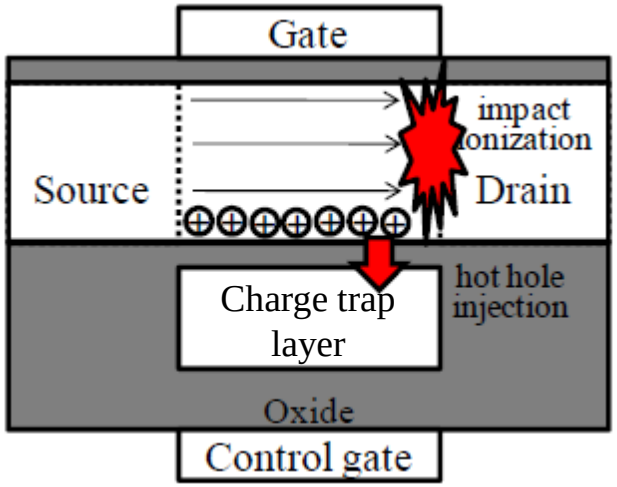
Floating-body Synaptic Transistor (2)

- Short-term memorization



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

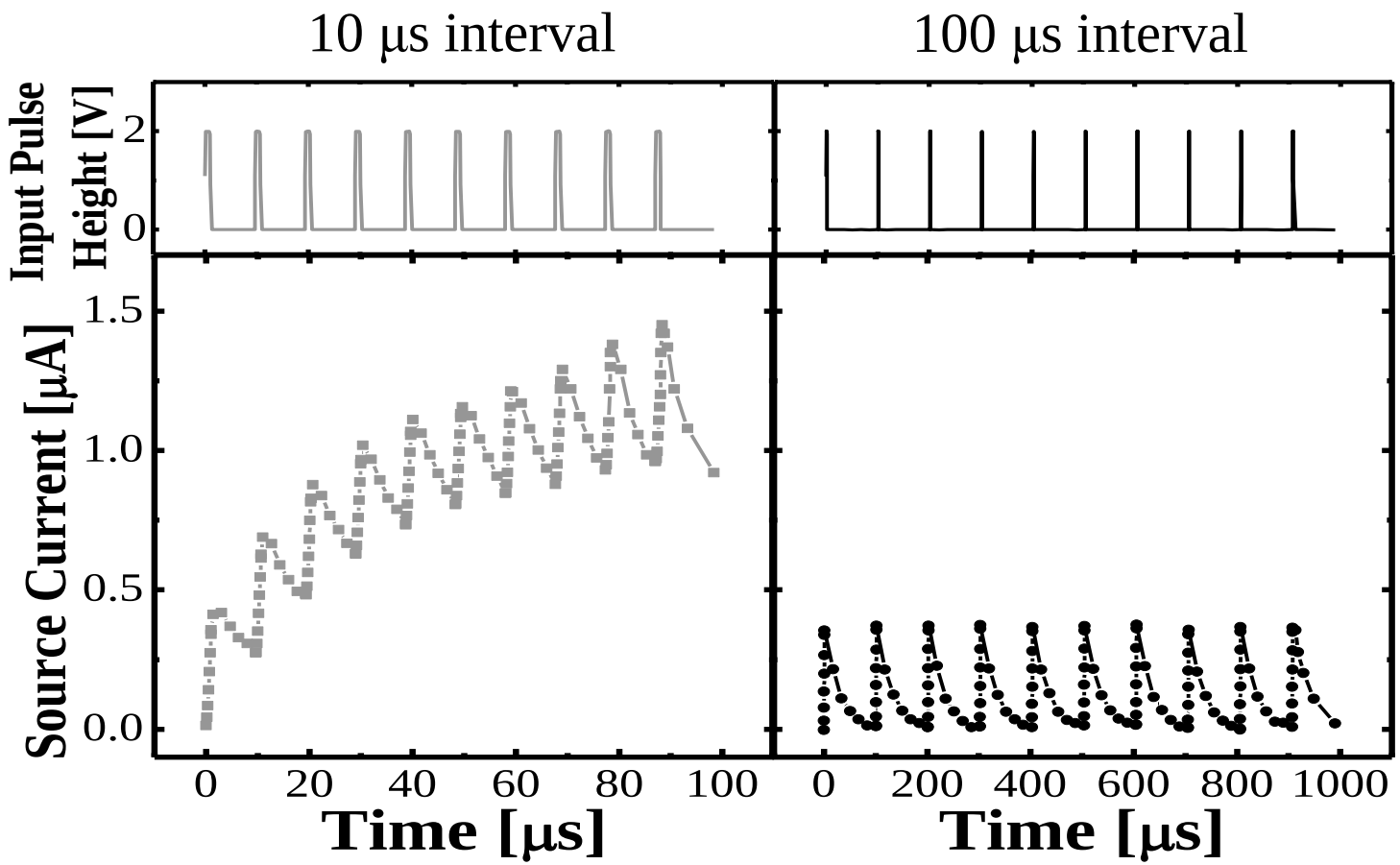
- Long-term memorization



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

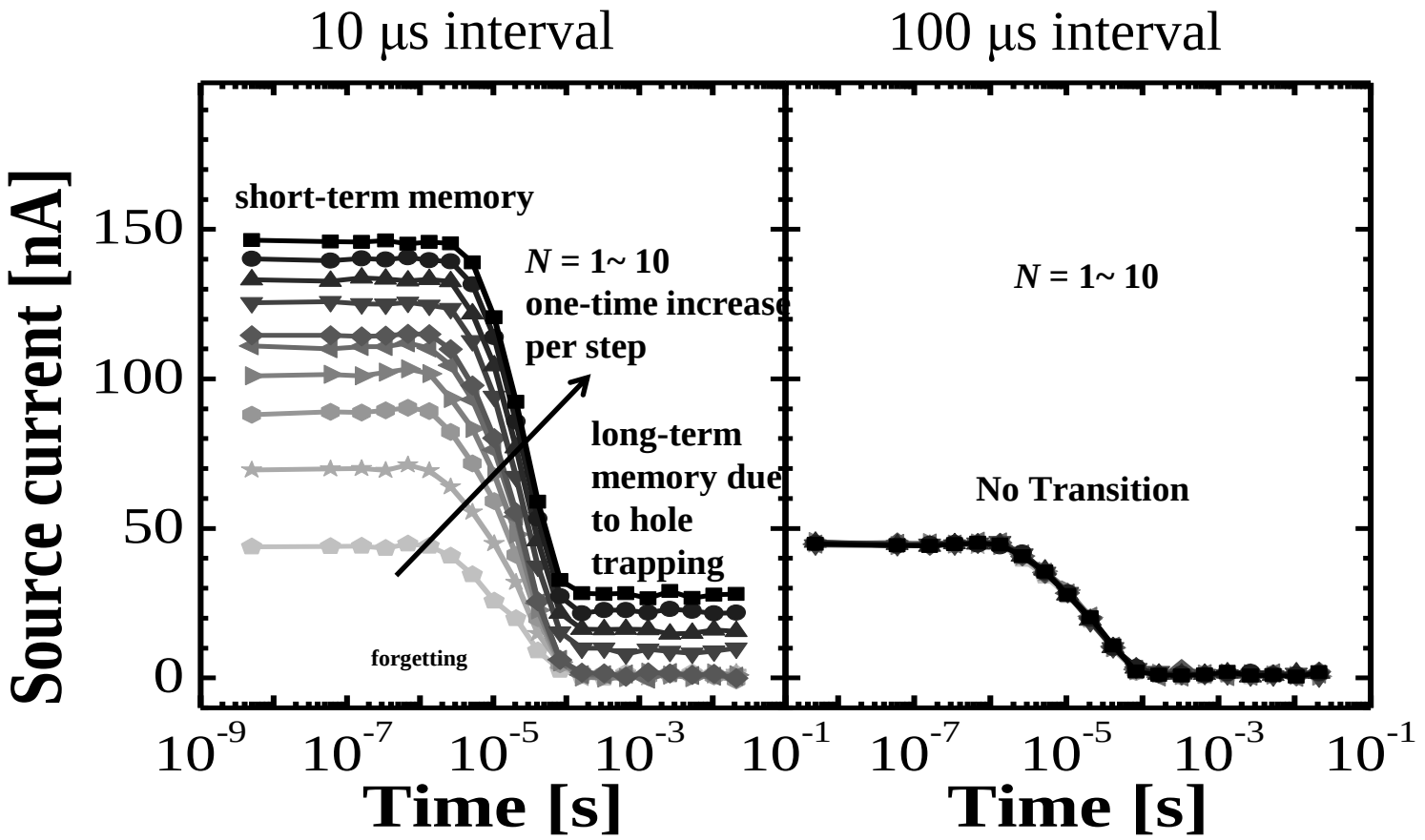
Floating-body Synaptic Transistor (3)

- Transient response of FST to spikes



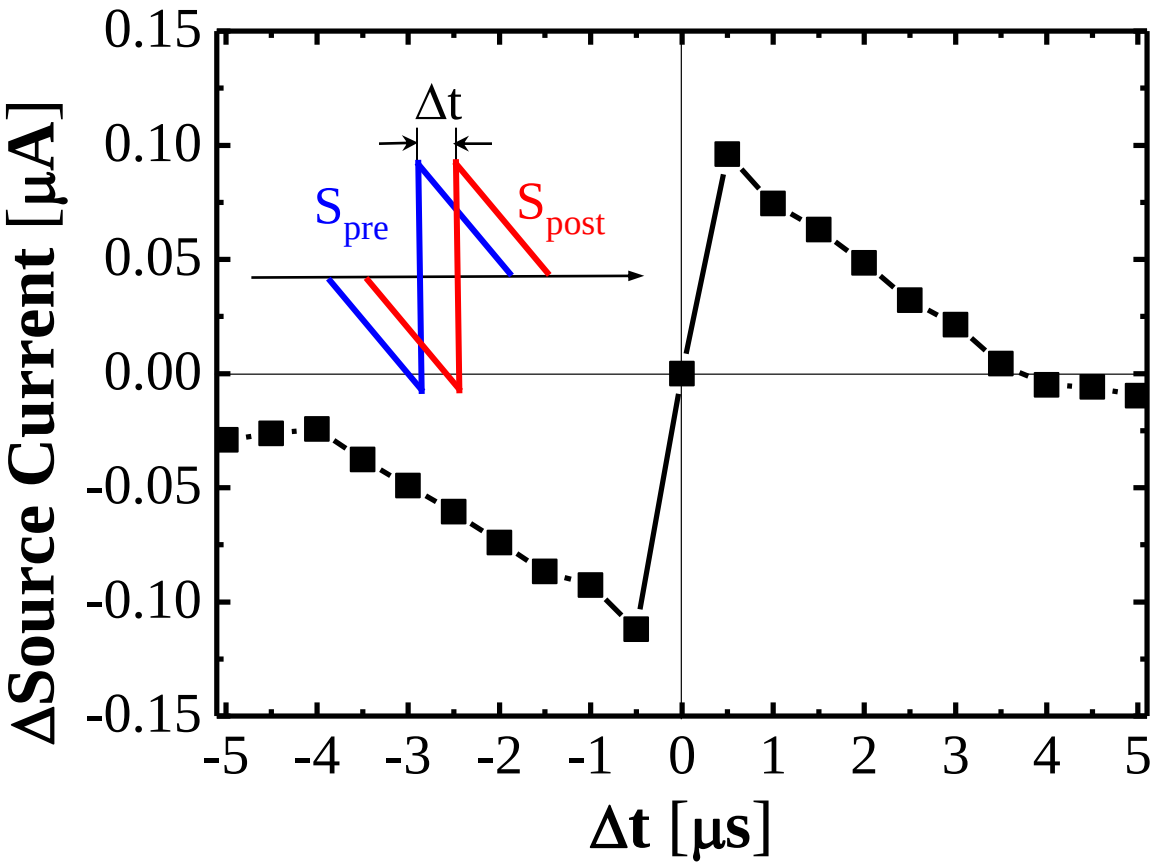
Floating-body Synaptic Transistor (4)

- Short-term to long-term memory transition



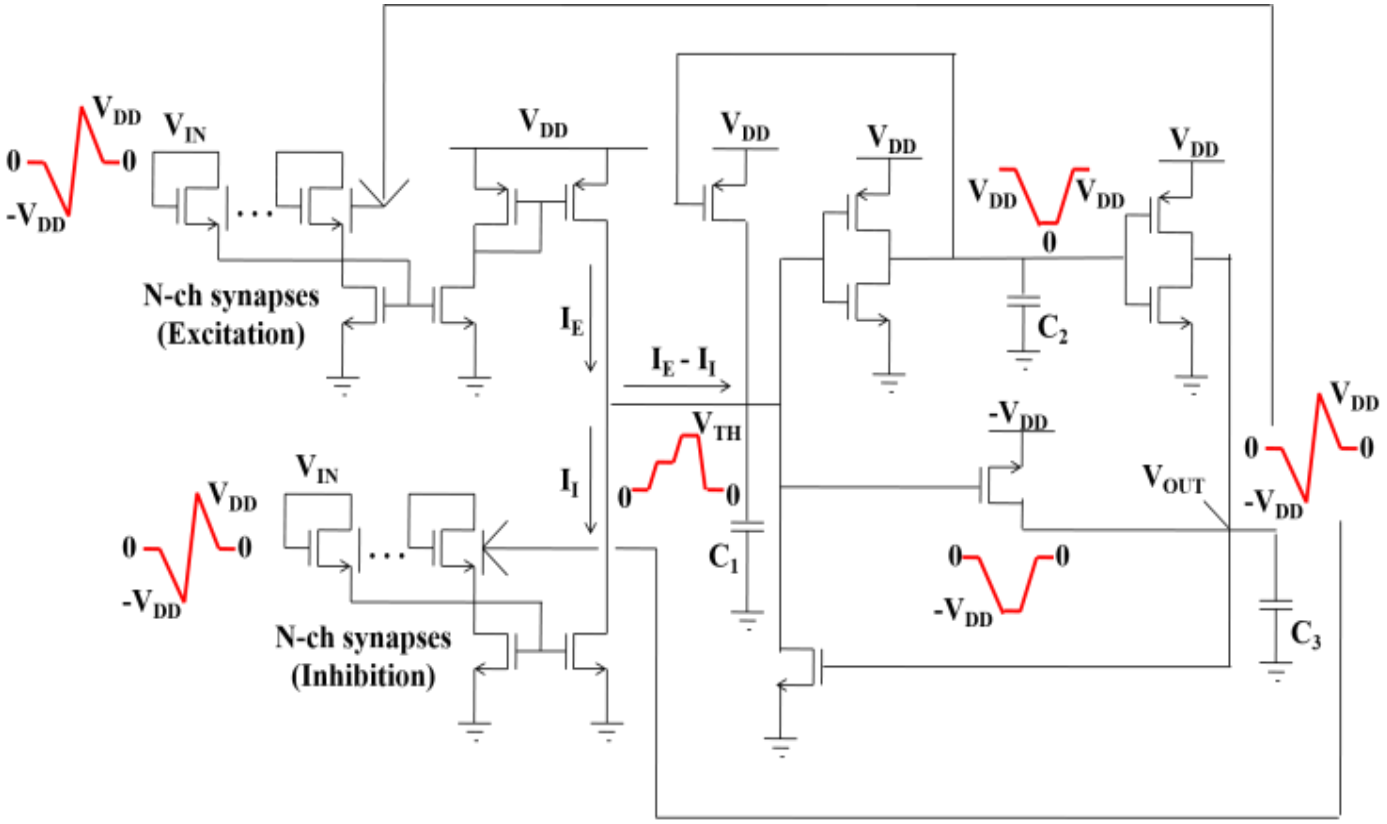
Floating-body Synaptic Transistor (5)

- STDP characteristic



Neuron Circuits (1)

- Integrate-and-fire neuron circuit with capacitor integration

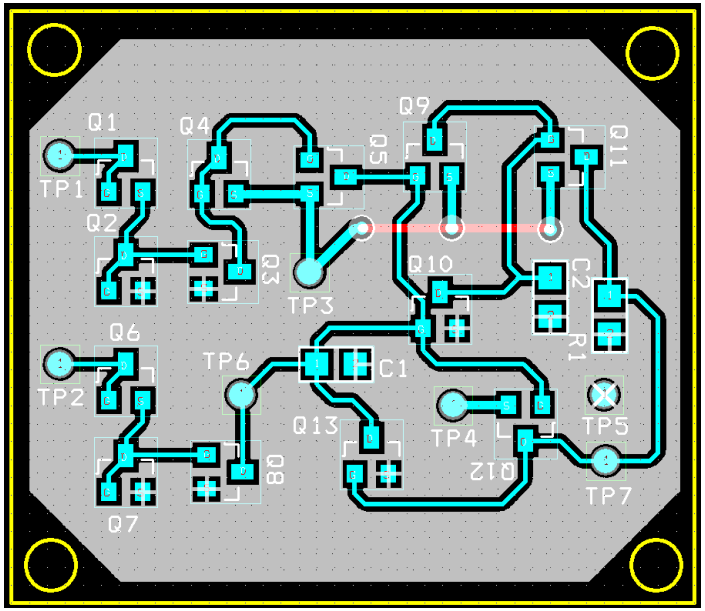


<synaptic integration part> <spike generation part>

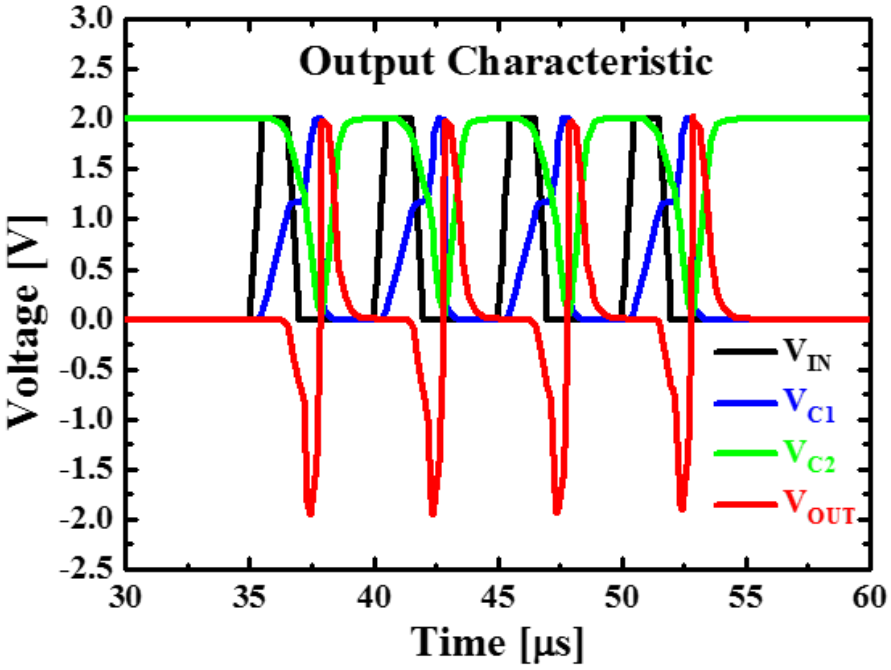
Neuron Circuits (2)

- Implementation with discrete devices

<PCB layout>



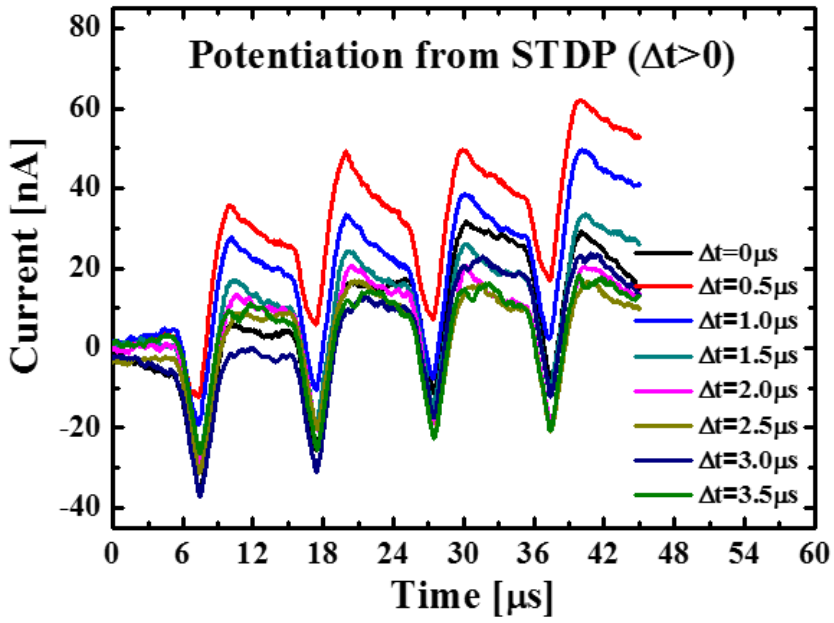
<Output of neuron>



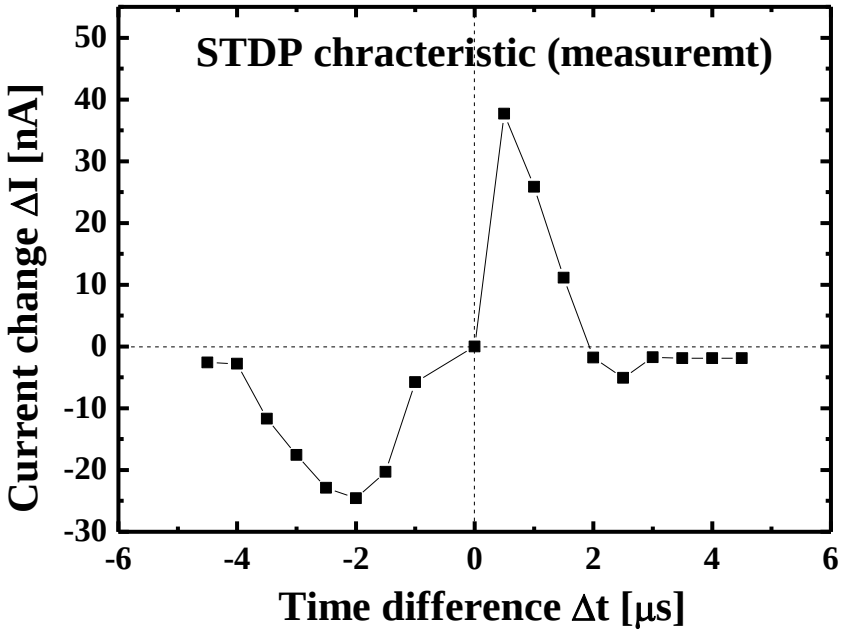
Neuron Circuits (3)

- Measured STDP characteristics

< Transient current >

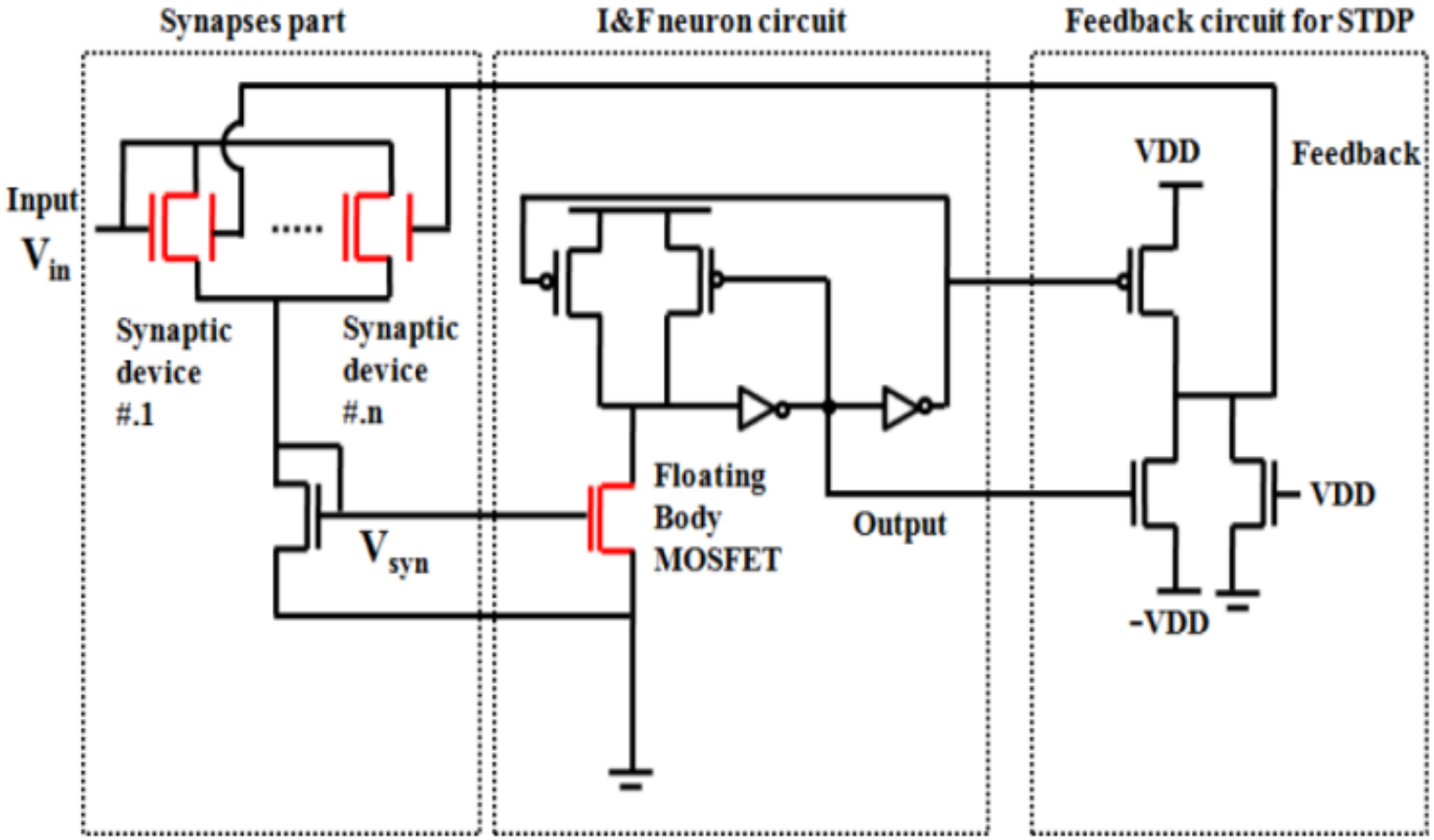


<STDP characteristic>



Neuron Circuits (4)

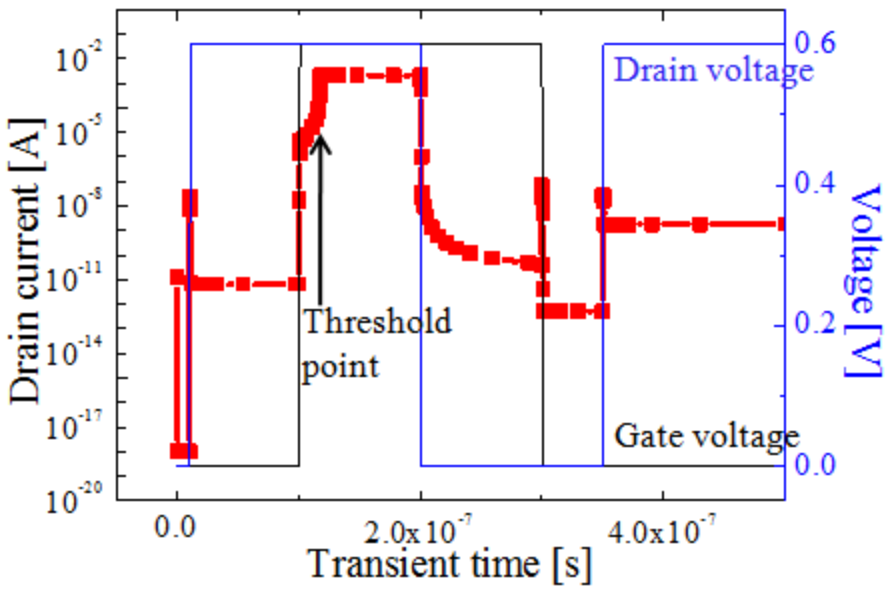
- Integrate-and-fire neuron circuit with a floating-body MOSFET



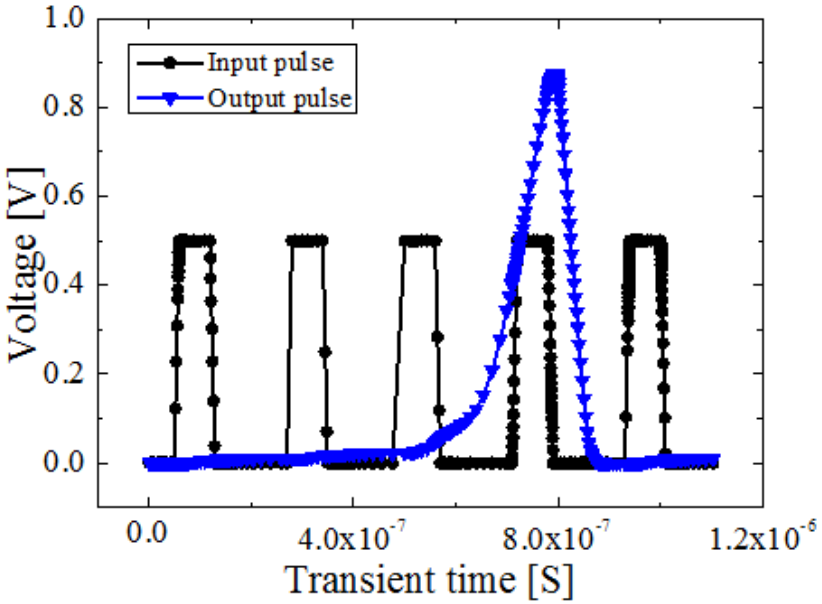
Neuron Circuits (5)

- Role of floating-body MOSFET in the circuit

<transient current>



<spike generation>

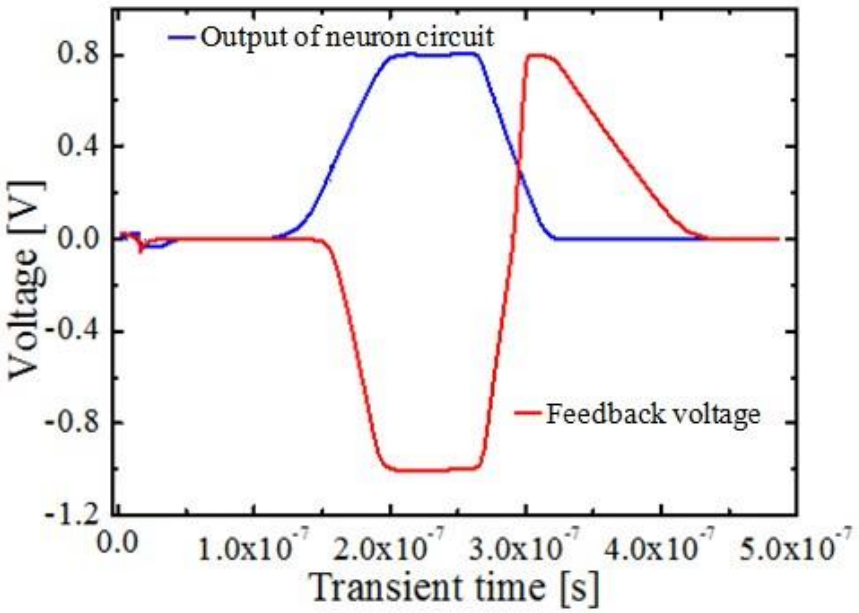


- Temporal integration of input spike signal by the floating body

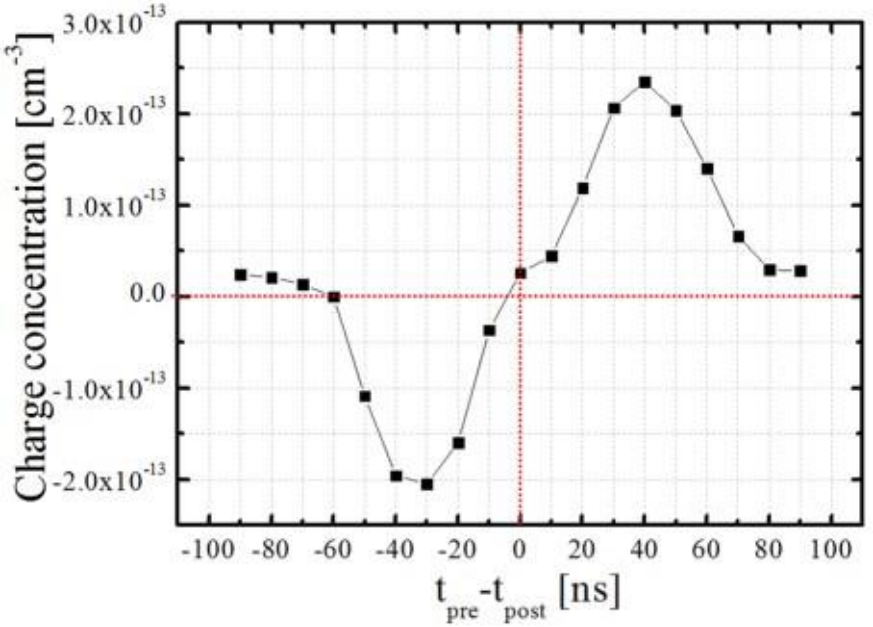
Neuron Circuits (6)

- Shapes of spikes and STDP characteristic

<spike shapes>



<STDP characteristic>

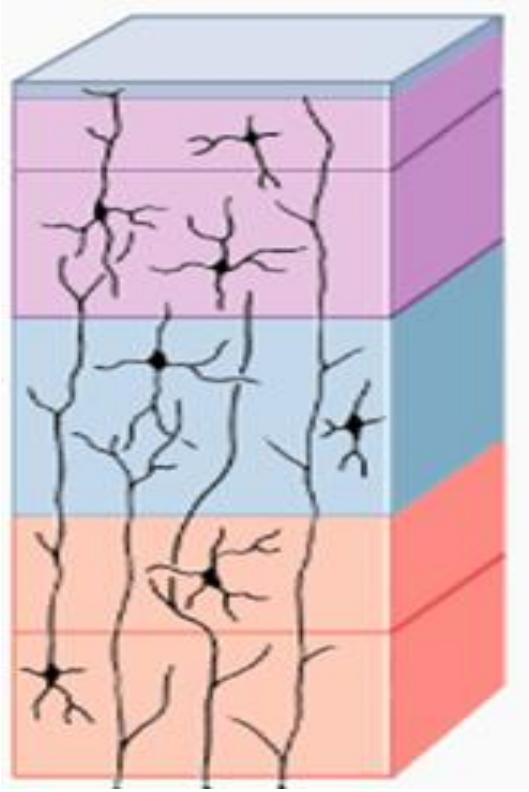


- Output and feedback spike shapes are different !!

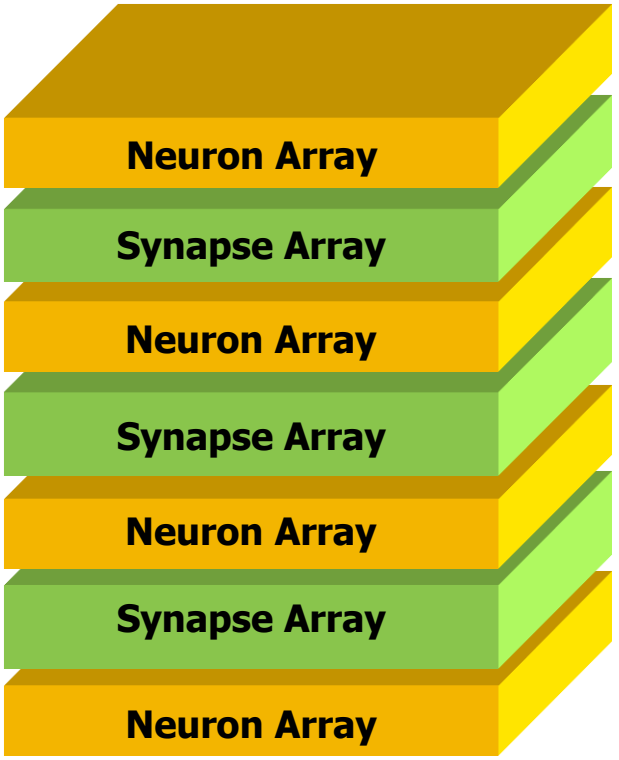
Integration of Neurons and Synapses

- Stacking of neuron and synapse arrays

<primary sensory cortex>



<neuromorphic system>



Conclusions

- 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.
- Integrate-and-fire neuron circuits with capacitor and floating-body integration are proposed and implemented.


Thank you for
your attention !