Time-domain analog computing and VLSI systems toward ultimately high-efficient brain-like hardware

Takashi Morie

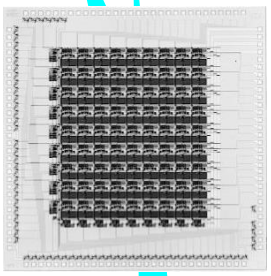
Kyushu Institute of Technology, Japan

Outline

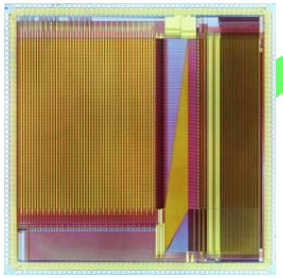
- Introduction
  - Our brain-like VLSI chips
  - My approach toward brain
- Time-domain analog computing and VLSI systems
  - Time-domain energy-efficient weighted sum calculation based on simple spiking neuron model
  - Chaotic Boltzmann machine circuit based on oscillator neuron model
- Conclusion

Our brain-like VLSI chips

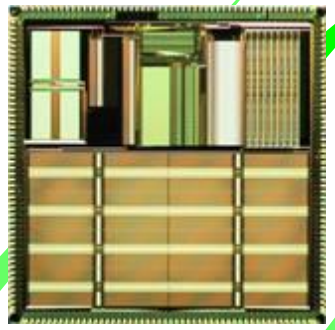
1995 2000 2005 2010 Year



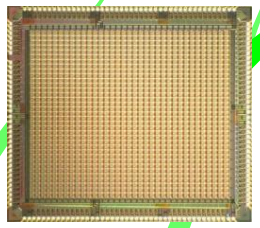
BP/DBM nets
JSC 1994



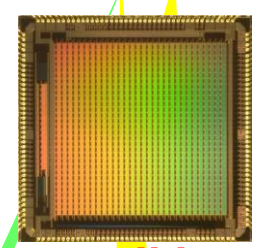
Nonlinear oscillator
ESSCIRC 2002



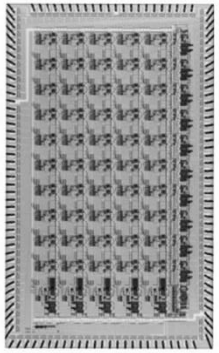
Conv net
VLSI Cir. 2005



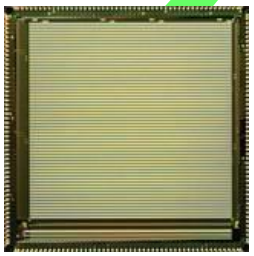
Anisotropic propagation
ISSCC2009



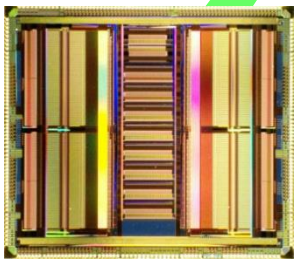
Spiking coupled MRF
ISSCC2012SRP



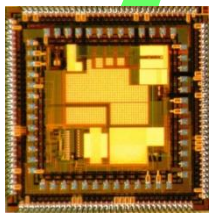
BP nets
Floating-gate memory
IEICE Trans. 1997



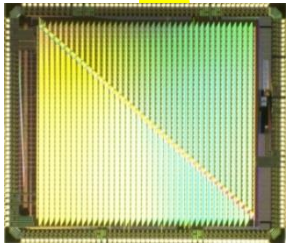
Gabor filter
VLSI Cir. 2004



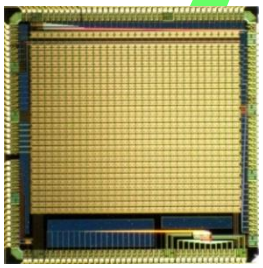
Matching processor
NCSP2007



Chaos circuit
ISCAS2008



Spiking neural net
ICONIP2011



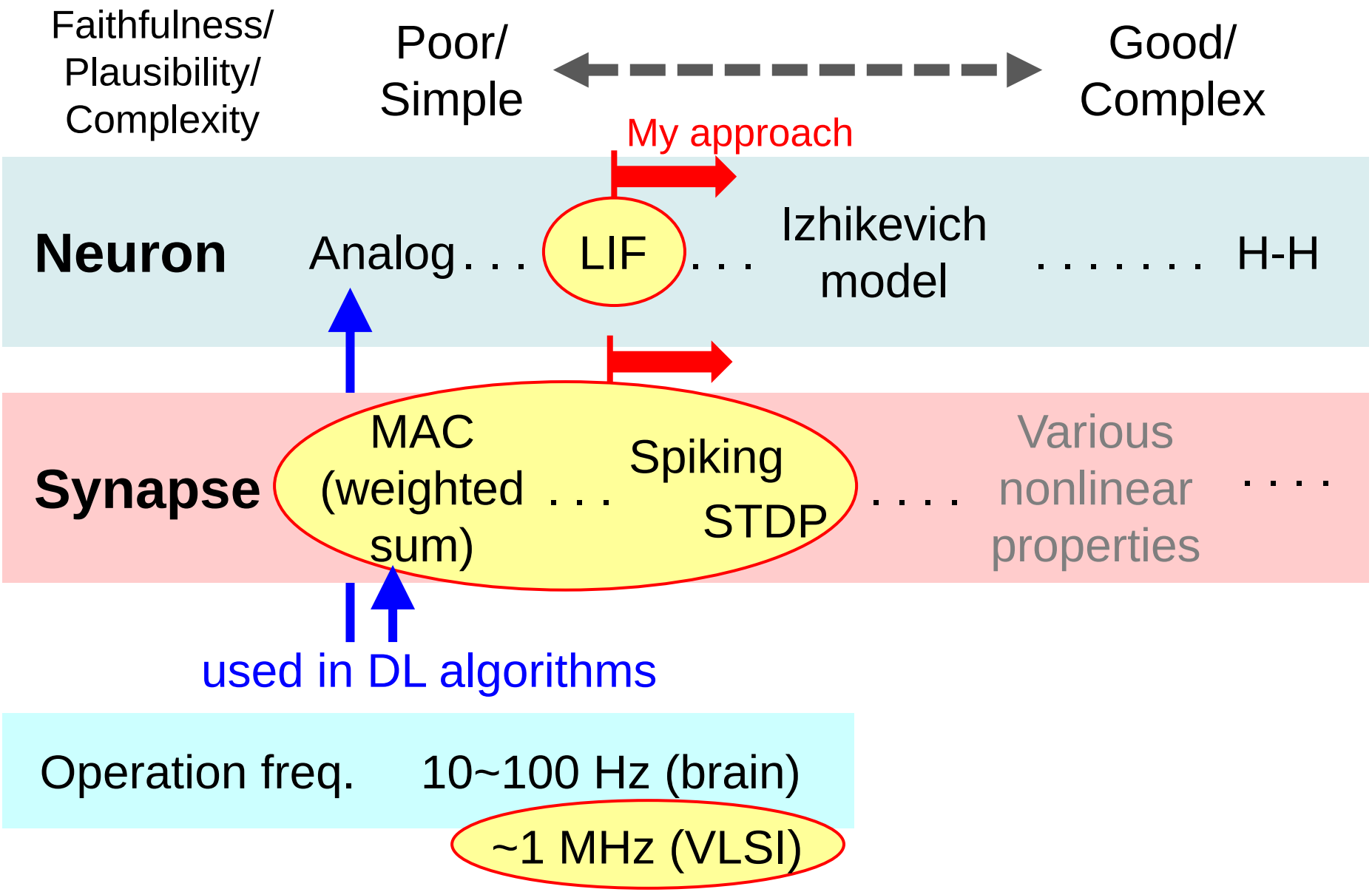
Coupled chaotic system
ECCTD2011

Analog

PWM/PPM

Spike based

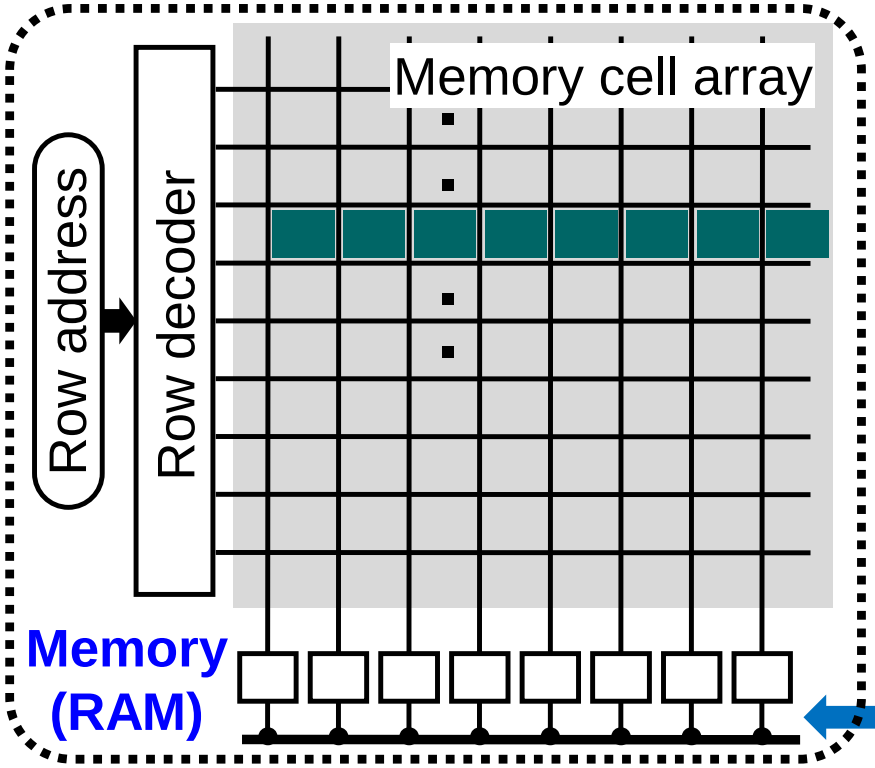
Different approaches to brain functions

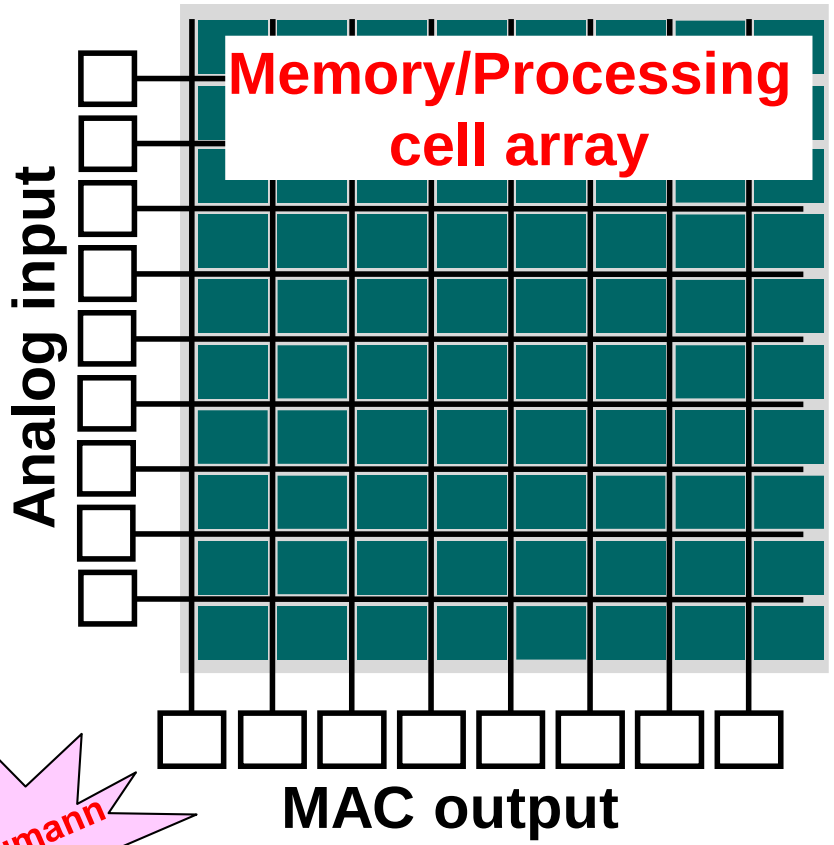


Outline

- Introduction
  - Our brain-like VLSI chips
  - My approach toward brain
- **Time-domain analog computing and VLSI systems**
  - **Time-domain energy-efficient weighted sum calculation based on simple spiking neuron model**
  - Chaotic Boltzmann machine circuit based on oscillator neuron model
- Conclusion

Digital and analog processor architectures





von Neumann bottleneck

Digital processing (von Neumann architecture)

RAM only accesses one row data, and no calculation can be performed in RAM.

Analog/pulse processing (Cross-bar architecture)

DL Processors in ISSCC 2017

SESSION 14 — Tuesday February 7th, 1:30 PM

Deep-Learning Processors

Session Chair: ***Takashi Hashimoto,*** *Panasonic, Kadoma, Japan*
Associate Chair: ***Mahesh Mehendale,*** *Texas Instruments, Bangalore, India*

1:30 PM

14.1 A 2.9TOPS/W Deep Convolutional Neural Network SoC in FD-SOI 28nm for Intelligent Embedded Systems
G. Desoli[1], N. Chawla[2], T. Boesch[3], S-P. Singh[2], E. Guidetti[1], F. De Ambroggi[4], T. Majo[1], P. Zambotti[4], M. Ayodhyawasi[2], H. Singh[2], N. Aggarwal[2]
[1]STMicroelectronics, Cornaredo, Italy; [2]STMicroelectronics, Greater Noida, India
[3]STMicroelectronics, Geneva, Switzerland; [4]STMicroelectronics, Agrate Brianza, Italy

2:00 PM

14.2 DNPU: An 8.1TOPS/W Reconfigurable CNN-RNN Processor for General-Purpose Deep Neural Networks
D. Shin, J. Lee, J. Lee, H-J. Yoo, KAIST, Daejeon, Korea

2:30 PM

14.3 A 28nm SoC with a 1.2GHz 568nJ/Prediction Sparse Deep-Neural-Network Engine with >0.1 Timing Error Rate Tolerance for IoT Applications
P. N. Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks, G-Y. Wei
Harvard University, Cambridge, MA

Latest digital DL processors ~10TOPS/W

14.4 A Scalable Speech Recognizer with Deep-Neural-Network Acoustic Models and Voice-Activated Power Gating
M. Price, J. Glass, A. Chandrakasan
Massachusetts Institute of Technology, Cambridge, MA

3:45 PM

14.5 ENVISION: A 0.26-to-10TOPS/W Subword-Parallel Computational Accuracy-Voltage-Frequency-Scalable Convolutional Neural Network Processor in 28nm FDSOI
B. Moons, R. Uytterhoeven, W. Dehaene, M. Verhelst, KU Leuven, Leuven, Belgium

Measure of energy efficiency of processors

FLOPS -> OPS (Fixed-point operations per sec.)
e.g. PC(Core i7) ~500GFLOPS

Operation performance: TOPS/GOPS
(Tera/Giga Operations Per Second)

Energy efficiency: TOPS/W
= Tera Ops. per sec. / Joule per sec.
= Tera ops. / Joule
Energy consumption per op.: 1/(TOPS/W) [pJ/op]
= 1 [pJ/op]

Latest digital **DL processors**:
~10TOPS/W

Synapse op. in **brain**: 0.1~1 fJ/op
1,000~10,000 TOPS/W
=1~10 POPS/W

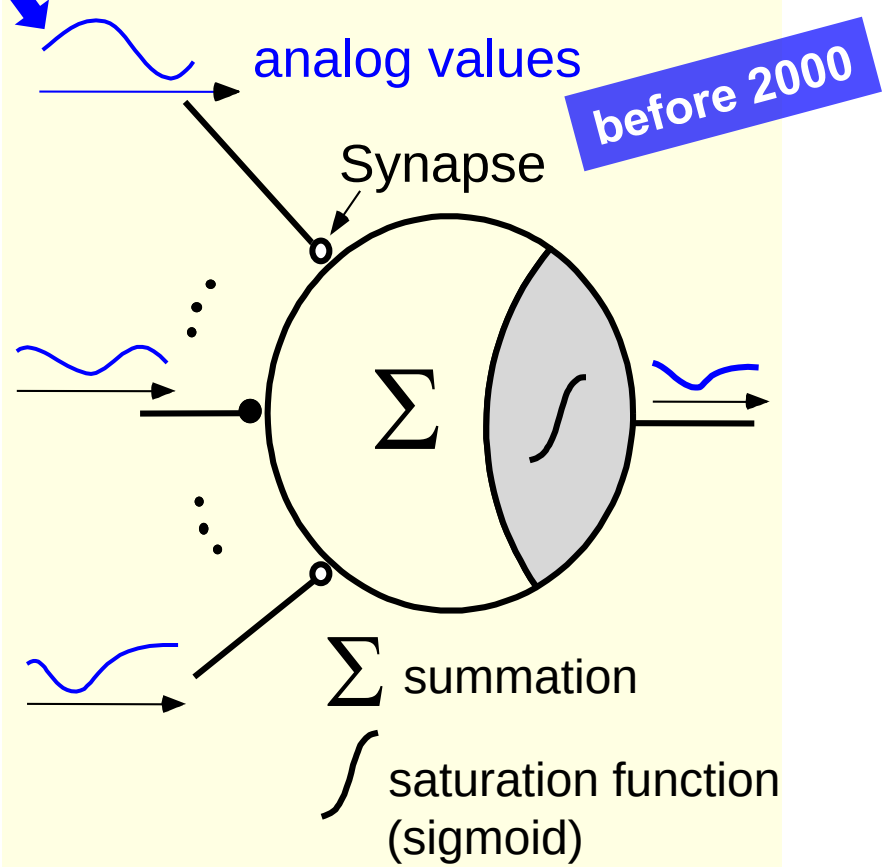
# of neurons: ~10^{11}
# of synapses: ~10^{15}
Power cons.: 20（~1）W
Op. freq.: 10~100 Hz
Activity: ~10 %

Two types of neuron models

Analog neuron models

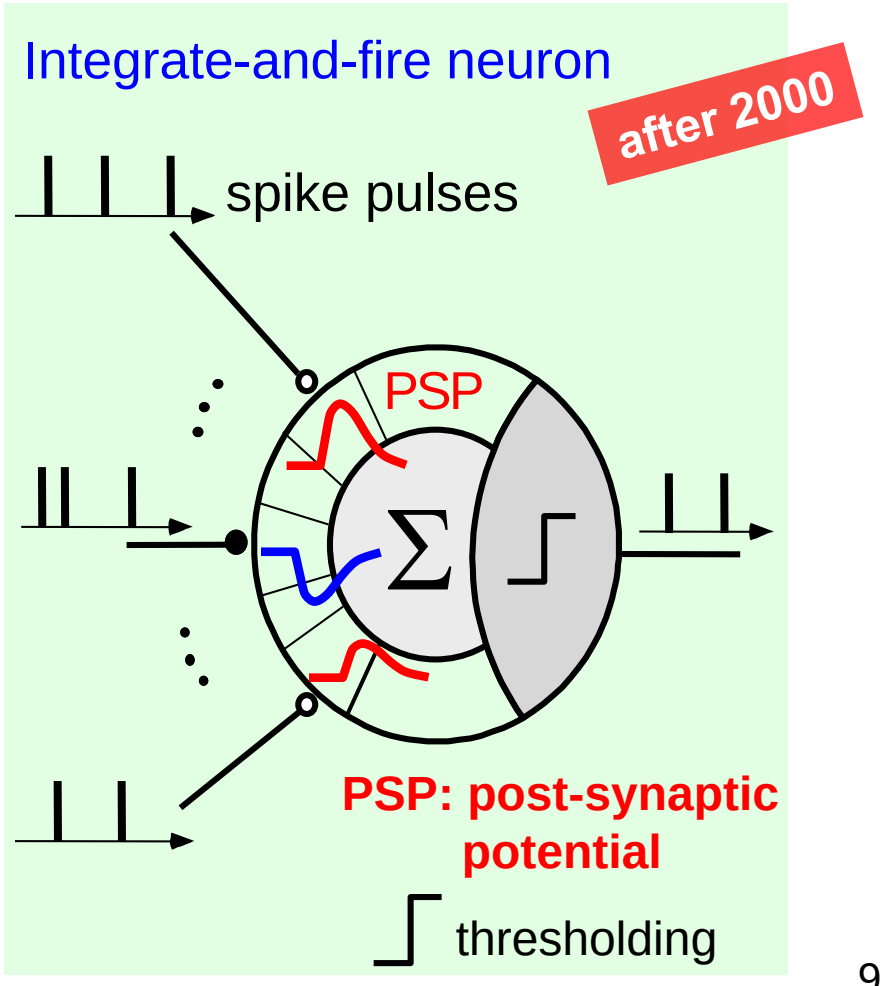
coded by analog values representing firing rate or population of spike pulses

biological neurons output “spikes”



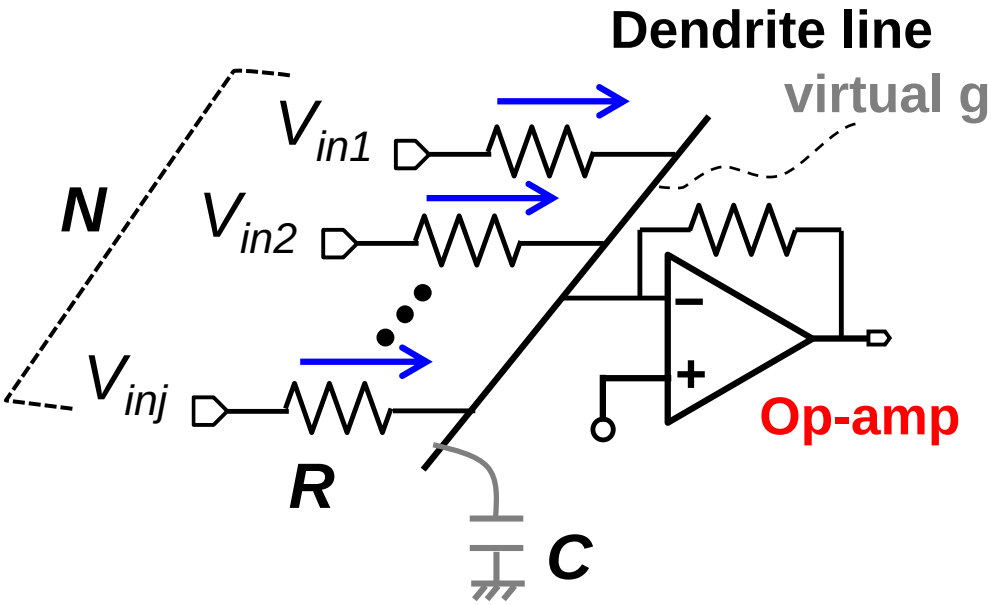
Spiking neuron models

coded by spatiotemporal patterns of spike pulses



Energy consumption using resistive elements

Voltage-domain circuits based on analog neuron model



uses DC-mode operation

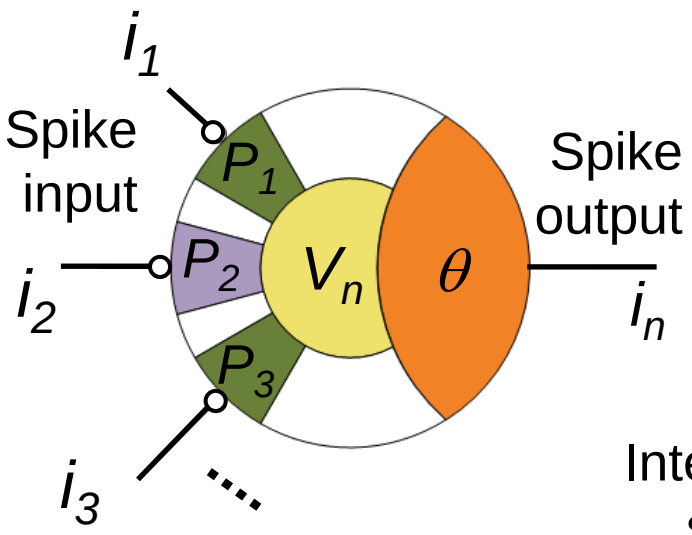
$$\frac{V_{out}}{R_f} = -\sum_{j=1}^{N} \frac{V_{inj}}{R_j}$$

Assuming R=100MΩ, Vin=0.1~1V, τ=1μs

$E_w = \tau V^2/R$ = 0.1～10fJ

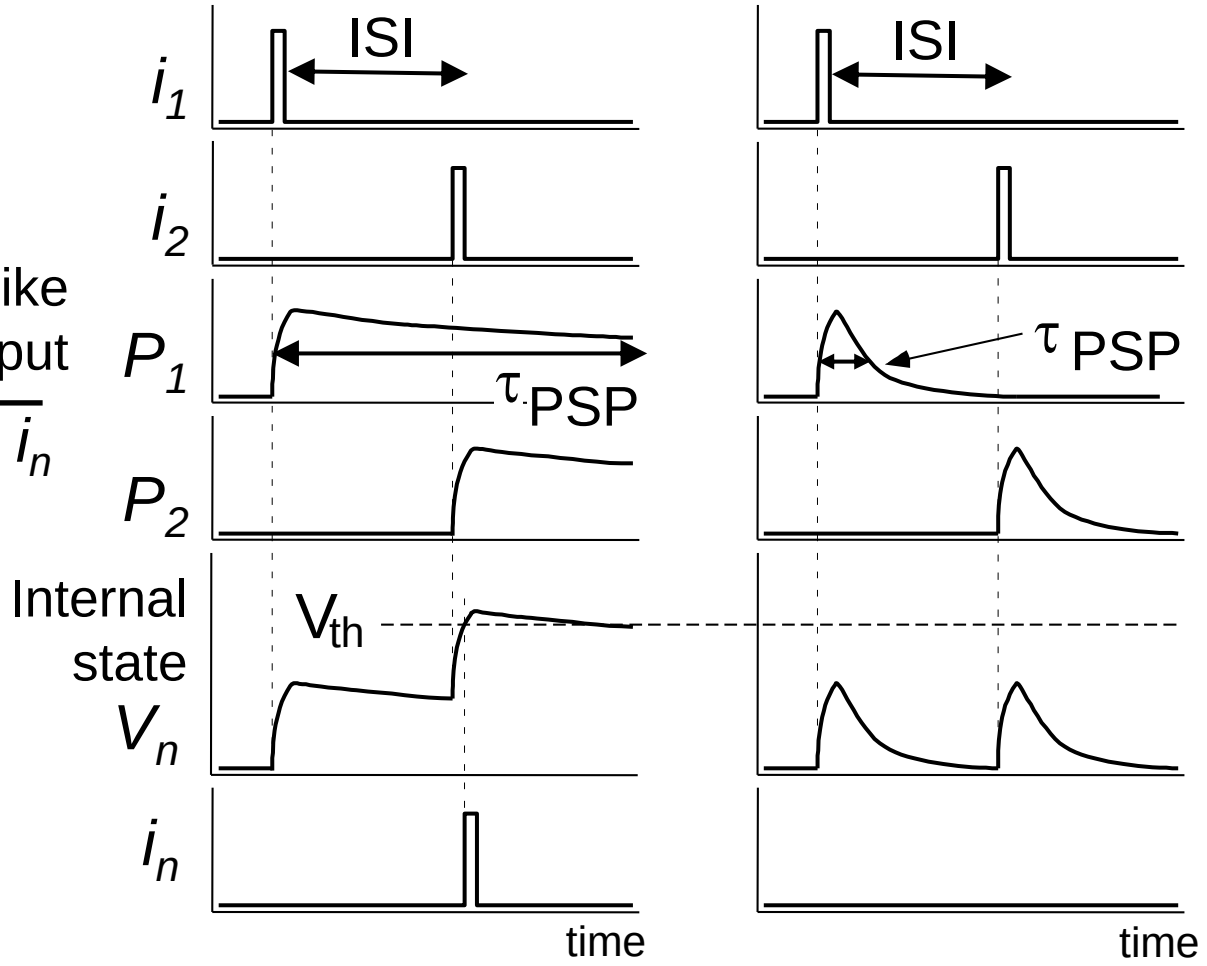
Op-amps consume much more energy than resistors.

Functions of PSPs in spiking neurons



ISI: Inter-spike interval

PSP: Post-synaptic potential



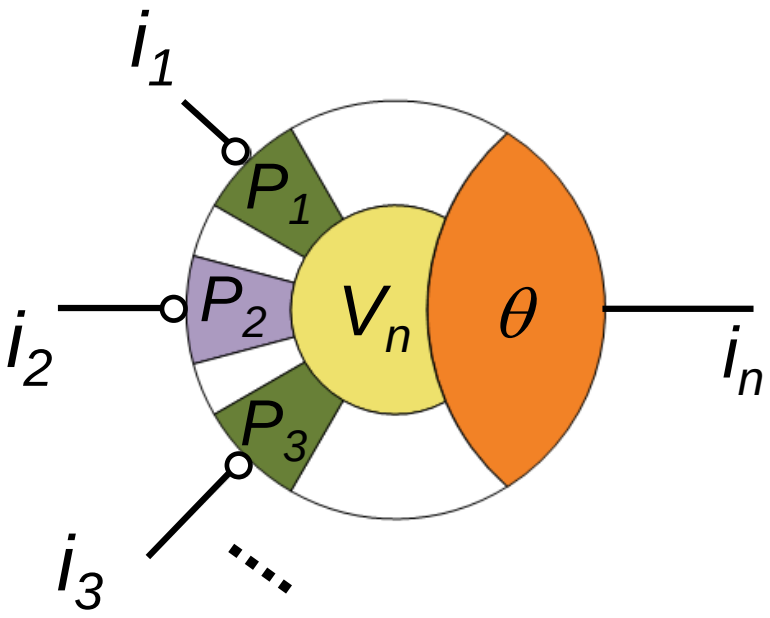
$ISI << \tau_{PSP}$

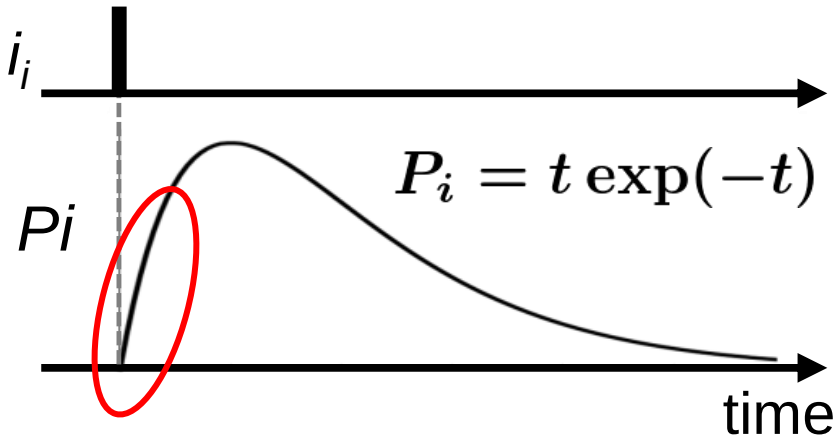
Integrator

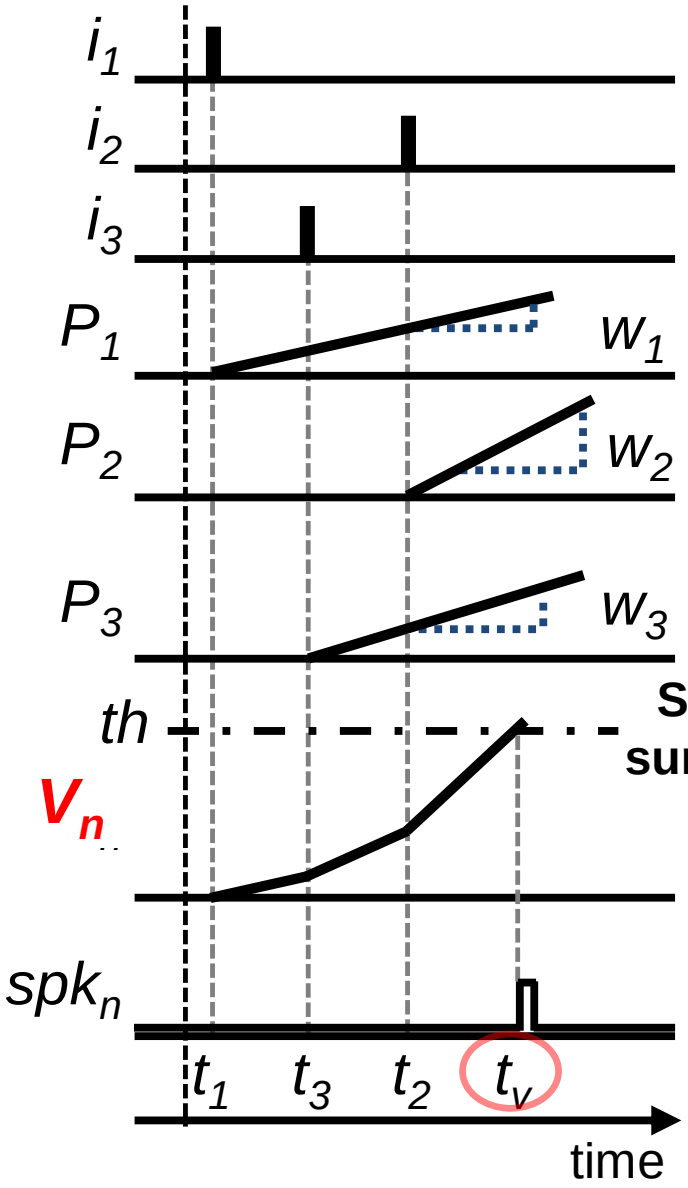
$ISI >> \tau_{PSP}$

Coincidence detector

Integrate & fire neuron





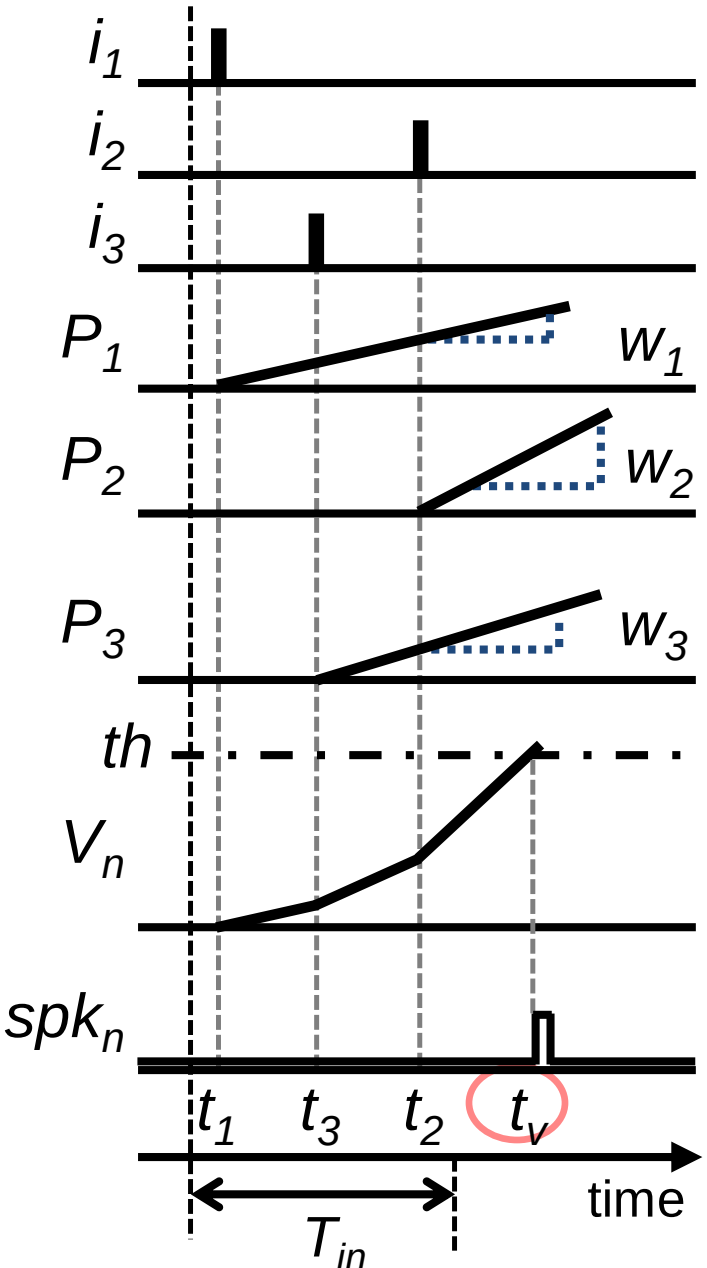


Spatiotemporal summation of all linear rising waveforms of PSPs

Typical time course of PSPs: α-function

W.Maass et al., 1999

Time-domain weighted-sum calculation model



$$y = \sum_i w_i x_i$$

weighted summation (MAC: multiply and accumulate)

x_i --> spike timing t_i

w_i --> slope of PSP

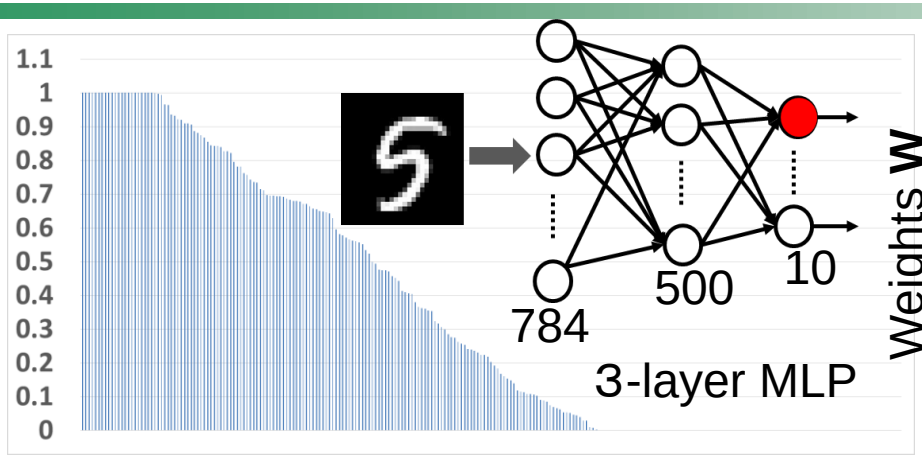
spike timing t_v --> y

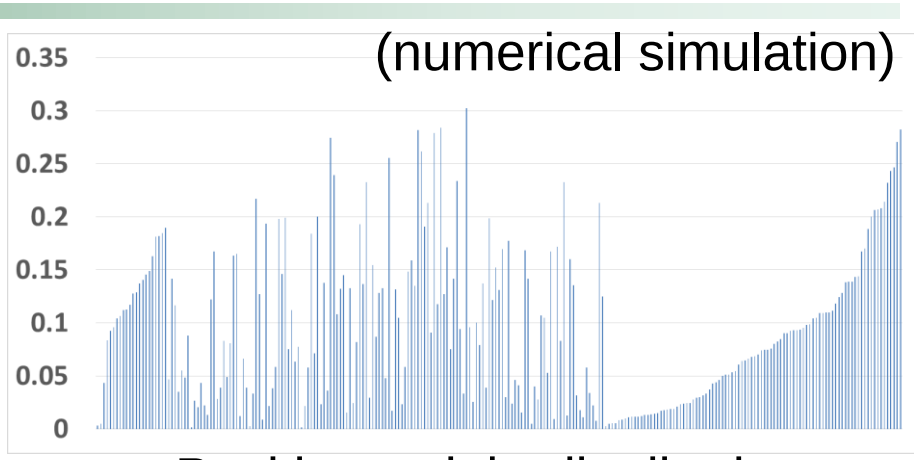
$$\text{given} \begin{cases} x_i \in [0,1], \quad t_i = T_{in}(1 - x_i) \\ \sum_i w_i = \beta \end{cases}$$

$$\sum_i w_i x_i = \frac{th + \beta(T_{in} - t_v)}{T_{in}}$$

T. Morie et al. IEEE NANO 2016, Q. Wang et al. ICONIP 2016

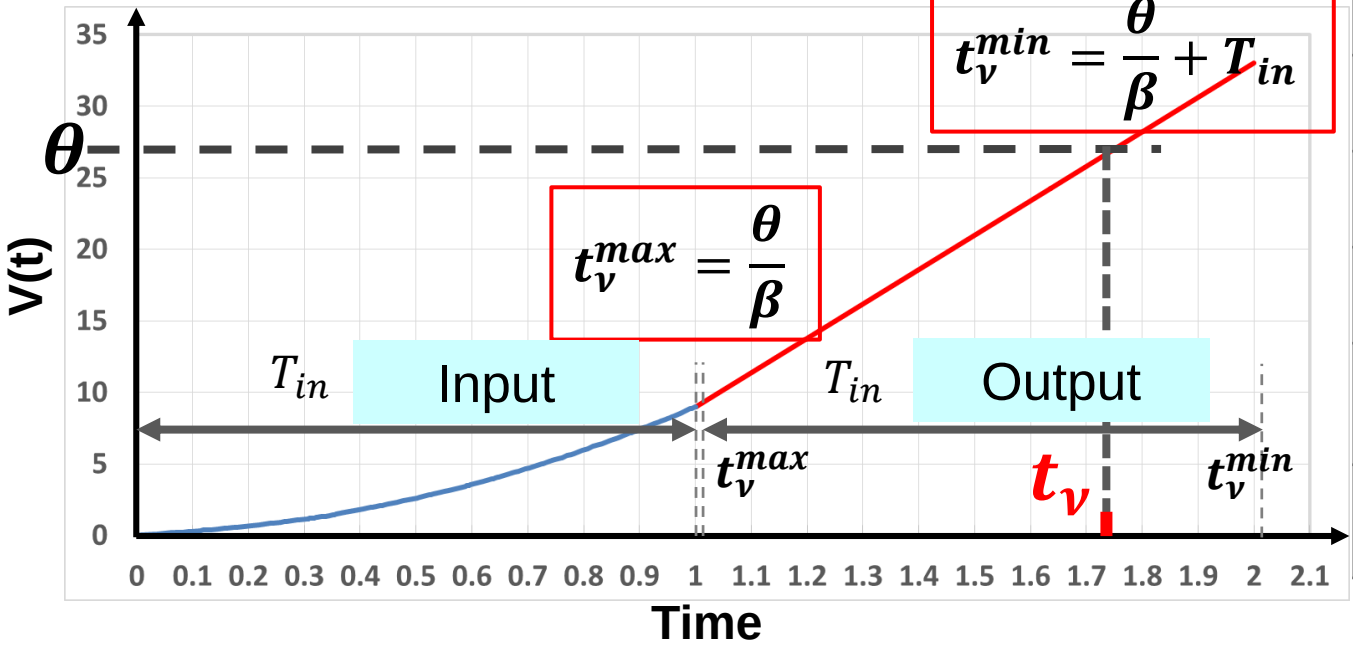
Time-domain MAC calculation





$x_i \in [0,1]$

$$\sum_{i=1}^{N} w_i x_i = \mathbf{8.99}$$

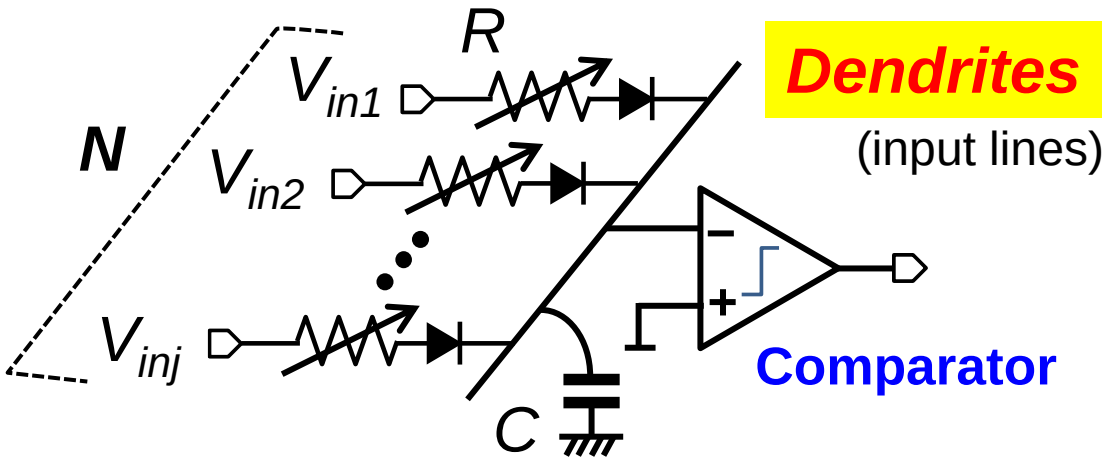


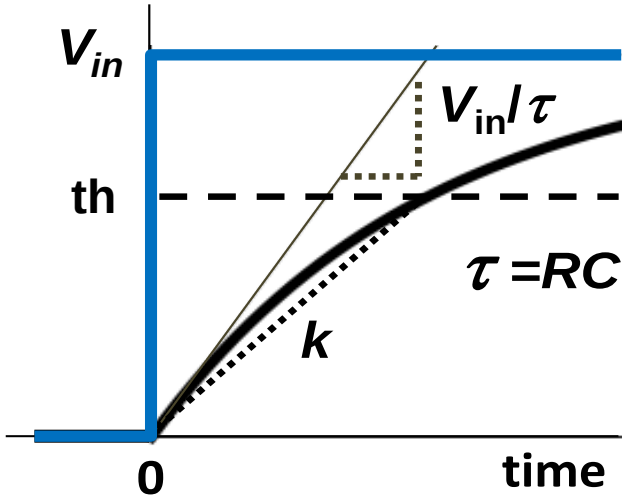
N	249
T_{in}	1
β	24.01
θ	24.25
t_v	**1.635**
$\frac{\theta/\lambda + \beta(T_{in} - t_v)}{T_{in}}$	**8.99**

Q. Wang et al. ICONIP 2016

Energy consumption using resistive elements

Time-domain analog circuits





transient state of RC circuit

$E_{wsum} = CV^2$

independent of resistance values

If C=1fF, V=1V, E_{wsum}=1fJ.

Assuming N=100, $E_w = E_{wsum}/N$=10aJ

Extremely low-energy computation due to parallelism, but very high resistance is needed.

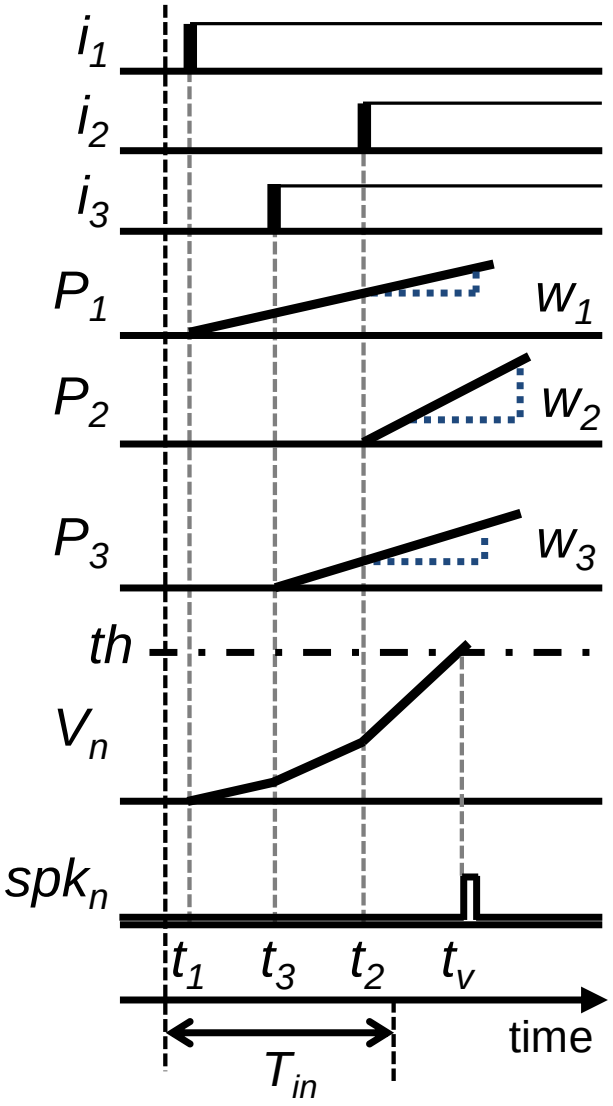
To guarantee time resolution, τ=1μs (1,000 steps with 1 ns)

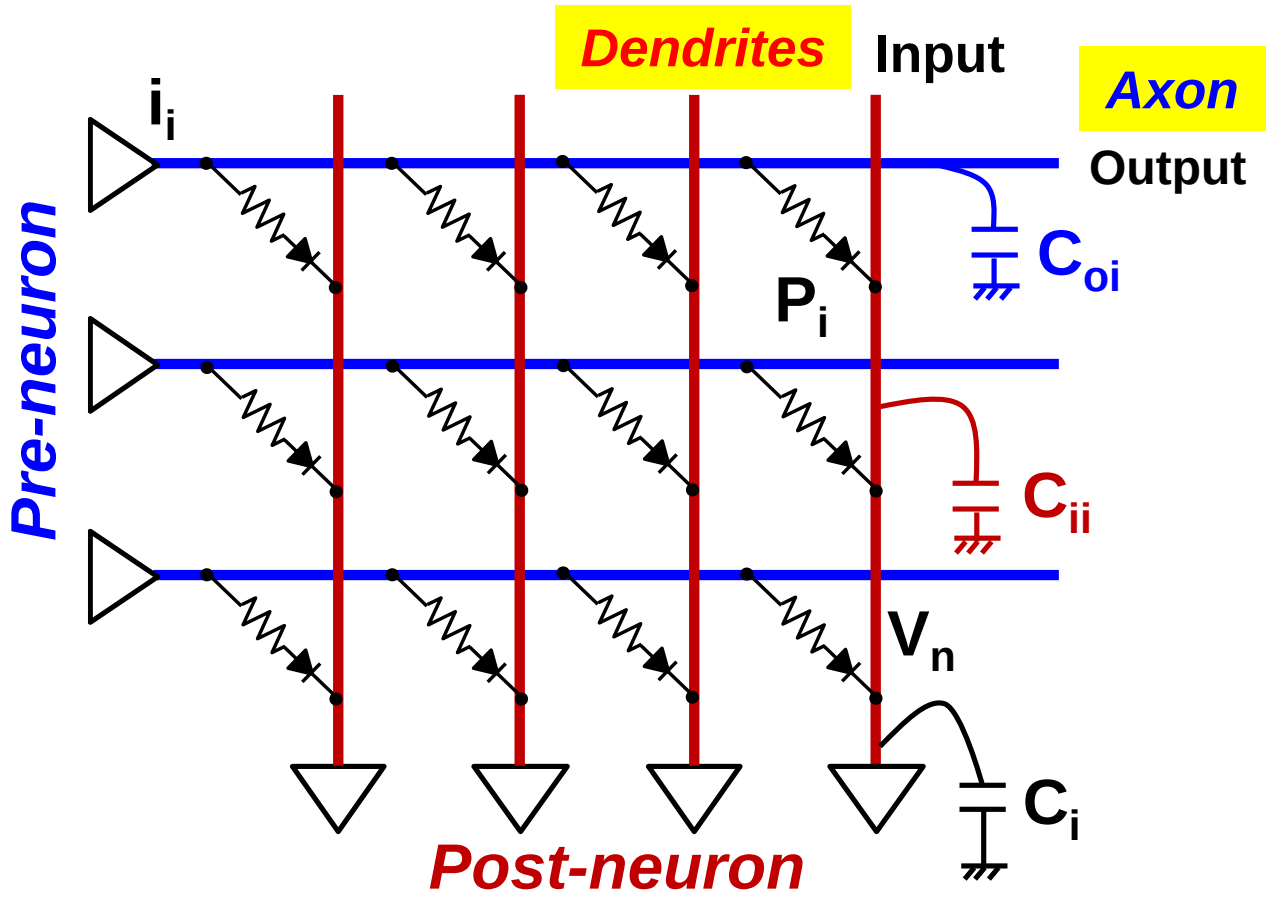
To reduce energy, reduce (parasitic) interconnect capacitance

R_{ON} ~ 1GΩ, (R_{OFF} ~ 1TΩ)

Q. Wang et al. ICONIP 2016

Time-domain analog cross-bar circuit





Resistance change memory

[ReRAM device]

Low parasitic (interconnect) capacitance

[Nanowire]

Low-power neuron circuit (comparator)

Comparison among different approaches

Approach		# of neurons/ synapses	Energy per synapse operation	Processing frequency	Power consump-tion
Biologi-cal	Human brain	~10^{11}/ ~10^{15}	10^{-16}~10^{-15} J (**0.1~1 fJ**)	10~100 Hz (10% active)	20 (~1) W
Digital	Super Comput. (Kei「京」) (*1)	1.7x10^{9}/ 1.0x10^{13}	(6.5x10^{-4}J) **(~1 mJ)**	Brain:1s = Kei: 2,400s (4.4 fires/s)	~12 MW
	Digital chips (TrueNorth~)	1x10^{6}/ 256x10^{6}	~10^{-13} J **(< 0.1 pJ)**	1 kHz	86 mW
Analog	Voltage/current -domain	[1x10^{6}/ 256x10^{6}] (*2)	>~10^{-15} J **(~ 1 fJ)** (*3)	<~MHz	[~300 mW] (*4)
	Time-domain	[1x10^{6}/ 256x10^{6}] (*2)	>~10^{-17} J **(~10 aJ)** (*3)	<~MHz	[~3 mW] (*4)

(*1) http://www.riken.jp/pr/topics/2013/20130802_2/

(*2) Assuming the same values as TrueNorth

(*3) Estimation when assuming ideal condition

(*4) Only for weighted summation

Outline

- Introduction
  - Our brain-like VLSI chips
  - My approach toward brain
- **Time-domain analog computing and VLSI systems**
  - Time-domain energy-efficient weighted sum calculation based on simple spiking neuron model
  - **Chaotic Boltzmann machine circuit based on oscillator neuron model**
- Conclusion

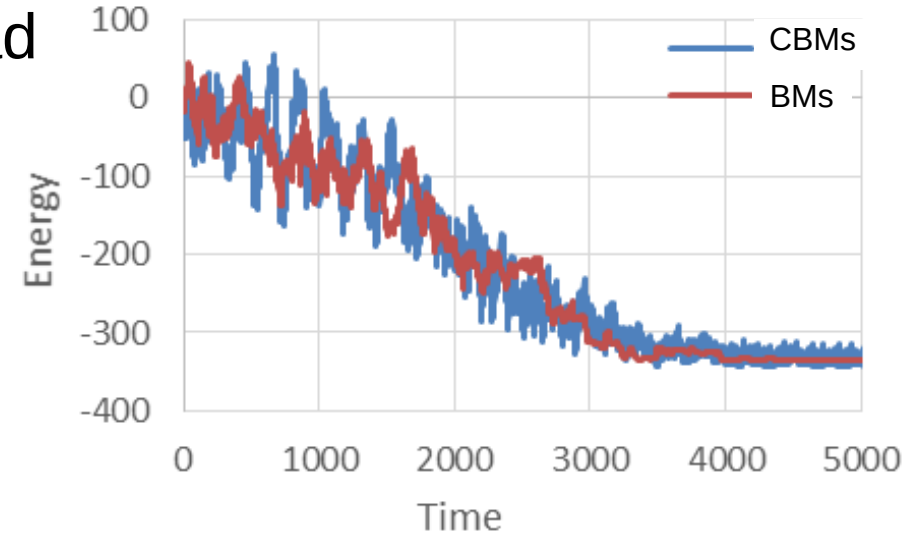
Original and chaotic Boltzmann machines

Boltzmann machines (BMs)[1]

- **Stochastic operation** of binary neurons
- Symmetrically connected networks
- Solving optimization problems using energy minimization
- Success of deep learning by restricted BMs

Chaotic Boltzmann machines (CBMs)[2]

- **Deterministic operation**
- Using chaotic dynamics instead of stochastic operation
- Computing ability comparable to BMs



Simulation results

[1] D. H. Ackley et al., Cog. Sci. 9, 1985.
[2] H. Suzuki et al., Sci. Rep., 2013.

Chaotic Boltzmann machines (CBMs)

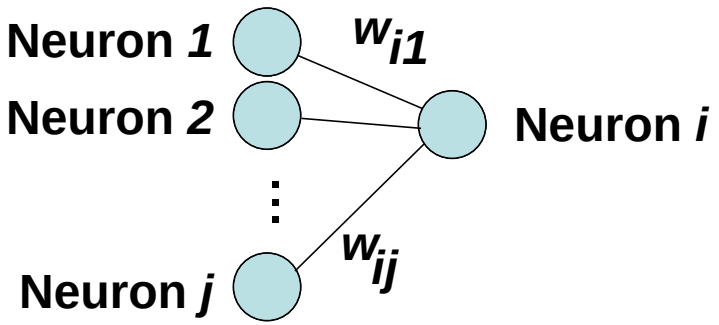
Dynamics of CBMs

$$\frac{dx_i}{dt} = (1-2S_i)\left(1+\exp\frac{(1-2S_i)z_i}{T}\right)$$

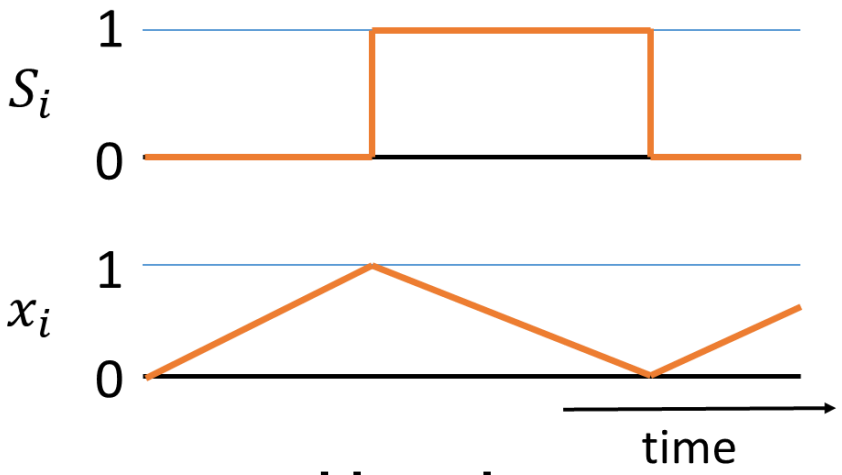
$$S_i \leftarrow 0 \quad (x_i = 0)$$

$$S_i \leftarrow 1 \quad (x_i = 1)$$

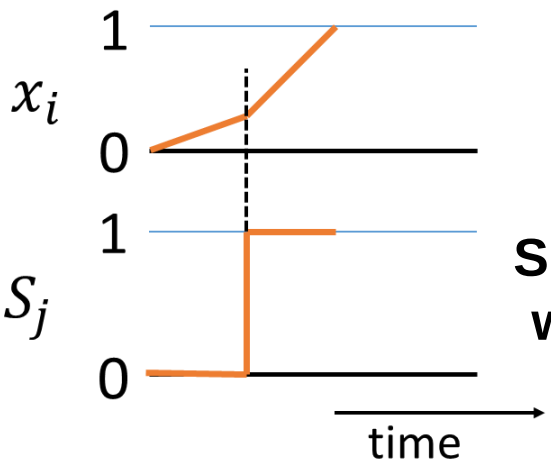
$$z_i = \sum_{j=1}^{N} w_{ij}S_j + \theta_i \quad (w_{ij} = w_{ji})$$



S_i : Binary output of *i*-th neuron
x_i : Internal state of *i*-th neuron
T : Temperature parameter
w_{ij} : Synaptic weight between *i* and *j*
θ_i : Constant bias of *i*-th neuron

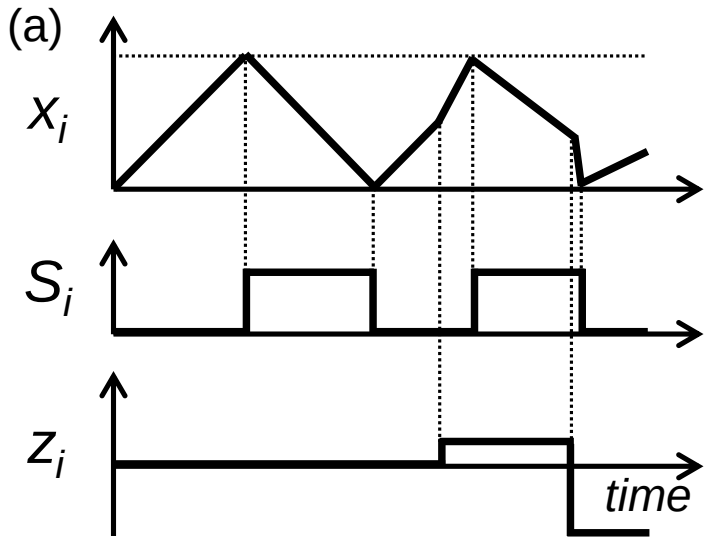


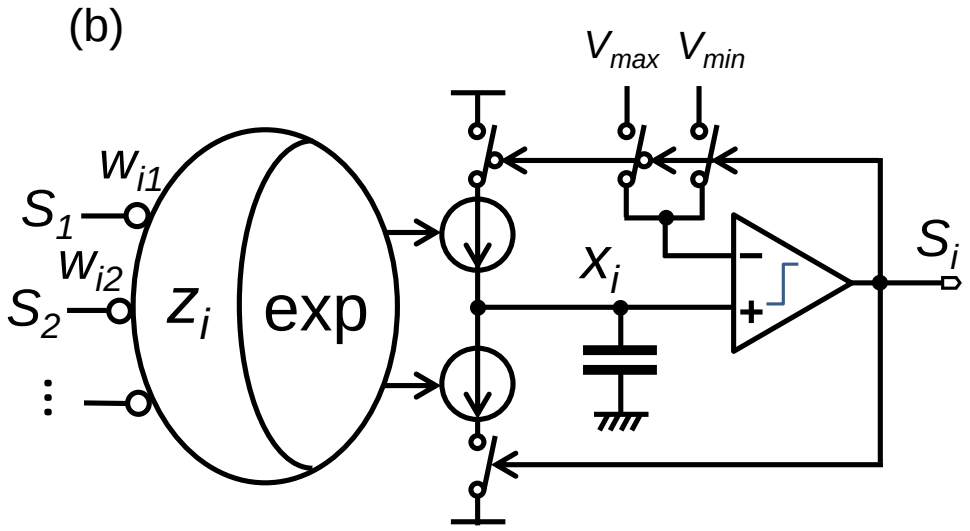
without input



Slope change in x_i when S_j changes

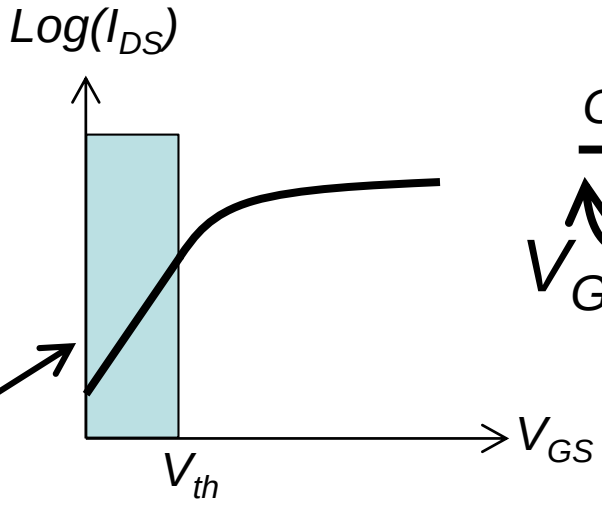
A CMOS unit circuit for CBMs

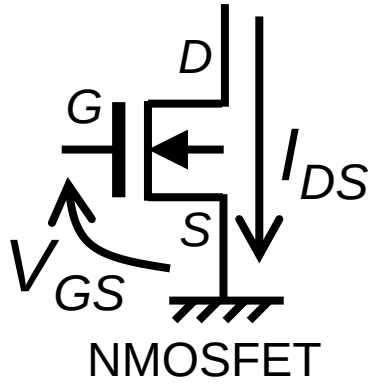




$$\frac{dx_i}{dt} = (1-2S_i)\left(1+\exp\frac{(1-2S_i)z_i}{T}\right)$$

Implemented using subthreshold region of MOSFET





V_{th}: Threshold voltage

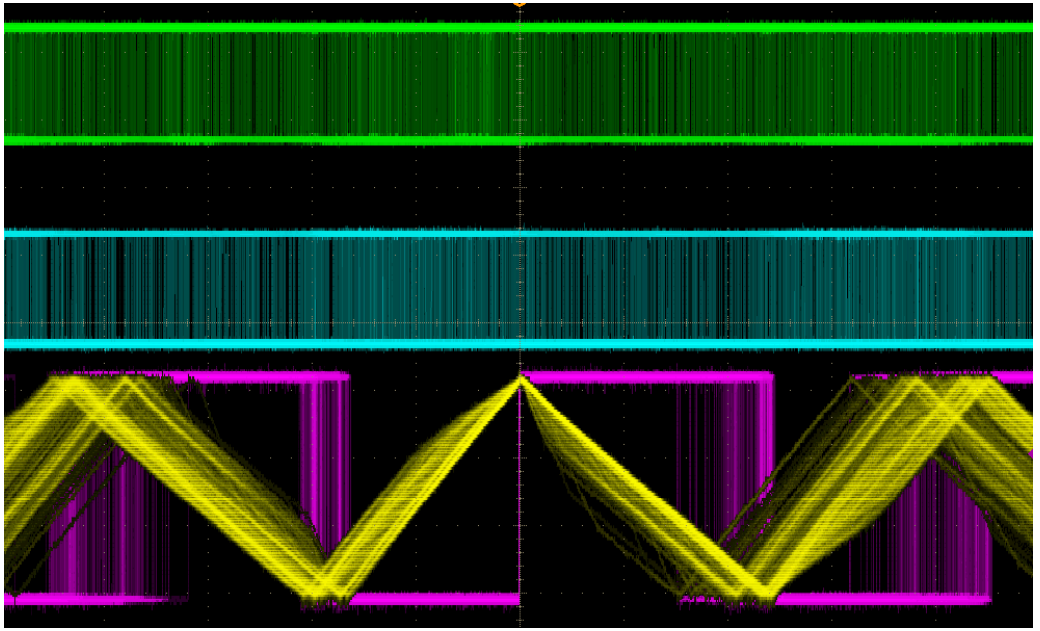
: Sub-threshold region

M. Yamaguchi et al. ICONIP 2016.

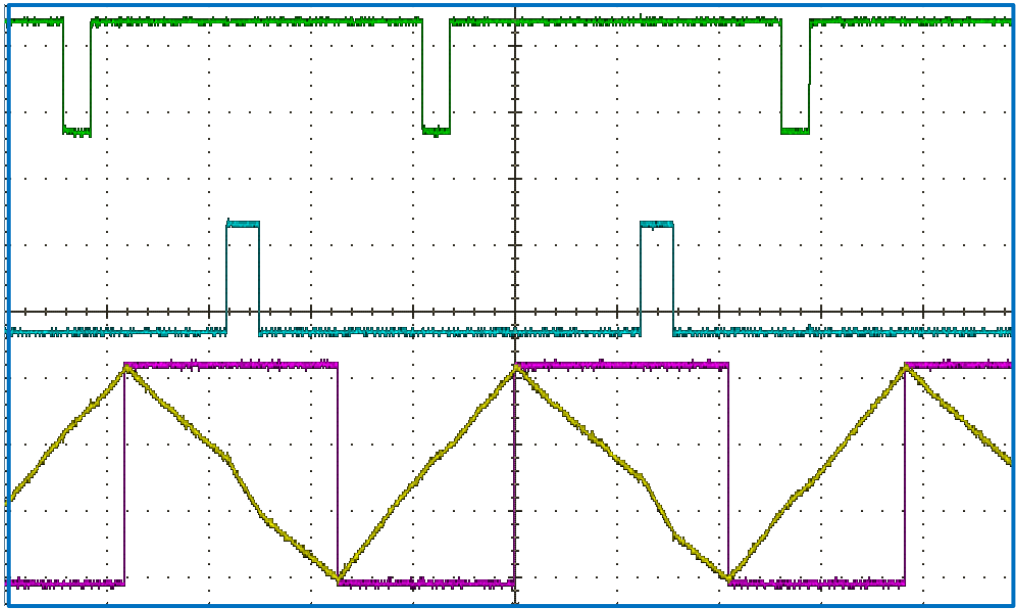
Measurement results of CBM VLSI chip

3 neurons
with 3 synapses

Continuous
scan



Single
scan



Conclusions

- ***Time-domain analog VLSI implementation can achieve extremely energy-efficient operation including nonlinear transforms, which is difficult for digital VLSI implementation.***
- ***Weighted-sum calculation as a simple synaptic function can be achieved with extremely low energy consumption based on time-domain operation of simple spiking neuron model, but high-resistance element is required. Also, to reduce parasitic interconnection capacitance is another challenge to achieve energy-efficient operation.***
- ***Nonlinear transforms performed in charging operation to a capacitor were successfully applied to implementation of nonlinear dynamical models, such as chaotic Boltzmann machines.***