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
BP/DBM nets JSC 1994
BP nets Floating-gate memory IEICE Trans. 1997

2000
Nonlinear oscillator ESSCIRC 2002
Gabor filter VLSI Cir. 2004
Matching processor NCSP2007

2005
Conv net VLSI Cir. 2005
Anisotropic propagation ISSCC2009
Chaos circuit ISCAS2008
Spiking coupled MRF ISSCC2012SRP

2010
Spiking neural net ICONIP2011

Year
ISPAS2008 NCSP2007 JSC 1994
ISSCC2009 ISSCC2012SRP
VLSI Cir. 2005 ECCTD2011
PWM/PPM

Analog
Coupled chaotic system ECCTD2011
Spike based
Different approaches to brain functions

Faithfulness/Plausibility/Complexity

Neuron
- Analog
- LIF
- Izhikevich model
- H-H

Synapse
- MAC (weighted sum)
- Spiking STDP
- Various nonlinear properties

Operation freq.
- 10~100 Hz (brain)
- ~1 MHz (VLSI)

My approach

My approach used in DL algorithms

- Poor/Simple
- Good/Complex
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

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)
14.1 A **2.9TOPS/W Deep Convolutional Neural Network** SoC in FD-SOI 28nm for Intelligent Embedded Systems

G. Desolli¹, N. Chawla², T. Boesch³, S-P. Singh⁴, E. Guidetti⁵, F. De Ambroggi⁶, T. Majo⁷, P. Zambotti⁸, M. Ayodhyawas², H. Singh³, N. Aggarwal⁹

¹STMicroelectronics, Cornell, Italy; ²STMicroelectronics, Greater Noida, India
³STMicroelectronics, Geneva, Switzerland; ⁴STMicroelectronics, Agrade Brianza, Italy

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

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


Harvard University, Cambridge, MA

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

**Spiking neuron** models
coded by spatiotemporal patterns of spike pulses

**Integrate-and-fire neuron**

Before 2000

- analog values
- Synapse
- summation
- saturation function (sigmoid)

After 2000

- spike pulses
- PSP: post-synaptic potential
- thresholding

Biological neurons output “spikes”
Energy consumption using resistive elements

Voltage-domain circuits based on analog neuron model

Dendrite line

virtual ground

Op-amp

Assuming $R=100\,\text{M}\Omega$, $Vin=0.1\sim1\,\text{V}$, $\tau=1\,\mu\text{s}$

$E_w=\tau V^2/R = 0.1\sim10\,\text{fJ}$

Op-amps consume much more energy than resistors.
Functions of PSPs in spiking neurons

ISI: Inter-spike interval
PSP: Post-synaptic potential

\[ i_1 \quad | \quad \text{ISI} \quad | \quad i_2 \]

\[ P_1 \quad | \quad \tau_{\text{PSP}} \quad | \quad P_2 \]

\[ V_{\text{th}} \quad i_n \quad | \quad \text{time} \]

\[ \text{ISI} << \tau_{\text{PSP}} \quad \text{Integrator} \]

\[ \text{ISI} >> \tau_{\text{PSP}} \quad \text{Coincidence detector} \]
Integrate & fire neuron

Spatiotemporal summation of all linear rising waveforms of PSPs

Typical time course of PSPs: $\alpha$-function

$P_i = t \exp(-t)$

$V_n$
Time-domain weighted-sum calculation model

Given:

\[ y = \sum_{i} w_i x_i \]

Weighted summation (MAC: multiply and accumulate)

- \( x_i \rightarrow \) spike timing \( t_i \)
- \( w_i \rightarrow \) slope of PSP

Spike timing \( t_v \rightarrow y \)

\[ \sum_{i} w_i = \beta \]

\[ \sum_{i} w_i x_i = \frac{th + \beta (T_{in} - t_v)}{T_{in}} \]

Time-domain MAC calculation

\[ x_i \in [0,1] \quad \sum_{i=1}^{N} w_i x_i = 8.99 \]

\[ t_{v_{min}} = \frac{\theta}{\beta} + T_{in} \]

\[ t_{v_{max}} = \frac{\theta}{\beta} \]

\[ t_v = 1.635 \]

\[ \frac{\theta/\lambda + \beta(T_{in} - t_v)}{T_{in}} = 8.99 \]

Q. Wang et al. ICONIP 2016
Energy consumption using resistive elements

\[ V_{in} = \frac{R}{C} V_{in} \]

\[ E_{wsum} = CV^2 \]

If \( C=1fF \), \( V=1V \), \( E_{wsum}=1fJ \).
Assuming \( N=100 \), \( E_w = E_{wsum}/N = 10aJ \)

To guarantee time resolution, \( \tau=1\mu s \) (1,000 steps with 1 ns)

\( R_{ON} \sim 1G\Omega \), \( R_{OFF} \sim 1T\Omega \)

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

To reduce energy, reduce (parasitic) interconnect capacitance

Q. Wang et al. ICONIP 2016
Time-domain analog cross-bar circuit

Pre-neuron

Post-neuron

**Dendrites**

Input

Axon

Output

$\begin{align*}
&i_1, i_2, i_3, P_1, P_2, P_3, w_1, w_2, w_3, \theta, V_n, s, spk_n, t_1, t_2, t_3, t_v, T_{in}
\end{align*}$

Resistance change memory

Low parasitic (interconnect) capacitance

Low-power neuron circuit (comparator)

$\begin{align*}
&\text{[ReRAM device]} \\
&\text{[Nanowire]}
\end{align*}$
# Comparison among different approaches

<table>
<thead>
<tr>
<th>Approach</th>
<th># of neurons/synapses</th>
<th>Energy per synapse operation</th>
<th>Processing frequency</th>
<th>Power consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biological</td>
<td>Human brain</td>
<td>$10^{11}$/10^{15}</td>
<td>10^{-16}<del>/10^{-15} J (0.1</del>1 fJ)</td>
<td>10~/100 Hz (10% active)</td>
</tr>
<tr>
<td>Digital</td>
<td>Super Comput. (Kei「京」) (*1)</td>
<td>1.7x10^9/1.0x10^{13}</td>
<td>(6.5x10^{-4}J) (~1 mJ)</td>
<td>Brain:1s = Kei: 2,400s (4.4 fires/s)</td>
</tr>
<tr>
<td></td>
<td>Digital chips (TrueNorth~)</td>
<td>1x10^6/256x10^6</td>
<td>~10^{-13} J (&lt;0.1 pJ)</td>
<td>1 kHz</td>
</tr>
<tr>
<td>Analog</td>
<td>Voltage/current-domain</td>
<td>[1x10^6/256x10^6] (*2)</td>
<td>&gt;~10^{-15} J (~1 fJ) (*3)</td>
<td>&lt;=~MHz</td>
</tr>
<tr>
<td>Time-domain</td>
<td>[1x10^6/256x10^6] (*2)</td>
<td>&gt;~10^{-17} J (~10 aJ) (*3)</td>
<td>&lt;=~MHz</td>
<td>[~3 mW] (*4)</td>
</tr>
</tbody>
</table>

(*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

Chaotic Boltzmann machines (CBMs)

Dynamics of CBMs

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

\[
S_i \leftarrow 0 \quad (x_i = 0)
\]

\[
S_i \leftarrow 1 \quad (x_i = 1)
\]

\[
z_i = \sum_{j=1}^{N} w_{ij} s_j + \theta_i \quad (w_{ij} = w_{ji})
\]

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

Slope change in \( x_i \) when \( S_j \) changes

without input
A CMOS unit circuit for CBMs

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

Implemented using subthreshold region of MOSFET

M. Yamaguchi et al. ICONIP 2016.
Measurement results of CBM VLSI chip

3 neurons with 3 synapses

Continuous scan

Single scan

$S_1$, $S_2$, $S_3$, $V_{x3}$

20μs
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**.