Hardware Acceleration for Neuromorphic Computing
– An Evolving View

Yiran Chen & Hai (Helen) Li
Electrical and Computer Engineering
University of Pittsburgh
www.ei-lab.org
Outline

• Motivations of Brain Inspired Computing
  • Research Spotlights
    – Probabilistic inference: RNNLM: FPGA
    – Pattern detection: BSB: Memristor Crossbar
    – Reconfigurable Neuromorphic Accelerator
  • Challenges and Perspectives
Evolution of Computer Power

The Age of Intelligent Machines
MIT Press, 1990
ISBN 0-262-11121-7

Dr. Raymond Kurzweil

“...A sufficiently advanced computer program could exhibit human-level intelligence”

Million instructions per second (MIPS)

10⁹

10⁶

10³

10⁰

10⁻³

10⁻⁹

Human
100G Neurons

Monkey
3G Neurons

Lizard
2M Neurons

Worm
300 Neurons

Bacteria
1 “Neuron”
Artificial Brain – Intelligent, Creative, Self-aware

Google Brain Simulator (2012)
- Unsupervised training
- Deep learning
- 16,000 processors
- 1B connections
- 10M YouTube videos

IBM Watson (2011)
- Defeated Humanity in “Jeopardy”
- 9032 core IBM servers
- 16TB memory

In Developing (2014)
- Unsupervised learning
- Largest cluster for deep learning
- 100B neural connections
- Heavy cluster of GPUs

K Computer (2014)
- Human brain activity
- The 4th most powerful computer in the world
- 40 minutes of simulation → 1-second of brain activity
- 700,000 processor cores and 1.4M GB RAM
What Are The Major Limitations?

- Stalled single-thread performance
- Limited data throughput
- Constrained power efficiency

![Diagram showing CPU components: Arithmetic Logic Unit, Control Unit, Memory]

- Turing Machine
- von Neumann Arch.

- "Taming the Power Hungry Data Center" by Fusion-IO.

D. Hammerstrom, Neucomp, 2013
Brain Inspired Information Processing

- Brain inspired information processing relies on two main operators
  - *Pattern detection*
  - *Probabilistic inference*
- Multiple stages in human sensory processing
  - Primary sensory cortex detects a specific input (i.e. contour, color, or pitch, etc.)
  - Association cortex combines information from primary sensory cortex to produce perception
  - Higher order association combines different sensory association areas
Key Features of Neuromorphic Computing

- Performs pattern detection and probabilistic inference
- Massive parallel
- Closely coupled storage and computation
- Distributed storage with high redundancy provides reliability
- Simple unified building blocks (i.e., neurons)
- Analog domain operation

Non-conventional hardware architecture is required
Outline

• Motivations of Brain Inspired Computing

• Research Spotlights
  – Probabilistic inference: RNNLM: FPGA
  – Pattern detection: BSB: Memristor Crossbar
  – Reconfigurable Neuromorphic Accelerator

• Challenges and Perspectives
Context Aware Intelligent Text Recognition

...but beginning to perceive that the handcuffs were not for me and that the military had so far got....

BSB Recognition

Perception based on neural network models

...but beginning to perceive that the handcuffs were not for me and that the military had so far got....

Knowledge Base (KB)

Word Level Confabulation

Prediction

Sentence Level Confabulation

Prediction
Sentence Confabulation on FPGA

I saw __ dog.
A. a      B. an      or

Statistics based language model

\[
P(I \text{ saw } a \text{ dog}) = P(I) \times P(\text{saw} | I) \times P(a | I \text{ saw}) \times P(\text{dog} | I \text{ saw } a)
\]

\[
P(I \text{ saw } a \text{ dog}. ) > P(I \text{ saw } \text{ an } \text{ dog}.)
\]

Complexity: \(V^n\)

\(V\) – vocabulary size (10~60K)
\(n\) – sentence length

- Knowledge base is large and sparse
- Only the short-term perspective of a sequence.

Neural Network based language model

- Record the long-term historical information
- Stronger learning ability

**RNNLM – Training**

**Feed forward**

\[ h(t) = f(W_{ih} \cdot x(t) + W_{hh} \cdot h(t-1) + b_h) \]

\[ y(t) = g(W_{ho} \cdot h(t) + b_o) \]

**Activation functions**

\[ f(x^i) = \frac{1}{1+e^{-x_i}} \]

\[ g(x^i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \]

**Back propagation through time (BPTT)**

\[ \delta_p(t) = t_p(t) - o_p(t) \]

**Weigh update**

\[ W_{hi} \leftarrow W_{hi} + \eta \sum_{t=1}^{B} \delta_j(t) \cdot x_i(t) \]

- **High computational cost**
  
  \(~10^7\) parameters between hidden/output layer

10~20 epochs for convergence

- **Computation resource utilization**

<table>
<thead>
<tr>
<th>Matrix-vector Multi.</th>
<th>Activation Function</th>
<th>Vector Sum</th>
<th>Vector Scaling</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>71.0%</td>
<td>21.4%</td>
<td>2.3%</td>
<td>1.9%</td>
<td>3.4%</td>
</tr>
</tbody>
</table>

- **Memory efficiency**

![Memory efficiency graph](image)

**T. Mikolov, et al., ASRU, 2011**

S. Li, et al., FCCM 2015
Extend Inherent Parallelism of RNNLM

Pipeline stages are not balanced

- **Hidden layer**: $O(H \times H)$
- **Output layer**: $O(H \times V)$

Output layer is more critical

**Vocabulary size**: $V = 10K$

**Hidden layer size**: $H = 0.1K$

**BPTT time step**: $B = 4$

**Speed-up** (pipeline) $= 1.07 \times$

**Speed-up** (parallel) $= 3.86 \times$
System Overview

- **Resource Utilization**
  - Inherent parallelism
  - Data format conversion
  - Approximation of activation functions

- **Design Scalability**
  - Customized processing element (PE)

- **Memory Efficiency**
  - Multi-thread management unit (TMU)
  - Extensive data reuse

---

Diagram:
- Off-chip RNN Memory
- CE: Hidden
- CE: Output
- PE
- Global Control Unit
- Application Engine Hub
- Softmax
- Initializer
- CPU
Neural Acceleration – FPGA Exploration

Microsoft Research Sentence Completion (MRSC)

1. I have seen it on him, and could _____ to it.
   (a) write       (b) migrate         (c) climb
   (d) swear      (e) contribute

2. …

Training corpus:
19th and 20th Century novels (38M)

Vocabulary size: 10,583
Hidden layer size: 1024
BPTT: 4
## Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>20%</td>
</tr>
<tr>
<td>Smoothed 3-gram *</td>
<td>37%</td>
</tr>
<tr>
<td>RNN-100 with 100 classes</td>
<td>40%</td>
</tr>
<tr>
<td>RNNLM (this work)</td>
<td>46.2%</td>
</tr>
<tr>
<td>vLBL+NCE5 *</td>
<td>60.8%</td>
</tr>
<tr>
<td>Human *</td>
<td>91%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intel Eeon E5-2630</th>
<th>Nvidia GeForce GTX580</th>
<th>Convey HC-2ex (CPU+FPGA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core (#)</td>
<td>12</td>
<td>512</td>
<td>8*4</td>
</tr>
<tr>
<td>Clock</td>
<td>2.3 GHz</td>
<td>772 MHz</td>
<td>150 MHz</td>
</tr>
<tr>
<td>Memory BW (GB/s)</td>
<td>42.6 GB/s</td>
<td>192.4 GB/s</td>
<td>76.8 GB/s</td>
</tr>
<tr>
<td>Runtime (s)</td>
<td>2,566.80 (1 × )</td>
<td>130.01 (19.7 × )</td>
<td>160.72 (16.0 × )</td>
</tr>
<tr>
<td>Power-TDP (W)</td>
<td>95</td>
<td>244</td>
<td>25</td>
</tr>
<tr>
<td>Energy (J)</td>
<td>243,846 (1 × )</td>
<td>31,722 (7.7 × )</td>
<td>4,018 (60.7 × )</td>
</tr>
</tbody>
</table>
Outline

• Motivations of Brain Inspired Computing

• Research Spotlights
  – Probabilistic inference: RNNLM: FPGA
  – Pattern detection: BSB: Memristor Crossbar
  – Reconfigurable Neuromorphic Accelerator

• Challenges and Perspectives
Memristor – Rebirth of Neuromorphic Circuits

Memristor

Synapse Network

Memristor Crossbar

Programmable resistor w/ analog states

Nature, 2015

\[ [x_1 \ x_2 \ldots \ x_m] \]

\[ [y_1 \ y_2 \ldots \ y_n] \]

Natural matrix operation

\[ y_1 = \sum x_i \cdot g_{il} \]

\[ I_1 = \sum_{i=1}^{M} g_{il} V_i \]

High density

Figure 1 | Memristor crossbar. a. Integrated 12 × 12 crossbar.
Perception: Brain-State-in-a-Box (BSB)

**BSB Training Process**

\[ \Delta A = lr \ast (X - AX) \otimes X \]
\[ A = A + \Delta A \]

- \( X \) is the normalized input training vector;
- \( lr \) is the “Learning Rate”;
- \( \otimes \) is the outer product of two vectors;

**BSB Recall Process**

\[ X(t+1) = S(\alpha \cdot A \cdot X(t) + \lambda \cdot X(t)) \]

- \( X(t+1) \) and \( X(t) \) are \( N \) dimensional real vectors;
- \( X(0) \) is the input pattern (vector);
- \( A \) is the \( NxN \) connection matrix (memory);
- \( S() \) is a linear output-limiting function

**BSB Recall Convergence Criteria**

\[ X(t + 1) = X(t) \]
Different Design Approaches

**Level-base Design**
- Analog computation
- High speed
- Large design cost at DAC/ADC
- Susceptible to signal noise

**Spike-base Design**
- Mimicking biological systems
- High power efficiency
- High robustness and reliability
- Compatible to digital I/O
BSB Circuit: Recall Only

The recall function: \( x(t + 1) = S(\alpha \cdot A \times x(t) + \lambda \cdot x(t)) \)

Comparers detect the converge status.

Next iteration

We need two memristor arrays since memristor can only represent positive weights.

Input vector \( V(0) \)

Summing op-amps perform analog voltage signal addition/subtraction
BSB Circuit: Training

- **Original Delta rule:** \( \Delta w_{ij} = \alpha \cdot (t_j - y_j) \cdot x_i \)
- **Modification for hardware implementation:**
  \[
  \Delta g_{ij} \propto V_t \cdot T_t \cdot \text{Sign}(V_{\text{ref},j} - V_{\text{out},j}) \cdot \text{Sign}(V_{\text{in},i})
  \]

Minimize the design complexity meanwhile ensuring the weight change in the same direction as that of the Delta rule
Racing-BSB Model for Pattern Recognition

- BSB Model
  - Simple
  - Good noise resistibility
  - High correlation between convergence speed and pattern similarity

Each BSB model remembers one pattern (and its variations)

Recall against ALL BSB models

Compare convergence speed

Multiple matching patterns for each image.

Q. Qiu, DAC-DASS 2015
Racing-BSB Model for Pattern Recognition

- BSB Model
  - Simple
  - Good noise resistibility
  - High correlation between convergence speed and pattern similarity

Successfully adopted in various applications

EI-Lab, DAC’13

(a)

5 lowest average convergence 0-75 scaled to 0-255

character set

d g j m p s y BEK N O T W Z 2 5 8 - | @ | |' | : ?
Outline

• Motivations of Brain Inspired Computing

• **Research Spotlights**
  – Probabilistic inference: RNNLM: FPGA
  – Pattern detection: BSB: Memristor Crossbar
  – **Reconfigurable Neuromorphic Accelerator**

• Challenges and Perspectives
RENO Overview

- **An efficient memristor-based mixed-signal accelerator** is designed to speed up neuromorphic computing and support the implementations of a variety of neural network topologies;

- **A mixed-signal interconnection network (M-Net)** is proposed to assist the communication of computational signals among the MBCs;

- **An optimized configuration** is discussed and established by analyzing the impact of various design parameters on the system performance/accuracy.
Neuromorphic Computing Acceleration (NCA)

NCA Hardware

NCA Software

bool Recall(float *vec, float *wm)
{ /* simulate the synapse network*/
    for(i=0;i<BsbSize;++i) wx[i] +=
    wm[i*BsbSize+j] * vec[j];
    ......  
}

Find the candidate codes

Source-to-source translation

bool Recall(float *vec)
{   Send(NCA.id, vec);
    return Receive(NCA.id)
    ......
}

The neural topology

NCA-aware compilation

MOV D NCA.id, R1
......
SET NCA.id, #VAL
LAUNCH
DEQ R1, NCA.id

The NCA-aware executable
System Level Evaluation

- Two implementations representing tradeoffs between computation performance and accuracy
  - Multi-layer perception (MLP)
  - Auto-associative memory (AAM)

- 7 classification benchmarks
- Classification rate is used as reliability metric

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer</td>
<td>breast cancer diagnose</td>
</tr>
<tr>
<td>connect-4</td>
<td>connect-4 game</td>
</tr>
<tr>
<td>gene</td>
<td>nucleotide sequences detection</td>
</tr>
<tr>
<td>lymphography</td>
<td>lymph diagnose</td>
</tr>
<tr>
<td>MNIST</td>
<td>digit recognition</td>
</tr>
<tr>
<td>mushroom</td>
<td>poisonous mushroom discrimination</td>
</tr>
<tr>
<td>thyroid</td>
<td>thyroid diagnose</td>
</tr>
</tbody>
</table>
Experimental Setup

The Design Parameters of NCA Components

<table>
<thead>
<tr>
<th>Memristor</th>
<th>( R_L = 200\Omega, R_H = 160k\Omega, V_{th} = 2V )</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>MBC Array &amp; M-Net</th>
<th>Op amp</th>
<th>Network</th>
<th>Sigmoid</th>
<th>MBC</th>
<th>DAC</th>
<th>ADC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>100(\mu)W</td>
<td>0.72(\mu)W</td>
<td>10(\mu)W</td>
<td>0.69(\mu)W</td>
<td>5.2mW</td>
<td>3.8mW</td>
</tr>
<tr>
<td>Speed</td>
<td>0.60ns</td>
<td>4.2ns</td>
<td>0.24ns</td>
<td>3ns</td>
<td>333MHz</td>
<td>333MHz</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Area Estimation</th>
<th>NCA area ((mm^2))</th>
<th>NoC ((mm^2))</th>
<th>DAC/ADC ((mm^2))</th>
<th>MBC ((mm^2))</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-Net</td>
<td>0.943</td>
<td>0.598</td>
<td>0.014</td>
<td>0.072</td>
</tr>
<tr>
<td>D-Net</td>
<td>1.793</td>
<td>0.268</td>
<td>0.065</td>
<td>0.301</td>
</tr>
</tbody>
</table>

The Benchmark Implementation Details

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Description</th>
<th>Training error</th>
<th>MLP Topology</th>
<th>MBC array usage</th>
<th>Training error</th>
<th>AAM MBC array usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>cancer</td>
<td>breast cancer diagnose</td>
<td>0.02%</td>
<td>36→16→2</td>
<td>2 arrays in 1 group</td>
<td>0.07%</td>
<td>2 arrays in 1 group</td>
</tr>
<tr>
<td>connect-4</td>
<td>connect-4 game</td>
<td>0.02%</td>
<td>42→30→3</td>
<td>2 arrays in 1 group</td>
<td>0.08%</td>
<td>3 arrays in 1 group</td>
</tr>
<tr>
<td>gene</td>
<td>nucleotide sequences detection</td>
<td>0.09%</td>
<td>120→100→3</td>
<td>6 arrays in 2 groups</td>
<td>0.03%</td>
<td>12 arrays in 3 groups</td>
</tr>
<tr>
<td>lymphography</td>
<td>lymph diagnose</td>
<td>0.05%</td>
<td>29→19→4</td>
<td>2 arrays in 1 group</td>
<td>0.02%</td>
<td>4 arrays in 1 group</td>
</tr>
<tr>
<td>MNIST</td>
<td>digit recognition</td>
<td>0.35%</td>
<td>64→128→32→10</td>
<td>5 arrays in 2 groups</td>
<td>0.02%</td>
<td>10 arrays in 3 groups</td>
</tr>
<tr>
<td>mushroom</td>
<td>poisonous mushroom discrimination</td>
<td>0.01%</td>
<td>125→32→2</td>
<td>3 arrays in 1 group</td>
<td>0.01%</td>
<td>8 arrays in 2 groups</td>
</tr>
<tr>
<td>thyroid</td>
<td>thyroid diagnose</td>
<td>0.15%</td>
<td>21→32→3</td>
<td>2 arrays in 1 group</td>
<td>0.11%</td>
<td>3 arrays in 1 group</td>
</tr>
</tbody>
</table>
Impact of Deficient Hardware

- **Programming precision** due to limited device resolution

- **Device variations and signal fluctuations**

- AAM is more robust than MLP
Optimal MBC Size

- Performance: large size is preferable
- But... the classification rate decreases due to the aggravated variations
- The size of 64x64 obtains the best tradeoff
Compare to Other Designs

Example: Multilayer Perception (MLP)

Seven representative learning benchmarks.
All the results are normalized to the baseline CPU.

- Digital NPU + Digital NoC [1]
- MBC + Digital NoC
- RENO (MBC + Mixed-signal NoC)

[1] H. Esmaeilzadeh et al., MICRO’12
Outline

• Motivations of Brain Inspired Computing
• Research Spotlights
  – Probabilistic inference: RNNLM: FPGA
  – Pattern detection: BSB: Memristor Crossbar
  – Reconfigurable Neuromorphic Accelerator

• Challenges and Perspectives
Challenges I – Understanding

- Unfortunately we still do not know much about human brains.
- The artificial neural network models also evolves over years.
  - Representation of neuron: 1943, McCulloch (Pitt)
  - The 1st learning rule: 1949, Hebb
  - Neuron nets: 1955, Dartmouth Summer Research Project on AI
  - STDP (Spike-timing-dependent plasticity): 1973, Taylor
- Do we really need to understand brains before designing a useful N.C. system?
  - No. Many useful systems have been prototyped, e.g., IBM TrueNorth.
  - The debates on “Emulative vs. Simulative”...
Challenges II – Platform

- **Application Specific IC**
  - Misra *et al.*, *Neurocomputing*, 2010

- **Programmable Hardware**
  - Graf *et al.*, *NIPS*, 2009
  - Misra *et al.*, *Neurocomputing*, 2010

- **General Purpose Platform**
  - Graf *et al.*, *NIPS*, 2009

- **eNVM-Based Reconfigurable Design**
Challenges III – Technologies

- Are conventional CMOS and EDA technologies capable to support long-term research and development of N.C. systems?
  - Debates
    - Analog or Digital?
    - Spiking-based or level-based?
    - Synchronous or asynchronous?
    - CMOS or Post-Silicon?
  - Other Challenges
    - Programmability
    - Reliability
    - Scalability
    - Security

J. Hsu, IEEE Spectrum, 2014
B. Benjamin, Neurogrid, 2014
S. Miller, ESANN, 2012
F. Samarrai, UVAToday, 2014
Summary

• Non-conventional hardware architectures become critical for cognitive applications.

• A holistic scheme integrating the efforts on device, circuit, architecture, algorithm, etc. is necessary.

• There are many challenges and opportunities in circuit-architecture co-designs.

“I imagine a world where the difference between man and machine blurs, where the difference between humanity and technology fades, where the soul and silicon chip unite.”

Raymond Kurzweil

_The Age of Intelligent Machines_
Evolutionary Intelligence Lab (EI-Lab)

Our objectives are to enhance conventional systems and to explore new computing diagram by leveraging emerging technologies.

http://www.ei-lab.org
Thanks to EI-LAB members