NeoNexus: The Next-generation Information Processing System across Digital and Neuromorphic Computing Domains

Qinru Qiu
Dept. of Electrical Engineering and Computer Science, Syracuse University

Hai Li, Yiran Chen
Dept. of Electrical and Computer Engineering, University of Pittsburgh
Brain Inspired Computing

- The performance of traditional Von Neumann machine is reaching to a limit
- Human neocortex system has unprecedented performance and power efficiency
  - Particularly in language understanding, image recognition and situation awareness
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/mixed signal domain operation

We need non-conventional solutions for both hardware architectures and software computation models
Brain Inspired Cognitive Architecture

Analysis and Decision-Making Outputs

Raw Inputs

Abstraction 1
Lower level association (Statistical Inference)

- Simple associative memory with fuzzy output (i.e., ambiguity)
- Low complexity and retains maximum information

Abstraction 2
Higher level association (Statistical Inference)

- massive parallel pattern detection
- primary sensory cortex

Information association ≈ sensory association cortex

- Resolves ambiguity using probabilistic inference
Computation Models

- **Bottom layer: BSB (Brain-State-in-a-Box) model**
  - Convergence speed gives fuzzy information about pattern similarity

- **Upper layer: probabilistic inference**
  - Features and attributes represented as lexicons and symbols
  - Association among features represented by knowledge links
    - Captures \( \log[p(s_i | t_j)] \) between source and target symbols

- **Analogies to neocortical system**
  - Symbols \( \Leftrightarrow \) neurons
  - Knowledge link \( \Leftrightarrow \) synapses
  - Knowledge link values \( \Leftrightarrow \) Hebbian plasticity
  - Symbols in same lexicon \( \Leftrightarrow \) neurons with inhibition link
  - Symbols in different lexicon \( \Leftrightarrow \) neurons with excitation link
  - Likelihood calculation and belief propagation
    - Integration-and-fire with soft-winner-takes-all
Context Aware Intelligent Text Recognition

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

Perception based on neural network models

Prediction

Word Level Confabulation

BSB Recognition

Sentence Level Confabulation

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

- Recognizing skew and distorted text
- Recovers scratched words (60%)
- Separate connected characters (80%)

Word Accuracy

- Scanned w/o. scratch: ITRS 99%, Tessaract 100%
- Scanned w. scratches: ITRS 95%, Tessaract 93%
- Camera w. scratches: ITRS 92%, Tessaract 89%

Error due to scratch and error due to image distortion

In a town in Persia there dwelt two brothers, one named Ali Baba. Cassim was married to a rich wife and lived in a forest and selling it in the town.

One day, when Ali Baba was coming toward him in a diligence, a man

collapsed right in the middle of a packed courtroom. He was one of this country's most distinguished trial lawyers. He

was also a man who was as well-known for the three-thousand-dollar Italian suit as for his remarkable string of legal

victories. I simply stood there, paralyzed by the shock of what I had just witnessed. The great Julian M"anle had been reduced to a victim and was now squirming.

He said, Open, Sesame, and the door opened and shut behind him. He could have feasted his eyes all day on the treasures, but he now had to go together as much as it was possible. But when he was ready to go he could not remember what to say for thinking of his great riches. Instead of Sesame, he said, Open, Barley, and the door remained fast.

He named several different sorts of grain, all but the right one, and the door still stuck fast. He was so frightened at the danger he was in that he had much

The Wake-Up Call...
On Multicore Heterogeneous Architecture

Character images

Character candidates

Word candidates

Sentences

Result Gather

Image Processing

BSB Xeon/Cell/MIC/GPU

WordConfab Xeon/MIC

Sentence Confabulation

Result Gather
• Configuration: BSB (MIC0), Word (MIC1), Sentence (Xeon)
• The processing time of Tesseract rises rapidly as the image size increases and image quality reduces
• The processing time of ITRS remains stable

• Sentence and word confab time increases as the image quality reduces
• BSB processing is the bottleneck in ITRS
### Performance Comparison

<table>
<thead>
<tr>
<th></th>
<th>Xeon</th>
<th>Cell</th>
<th>Phi</th>
<th>GPGPU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clock Frequency (GHz)</strong></td>
<td>3.1</td>
<td>3.2</td>
<td>1.1</td>
<td>0.575</td>
</tr>
<tr>
<td><strong>Number of Physical Cores</strong></td>
<td>8</td>
<td>7</td>
<td>61</td>
<td>14</td>
</tr>
<tr>
<td><strong>Number of Logical Cores</strong></td>
<td>32</td>
<td>7</td>
<td>244</td>
<td>448</td>
</tr>
<tr>
<td><strong>Peak Performance (TFLOPS)</strong></td>
<td>~0.5</td>
<td>~0.2</td>
<td>~2</td>
<td>~1.0</td>
</tr>
<tr>
<td><strong>Sustained Performance (GFLOPS)</strong></td>
<td>116</td>
<td>96</td>
<td>128</td>
<td>83</td>
</tr>
<tr>
<td><strong>Utilization</strong></td>
<td>23%</td>
<td>48%</td>
<td>6.4%</td>
<td>8.3%</td>
</tr>
</tbody>
</table>

1 workload = checking 96 images against 93 patterns ≈ 58×10⁹ floating point operations
Brain-inspired Anomaly Detection

- An anomaly is a surprise
  - Something different from expectation
  - An attribute with low likelihood

- Likelihood-ratio test for anomaly detection
  - \( x \) is abnormal if it is less likely to be observed than \( \exists a_i \in A \)
    - \( x \): observed attribute, \( A \): the set of all potential attributes
  - Anomaly score: \( \frac{\max_i[el(a_i)] - el(x)}{\max_i[el(a_i)]} \)
    - A high anomaly score means relatively less likely event
    - Successfully applied to vehicle monitoring and cyber security detection
Observations

- Matrix-vector multiplication is the dominant operation in both layers:
  - Pattern matching layer: dense matrix, dense vector, consistent in matrix size
  - Inference layer: sparse matrix, sparse vector, large variations in size
- No intra-layer communication within pattern matching layer
- Frequent intra-layer communication is needed in association layer for belief propagation/likelihood estimation
  - Delay insensitive
  - Lexicons can work asynchronously
- Computation complexity of inference layer reduces as more features are considered
  - Example:
    - Sentence completion based on only language features requires at least 12-bit fix-point representation of knowledge value
    - Sentence reconstruction in ITRS, binary representation of knowledge value gives good results
  - Use additional knowledge / sensory information to reduce computation
  - Input specific computing kernel
Memristor – Rebirth of Neuromorphic Circuits

Memristor

Synapse Network

Memristor Crossbar

Programmable resistor w/ analog states

[\mathbf{x} = x_1 \ x_2 \ \ldots \ \ x_m]

Natural matrix operation

$$y_1 = \sum x_i \cdot g_{il}$$

$$\begin{bmatrix}
g_{11} & g_{12} & \cdots & g_{1n} \\
g_{21} & g_{22} & \cdots & g_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
g_{m1} & g_{m2} & \cdots & g_{mn}
\end{bmatrix}$$

[\mathbf{y} = y_1 \ y_2 \ \ldots \ y_n]

High density

$$I_i = \sum_{i=1}^{M} g_{il} V_i$$

$$I_{out}$$

El lab & HP lab

TiN-TaOx device

El lab & APL’13

TaN1+x

EI lab

APL’13

Programmable
resistor w/
analog states

HP lab, 2012

El lab, 2012

DAC’12
Two Design Approaches

Level-base Design

- Compatible to existing signal processing
- High speed computation

DAC

010

\(0.34\text{V} (V_{dd}=1.2\text{V})\)

101

\(0.86\text{V} (V_{dd}=1.2\text{V})\)

Op-amp

0.69V

ADC

100

DAC

010

Integrate and Fire (I&F)

010

101

Spike-base Design

- Closer to biological system
- Extremely high power efficiency

ADC

100
A Cross-Optimization Design Flow

Algorithm Evaluation & Test

- Applied Algorithm
- Trained Matrix
- Function Verification

Off-chip Training

Noise Injection

Circuit Design & Verification

- Training Circuit
- Programmed Crossbar
- Simulation Results

On-chip Training

Noise Injection
Neuromorphic Computing Acceleration (NCA)

NCA Hardware

NCA Software

Find the candidate codes

Source-to-source translation

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

  ......
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
NCA (MBC + Mixed-signal NoC)

[1] H. Esmaeilzadeh et al., MICRO'12
Neuron Clustering

Assume the maximum available crossbar size is 4x4

16 crossbars -> 6 crossbars + 2 discrete memristors
Summary

- **Selected publications**
  - ICCAD’13, TNNLS’14, ASP-DAC’14, ISCAS’14, IJCNN’14, CogSIMA’14, SSCI’14, SiPS’14, FCCM’15, DAC’15

- **Future works**
  - HW/SW co-design platform
  - SW: Design a smaller scale representative application for hardware prototyping
  - HW: Improve the scale of NCA design and evaluate its use in larger applications