Neuro-Inspired Adaptive Manycore SoCs and Applications

Abderazek Ben Abdallah
University of Aizu
Division of Computer Engineering
Adaptive Systems Laboratory
benab@u-aizu.ac.jp
Computing Trends: Opportunities & Challenges

- Huge progress in IC technology
- Emergence of IoT, Embedded/Ubiquitous/Pervasive applications.
- Large bandwidth and low power requirements.
- Data become more knowledge intensive (unstructured).
- Performance is limited by the Communication Network rather than the Computation Logics.
Neural Networks

1954
- F-Forward (FFNN)

1982
- Hopfield (HN)

1986
- Boltzmann, Restricted Boltzmann (RBM)

1988
- Radial Bas. Func, Autoencoder (AE)

1990
- Recurrent Neural (RNN)

2007
- Deep Belief Net (DBN), SAE

2004
- Echo State Network (ESN)

2008
- Conv. Neural Network (CNN)

2010
- Deconvolutional Network (DN)

2013
- Markov Chain (MC), VAE, ELM

2014
- Generative adversarial net(GAN, GRU, NTM)

2015
- Deep Residual Net (DRN), DCIGN,

2015
- Deep Convol. Inverse graphics Net (DCIGN)

Ex.
I: red flower
O: red flower image

[Kulkarni, 2015]

[Hopfield,1982]

[Hinton,1986]

[Hochreiter,1997]

[Benjamin, 2007]

[Jaeger, 2004]

[LeCum, 1998]

[He, 2015]

[Kulkarni, 2015]

April 23, 2017

benab@u-aizu.ac.jp
System-on-Chip

- **Multi/Many cores integrated on a single chip:**
  - Technology scaling
  - 3D integration
  - Many examples
    - STMicro
    - picoChip
    - Tilera Tile GX, Tile Pro
    - Intel Polari, ...

- **Complex apps, strict constraints**
  - Massively parallel applications
    - Big-data Analysis, Pattern Recognition, Deep Learning

---

The Soccer Robot based on SoCs.
https://www.intechopen.com/
Adaptive Systems-on-Chip (AsoC)

- **Compute** (High-performance, Real-Time)
- **Adapt** (needs to run in dynamic environments, where physical context, network topologies, and workloads are always changing)
- **Learn** (No programming!)
ASoC - Approach I: Scalable Packet-Switched SoC/NoC
Technology: 45nm CMOS Process
Chip Size: 2.205X2.220 (micron)
Frequency: 0.91 GHZ Confirmed
Supply voltage: 1.1V
Power Dissipation: 222.387 uW
Number of Pins: 557

Figure 10: Average packet latency evaluation of the realistic benchmarks.
ASoC - Approach II: Photonic Packet-Switched SoC/NoC

Electronics is not Good @ High bit/s Communication.

![Diagram](image)

Photonics is used for high-speed communication. Wavelengths are modulated and detected properly. Rings' resonance shifted by Δr due to temperature variation. MR1 is permanently off due to process variation.

Photonics is used for high-speed communication. Wavelengths are modulated and detected properly. Rings' resonance shifted by Δr due to temperature variation. MR1 is permanently off due to process variation.
Microring fault-tolerant

- Can support enormous intrinsic data bandwidths.
- Immune to the electrical interference.
- Bit-rate transparency
- Transmission latency is very small – Depends on the group velocity of light
ASoC - Approach III: Neuro-inspired Manycore SoC

• Neurons accumulating charge and firing over periods of time.

• Different from traditional ANN models, which are static organizations of mathematic formulas.

• Adaptive, spike-based (time dependent) and its structure can change over time:
  – Neurons and synapses may be created or destroyed over time.
The return to the neuro-inspired computing – Why Now?

- Huge progress in IT technologies
  - \( \rightarrow \) Number of PCs, cell phones doubles every 5 years
- Emergence of nano-devices
- Better understanding of neural functions
- IC power consumption is reaching its limit
- Brain is extremely energy efficient
  - \( \rightarrow \) \( \sim 10^{-16} \text{j/op/s} \) vs. \( 10^{-16} \text{j/op/s} \) for the best computers today

[NCLJ,A.Bullineria2015]
Deep Neural Network for DL
Parallel and Deep

(1) McCulloch-Pitts Model, 1943
Deep Neural Network for DL
Parallel and Deep

(2) The Perceptron
(Frank Rosenblatt, 1950s)

\[
y = \frac{1}{1 + e^{(-x)}}.
\]

\[
Sum = \sum_{i=1}^{N} I_i W_i + b_i,
\]
Brain Features

- Ten billion \((10^{10})\) neurons
- Neuron switching time > \(10^{-3}\) secs
- On average, each neuron has several thousand connections
- Hundreds of operations per second
- Face Recognition \(~0.1\) secs
- High degree of parallel computation, Distributed representations
- Each neuron is connected to the others through 10000 synapses
- It can learn, reorganize itself → adaptive
Computer Vs. Brain

- **Computer**
  - Calculation
  - Precision
  - Logic

- **Brain**
  - Pattern Recognition
  - Noise Tolerance
  - Complexity
Computer Vs. Brain

- High Power
- Storage and computation are separated
- Poor at recognition

- Low Power
- Storage and computation are not separated
- Good at recognition
Simplistic Neural Network Models

(1) Linear threshold gate
(McCulloch-Pitts Model, 1943)

\[ \text{Sum} = \sum_{i=1}^{N} I_i W_i, \]
\[ y = f(\text{Sum}). \]

W_1, W_2 \ldots W_m are weights normalized in (0,1) or (-1,1), \text{Sum} is the weighted sum, and \( T \) is a threshold constant, the function \( f \) is a linear step function at threshold \( T \).

This model is so simplistic that it only generates a binary output and also the weight and threshold values are fixed.

(2) The Perceptron
(Frank Rosenblatt, 1950s)

\[ \text{Sum} = \sum_{i=1}^{N} I_i W_i + b, \]

- Merging between McCulloch-Pitts model and Hebbian learning rule of adjusting weights.
- In addition to the variable weight values, the perceptron model added an extra input that represents bias.
DNN/DL systems have been highly successful in the areas of image classification and customer preference determination.

They are not designed for applications that are time-dependent/dynamic, which is our focus.
Spike Time Dependent Plasticity

- Large number of highly interconnected neurons with small local memory, communicating via spike timing.
- Connections (synapses) hold knowledge (weights)
- Need training on examples (supervised/un-supervised learning) to adjust weights (learning rule)

\[ s_{ij} = w_{ij} x_j \]

April 23, 2017
benab@u-aizu.ac.jp
Short term and long term plasticity for adaptive architecture

- Time multiplexing → Dynamically reconfigurable interconnect
- Short term and long term Plasticity → Adaptive architecture
- Time/space summation → Memory/logic operations at interconnect
- Time/space synchronization → Noise/defect tolerant signal processing
Neuron Action Potential

Memristor

Action potential

Excitatory postsynaptic potential

Action potential

Excitatory state required for reexcitation

Threshold for excitation

Hyperpolarization

Excitatory postsynaptic potential

Action potential

16 - 16 synapses firing
8 - 8 synapses firing
4 - 4 synapses firing

Resting membrane potential

Spike reception: EPSP
\[
\varepsilon(t - t_j^f)
\]

Spike emission: AP
\[
\eta(t - t_i^\wedge)
\]

\[
u_i(t) = \eta(t - t_i^\wedge) + \sum_j \sum_f w_{ij} \varepsilon(t - t_j^f)
\]

\[
u_i(t) = \mathcal{J} \Rightarrow \text{Firing:} \quad t_i^\wedge = t
\]
...but great opportunities come with great challenges

How do we interconnect a large number of SNs in a networked many-core fashion?

We should consider:

- Scalability
- Area utilisation
- Power consumption
- Throughput
- Synapse/neuron ratio

The brain is a network of neurons
• NASH processes data coming from sensors via 3D-TSV.

• Each neuro core processes a collection of N neurons each with M synaptic weights $W_{ij}$. 
Clustered Neuro-inspired Architecture

Neuro-Packet

LUT

TSV-Sharing Mechanism
Neuro Processing Core (NPC)

(a) MEMRISTOR based Neuro Processing Core (NPC)

(b) SRAM based Neuro Processing Core (NPC)
Neuro-inspired Architectures in Hardware - NASH

Structure of one PE

Mimicking the bio-neuron
Neuro-inspired Architectures in Hardware - NASH

Structure of one PE

Processing Stages in one PE

Decoder LUT

Encoder

Controller

Router

N1 N2 N3 N4...

C1 C2 C3 C4 C5 C6 C7

Decode Stage

Compute Stage

LUT Stage

Compute Stage

Code Stage

LUT Stage

Code Stage

Code Stage

LUT Stage
Some ASoC Applications

Demo 1

Character Recognition based on BP training

BANSMOM System (Conventional)

‘O’ letter
Some ASoC Applications

Demo 2

BANSMOM System (Conventional)

Character Recognition based on BP training

‘O’ letter

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

December 19, 2016

benab@u-aizu.ac.jp
ASL Adaptive SoCs

- **2006**: OASIS-1 – Scalable Packet-Switched Network-on-Chip
  - JASSST06, MCSOC12, JPDC14, SUP14

- **2013**: OASIS-2 - Fault-Tolerant Network-on-Chip
  - MCSOC14, JPDC14, SUP16

- **2014**: BANSMOM - Bio-Chip for Elderly Monitoring
  - ES2016, ACHRAF-MS1, KIMEZAWA-MS

- **2015**: PHENIC- High-bandwidth Photonic NoC
  - SUP16, MCSOC15, CANDAR16,

- **2016**: Adaptive Neuro-inspired SoC and Platform, NASH
  - TR-OASIS-NP-042015
Thank you!

Abderazek Ben Abdallah
benab@u-aizu.ac.jp