The N3XT Technology for Brain-Inspired Computing

H.-S. Philip Wong
Stanford University
1988 Winter Olympic Games in Calgary, Canada
100’s of kW

Source: Google
## Scale Up Requires Energy Efficiency

<table>
<thead>
<tr>
<th>Application</th>
<th>Hardware used</th>
<th>Estimated power consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Large scale</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emulating 4.5% of human brain</td>
<td>Blue Gene/P: 36,864 nodes, 147,456 cores</td>
<td>2.9 MW (LINPACK)</td>
</tr>
<tr>
<td>synapses, $10^9$ neurons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Deep sparse autoencoder:</td>
<td>1,000 CPUs (16,000 cores)</td>
<td>~100 kW (cores only)</td>
</tr>
<tr>
<td>$10^9$ synapses, 10M images</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Small to moderate scale</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Convolutional neural net</td>
<td>2 GPUs</td>
<td>1,200 W</td>
</tr>
<tr>
<td>with 60M synapses, 650K neurons</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Restricted Boltzmann Machine:</td>
<td>GPU</td>
<td>550 W</td>
</tr>
<tr>
<td>28M synapses; 69,888 neurons</td>
<td>CPU</td>
<td>65 W</td>
</tr>
<tr>
<td>Processing 1 s of speech using</td>
<td>GPU</td>
<td>238 W</td>
</tr>
<tr>
<td>deep neural network</td>
<td>CPU (4 cores)</td>
<td>80 W</td>
</tr>
<tr>
<td>S. B. Eryilmaz et al., IEDM 2015</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
These nanotechnology innovations will have to be developed in close coordination with new computer architectures, and will likely be informed by our growing understanding of the brain—a remarkable, fault-tolerant system that consumes less power than an incandescent light bulb.
Approaches of Neuromorphic Hardware
Approaches of Neuromorphic Hardware

- Biology-based models/ algorithms
- Conventional ML algorithms
Approaches of Neuromorphic Hardware

- Neuromorphic hardware
- Conventional hardware (CPU, GPU, supercomputers, etc)
Approaches of Neuromorphic Hardware

Neuromorphic hardware with analog non-volatile memory synapses

Conventional hardware (CPU, GPU, supercomputers, etc)
Approaches of Neuromorphic Hardware

- **Neuromorphic hardware**
  - with analog non-volatile memory synapses

- **Biology-based models / algorithms**

- **Conventional hardware (CPU, GPU, supercomputers, etc)**
  - Brain emulation on BlueGene
  - HTM
Approaches of Neuromorphic Hardware

- **Biology-based models / algorithms**
  - Neuromorphic hardware
    - with analog non-volatile memory synapses

- **Conventional hardware (CPU, GPU, supercomputers, etc)**
  - Brain emulation on BlueGene
  - HTM

- **Conventional ML algorithms**
  - “Cats on YouTube”
  - ANNs: ConvNets, DNNs, DBNs
Approaches of Neuromorphic Hardware

- **Neuromorphic hardware** with analog non-volatile memory synapses
  - TrueNorth
  - SpiNNaker
  - Human Brain Project
  - HTM
  - Brain emulation on BlueGene
  - Human Brain Project
  - TrueNorth
  - SpiNNaker
  - Human Brain Project
  - HTM

- **Conventional hardware (CPU, GPU, supercomputers, etc)**
  - "Cats on YouTube"
  - ANNs: ConvNets, DNNs, DBNs
Approaches of Neuromorphic Hardware

- **Neuromorphic hardware with analog non-volatile memory synapses**
  - Biology-based models / algorithms:
    - TrueNorth
    - SpiNNaker
    - Human Brain Project
  - Conventional ML algorithms:
    - Hebbian learning
    - Spike-based ANN
    - PCM, RRAM, CBRAM
  - Conventional hardware (CPU, GPU, supercomputers, etc):
    - Brain emulation on BlueGene
    - HTM

- "Cats on YouTube"
  - ANNs: ConvNets, DNNs, DBNs
Approaches of Neuromorphic Hardware

**Biology-based models / algorithms**
- TrueNorth
- SpiNNaker
- Human Brain Project

**Conventional ML algorithms**
- ANN, RBM, sparse learning
- PCM, RRAM

**Neuromorphic hardware with analog non-volatile memory synapses**
- Hebbian learning
- Spike-based ANN
- PCM, RRAM, CBRAM

**Conventional hardware (CPU, GPU, supercomputers, etc)**
- Brain emulation on BlueGene
- HTM

- “Cats on YouTube”
- ANNs: ConvNets, DNNs, DBNs
Many of these breakthroughs will require new kinds of nanoscale devices and materials integrated into three-dimensional systems and may take a decade or more to achieve.
N3XT Nanosystems
Computation immersed in memory

Memory

Ultra-dense vertical connections

Computing logic
N3XT Nanosystems

Computation immersed in memory

Impossible with today’s technologies
N3XT: Computation Immersed in Memory

3D Resistive RAM
Massive storage

1D CNFET, 2D FET
Compute, RAM access

MRAM
Quick access

1D CNFET, 2D FET
Compute, RAM access

1D CNFET, 2D FET
Compute, Power, Clock

thermal

Ultra-dense, fine-grained vias

Silicon compatible

Not TSV
Energy-Efficient Abundant-Data Computing: The N3XT 1,000×

Aly et al., IEEE Computer, 2015

Next-generation information technologies will process unprecedented amounts of loosely structured data that overwhelm existing computing systems. N3XT improves the energy efficiency of abundant-data applications 1,000-fold by using new logic and memory technologies, 3D integration with fine-grained connectivity, and new architectures for computation immersed in memory.
Non-Volatile Memory (NVM)

Phase change memory (PCM)

Metal oxide resistive switching memory (RRAM)

Conductive bridge memory (CBRAM)

D. Kuzum et al., *Nano Lett.* 2013, Y. Wu et al., *IEDM* 2013; A. Calderoni et al., *IMW* 2014
Non-Volatile Memory (NVM) → Synapse

- Analog programmable
- Scalable to a few nm
- Stack in 3D

D. Kuzum et al., *Nano Lett.* 2013, Y. Wu et al., *IEDM* 2013; A. Calderoni et al., *IMW* 2014
Nanoscale Memory as Synaptic Weights

Synaptic updates in the brain: basis for learning
Requirement: analog resistance change

100-step grey scale (1% resolution)

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)

Various STDP kernels

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)

Various STDP kernels

Various time constants

Nanoscale Memory Can Emulate Biological Synaptic Behavior

STDP (spike-timing-dependent plasticity)

Various STDP kernels
Various time constants
Weight update saturation

Hyper Dimensional (HD) Computing

Elements of Hyper dimensional Computing:
Also known as vector symbolic architectures or holographic reduced representation (Kanerva, *Cognitive Computation*, 1(2):139-159, 2009)

- Information is represented by High-dimensional representation (e.g., $D = 10,000$)
- Variables and values are combined into a “holistic” record using vector algebra:
  - Multiplication for Binding
  - Addition for Bundling
- Composed vector can in turn become a component in further composition
- Holistic record is decoded with (inverse) multiplication
- Approximate results of vector operations are identified with exact ones using content-addressable memory
HD Computing layers

Projected into hyper-dimensional space

Letters/Image features/
Letters/Image features/
Letters/Image features/
Phonemes/DNA sequences...

Random vectors: 1k ~ 10k bits

MAP Kernels
(Multiplication-Addition-Permutation)
\(v_1 \oplus v_2, \text{sum}(v_1, v_2, \ldots), \text{sum}(v_1), \text{perm}(v_1)\)

unknown 'closest'? Measure 'distance': learned & unseen vectors
(recognition/classification/reasoning/etc.)

Application

Representation

Computation

Inference
HD Computing layers

Projected into hyper-dimensional space

Random vectors: 1k ~ 10k bits

MAP Kernels
(Multiplication-Addition-Permutation)
$v_1 \oplus v_2, \text{sum}(v_1, v_2, \ldots), \text{sum}(v_1), \text{perm}(v_1)$

Unknown ‘closest’?

Measure ‘distance’: learned & unseen vectors
(recognition/classification/reasoning/etc.)

Application

Representation

Computation

Inference

Algorithms
HD Computing layers

Projected into hyper-dimensional space

MAP Kernels
(Multiplication-Addition-Permutation)

\[ v_1 \oplus v_2, \text{sum}(v_1, v_2, \ldots), \text{sum}(v_1), \text{perm}(v_1) \]

Apply Algorithms

System and Circuit Design

Letters/Image features/
Phonemes/DNA sequences...

Random vectors: 1k ~ 10k bits

"closest"?

Measure ‘distance’: learned & unseen vectors
(recognition/classification/reasoning/etc.)

Application

Representation

Computation

Inference
HD Computing layers

Projected into hyper-dimensional space

MAP Kernels
(Multiplication-Addition-Permutation)
$\mathbf{v}_1 \oplus \mathbf{v}_2, \text{sum} (\mathbf{v}_1, \mathbf{v}_2, \ldots), \text{sum} (\mathbf{v}_1), \text{perm} (\mathbf{v}_1)$

Random vectors: 1k ~ 10k bits

Unknown 'closest'?

Measure 'distance': learned & unseen vectors
(recognition/classification/reasoning/etc.)

Application
Representation
Computation
Inference

Algorithms
System and Circuit Design

Associative memory enabled by novel device technologies
Hyperdimensional (HD) Computing

- Monolithic 3D enables
  - Energy-efficient classification
  - Area efficient HD projection
  
  ➔ use RRAM variability & stochastic switching

![Diagram of 3D RRAM structure with labels: 3D RRAM + low power access transistors + address decoders.]

- High-density
- Inter-layer vias

- Low power computation
3D Enables In-Memory Computing

3D RRAM with FinFET BL select

TiN/Ti (TE)
HfOx
TiN
TiN
TiN (BE)

Layer 4 (L4)
Layer 3 (L3)
Layer 2 (L2)
Layer 1 (L1)

50 nm
TiN/Ti (50 nm)

HfO+ (5 nm)

TiN (20 nm)

50 nm

H. Li et al., Symp. VLSI Tech., 2016
MAP Kernels: 3D RRAM Approach

Key HD operations: multiplication, addition, permutation

- **Multiplication**
- **Addition**
- **Permutation**

Measured data on 4-layer 3D vertical RRAM
3D Integration of Memory and Logic Circuits within the Same Layer & Across Layers

Logic (Si CMOS)

Neuron circuits

Communication

Synapses/Weights (CNFET/2D FET) + RRAM

Synapses/Weights (3D RRAM)
Nano-Engineered Computing Systems Technology

Aly et al., IEEE Computer, 2015
Students and Post-Docs
Collaborators

Gert Cauwenberghs
Siddharth Joshi
Emre Neftci
(UC San Diego)

Jinfeng Kang
(Peking U)

Chung Lam
SangBum Kim
Matt Brightsky

K.S. Lee, J.M. Shieh, W.K. Yen...
(NDL, Taiwan)

NDL A Member of NARLabs
National Nano Device Laboratories
Sponsors

STARnet

SONIC

E2CDA “Engima”

Visual Cortex on Silicon
http://www.cse.psu.edu/research/visualcortexonsilicon.expedition/
Supported by National Science Foundation Expeditions in Computing Program

Stanford | SystemX Alliance

Stanford | Non-Volatile Memory Technology Research Initiative
Non-Volatile Memory Technology Research Initiative (NMTRI) @ Stanford University
End of Talk

Questions?

poly c - GST

amorphous

SiO₂

TiN

fully set state

partially reset state

fully reset state
Open Research Questions

1. Functionality $\rightarrow$ performance/Watt, performance/m² $\rightarrow$ variability $\rightarrow$ reliability
2. Scale up (system size), scale down (device size)
3. Role of variability (functionality, performance)
4. Fan-in / fan-out, hierarchical connections, power delivery
5. Low voltage (wire energy $\cong$ device energy)
6. Stochastic learning behavior $\rightarrow$ statistical learning rules
7. Meta-plasticity (internal state variables)
8. Timing as an internal variable
10. Algorithm-device co-design
11. **Materials/fabrication:** monolithic 3D integration **a must**, **MUST** be low temperature