Advances in Energy Efficient Neuromorphic Computing: Ready for Artificial Intelligence at the Edge?

Elisabetta Corti\textsuperscript{1,2} & Adrian M. Ionescu\textsuperscript{2}
\textsuperscript{1}IBM Research Zurich, Switzerland
\textsuperscript{2}EPFL, Lausanne, Switzerland
Outline

• Digitalization Era: from smartphone to cloud and edge computing

• Neuromorphic computing
  □ What is and is not
  □ Digital and analog implementation
  □ Disruptive computing: pattern recognition with coupled oscillators and stochastic neuromorphic computing

• Conclusion
Nanoelectronics: ~10 nm 3D transistors

Today: 10 nm:
- 100 millions transistors/mm²

Next: 7 nm

Virus
Negatively stained Influenza Virus, usually spherical or ovoid in shape, 80 to 150 nm.
2007: First wireless computer with sensors

2011: Cloud technologies

Future: Edge computing of the cloud technologies

2018: 30% revenue growth for cloud computing services. 183 billion $ revenue

2019: 80% of US citizens own a smartphone

Source: IDC
Why is edge computing needed?

- Application for which latency is a problem
- When security is concerned
- When dimensions and power consumption are limited
Future: billions of energy efficient IoT edge & fog devices...

**CLOUD:** Data Centers

**FOG:** Nodes

**EDGE:** Devices

Source: IDC
What is required to compute AI?

ImageNet 1-k benchmark:

90 epochs – with ResNet-50
→ $10^{18}$ single precision ops
→ NVIDIA M40 GPU = 14 days!

Audio Analysis:

FFT(1024) $\sim 10^6$ operations ($5 \cdot N \log_2(N)$)
Audio Spectrogram/s: $\sim 2 \times 10^6$ Ops.

The Cloud

Raw compute power is one thing, **Computational efficiency**, i.e. (F)lops per Watt is equally important!

Wearable Devices

$\sim 10^{17}$ FLOPS $\sim x 10^9$ $\sim 10^8$ IPS

Deep Learning is energy expensive!
Hardware accelerators are needed

Source: J. Weiss, IBM.
2025: AI-related semiconductors will account for 20% percent of all demand, $67 billion in revenue.

Growth for semiconductors related to artificial intelligence (AI) is expected to be five times greater than growth in the remainder of the market.

<table>
<thead>
<tr>
<th>AI semiconductor total available market, $ billion</th>
<th>AI semiconductor total available market, %</th>
<th>Estimated AI semiconductor total available market CAGR, 2017–25, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>240 17</td>
<td>256 32</td>
<td>382 65</td>
</tr>
</tbody>
</table>

1. Total available market includes processors, memory, and storage; excludes discretes, optical, and micro-electrical-mechanical systems.
2. Compound annual growth rate.
3. E = estimated.

Source: Bernstein; Cisco Systems; Gartner; IC Insights; IHS Markit; Machina Research; McKinsey analysis
How do we solve this problem?
How do we make computing more efficient?
A neuromorphic computer is not a brain but a **brain-like energy efficient system** to do machine learning & AI.

In standard Von Neumann architecture:
- Separation CPU/Memory
- Slower computation
- High power consumption (GPU)

In dedicated architecture:
- CPU/Memory in the same place
- Faster computation
- Reduced power consumption
- Reconfigurable, fault-tolerant
Technologies for Neuromorphic computing

1. Standard CMOS based solutions, but bringing memory near the computation

2. Analog computing promises 100x improvements – example of multiply-accumulation operations with memristors

3. Other solution: disruptive architectures

We require **NEW ARCHITECTURES** combined with **NEW TECHNOLOGIES**
1. Tremendous recent progress in CMOS neuromorphic computers

Standard CMOS technology but architecture reorganization

- **IBM’s TrueNorth** (DARPA’s SyNAPSE project)
  - 65 mW real-time neurosynaptic processor, 4096 neurosynaptic cores tiled in 2-D array, 1 million digital neurons and 256 million synapses, with computational energy efficiency = 400 GSOPS/Watt.

- **Intel’s Liohi** (September 2017)
  - 130,000 neurons, 130 million synapses

- **Spinnaker 2**
  - 64 KB SRAM- 18 cores
  - 16,000 neurons with eight million plastic synapsis per chip

Potential future applications: cognitive prosthetics, BMI, wearables, smart in situ imaging facilities.
2. Why analog computing

- Network with inputs $N_i$
- Each multiplied by weight $w_{ij}$
- Sum all products $M_j = \sum_i N_i \times w_{ij}$
- Apply filter function (activation, threshold, ...)

Multiply- accumulating instruction
2. Why analog computing

Computational efficiency of various technologies

Power Efficiency Scaling

Sum all products \( M_j = \sum_i N_i \times w_{ij} \)

Multiply-accumulating instruction

Digital design space

Brain

2. Why analog computing

### Computational efficiency of various technologies

<table>
<thead>
<tr>
<th>Power Efficiency Scaling</th>
<th>Three Orders of Magnitude</th>
<th>Region of power efficient algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>10MMAC(s)/W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100MMAC(s)/W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MMAC(s)/mW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10MMAC(s)/mW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MMAC(s)/uW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10MMAC(s)/uW</td>
<td></td>
<td>(i.e. Analog VMM)</td>
</tr>
<tr>
<td>100MMAC(s)/uW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MMAC(s)/pW</td>
<td>(c) Biological Neuron</td>
<td></td>
</tr>
<tr>
<td>10MMAC(s)/pW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100MMAC(s)/pW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1MMAC(s)/nW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10MMAC(s)/nW</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100MMAC(s)/nW</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Digital design space**

Multiplication and summation in analog circuit

**J. Hasler, B. Marr, Frt. Neurosc., 2013.**

- Multiplication: \( I_i = V_i \times G_{ij} \) (Ohm’s law)
- Summation: \( I_j = \sum_i I_i \) (Kirchhoff’s law)
Device Candidates for Cross-Bar Arrays

Synaptic Devices & Materials – Memristors

FeFET-based memristors memristive functions with FeFET with doped high-k are transferable to advanced scaling.

ReRAM (Resistive Ram)
- Inline oxides
- Small cell size
- Variability
- Asymmetry

PCM (Phase Change Memory)
- Commercial SCM
- Abrupt reset
- Conductance drift

MRAM (Magnetic RAM)
- High endurance
- Low power
- Small on/off ratio
- Limited # of states

ECRAM (Electro-Chemical RAM)
- Low power
- Symmetry
- Scalability
- No CMOS comp

Source: V. Bragaglia, IBM
Additional ref.: I. Boybat, Nat. Comm. 2018

Halid Mulaosmanovic et al, Applied Mat. & Interfaces, 2017.
Disruptive architecture (I): computing with time in coupled oscillators systems

- Idea: timing rather than amplitude information is used for computation.
- Coupled oscillators lock in frequency and the phase relation can be adjusted by the coupling resistance. Applications in pattern recognition.

With VO$_2$:
Single oscillator footprint: 200x200nm$^2$
Power consumption -> decreases with scaling of the device (now around 20µW)
Disruptive architecture (II): computing with time in coupled oscillators systems


Image filter with VO2 capacitively coupled oscillators

Corti et al. ICRC 2018.

Associative memory computation with VO2 coupled oscillators.
Resistively coupled for compatibility with RRAM or PCM technology.
Spiking neuromorphic computation: towards probabilistic MIT neurons for SSM


Benchmarking challenge for artificial neuron technology:

- **Energy efficiency:**
  - $> 10^{13}$ spikes/Joule
  - (< 0.1pJ/spike)

- **Footprint:** < 100um²

- **RT to 100°C**

  **Human brain:**
  - $1.8 \times 10^{14}$ spikes/Joule
  - **Neuron < 3um²**
  - ~36-37°C (deep brain temperature is less than 1°C higher than body temperature)

**Source:** A.M. Ionescu, EPFL.

**II. HOW CAN ONE COMPUTE WITH STOCHASTICALLY SPIKING NEURONS?**

The human brain employs about $10^{14}$ spiking neurons, a number which is in the same range as the number of transistors in a modern supercomputer. But whereas the power consumption of supercomputers lies in the range of megawatts (for example, the K computer consumes almost as much energy as 10 000 suburban homes), the brain only consumes 30 W. Hence the brain provides an intriguing paradigm for energy-efficient computing. But in...
VO$_2$ probabilistic neurons

Biological plausibility and stochasticity in scalable VO$_2$ active memristor neurons

Wei Yi$^1$, Kenneth K. Tsang$^1$, Stephen K. Lam$^1$, Xiwei Bai$^1$, Jack A. Crowell$^1$ & Elias A. Flores$^1$
**VO₂ probabilistic neurons**

**Stochastic VO₂ neuron**

S. Datta group @ VLSI 2017:
- probabilistic hardware for stochastic IMT neuron due to nucleation (filament formation) IMT/MIT process.
- implementation of SSM for pattern recognition
Recent advances in Ge-doped VO$_2$ material @ EPFL

- Phase-change transition temperature near 90°C, enabling future implementations of advanced neuromorphic hardware.

Further grain engineering & control needed

Grains ~ 50 – 200nm
Sputtered Ge-VO$_2$
The Future of Energy Efficient Computing will be Hybrid: CMOS + Neuromorphic + Quantum

Source: A.M. Ionescu @ IEDM 2017
Conclusions

• **Energy efficiency technologies** form the next driver in the zettabyte era.
• Paradigm change of **distributed computation from Cloud to Edge**.
• **Future neuromorphic hardware needed**: CMOS + new emerging material, device and architecture concepts.
• The **future of computing will probably be hybrid**, with CMOS, neuromorphic and quantum computing serving from edge-to-cloud.
Thank you!

Questions?
Efficiency is also a lot about (not) Moving Data

Compute Energy

<table>
<thead>
<tr>
<th>Operation</th>
<th>Energy (pJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8b Add</td>
<td>0.03</td>
</tr>
<tr>
<td>16b Add</td>
<td>0.05</td>
</tr>
<tr>
<td>32b Add</td>
<td>0.1</td>
</tr>
<tr>
<td>16b FP Add</td>
<td>0.4</td>
</tr>
<tr>
<td>32b FP Add</td>
<td>0.9</td>
</tr>
<tr>
<td>8b Mult</td>
<td>0.2</td>
</tr>
<tr>
<td>32b Mult</td>
<td>3.1</td>
</tr>
<tr>
<td>16b FP Mult</td>
<td>1.1</td>
</tr>
<tr>
<td>32b FP Mult</td>
<td>3.7</td>
</tr>
<tr>
<td>32b SRAM Read (8KB)</td>
<td>5</td>
</tr>
<tr>
<td>32b DRAM Read</td>
<td>640</td>
</tr>
</tbody>
</table>

(Values are trends)

(Energy/Performance)

For better:
- Power Efficiency
- Performance

We need to:
- Not move data
- Bring «Stuff» closer together

Source: J. Weiss IBM
Anatomy of “Heavy” Workloads – The Actual Problem

Scientific Workloads

- (electro) Chemistry
- Drug Discovery
- Weather/Climate

AI & Machine Learning

- Backpropagation
- PDEs
- Clustering Algorithms
- Image Classification

Matrix-Vector Multiplication

\[ Av = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \times (x, y) \]

Is Common to all Workloads! and computationally Expensive!
RRAM Development in IBM Zurich

**Goal** (symmetry & linearity)

![Conductance vs Potential and Depr.](image)

Baseline Stack

- **Substrate**
- **TiN**
- **Ti**
- **HfO$_2$**
- **TiN**

Tuned Stack

- **Substrate**
- **TiN**
- **Metal-Oxides**
- **HfO$_2$**
- **TiN**

Replace scavenging layer with intercalation metal-oxides

---

**Current Status**

In-house:
- Material growth
- Device fabrication
- Characterization

- **60 x 60 µm$^2$**
- **.1 x .1 µm$^2$**

---

**Device DC-Characteristics**

![Device Characteristics](image)

**Optimized Device Characteristics**

New Results to be presented at MEMRISYS 2019 Dresden