Mixed-signal neuromorphic VLSI devices for spiking neural network

Ning Qiao
Institute of Neuroinformatics
University of Zurich and ETH Zurich

Jun 20, 2018
ADAC6 Workshop
Outlines

- Brain inspired computing
- Neuromorphic engineering
- Analog synapse and neuron circuits
- Multi-core Neuromorphic architectures
- Applications
The cost of current computing technologies is not sustainable

- In 2017 > 10 zettabytes of data were produced.
- IT infrastructures and consumer electronics absorbed > 10% of the global electricity supply.
- By 2025, over 50 billion of Internet-of-Things (IoT) devices will be interconnected.
- Over 180 zettabytes of data will be generated annually, potentially leading to a consumption of one-fifth of global electricity.
Current trends in computing HWs

40 Years of Microprocessor Trend Data

- Transistors (thousands)
- Single-Thread Performance (SpecINT x 10^3)
- Frequency (MHz)
- Typical Power (Watts)
- Number of Logical Cores

Brain-Inspired computing

- Slow, noisy and variable processing elements
- Massively parallel distributed computation, local connectivity
- Real-time interaction with the environment
- Complex spatio-temporal pattern recognition
- Foraging, navigation, language, and social behavior

1mg weight
1mm$^3$ volume
960,000 neurons
10e-15 J/spike
<100 μW
Neuromorphic Computing vs. Neuromorphic Engineering

**Neuromorphic “computing”**
- Dedicated VLSI hardware.
- High performance computing.
- Application driven.
- Conservative approaches.

**Neuromorphic engineering**
- Fundamental research.
- Deeply rooted in biology.
- Emulation of neural function.
- Subthreshold analog and asynchronous digital.
Current trends in neuromorphic processors
Not so radically different after-all (not solving the von Neumann bottleneck problem)
Current trends in neuromorphic processors
Not so radically different after-all (not solving the von Neumann bottleneck problem)
“Listen to the silicon” (original approach)
Mixed-signal analog/digital neuromorphic systems

- Analog/digital computation, digital asynchronous communication.
- Directly emulate the physics of neural systems.
- Massively parallel collections of non-linear circuits.
- Realistic neural and synaptic dynamics
- Distributed memory
- Co-localized memory and computation
Channel current-voltage relationships

![Graph showing channel current-voltage relationships. The graph plots logarithmic current (Ids) against gate voltage (Vgs). It highlights subthreshold and above threshold states.]
Analog circuits
Direct emulation of synaptic dynamics

\[ \tau \frac{d}{dt} I_{syn} + I_{syn} = \frac{I_{thr} I_w}{I_\tau} \]

[Bartolozzi and Indiveri, 2007]
Analog circuits
Direct emulation of neuron dynamics

\[ \tau \frac{d}{dt} I_{mem} + I_{mem} \approx \frac{l_{th} l_{in}}{l_{r}} - I_g + f(I_{mem}) \]

\[ \tau_{ahp} \frac{d}{dt} I_g + I_g = \frac{l_{thr} l_{ahp}}{l_{r}} \]
Spike-based plasticity VLSI implementation
Cortical networks: a high degree of clustering

Pyramidal Cell of Layer 3 of Cat Visual Cortex. Dendrites (Green), Axon (Red), Clusters of Boutons (Black) in Layer 3 and 5. Scale bar, 500 µm

[R.J. Douglas and K.A.C. Martin, Neuron, 2007]

Minimize memory requirements:
two-stage routing

\[2\sqrt{F \times \log_2(C) \times \log_2(N)}\] bits/neuron

[Moradi and Indiveri 2014]
Memory optimized multi-core neural architecture
Hierarchical routing with heterogeneous memory structures

- Two-stage + 2D tree + 2D mesh multi-cast routing schemes using both source-address and destination-address encoding.
- Fully asynchronous hierarchical routers for intra-core (R1), inter-core (R2) and inter-chip (R3) connectivity.
- Embedded asynchronous CAM and SRAM cells distributed across and within cores.
Co-localized memory and computation
FD-SOI design, ready for beyond CMOS technology

- Multiple parallel I/O pathways
- Multiple distributed asynchronous SRAM LUTs
- Distributed multi-bit TCAM cells
- Capacitors for state dynamics and learning

- Ideal for integration with (binary) resistive memories
- Ideal for integration with (learning) memristive devices
- Ideal for integration in 3D VLSI technology
Latest NP chip specs

A large-scale, multi-core, neuromorphic processor DynapSEL in 28 nm FDSOI, is reported in ISSCC 2018.

<table>
<thead>
<tr>
<th></th>
<th>IBM TrueNorth</th>
<th>DynapSEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology</td>
<td>28nm CMOS</td>
<td>28 nm FDSOI</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>0.7V</td>
<td>0.73 V</td>
</tr>
<tr>
<td>Neuron Type</td>
<td>Digital</td>
<td>Analog</td>
</tr>
<tr>
<td>Neurons per core</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>Core Area</td>
<td>0.094 mm²</td>
<td>0.36 mm²</td>
</tr>
<tr>
<td>Computation</td>
<td>Time multiplexing</td>
<td>Parallel processing</td>
</tr>
<tr>
<td>Fan In/Out</td>
<td>256/256</td>
<td>2k/8k</td>
</tr>
<tr>
<td>Synaptic Operation / Second / Watt</td>
<td>46 GSOPS/W</td>
<td>300 GSOPS/W*4</td>
</tr>
<tr>
<td>Energy per synaptic event</td>
<td>10 pJ</td>
<td>&lt;2 pJ*2</td>
</tr>
<tr>
<td>Energy per spike</td>
<td>3.9 nJ</td>
<td>&lt;1.68 nJ*3</td>
</tr>
</tbody>
</table>

- 8X Fan-in / 32X Fan-out for more complex spiking networks
- 13X more power efficient

© aiCTX. Confidential
Neural dynamics
with appropriate time constants

Paradigm shift

• Radically different from von Neumann architectures.
• Co-localized memory and computation.
• No virtual time (time represents itself).
• Data/event driven computation.

“Slow” (biologically plausible) time constants

• For interacting with the environment in real-time.
• Inherently synchronized with the real-world “natural” events.
• To process “natural” sensory signals efficiently (low bandwidth/power).
Event-based convolutional network
Real-time autonomous behaving agents
Connecting neuromorphic processors to neuromorphic sensors and robots
Hardware preliminary (state-of-the-art) results

![Diagram of neuromorphic cognitive systems](image)

- **Input Signal Shape**
- **Reservoir Raster Plot**
- **Out Neurons Predicted Activity**

Source: Neurons, Input Signal, Reservoir Neurons, Output Layer

INN | UZH | ETH | Zürich
Distributed Artificial Intelligence

Autonomous sensory-motor systems

embedded systems & emerging memory technologies

Brain Machine Interfaces & prosthetics
Team Work: Institute of Neuroinformatics

- Ning Qiao (INI)
- Yulia Sandamirskaya (INI)
- Lorenz Müller (INI)
- Melika Payvand (INI)
- Elisa Donati (INI)
- Dongchen Liang (INI)
- Raphaela Kreise (INI)
- Moritz Milde (INI)
- Marc Osswald (inSightness)
- Dora Sumislawksa (GeorgiaTech, USA)
- Fabio Stefanini (Columbia Univ., USA)
- Jonathan Binas (Univ. Montreal, CA)
- Emre Neftci (UC Irvine, USA)
- Saber Moradi (Yale, USA)
- Hesham Mostafa (UCSD, USA)
- Chiara Bartolozzi (IIT, Italy)
- Elisabetta Chicca (Univ. Bielefeld, DE)
- Stefano Fusi (Columbia Univ., USA)
Technology-transfer effort to commercialize

- Dr. Ning Qiao
- Prof. Giacomo Indiveri
- Dr. Kynan Eng
- Dr. Dylan Muir
- Dr. Sadique Sheik
- Dr. Qian Liu
- Felix Bauer
- Carsten Nielsen
- Ole Richter
- Anita Tuomi
Thank you for your attention
Spiking Neuron Network (SNN)
Spiking Neuron Network (SNN)

Communication

Computation

Learning
Co-localized memory and computation
FD-SOI design, ready for beyond CMOS technology

- No I/O bottleneck
- No memory bottleneck
A closed-loop bi-directional BMI