DYNAP-SEL: An ultra-low power mixed signal Dynamic Neuromorphic Asynchronous Processor with SElf Learning abilities

Giacomo Indiveri
Institute of Neuroinformatics
University of Zurich and ETH Zurich

NICE 2019
March 26, 2019
Credits

- Ning Qiao
- Elisa Donati
- Dongchen Liang
- Saber Moradi (Silicon Valley)
- Fabio Stefanini (Columbia University)
Neuromorphic computing origins

Basic research
- Fundamental research.
- Emulation of neural function.
- Subthreshold analog
- Asynchronous digital.

Recent developments
- Dedicated VLSI hardware.
- High performance computing.
- Application driven.
- Conservative approaches.

Carver Mead
Misha Mahowald
Highly interdisciplinary basic research rooted on neuroscience, non-linear dynamical systems theory, device physics, microelectronics,…

Exploit the physics of silicon and emerging nano-technologies to reproduce the bio-physics of neural systems.

Develop distributed multi-core spiking architectures using mixed signal analog/digital VLSI circuits.

Build real–time autonomous cognitive agents able to carry out behavioral tasks in complex environments.
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits (ultra low-power).
- Re-programmable network topology and connectivity.
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
  - Massively parallel, distributed computation.
  - Time represents itself (no time-multiplexing)
  - Biologically plausible temporal dynamics
  - Adaptation and learning at multiple time scales.
  - No clock and no active circuits (ultra low-power).
  - Re-programmable network topology and connectivity.
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
  - Time represents itself (no time-multiplexing)
  - Biologically plausible temporal dynamics
  - Adaptation and learning at multiple time scales.
  - No clock and no active circuits (ultra low-power).
  - Re-programmable network topology and connectivity.
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
  - Biologically plausible temporal dynamics
  - Adaptation and learning at multiple time scales.
  - No clock and no active circuits (ultra low-power).
  - Re-programmable network topology and connectivity.
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
  - Adaptation and learning at multiple time scales.
  - No clock and no active circuits (ultra low-power).
  - Re-programmable network topology and connectivity.
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits (ultra low-power).
- Re-programmable network topology and connectivity.

![Neuromorphic Processor Diagram]

Spiking inputs
Homeostatic adaptation
Spiking output
Synapse and learning block
Soma block
Synaptic scaling block

Inputs
Encode Decode
Address Event Bus
Source Chip
Outputs
Destination Chip
Action Potential
Address-Event representation of action potential
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits (ultra low-power).
- Re-programmable network topology and connectivity.
Spikes in, and spikes out.
Analog subthreshold & digital asynchronous circuits.
Massively parallel, distributed computation.
Time represents itself (no time-multiplexing)
Biologically plausible temporal dynamics
Adaptation and learning at multiple time scales.
No clock and no active circuits (ultra low-power).
Re-programmable network topology and connectivity.
DYNAP-SEL: Dynamic Neuromorphic Asynch Processor with Self Learning

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chip Name</td>
<td>DynapSEL</td>
</tr>
<tr>
<td>Process</td>
<td>ST28FD-SOI</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>1V</td>
</tr>
<tr>
<td>IO Number</td>
<td>176 + (internal 59)</td>
</tr>
<tr>
<td>Chip area</td>
<td>2.8mm x 2.6mm</td>
</tr>
<tr>
<td>Core Numbers</td>
<td>4 non-plastic cores 1 plastic core</td>
</tr>
<tr>
<td>Neuron Type</td>
<td>Analog AExp I&amp;F</td>
</tr>
<tr>
<td>Non-plastic Synapse Type</td>
<td>TCAM based 4-bit</td>
</tr>
<tr>
<td>Plastic Synapse Type</td>
<td>Linear 4-bit digital</td>
</tr>
<tr>
<td>Throughput of Router</td>
<td>1G Events/second</td>
</tr>
<tr>
<td>Scalability</td>
<td>16 x16 chips non-plastic core 4 x4 chips (plastic cores)</td>
</tr>
</tbody>
</table>
Ready for future emerging nano-technologies

- Distributed SRAM and TCAM memory cells
- Capacitors for state dynamics
- Ideal for co-integration with binary, non-volatile resistive memory devices
- Ideal for co-integration with multi-level volatile/non-volatile memristive devices
- Ideal for integration in 3D VLSI technology

[Qiao and Indiveri, 2016],[EU ICT NeuRAM3 (687299) project]
Analog synapse circuits

Synapses x 64

Leakage Canceling

DPI

NMDA
Analog neuron circuits
Synapse and neuron circuit response properties

Graph showing the response of synaptic current ($I_{syn}$) and membrane current ($I_{mem}$) over time ($T$) with a threshold of $20 \text{nA}$.
Synapse and neuron circuit response properties

![Graph showing the relationship between synapse activity and synaptic events.](image-url)

- **dpi\_tau! = 4pA**
- **dpi\_tau! = 8pA**
- **dpi\_tau! = 12pA**

**Ax** and **Neuron Activity** (Hz)
Synapse and neuron circuit response properties

![Graph showing neuron activity vs. synaptic events](image)
Possible routing schemes

**Shared bus**
- Length: \( l \)
- SCX project

**1D Grid, Tree**
- Neurogrid

**2D Mesh**
- SpiNNaker, Tianji, TrueNorth

**2D Mesh** gives us maximum flexibility, but it is very expensive in terms of resources required: all-to-all connectivity for \( N \) neurons with a fan-out of \( F \) require

\[
F \log_2(N) \text{ bits/neuron}
\]
Cortical networks: a high degree of clustering

Pyramidal Cell of Layer 3 of Cat Visual Cortex.
Dendrites (Green), Axon (Red), Clusters of Boutons (Black).

Minimize memory requirements:
two-stage routing

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

[Douglas and Martin, 2007] [Moradi and Indiveri 2014]
Cortical network example

<table>
<thead>
<tr>
<th>Routing</th>
<th>bits/neuron</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard</td>
<td>$F \log_2(N)$</td>
</tr>
<tr>
<td>two-stage</td>
<td>$\sqrt{F \log_2(N)} \cdot 2\sqrt{\log_2(C)}$</td>
</tr>
</tbody>
</table>

2-stage Routing

Standard Routing

Required Memory (Kbits)

- For standard routing, the bits per neuron is $F \log_2(N)$.
- For two-stage routing, the bits per neuron is $\sqrt{F \log_2(N)} \cdot 2\sqrt{\log_2(C)}$. 
Multi-core neural architecture
with heterogeneous memory structures

- Combines best of 2D mesh, 2D tree, and multi-cast schemes, with combination of source-address and destination-address routing.
- Fully asynchronous hierarchical routers for intra-core (R1), inter-core (R2) and inter-chip (R3) connectivity.
- Distributed asynchronous CAM memory cells within the core and SRAM cells in the routers.

[Moradi et al. 2018]
Supports arbitrary large numbers of cores but assumes networks with structured connectivity
On-chip spike-based learning
moving beyond plain STDP

Recent spike-driven learning algorithm

Spike-driven weight change depends on the timing of the pre-synaptic input, and on the value of the post-synaptic neuron’s state variables.

W. Senn, S. Fusi, N. Brunel, S. Sheik, E. Neftci, R. Zecchina, M. Memmesheimer, etc.

Requirements for efficient implementation

- low resolution: use a small number of stable synaptic states;
- redundancy: implement many synapses that see the same pre- and post-synaptic activity
- stochasticity & inhomogeneity: induce LTP/LTD only in a subset of stimulated synapses.
Spike-driven learning rule in SW simulations

Pre-synaptic spikes

Synaptic internal variable $X(t)$

Post-synaptic depolarization $V(t)$

[Brader et al., 2007]
On-chip spike-based weight update measurements
On-chip spike-based weight update measurements
On-chip neural dynamics measurements

Experimental results

![Graph showing current (A) and time (s) for DPI and mem currents](image-url)
On-chip neural dynamics measurements

Experimental results

![Graph showing neuron activity and synaptic events](image)

- $dpi_{\tau} = 23.4 \text{pA}$
- $dpi_{\tau} = 16.4 \text{pA}$
- $dpi_{\tau} = 11.7 \text{pA}$
On-chip neural dynamics measurements

Experimental results
On-chip spike-based learning examples

CIFAR10

Original image  Weight Matrix (10ms)  Weight Matrix (20ms)

Weight evolution

MNIST

Weight Matrices
## Comparison to the state-of-the-art

<table>
<thead>
<tr>
<th></th>
<th>[2]</th>
<th>[3]</th>
<th>[4]</th>
<th>[1]</th>
<th>[5]</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Implementation</strong></td>
<td>Mixed-signal</td>
<td>Mixed-signal</td>
<td>Digital</td>
<td>Digital</td>
<td>Digital</td>
<td>Mixed-signal</td>
</tr>
<tr>
<td><strong>Technology</strong></td>
<td>180 nm</td>
<td>180 nm</td>
<td>28 nm</td>
<td>28 nm</td>
<td>14 nm</td>
<td>28 nm</td>
</tr>
<tr>
<td><strong>Supply voltage</strong></td>
<td>1.8V</td>
<td>1.8V</td>
<td>0.55V-1V</td>
<td>0.7V-1.05V</td>
<td>0.5V-1.25V</td>
<td>0.73V-1V</td>
</tr>
<tr>
<td><strong>Neuron type</strong></td>
<td>Analog</td>
<td>Analog</td>
<td>Digital</td>
<td>Digital</td>
<td>Digital</td>
<td>Analog</td>
</tr>
<tr>
<td><strong>Core area [mm²]</strong></td>
<td>51.4</td>
<td>7.5</td>
<td>0.086</td>
<td>0.095</td>
<td>0.4</td>
<td>0.36 (Core&lt;x&gt;) 1.01 (Core&lt;L&gt;)</td>
</tr>
<tr>
<td><strong>Neurons per core</strong></td>
<td>256</td>
<td>256</td>
<td>256</td>
<td>256</td>
<td>max 1k</td>
<td>256 (Core&lt;x&gt;) 64 (Core&lt;L&gt;)</td>
</tr>
<tr>
<td><strong>Synapses per core</strong></td>
<td>128k</td>
<td>16k</td>
<td>64k</td>
<td>64k</td>
<td>1M-114k</td>
<td>16k (Core&lt;x&gt;) 20k (Core&lt;L&gt;)</td>
</tr>
<tr>
<td><strong>Fan-in/Fan-out</strong></td>
<td>256/256</td>
<td>64/4k</td>
<td>256/256</td>
<td>256/256</td>
<td>16/4k</td>
<td>2¹¹/₈k (Core&lt;x&gt;) 1k/8k (Core&lt;L&gt;)</td>
</tr>
<tr>
<td><strong>Reconfigurable dendritic tree</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Synaptic weight</strong></td>
<td>Capacitor</td>
<td>(1+1)-bit</td>
<td>(3+1)-bit</td>
<td>1-bit</td>
<td>1- to 9-bit</td>
<td>(4+1)-bit</td>
</tr>
<tr>
<td><strong>On-line learning</strong></td>
<td>STDP</td>
<td>No</td>
<td>STDP</td>
<td>No</td>
<td>Programmable</td>
<td>STDP</td>
</tr>
<tr>
<td><strong>Operation mode</strong></td>
<td>Parallel processing</td>
<td>Parallel processing</td>
<td>Time multiplexing</td>
<td>Time multiplexing</td>
<td>Time multiplexing</td>
<td>Parallel processing</td>
</tr>
<tr>
<td><strong>Energy per SOP</strong></td>
<td><a href="mailto:77fJ@1.8V">77fJ@1.8V</a></td>
<td><a href="mailto:17pJ@1.3V">17pJ@1.3V</a>@1.8V</td>
<td><a href="mailto:9.8pJ@0.55V">9.8pJ@0.55V</a></td>
<td><a href="mailto:26pJ@0.775V">26pJ@0.775V</a></td>
<td><a href="mailto:23.6pJ@0.75V">23.6pJ@0.75V</a></td>
<td><a href="mailto:2pJ@0.73V">2pJ@0.73V</a></td>
</tr>
</tbody>
</table>

# Neuromorphic processors vs. standard processors

<table>
<thead>
<tr>
<th>What are they good for?</th>
<th>What are they bad at?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real-time processing of low-dimensional data</td>
<td>High accuracy pattern recognition</td>
</tr>
<tr>
<td>Ultra-low-power classification of sensory signals</td>
<td>High precision number crunching</td>
</tr>
<tr>
<td>Low-latency decision making</td>
<td>Batch processing of data sets</td>
</tr>
</tbody>
</table>

[Image of neuromorphic processor]

[Image of standard processor with labeled components]
Killer Apps

Technology transfer and applications

We are now entering the era of *neuromorphic intelligence* in which dedicated cognitive “chiplets” will be used to provide intelligence to a multitude of edge-computing devices.

- Health monitoring
- Prosthetic controllers
- Human body area networks
- Intelligent “watchdogs”
- Auditory scene analysis
- Environmental sensing

[https://techoverlook.com/]
Conclusions

Basic research

Objectives and preliminary results

- We aim to understand the **principles of computation** of cortical circuits for building neuromorphic agents that can interact intelligently with the environment.

- We developed **neuromorphic electronic circuits** that support neural computational primitive with synaptic plasticity and adaptation mechanisms.

- We can build (and program) **scalable neural processing systems** that can be interfaced to sensors and robotic platforms and (learn to) interact with the environment in real time.
Thanks to the funding bodies

erc
ICT H2020
EU Flag
FNS SNSF
Swiss National Science Foundation
aiCTX
Institute of Neuroinformatics
The end

Thank you for your attention
Backup slides
Orchestrating adaptation at multiple time-scales

Homeostatic plasticity

Learning performance improves when combining learning and adaptation mechanisms at multiple time scales.

Homeostatic plasticity in neuromorphic hardware

Spiking inputs

Synapse

\[ g \Delta w_i \]

\[ l_w^1 \]

\[ l_w^2 \]

\[ l_w^n \]

Injection current

Soma

Integrator

\[ \int \]

DPI

I&F neuron

STDP

HOMEEO

\[ l_{syn} \]

\[ I_g \sum_{i=1}^{n} l_w^i \]
Homeostatic plasticity in neuromorphic hardware

Process Technology: AMS 0.18 µm 1P6M CMOS
Silicon Area of DPI: 84 µm × 22 µm
Size of LLC (W/L): 0.5 µm / 1 µm
Power Consumption: 10.8 nW
Leakage Slope (1pF): 0.45 µV/s
Controllable Leakage Current: 0.45 aA (2.8 Electrons/sec)

Q. Ning, C. Bartolozzi, G. Indiveri, IEEE TBCAS, 2017
Neural processing computational primitives
A basic building block for both the brain and neuromorphic systems

Pyramidal Cell of Layer 3 of Cat Visual Cortex Showing Dendrite (Green) and Axon (Red) Forming Multiple Clusters of Boutons (Black) in Layer 3 and 5.

Canonical Cortical Circuit Based on Electrophysiological and Modeling Studies in the Cat Visual Cortex (from [Douglas and Martin, 1989]).

**Winner-Take-All networks**


**Hence we propose that the ubiquitous microcircuit motif [...] provides an important atomic computational operation to large-scale distributed brain computations.**

Neuromorphic processors for sensory processing

Simple event-based vision processors

DVS

3x3 cxQuad PCB

ROLLS

128x128 Input

32x32 Pooling

4@8x8 Convolution

4@16x16 Pooling

4@8x8

8@32x1 On-line Learning
Real-time low-latency convolutional neural networks
Real-time low-latency convolutional neural networks

Experimental setup
ECG anomaly detection using reservoir computing

[H. Jaeger, 2003] [W. Maass et al., 2002] [F. Bauer and D. Muir, aiCTX]
ECG anomaly detection using reservoir computing

[H. Jaeger, 2003] [W. Maass et al., 2002] [F. Bauer and D. Muir, aiCTX]
ECG anomaly detection using reservoir computing
preliminary results

- Generic, single-led ECG
- Six different anomaly types
- One read-out unit per anomaly

Detection accuracy: 84.4% (per anomalous heartbeat)
False positives: 1.8% (per nominal heartbeat)

[F. Bauer and D. Muir, aiCTX]
ECG anomaly detection using reservoir computing setup

Mean neural event rate: $14.8 \cdot 10^3 / s$
Mean synaptic event rate: $787.6 \cdot 10^3 / s$
Energy per neural event: 100 pJ
Energy per synaptic event: 40 pJ
Mean power consumption: 32.7 $\mu$W
The CapoCaccia Cognitive Neuromorphic Engineering Workshop

http://capocaccia.cc/

- Interdisciplinary, international, inter EU-US project
- Morning lectures, afternoon hands-on work-groups
- Active and lively discussions (no powerpoint)
- Concrete results, establishment of long-term collaborations

Capo Caccia, Sardinia, Italy. April 23 - May 5, 2019
A new start-up company

ultra-Low-Power Neuromorphic Processing

We develop dedicated brain-inspired ultra-low power mixed-signal Neuromorphic Processors with advanced scalable neural routing architectures and on-chip learning neural circuits.

aiCTX AG

www.ai-ctx.com
info@ai-ctx.com