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

institute of
neuroinformatics

University of Zurich^{UZH} | ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

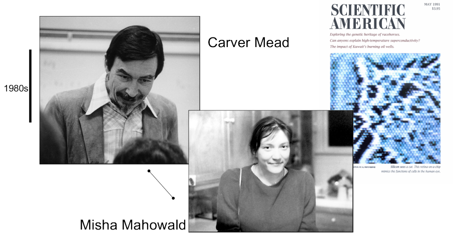
Neuromorphic computing origins

Basic research

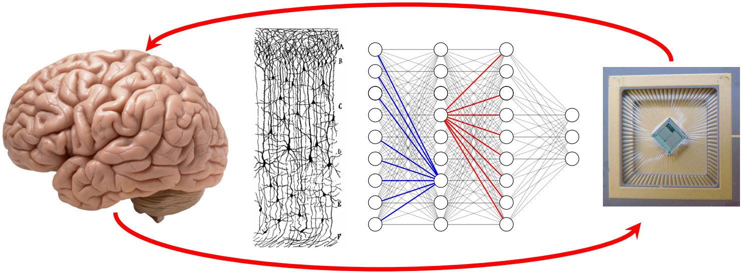
- Fundamental research.
- Emulation of neural function.
- Subthreshold analog
- Asynchronous digital.

Recent developments

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



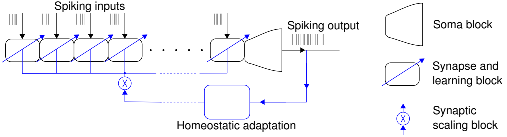
Neuromorphic computing approach at INI

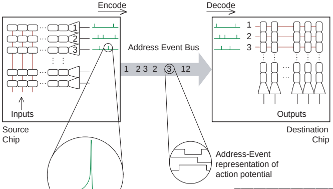


- Highly interdisciplinary basic research rooted on neuroscience, non-linear dynamical systems theory, device physics, microelectronics,. . .
- Exploit the physics of silicon and emerging nano-technologies to reproduce the *bio*-physics of neural systems.
- Develop distributed multi-core spiking architectures using mixed signal analog/digital VLSI circuits.
- Build real–time autonomous cognitive agents able to carry out behavioral tasks in complex environments.

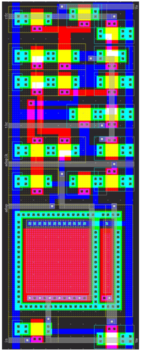
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead



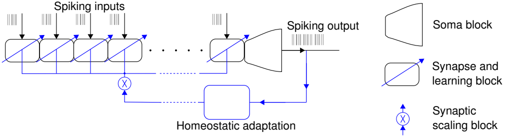


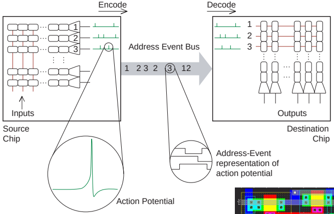
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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

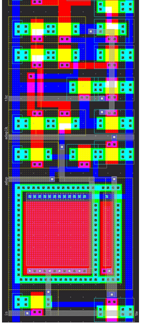


Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead



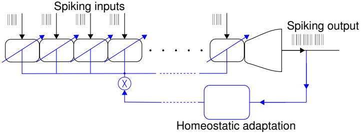


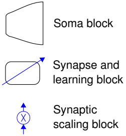


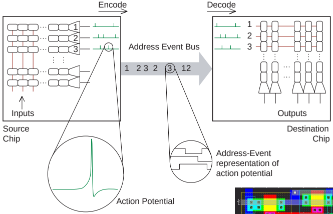
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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

Neuromorphic processor design choices

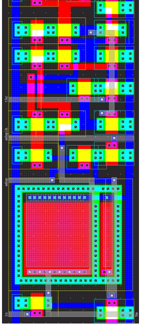
"Listen to the Silicon" - C. Mead





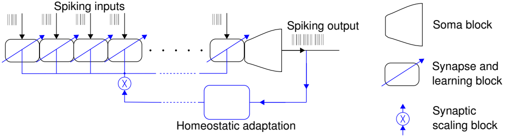


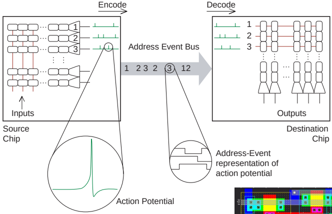
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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



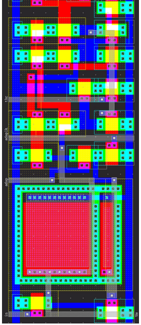
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead



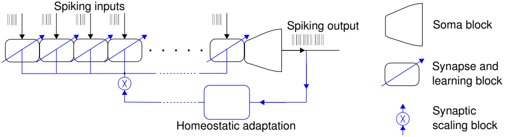


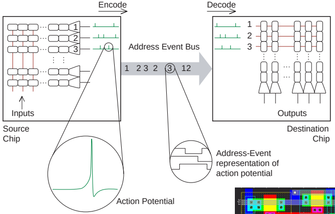
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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



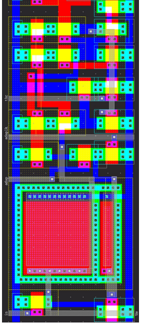
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead



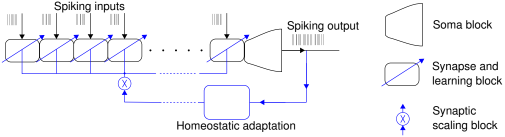


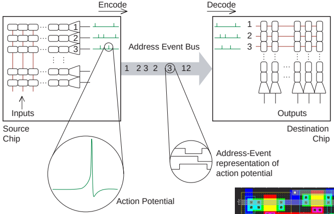
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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



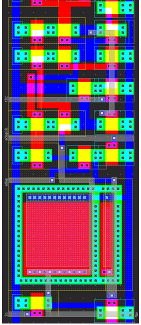
Neuromorphic processor design choices

"Listen to the Silicon" - C. Mead



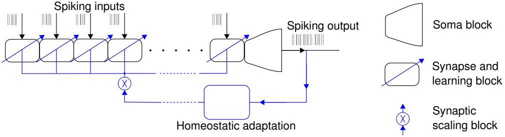


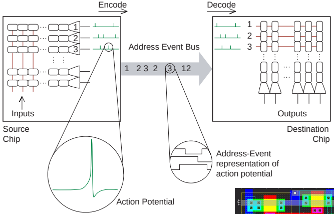
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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



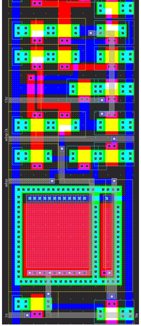
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead



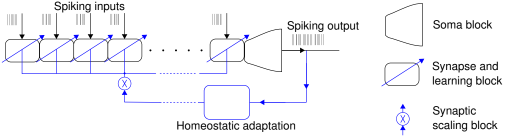


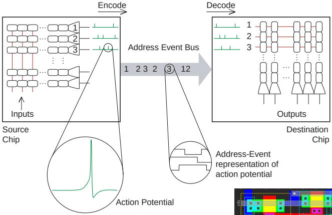
- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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



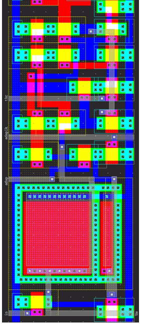
Neuromorphic processor design choices

“Listen to the Silicon” - C. Mead

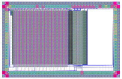
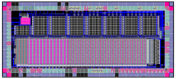
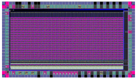
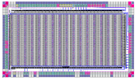
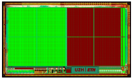


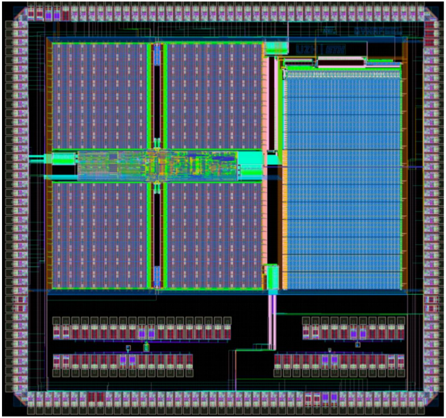


- Spikes in, and spikes out.
- Analog subthreshold & digital asynchronous circuits.
- Massively parallel, distributed computation.
- Time represents itself (no time-multiplexing)
- Biologically plausible temporal dynamics
- Adaptation and learning at multiple time scales.
- No clock and no active circuits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



DYNAP-SEL: Dynamic Neuromorphic Asynch Processor with Self Learning

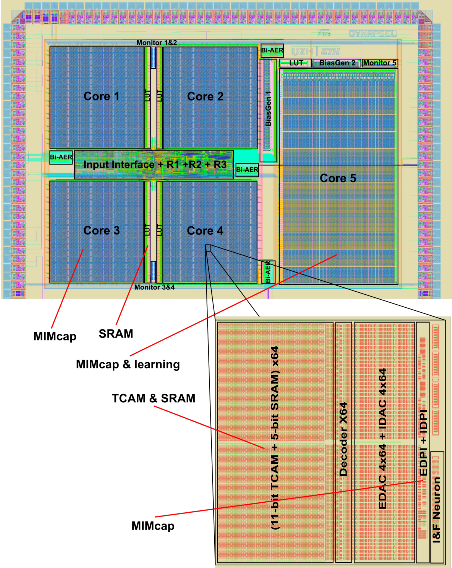








Chip Name	DynapSEL
Process	ST28FDSOI
Supply Voltage	1V
IO Number	176 + (internal 59)
Chip area	2.8mm x 2.6mm
Core Numbers	4 non-plastic cores 1 plastic core
Neuron Type	Analog AExp I&F
Non-plastic Synapse Type	TCAM based 4-bit
Plastic Synapse Type	Linear 4-bit digital
Throughput of Router	1G Events/second
Scalability	16 x16 chips non-plastic core) 4 x4 chips (plastic cores)

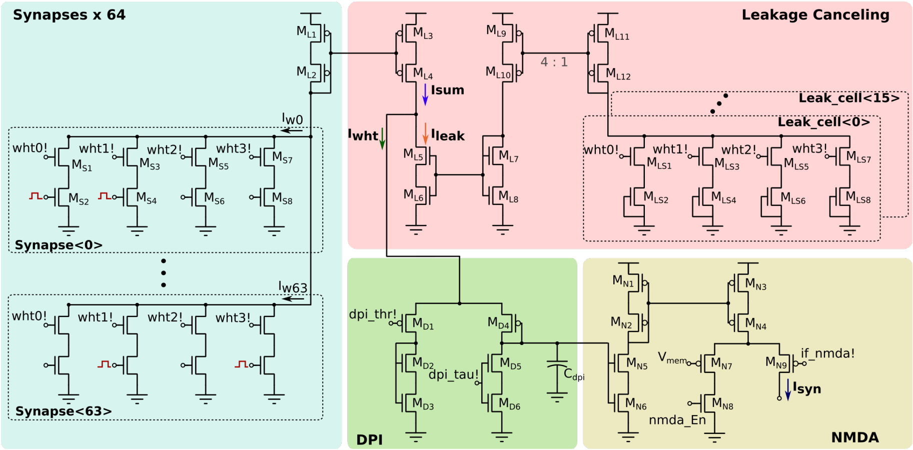
Ready for future emerging nano-technologies

- 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

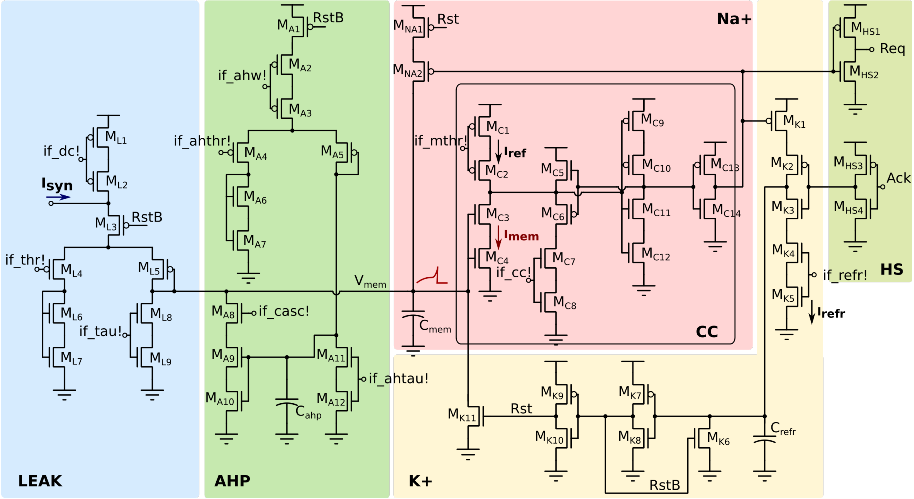


[Qiao and Indiveri, 2016],[EU ICT NeuRAM3 (687299) project]

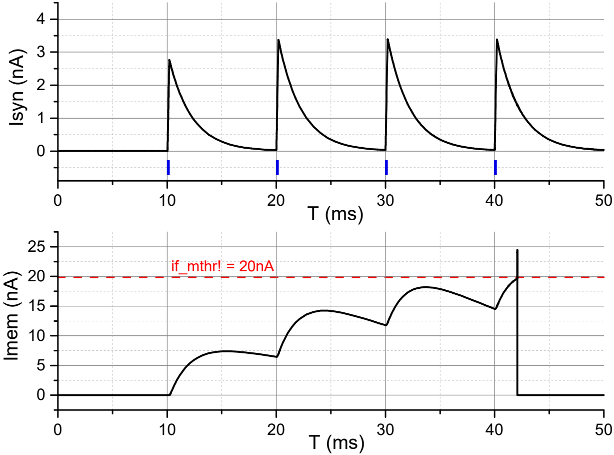
Analog synapse circuits



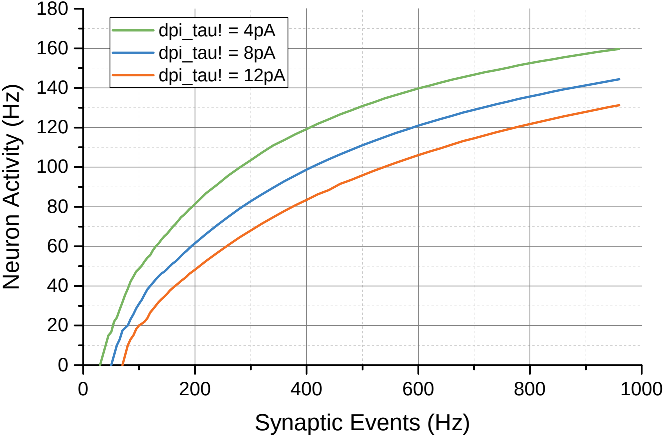
Analog neuron circuits



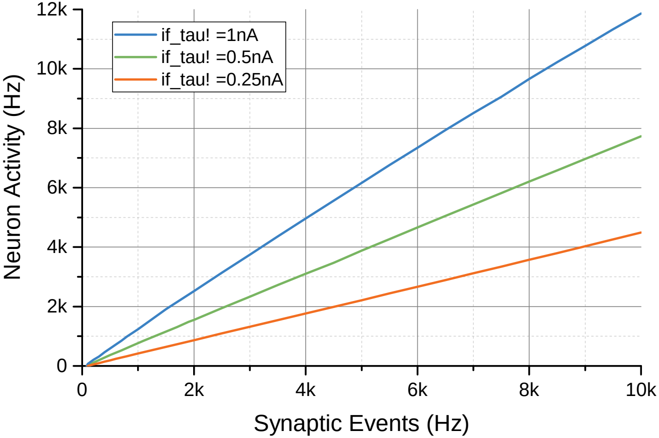
Synapse and neuron circuit response properties



Synapse and neuron circuit response properties

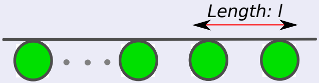


Synapse and neuron circuit response properties



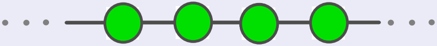
Possible routing schemes

Shared bus



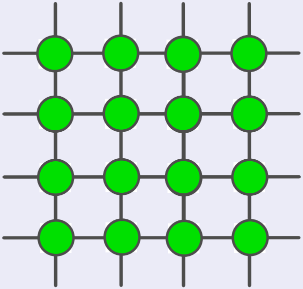
SCX project

1D Grid, Tree



Neurogrid

2D Mesh

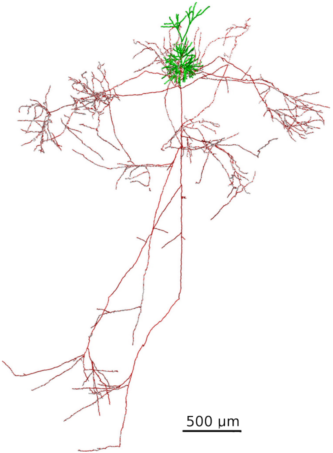


SpiNNaker, Tianji, TrueNorth

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

$$\boxed{F \log_2(N)} \ \textbf{bits/neuron}$$

Cortical networks: a high degree of clustering

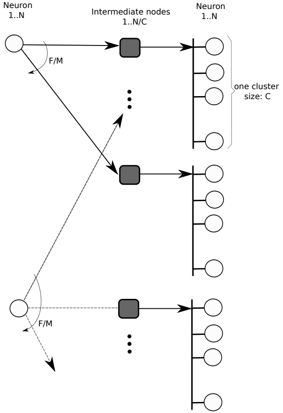


Pyramidal Cell of Layer 3 of Cat Visual Cortex.

Dendrites (Green), Axon (Red), Clusters of Boutons (Black).

Minimize memory requirements:

two-stage routing



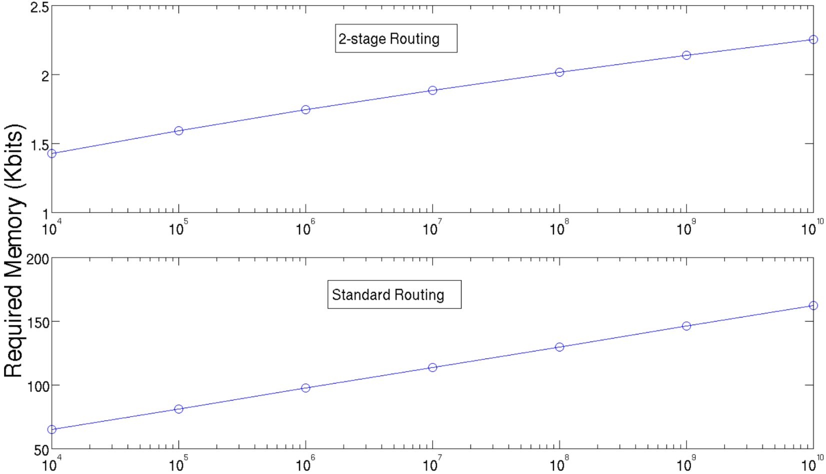
$$\boxed{2\sqrt{F \times log_2(C) \times log_2(N)}}$$ **bits/neuron**

[Douglas and Martin, 2007]

[Moradi and Indiveri 2014]

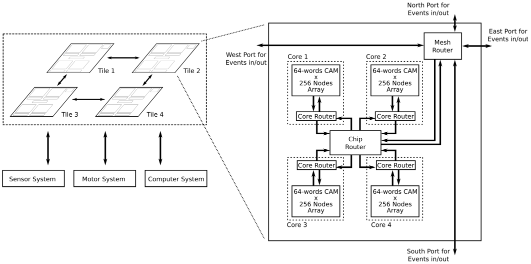
Cortical network example

Routing	bits/neuron
standard	$F\log_2(N)$
two-stage	$\sqrt{F\log_2(N)}\cdot 2\sqrt{\log_2(C)}$

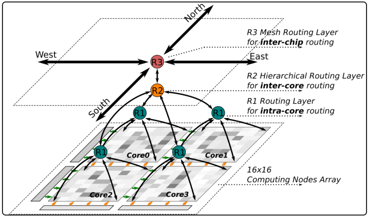


Multi-core neural architecture

with heterogeneous memory structures



Tile 1
Tile 2
Tile 3
Tile 4
Sensor System
Motor System
Computer System
North Port for Events in/out
East Port for Events in/out
West Port for Events in/out
South Port for Events in/out
Mesh Router
Core 1
Core 2
Core 3
Core 4
64-words CAM x 256 Nodes Array
Core Router
Chip Router



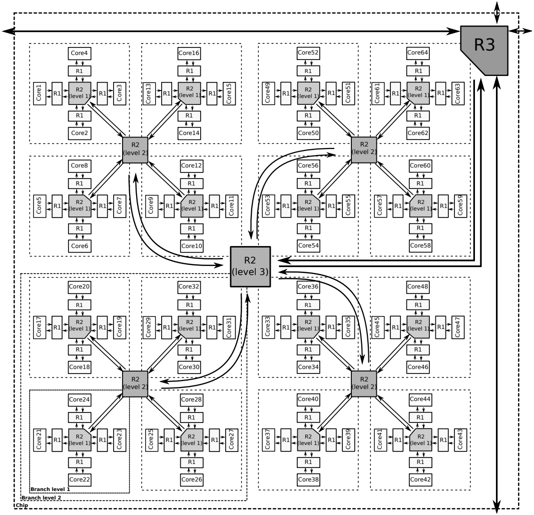
North
West
East
South
R3 Mesh Routing Layer for inter-chip routing
R2 Hierarchical Routing Layer for inter-core routing
R1 Routing Layer for intra-core routing
R3
R2
R1
Core0
Core1
Core2
Core3
16x16 Computing Nodes Array

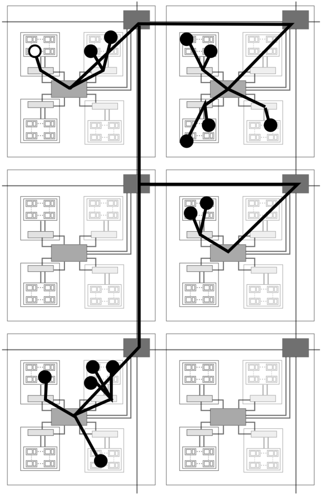
- Combines best of *2D mesh*, *2D tree*, and *multi-cast* schemes, with combination of *source-address* and *destination-address* routing.
- Fully asynchronous hierarchical routers for intra-core (R1), inter-core (R2) and inter-chip (R3) connectivity.
- Distributed asynchronous CAM memory cells within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

[Moradi et al. 2018]

Supports arbitrary large numbers of cores

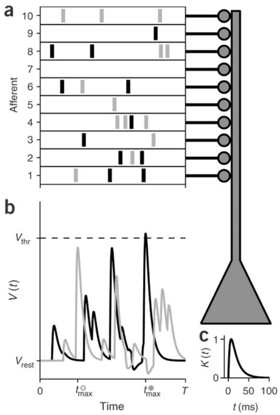
but assumes networks with structured connectivity





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.

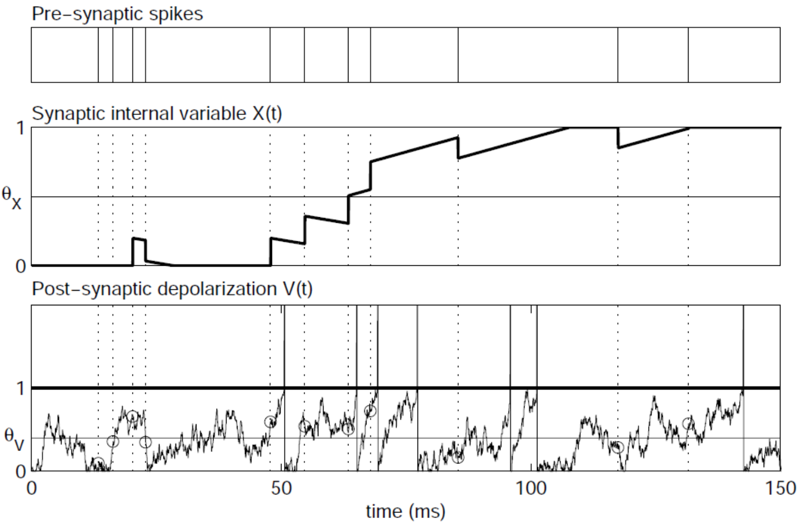
W. Senn, S. Fusi, N. Brunel, S. Sheik, E. Neftci, R. Zecchina, M. Memmesheimer, etc.

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.

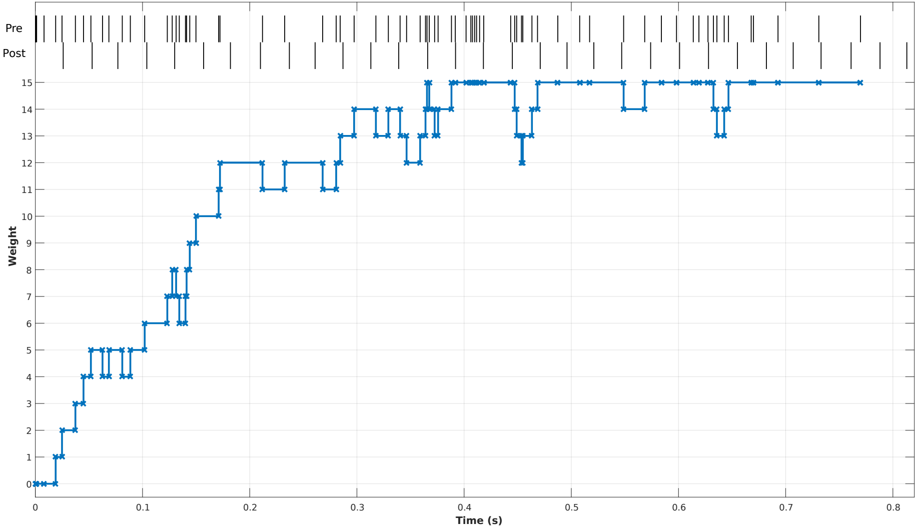
Spike-driven learning rule

in SW simulations

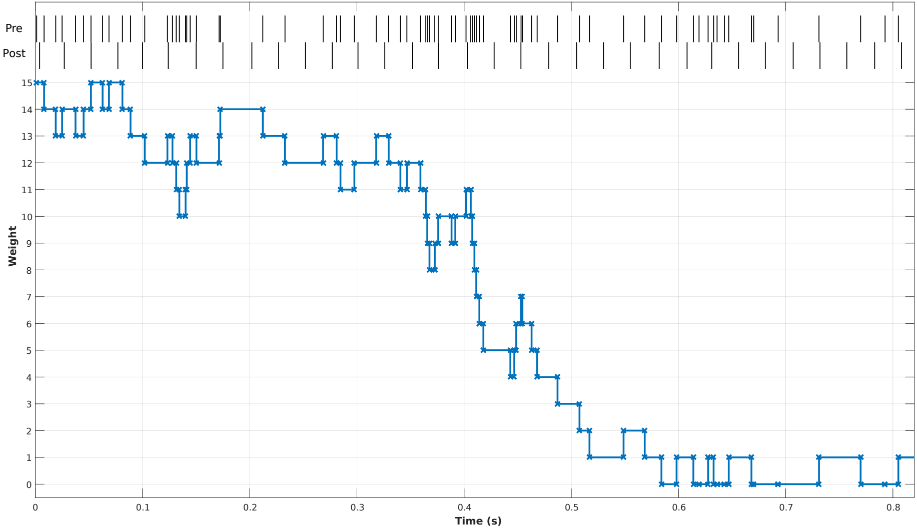


[Brader et al., 2007]

On-chip spike-based weight update measurements

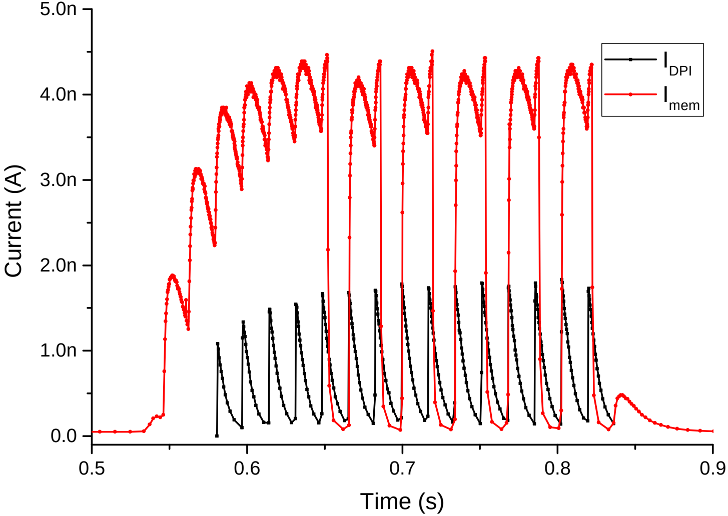


On-chip spike-based weight update measurements



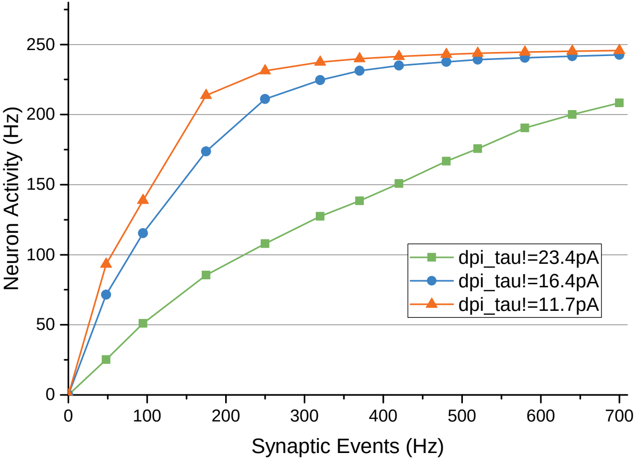
On-chip neural dynamics measurements

Experimental results



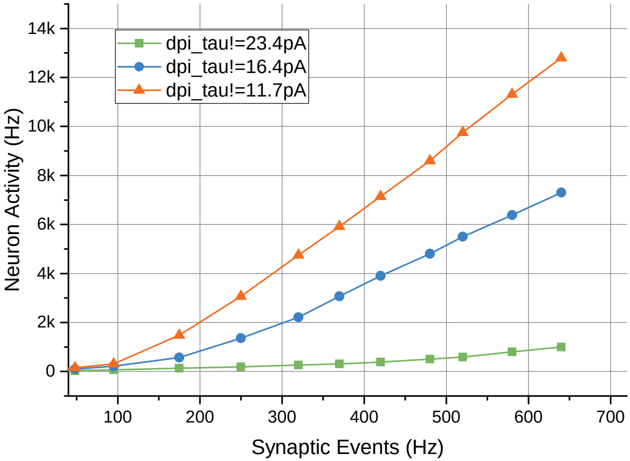
On-chip neural dynamics measurements

Experimental results

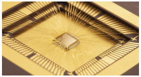


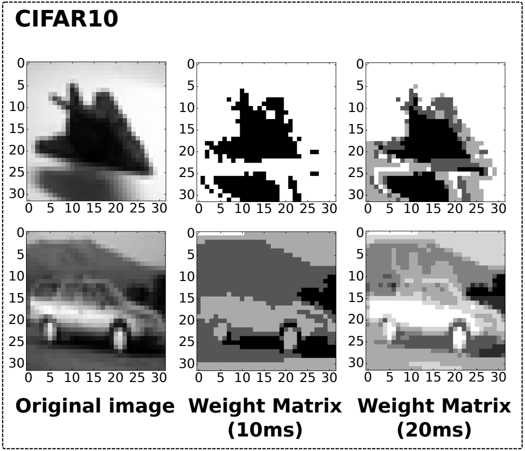
On-chip neural dynamics measurements

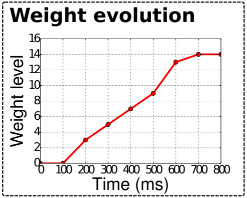
Experimental results

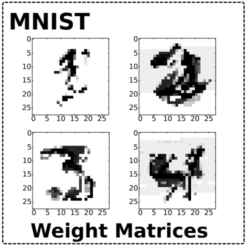


On-chip spike-based learning examples









Comparison to the state-of-the-art

	[2]	[3]	[4]	[1]	[5]	**This work**
Implementation	Mixed-signal	Mixed-signal	Digital	Digital	Digital	Mixed-signal
Technology	180 nm	180 nm	28 nm	28 nm	14 nm	28 nm
Supply voltage	1.8V	1.8V	0.55V-1V	0.7V-1.05V	0.5V-1.25V	0.73V-1V
Neuron type	Analog	Analog	Digital	Digital	Digital	Analog
Core area [mm^2]	51.4	7.5	0.086	0.095	0.4	0.36 (Core<x>) 1.01 (Core<L>)
Neurons per core	256	256	256	256	max 1k	256 (Core<x>) 64 (Core<L>)
Synapses per core	128k	16k	64k	64k	1M-114k	16k (Core<x>) 20k (Core<L>)
Fan-in/Fan-out	256/256	64/4k	256/256	256/256	16/4k	2^{11}/8k (Core<x>) 1k/8k (Core<L>)
Reconfigurable dendritic tree	Yes	No	No	No	No	Yes
Synaptic weight	Capacitor	(1+1)-bit	(3+1)-bit	1-bit	1- to 9-bit	(4+1)-bit
On-line learning	STDP	No	STDP	No	Programmable	STDP
Operation mode	Parallel processing	Parallel processing	Time multiplexing	Time multiplexing	Time multiplexing	Parallel processing
Energy per SOP	77fJ@1.8V	17pJ@1.3V@1.8V	9.8pJ@0.55V	26pJ@0.775V	23.6pJ@0.75V	2pJ@0.73V

[1] P. A. Merolla et al. Science. 2014.

[2] N. Qiao et al. Frontiers in Neuroscience, 2015.

[3] S. Moradi et al. Biomedical Circuits and Systems, IEEE Trans. 2018.

[4] C. Frenkel et al., arXiv. 2018.

[5] M. Davies et al. IEEE Micro, 2018.

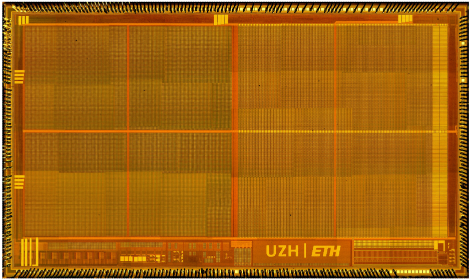
Neuromorphic processors vs. standard processors

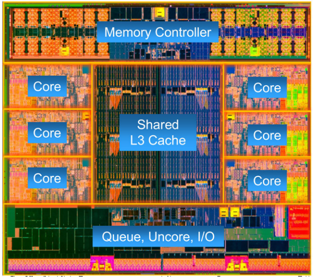
What are they good for?

- Real-time processing of low-dimensional data
- Ultra-low-power classification of sensory signals
- Low-latency decision making

What are they bad at?

- High accuracy pattern recognition
- High precision number crunching
- Batch processing of data sets

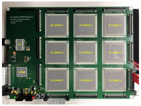


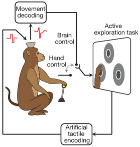


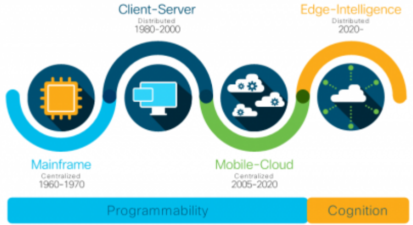
Killer Apps

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







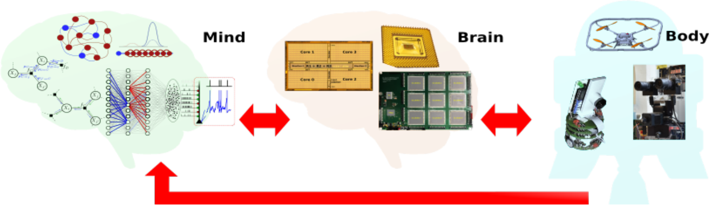
[https://techoverlook.com/]

- Health monitoring
- Prosthetic controllers
- Human body area networks

- Intelligent “watchdogs”
- Auditory scene analysis
- Environmental sensing

Conclusions

Basic research



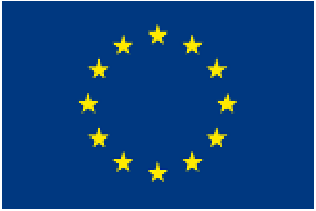
Objectives and preliminary results

- 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



FNSNF
Swiss National Science Foundation

aiCTX

institute of neuroinformatics

The end

Thank you for your attention

Appendix

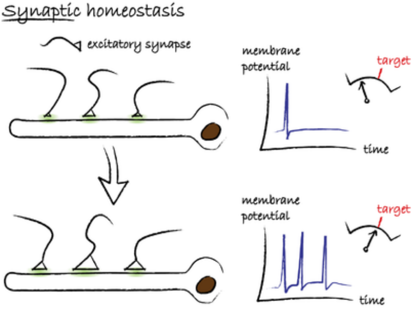
Backup slides

Orchestrating adaptation at multiple time-scales

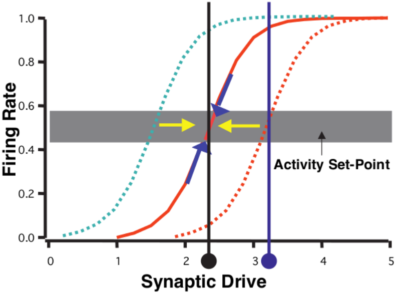
Homeostatic plasticity

Learning performance improves when combining learning and adaptation mechanisms at multiple time-scalse

[J. Lisman, G. Turrigiano, W. Gerstner, F. Zenke, S. Ganguli, S. Fusi, J. Triesh, W. Legenstein, W. Maass, . . .]

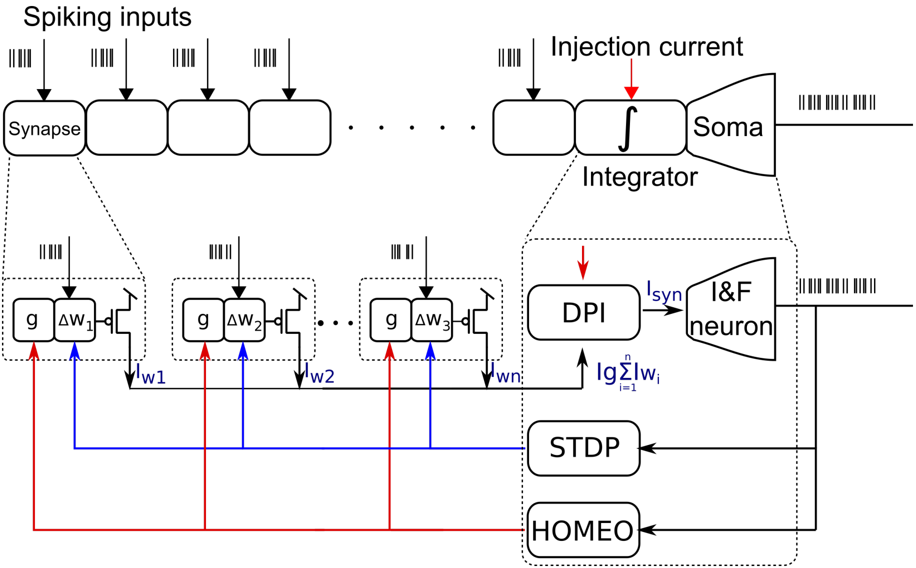


[Scholarpedia]

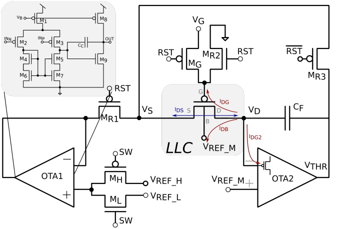


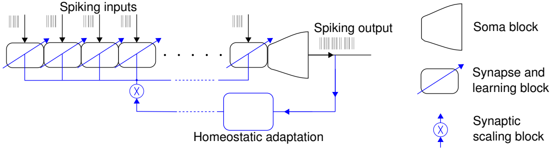
[G. Turrigiano, 2008]

Homeostatic plasticity in neuromorphic hardware

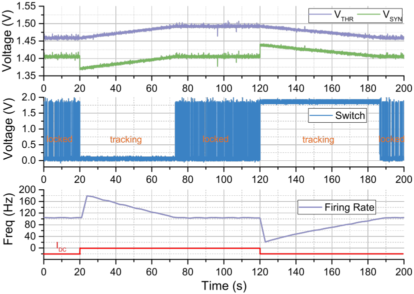


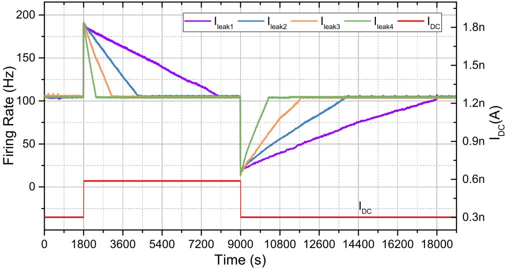
Homeostatic plasticity in neuromorphic hardware





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)

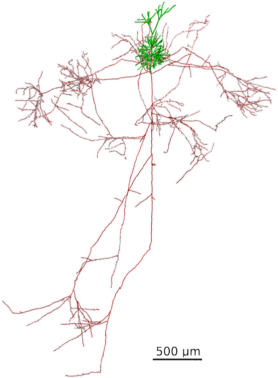




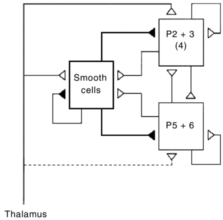
Q. Ning, C. Bartolozzi, G. Indiveri, IEEE TBCAS, 2017

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

[Marcus et al., "The Atoms of Neural Computation", Science 2014]

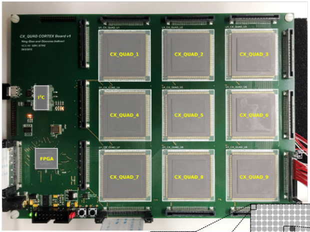
> *Hence we propose that the ubiquitous microcircuit motif [...] provides an important atomic computational operation to large-scale distributed brain computations.*
>
> *[Jonke et al. J. Neurosci. 2017]*

Neuromorphic processors for sensory processing

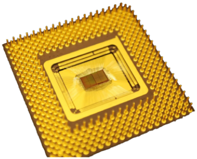
Simple event-based vision processors

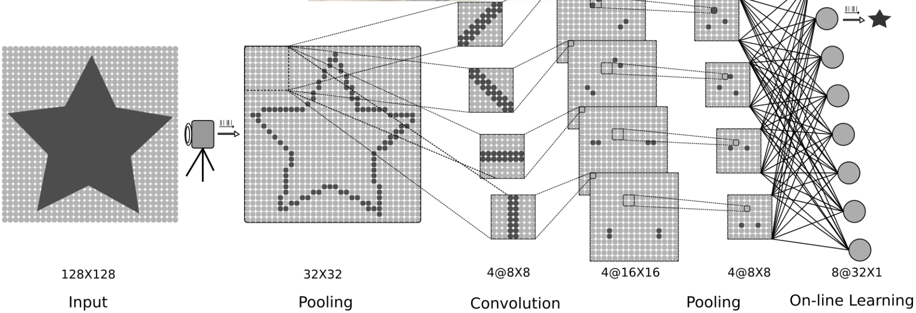
DVS

3x3 cxQuad PCB



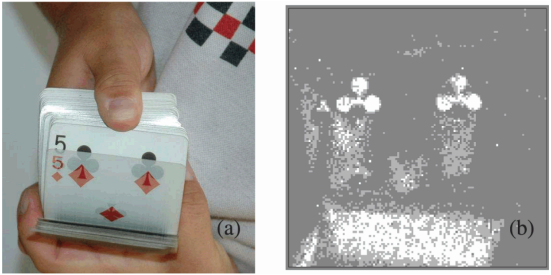
ROLLS

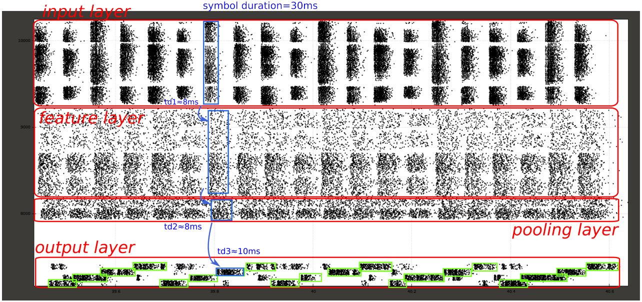




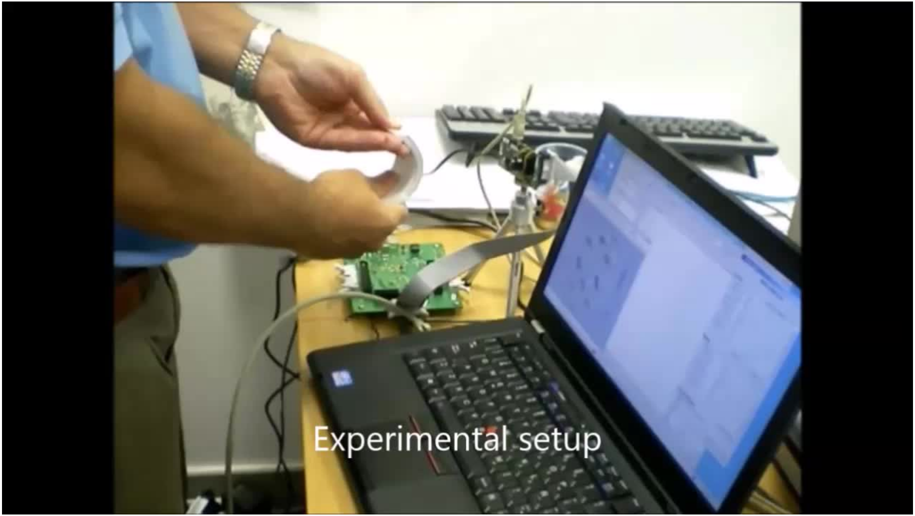
[Indiveri et al., IEDM 2015]

Real-time low-latency convolutional neural networks

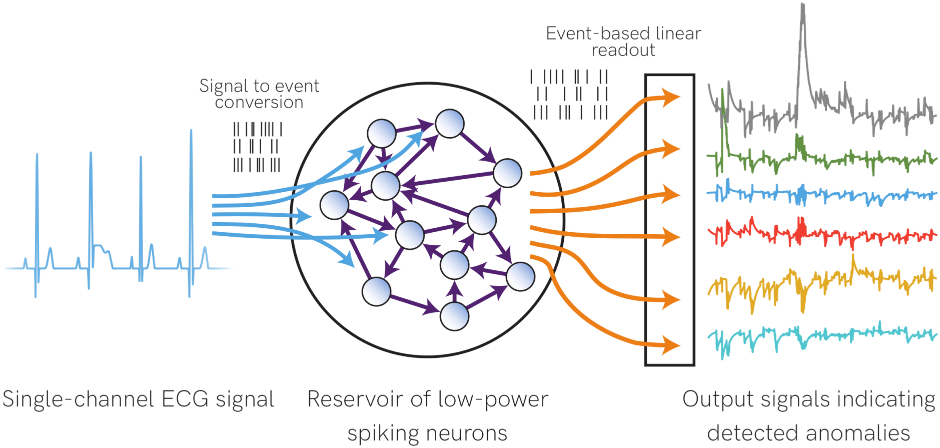




Real-time low-latency convolutional neural networks

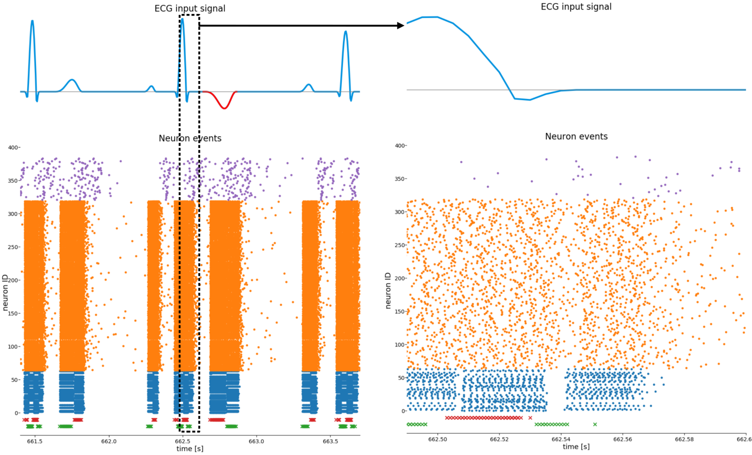


ECG anomaly detection using reservoir computing


Signal to event conversion
Event-based linear readout
Single-channel ECG signal
Reservoir of low-power spiking neurons
Output signals indicating detected anomalies

[H. Jaeger, 2003] [W. Maass et al., 2002] [F. Bauer and D. Muir, aiCTX]

ECG anomaly detection using reservoir computing



━ ECG input

🟩 Input channel UP

🟥 Input channel DN

🟦 Input layer neurons

🟧 Excitatory neurons

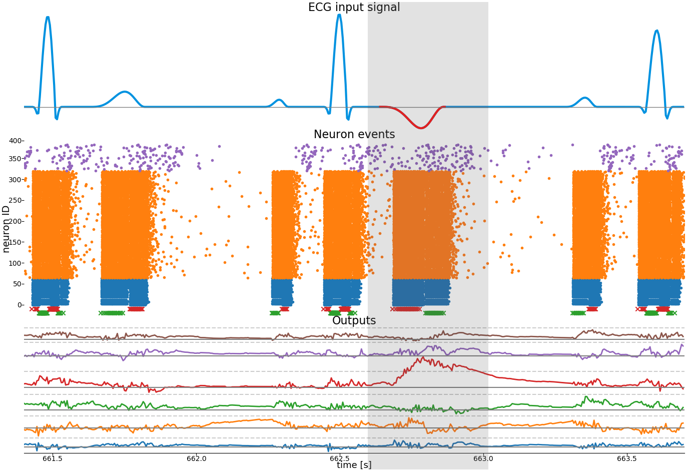
🟪 Inhibitory neurons

[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)

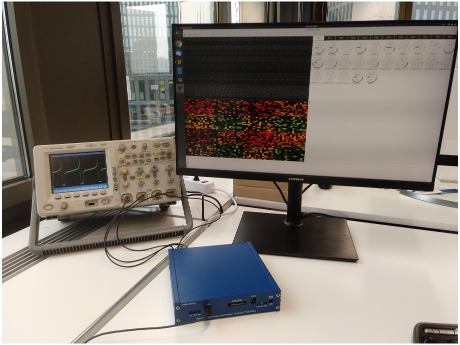
False positives: 1.8% (per nominal heartbeat)

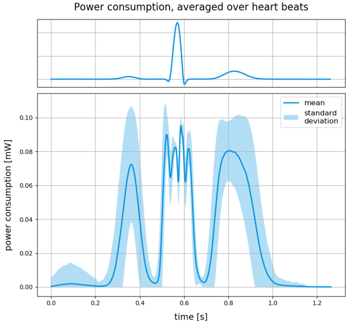
[F. Bauer and D. Muir, aiCTX]

ECG anomaly detection using reservoir computing

setup

Mean neural event rate:	$14.8\cdot10^3$/s
Mean synaptic event rate:	$787.6\cdot10^3$/s
Energy per neural event:	100 pJ
Energy per synaptic event:	40 pJ
Mean power consumption:	32.7 μW





aiCTX

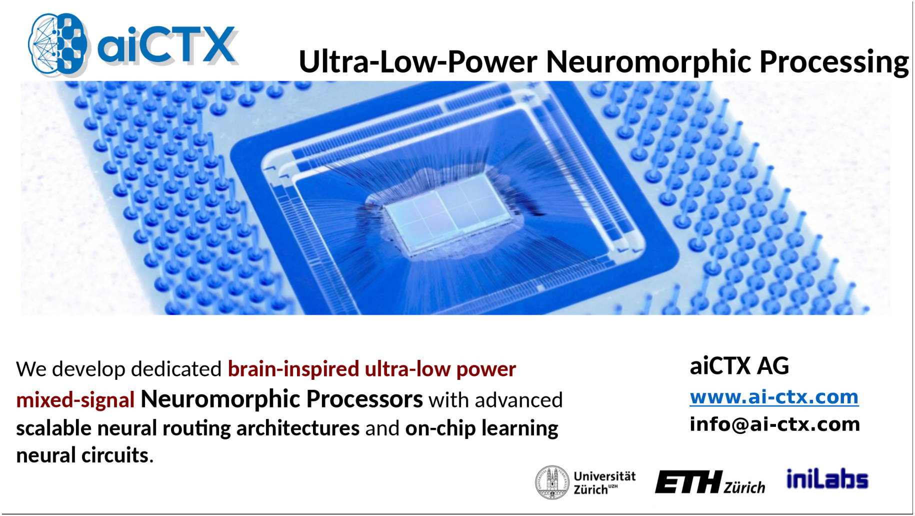
The CapoCaccia Cognitive Neuromorphic Engineering Workshop

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



aiCTX

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

Universität Zürich^{UZH} ETH Zürich iniLabs