Bio-inspired On-line Learning Circuits for Low-power Extreme-edge Spiking Neural Network Processing Systems

Giacomo Indiveri

Institute of Neuroinformatics
University of Zurich and ETH Zurich

19 September 2022
ESSDERC-ESSCIRC Workshop: Embedded Artificial Intelligence (EAI) – Devices, Systems, and Industrial Applications

University of Zurich[UZH] | ETH zürich

Problems and limitations of Artificial Intelligence

Energy Intensive

At this pace, by 2025 the ICT industry will consume 20% of the entire world's electricity

[International Renewable Energy Agency, Internet of Things innovation landscape brief]

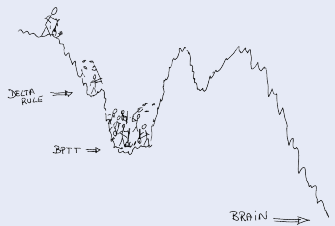
High cost of data movement

DRAM access is at least 1500x more costly than a MAC operation in CNN accelerators. [Tu et al., 2018]

Brittle and narrow AI

DNNs programmed to perform a limited set of tasks.They operate within a pre-determined, pre-defined range. [medium.com]

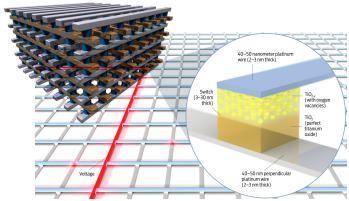
Algorithmic limitations

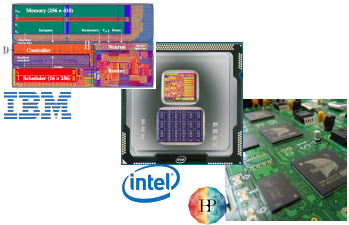


[I am] deeply suspicious of back-propagation.
I don't think it's how the brain works.
The future depends on some graduate student who is deeply suspicious of everything I have said. [Geoff Hinton]

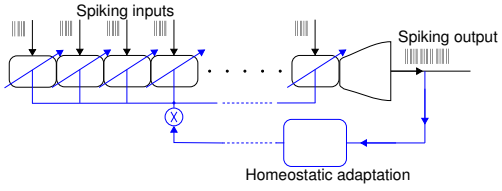
Bottom-up and top-down interdisciplinary approach

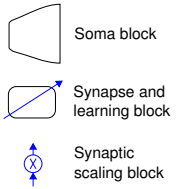
- Deeply rooted in neuro-biology and neuroscience
- Exploits the physics of CMOS and memristive devices to directly emulate neural computation
- Combines analog, asynchronous digital, and logic circuits
- Yields application-specific SNN solutions optimal for edge-computing

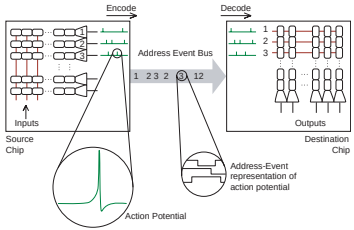




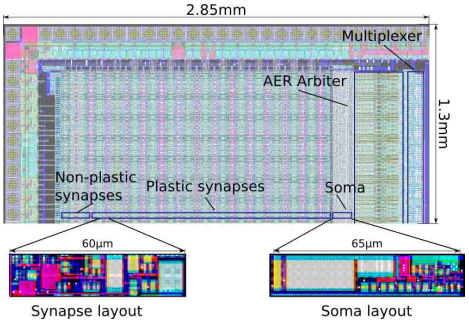
Neuromorphic SNN architectures







- Analog subthreshold circuits (neural computation).
- Spiking neuron and complex synapse circuits.
- Massively parallel arrays (no time multiplexing).
- Slow temporal dynamics (match data rates).
- Inhomogeneous, imprecise, and noisy.
- Fault tolerant and mismatch insensitive by design.
- Asynchronous digital circuits (event-based).
- Re-programmable network topology.



Neuromorphic vs conventional architectures

Pros

- Sub-ms latency
- Sub-mW power consumption

Cons

- Low resolution
- High variability, noisy

What are they good for?

- Real-time sensory-motor processing
- Sensory-fusion and on-line classification
- Low-latency decision making

What are they bad at?

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

How to “program” a neuromorphic processor?

- Define network structure and set its parameters
- Train the network with local learning rules
- Assemble neural computational primitives
- Combine multiple primitives for state dependent processing

Neuromorphic vs conventional architectures

University of Zurich | ETH zürich

Pros

- Sub-ms latency
- Sub-mW power consumption

Cons

- Low resolution
- High variability, noisy

What are they good for?

- Real-time sensory-motor processing
- Sensory-fusion and on-line classification
- Low-latency decision making

What are they bad at?

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

How to “program” a neuromorphic processor?

- Define network structure and set its parameters
- Train the network with local learning rules
- Assemble neural computational primitives
- Combine multiple primitives for state dependent processing

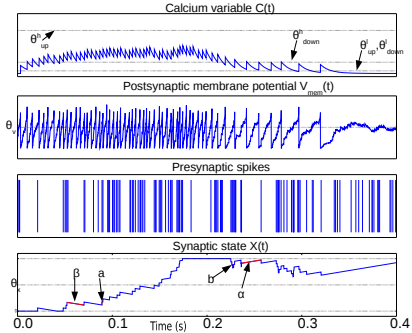
Spike-based learning

There are many spike-driven learning algorithms that are hardware friendly.

e.g., F. Zenke, E. Neftci, S. Bohte, W. Senn, S. Fusi, N. Brunel, R. Zecchina, M. Memmesheimer, etc.

These rule have the following properties in common:

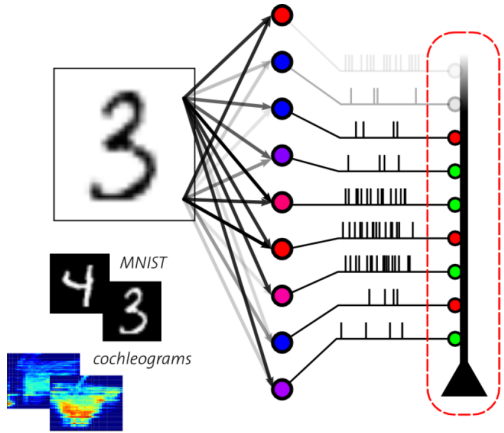
- Reduced resolution synaptic weights
- Redundancy
- Variability

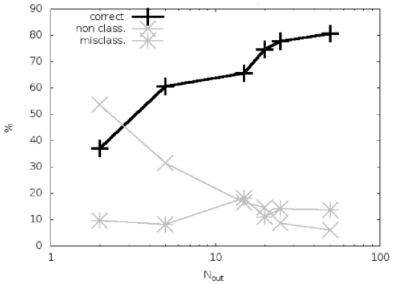


[Mitra et al., IEEE TBCAS 2009]

- Supervised learning, mean rates [Qiao et al., Front. Neurosci., 2015]
- Unsupervised learning, precise spike-timing [Sheik et al. Front. Neurosci., 2012]
- Ensemble learning (random forest, bagging) [Chicca et al. Proc. IEEE 2014]
- Hopfield/attractor networks [Indiveri and Liu, Proc. IEEE 2015]
- Reservoir computing, liquid state and perceptron [Corradi and Indiveri IEEE TBCAS, 2015]

Embracing variability and noise





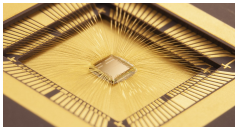
MNIST	deep/CNN (Hinton et al. 2012)	98.4%
	random + bistable synapses	$\sim$ 85%
	random + bistable synapses + (mod. protocol)	$\sim$ 96%
TIMIT	deep/CNN (Hinton et al. 2012)	77%
	VLSI cochlea + bistable synapses	$\sim$ 60%

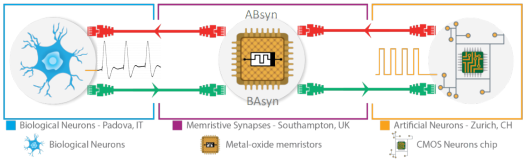
On-line ensemble techniques

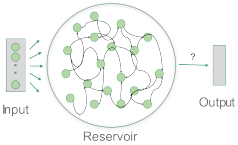
AdaBoost theorem: $1-\text{error}(H_{final}) \geq 1 - e^{-2\gamma^2 N}$

[Y. Freund And R. E. Schapire, 1995]

Edge-computing application examples

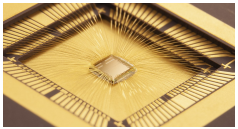


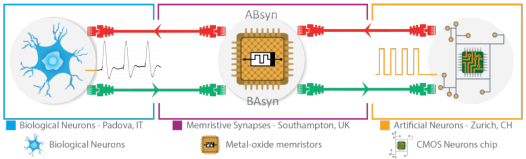


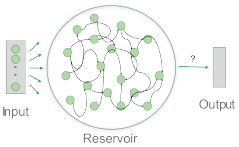


- Constraint Satisfaction Problem Solving [Mostafa et al., 2015, Mostafa et al., 2015b]
- Zebra-finch “Bird’s Own Song” classification [Corradi et al., 2015]
- Closed-loop bidirectional brain machine interfaces [Boi et al., 2016]
- Adaptive pace-maker with neuromorphic CPG network [Abu-Hassan et al., 2019]
- On-line ECG anomaly detection [Bauer et al., 2019]
- On-line classification of EMG signals [Donati et al., 2019]
- Closed-loop coupled biological-silicon neuron network [Serb et al. 2020]
- Closed-loop spiking control on the iCub humanoid robot [Zhao et al., 2020]
- Neuromorphic pattern generation circuits for bioelectronic medicine [Donati et al., 2021]
- Instantaneous stereo depth estimation of real-world stimuli [Risi et al., 2021]
- On-line detection of vibration anomalies using balanced spiking neural networks [Dennler et al., 2021]
- High-Frequency Oscillation (HFO) detection [Sharifhazileh, Burelo et al.,Nat. Comms. 202]

Edge-computing application examples



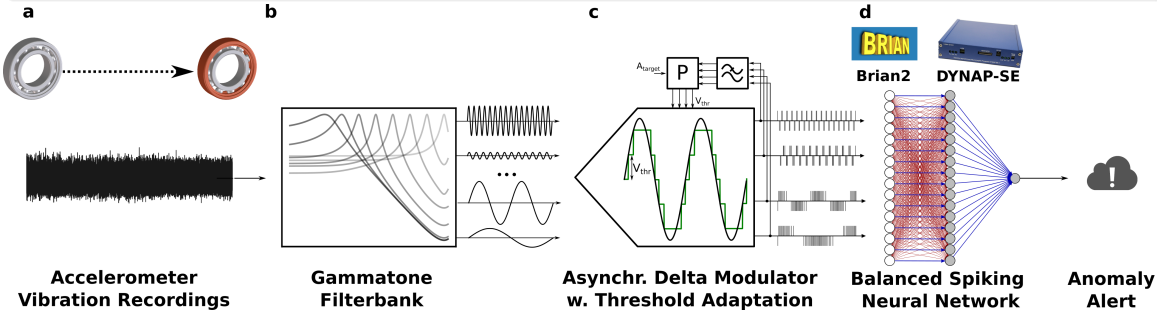




- Constraint Satisfaction Problem Solving [Mostafa et al., 2015, Mostafa et al., 2015b]
- Zebra-finch “Bird’s Own Song” classification [Corradi et al., 2015]
- Closed-loop bidirectional brain machine interfaces [Boi et al., 2016]
- Adaptive pace-maker with neuromorphic CPG network [Abu-Hassan et al., 2019]
- On-line ECG anomaly detection [Bauer et al., 2019]
- On-line classification of EMG signals [Donati et al., 2019]
- Closed-loop coupled biological-silicon neuron network [Serb et al. 2020]
- Closed-loop spiking control on the iCub humanoid robot [Zhao et al., 2020]
- Neuromorphic pattern generation circuits for bioelectronic medicine [Donati et al., 2021]
- Instantaneous stereo depth estimation of real-world stimuli [Risi et al., 2021]
- On-line detection of vibration anomalies using balanced spiking neural networks [Dennler et al., 2021]
- High-Frequency Oscillation (HFO) detection [Sharifhazileh, Burelo et al.,Nat. Comms. 202]

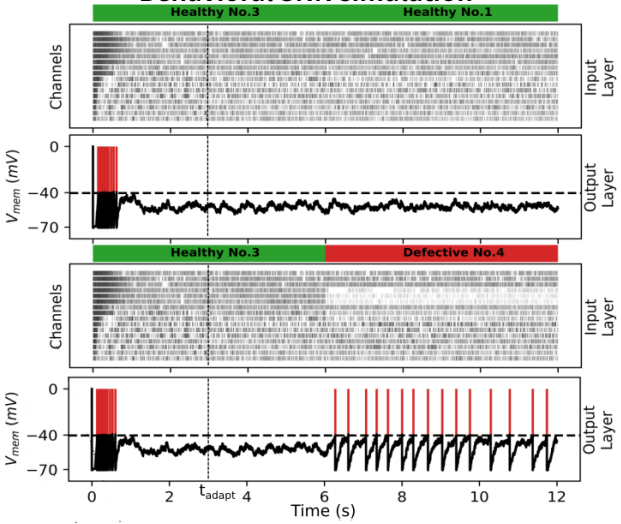
Industrial Predictive Maintenance (PM)

- Predictive Maintenance involves the health monitoring of a degrading system.
- Vibration patterns yield valuable information about the health state of a running machine.
- PM is typically applied to large industrial tasks, but could be useful for small appliances and robots as well.

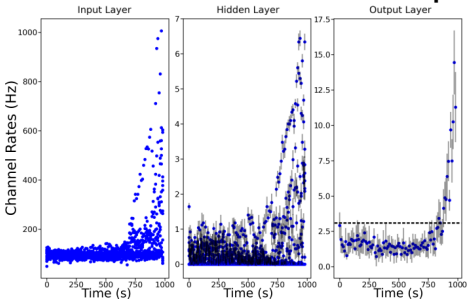


On-line fault prediction

Behavioral SNN simulation



Validation with the DYNAP-SE chip



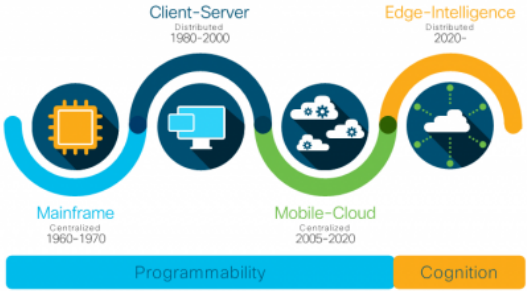
DETECTION TIMES (DATAPOINT) FOR RUN-TO-FAILUREDATASET

	b1	b2	b3	b4
LSSVM	**533**	**823**	893	700
AEC	547	-	-	-
This work	543	890	**873**	**683**

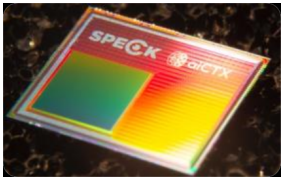
[Dennler et al., 2021]

Extreme edge-computing

We are now entering the era of *neuromorphic intelligence* in which dedicated cognitive "chiplets" will be used to provide intelligence to a multitude of *extreme* edge-computing use cases.



SynSense



- Health monitoring
- Wearable sensors
- Environmental sensing
- Industrial monitoring
- Intelligent machine vision
- MCMC sampling and CSP solving

Funding Sources

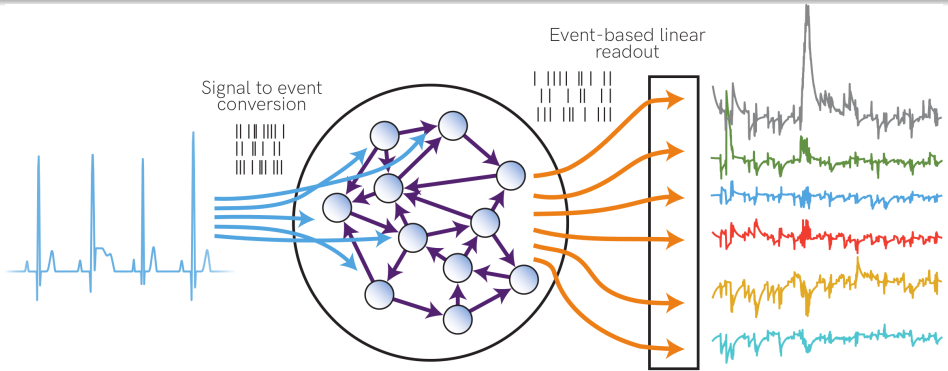
SWISS NATIONAL SCIENCE FOUNDATION

institute of **neuroinformatics**

Thank you for your attention

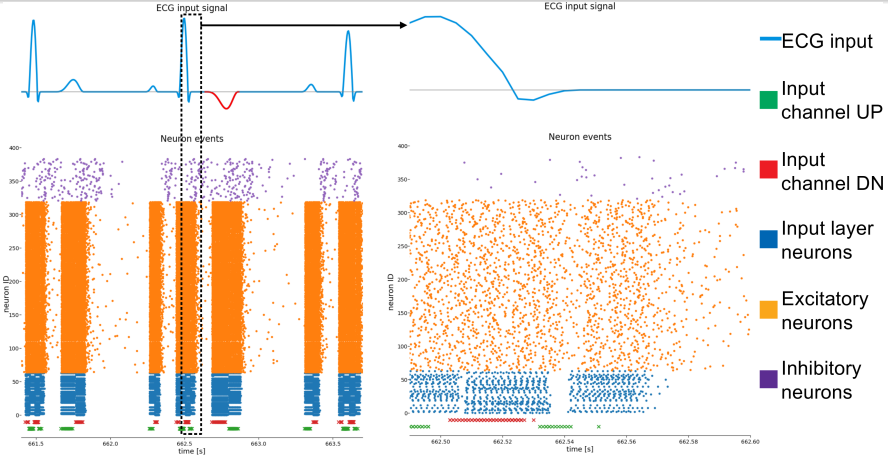
Backup slides

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

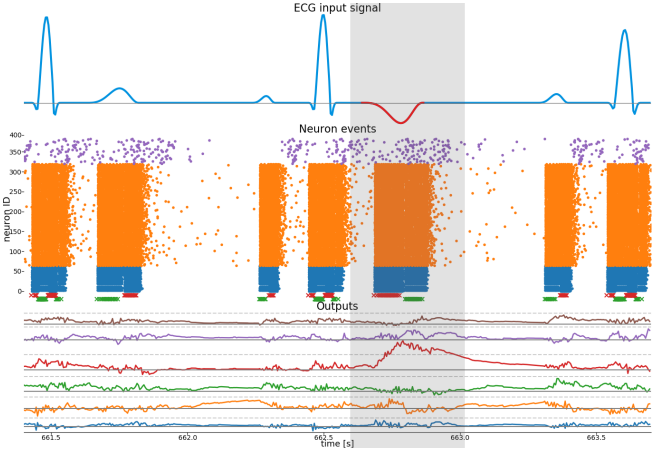


[H. Jaeger, 2003] [W. Maass et al., 2002] [F. Bauer and D. Muir, aiCTX]

ECG anomaly detection using reservoir computing

preliminary results

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



True positives rate (specificity): 91.3%

True negative rate (sensitivity): 97.6%

[F. Bauer et al. 2019]

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:	$< 500\ \mu$W