

Spiking Networks

(Neuromorphic computing)

Robert Forchheimer, prof. em.

Department of Electrical Engineering

Linköping University

Neuromorphic Computing

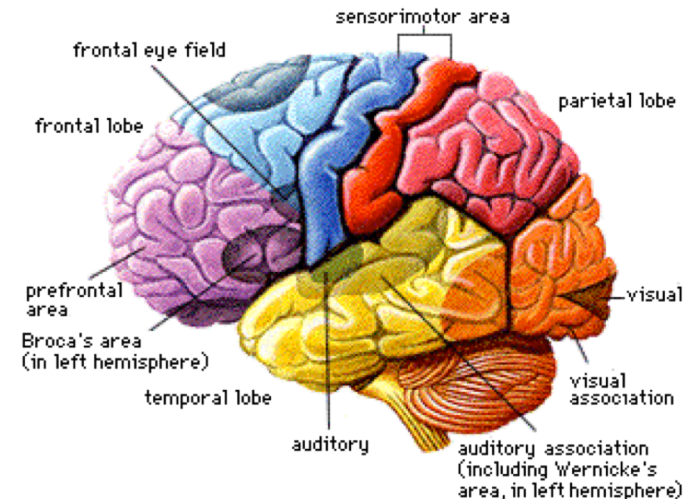
- Term introduced by Prof. Carver Mead, Caltech, 1990
- Devices/algorithms that mimic biological neurons and neural networks (cells, synapses, spike signaling...)
- Used to model processes in the brain
- Used in Machine Learning (“Spiking Networks”)

Content

- The human brain
- Neurons and synapses
- Signaling and signal representations
- Electronic models
- Spiking networks
- Large-scale neuromorphic designs
- Performance comparisons
- Home assignment
- References and publications

The Human Brain

- The brain contains about 86 billion neurons and 100 trillion synapses (approximately 1000 per neuron)
- The neocortex is a thin (2-4 mm) top layer. It contains 30 billion neurons and is responsible for our “intelligence” (cognition, sensory perception, language...)
- The operating frequency is 1 – 10 Hz and the power consumption 10-20 W (< 1 nW/neuron)
- Energy consumption per synaptic event is 1 – 10 fJ.



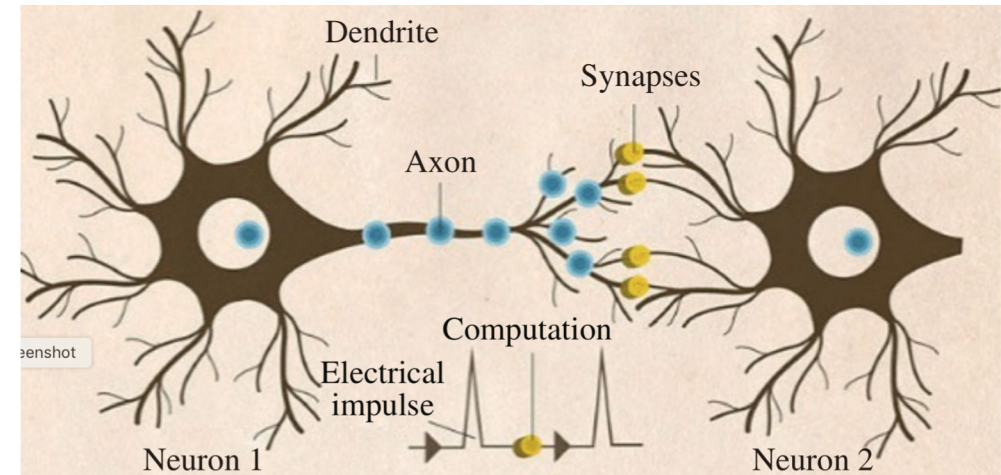
Looking inside – Neurons and connections



<https://www.psypost.org/2017/11/newborn-neurons-find-proper-place-adult-brain-50061>

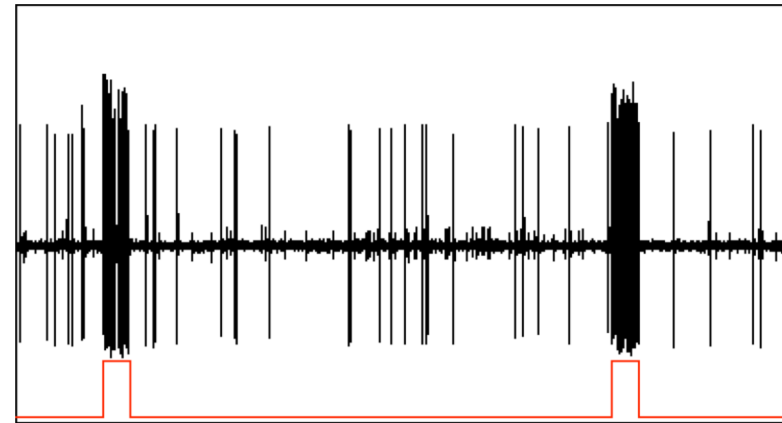
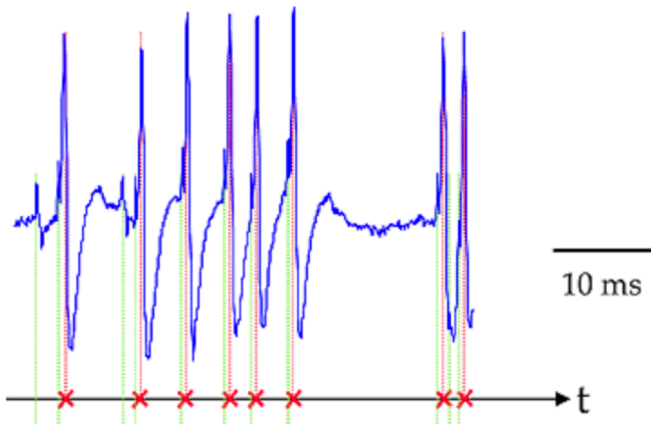
Neuron-to-neuron connection

- Neurons connect to other neurons via *axons* and *dendrites*
- The actual junction is called a *synapse*. It converts the electrical signal from the axon to a change in chemical concentration (of neuro-transmitters) in the receiving neuron.
- A synapse can have an excitatory or inhibitory effect on its neuron.
- When the summed concentration from all synapses exceeds a certain threshold, the receiving (post-synaptic) neuron generates an output pulse on its axon.



Reprinted from Versace et al, A mind made from memristors. IEEE Spectrum 2010

Signaling between neurons

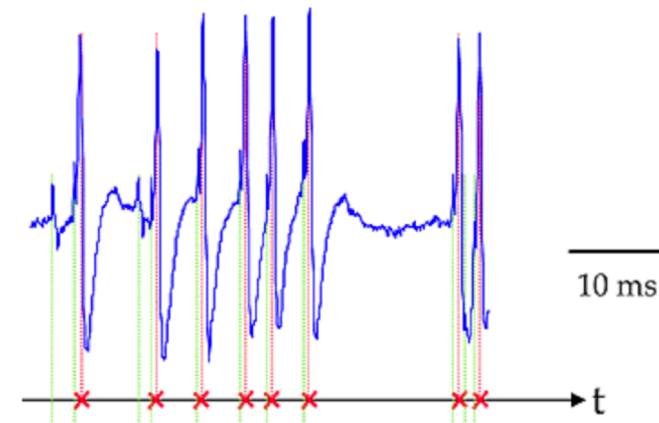


- Signaling between neurons is done by *action potentials* (spikes)
- A spike is about 1 ms long. Repetition rate varies from < 0.01 Hz to > 200 Hz.
- Spike rate is a measure of the “strength” of the neural signal
- Example: neuron in the auditory track reacting to bursts of sound



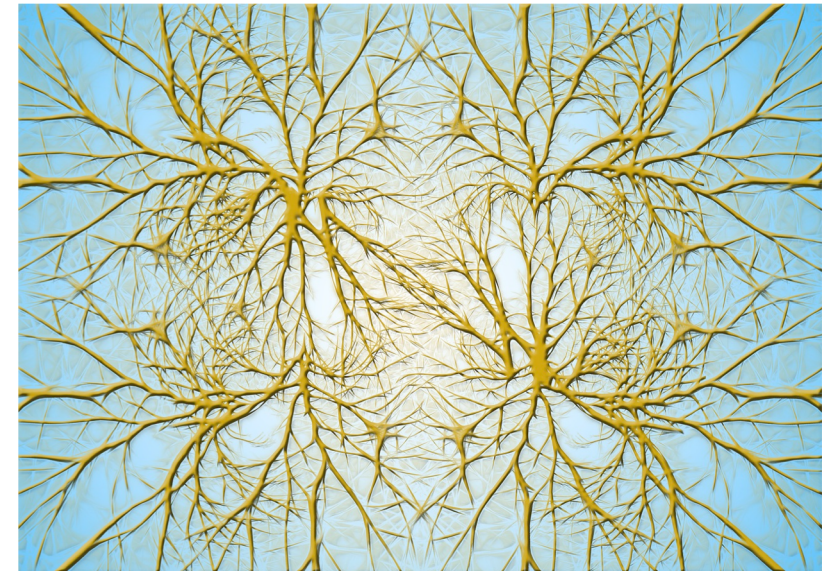
Signal representations

- Spike rate
- Time-to-first-spike
- Intra-spike distance
- Latency (related to other signals)
- Rank-order coding (within a population of neurons)



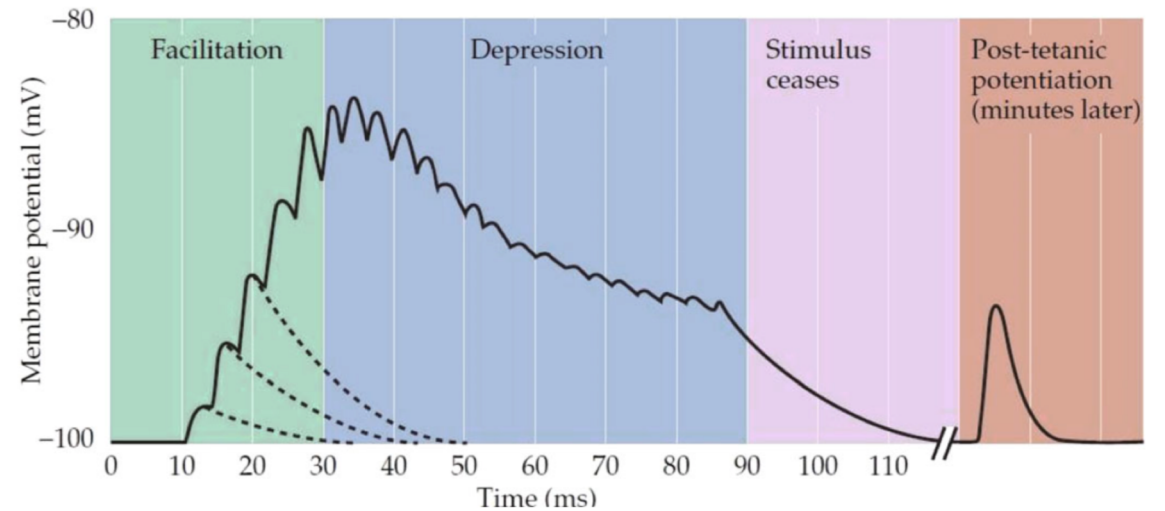
How does a network of neurons learn?

- Synapses change their response (“weights”) when spike rate is high (short term plasticity)
- Synapses also adapt over a longer timeframe (long term plasticity)
- 40% of synapses of a neuron are replaced with new ones every day



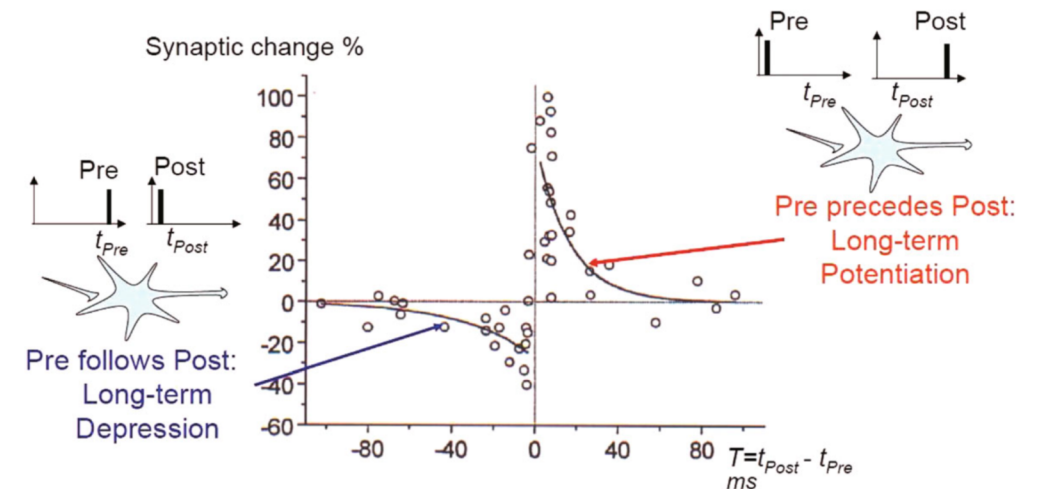
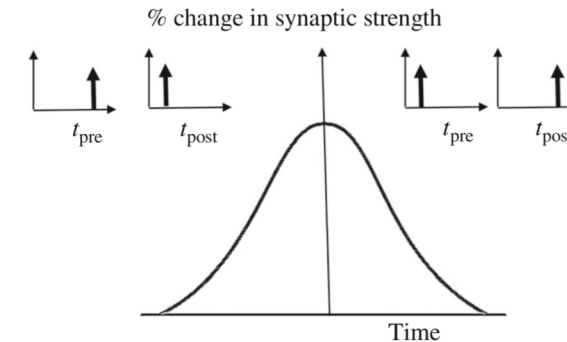
Short-term plasticity

- Short-term plasticity is a temporary increase in synaptic strength
- Appears when two or more action potentials on its input appear close in time (*rate dependent*)



Long-term plasticity

- Long-term plasticity refers to persistence changes in the synaptic strength
- Long-term plasticity is based on the time difference between the pre-synaptic and post-synaptic response
- Long-term plasticity is believed to be the main mechanism behind learning and memory. Two main models: *Hebbian vs STDP learning*.
- Short **absolute** time difference leads to increased strength (*Hebbian learning*)
- Short **positive** time difference leads to increased strength (potentiation), short **negative** time difference leads to decreased strength (depression) (*Spike Timing Dependent Plasticity, STDP learning*)



Reprinted from: Bi G Q, Poo M M. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *J Neurosci*, 1998, 18

Electronic synapse models

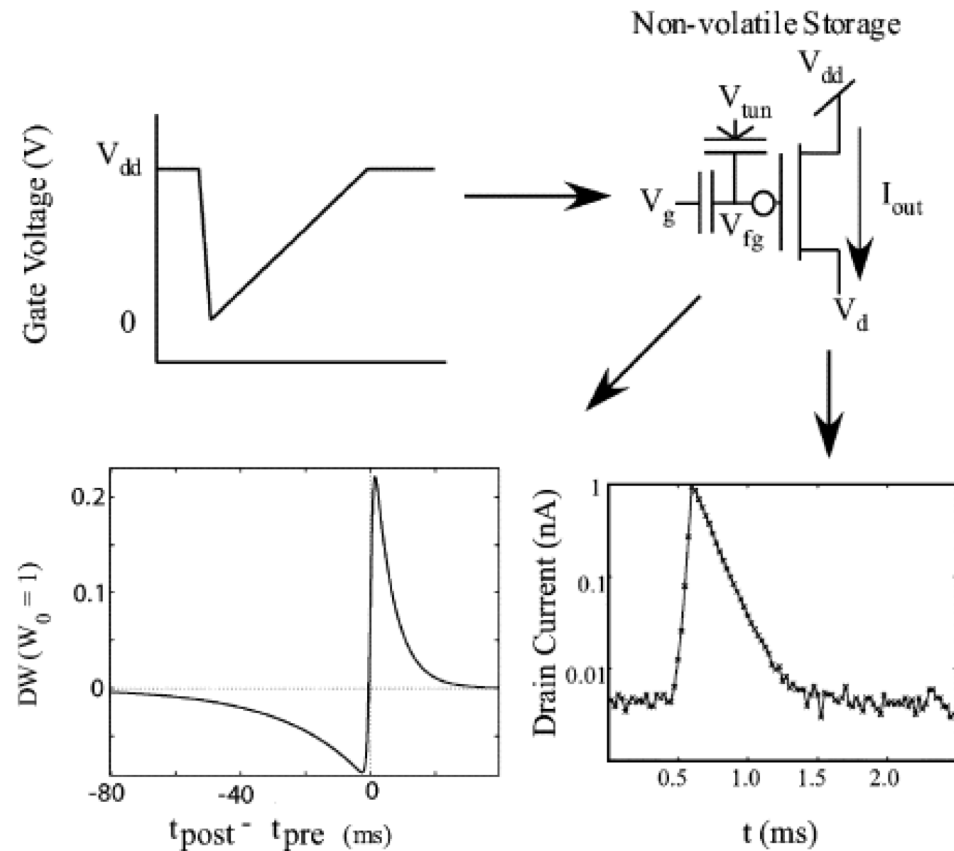
Spike representations

- Voltage/current pulses (analog models)
- Digital values, event times (digital models)
- Stochastic representations

Implementations

- Resistors (MOSFET channel, ionic conductance...)
- Digital implementations
- Hybrid analog/digital

C. Mead (1994) – The floating gate transistor synapse



Based on the same principle as EEPROM. Ions is injected into the gate oxide after which the gate is left floating => long-term plasticity

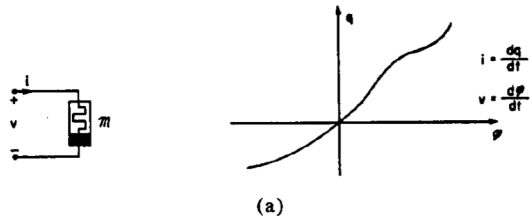
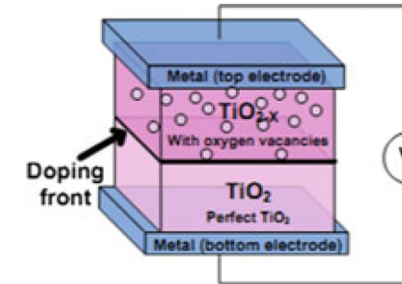
L.O. Chua (1971) – The memristor

IEEE TRANSACTIONS ON CIRCUIT THEORY, VOL. CT-18, NO. 5, SEPTEMBER 1971

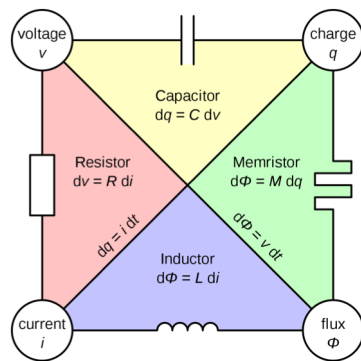
Memristor—The Missing Circuit Element

LEON O. CHUA, SENIOR MEMBER, IEEE

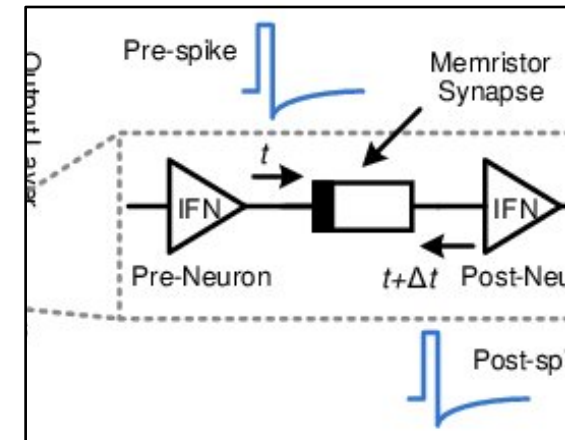
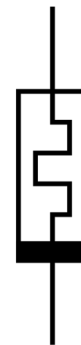
Abstract—A new two-terminal circuit element—called the *memristor*—characterized by a relationship between the charge $q(t) \equiv \int_{-\infty}^t i(\tau) d\tau$ and the flux-linkage $\phi(t) \equiv \int_{-\infty}^t v(\tau) d\tau$ is introduced as the *fourth basic circuit element*. An electromagnetic field interpretation of this relationship in terms of a quasi-static expansion of Maxwell's equations is presented. Many circuit-theoretic properties of memristors are derived. It is shown that this element exhibits some peculiar behavior different from that exhibited by resistors, inductors, or capacitors. These properties lead to a number of unique applications which cannot be realized with RLC networks alone.

HP Labs memristor 2008

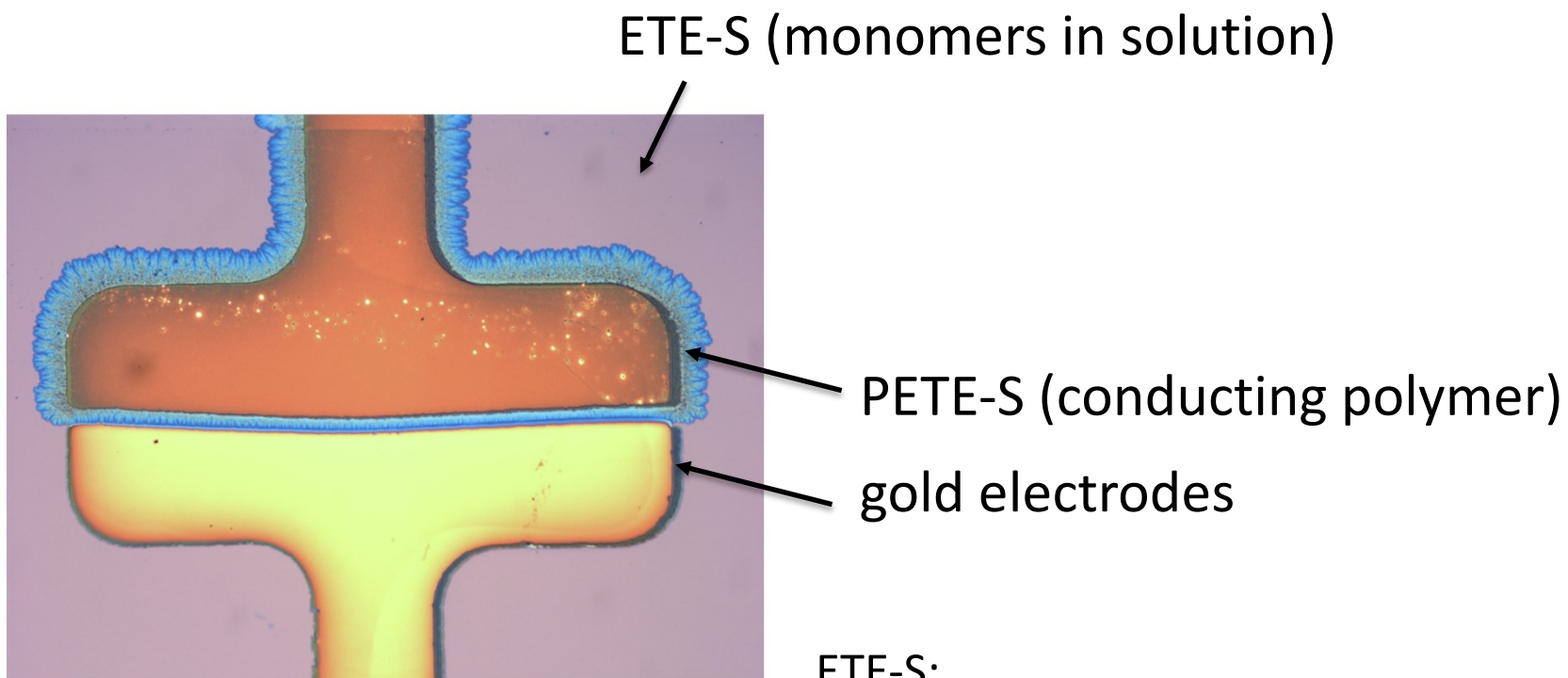


”resistor with memory”



Memristor synapse

J. Gerasimov (2019) – Evolvable organic electrochemical transistor



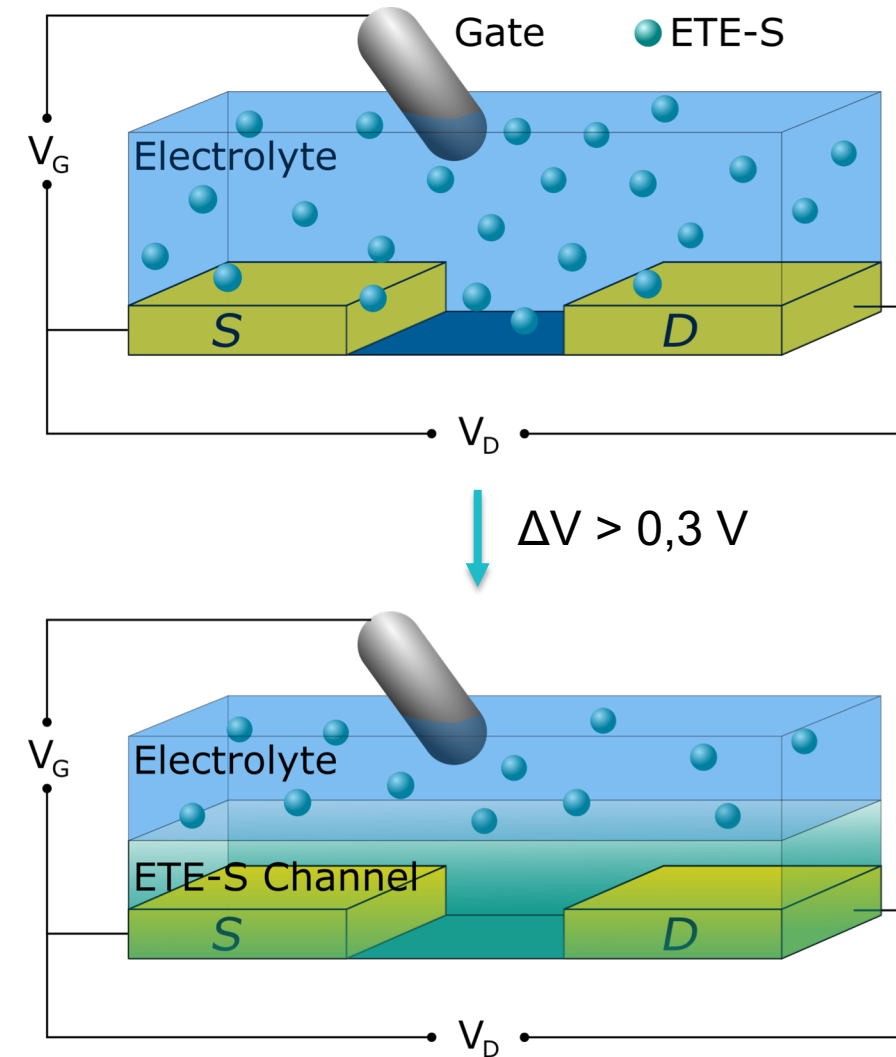
ETE-S:

sodium 4-(2-(2,5-bis(2,3-dihydrothieno[3,4-b][1,4]dioxin-5-yl)thiophen-3-yl)ethoxy)butane-1-sulfonate

The evolvable OECT synapse

- Long-term potentiation—channel growth
- Long-term depression—channel over-oxidation
- Short-term potentiation—channel doping
- Short-term depression—channel dedoping

Reprint from J. Gerasimov et al, An evolvable organic electrochemical transistor for neuromorphic applications. Adv. Sci. 2019.



Electronic models of synapses – other technologies

- RRAM (ReRAM)
- Electrochemical metallization
- Magnetoresistive RAM (MRAM)
- Phase change memory
- Carbon nanotubes
- Josephson junctions
- Digital implementations

Issues to consider

- Non-volatility
- Large dynamic range
- Multi-level
- Sustainability
- Short term/long term plasticity
- Small size
- Low energy consumption

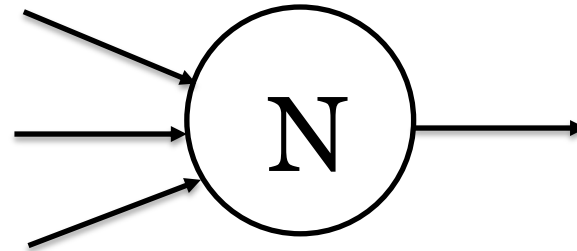
Energy efficiency – per synaptic event

Biological synapses: 20000 ATPs => 0.01 pJ

Analog synapses: 9 pJ (CMOS), 7.5 pJ (OFET)

Digital synapses: 26 pJ (TrueNorth)

Neuron models



Leaky Integrate-and-Fire model (LIF)

According to this model, the neuron has two internal state variables:

$u_i(t)$: Synaptic response current (weighted sum of filtered synaptic inputs)

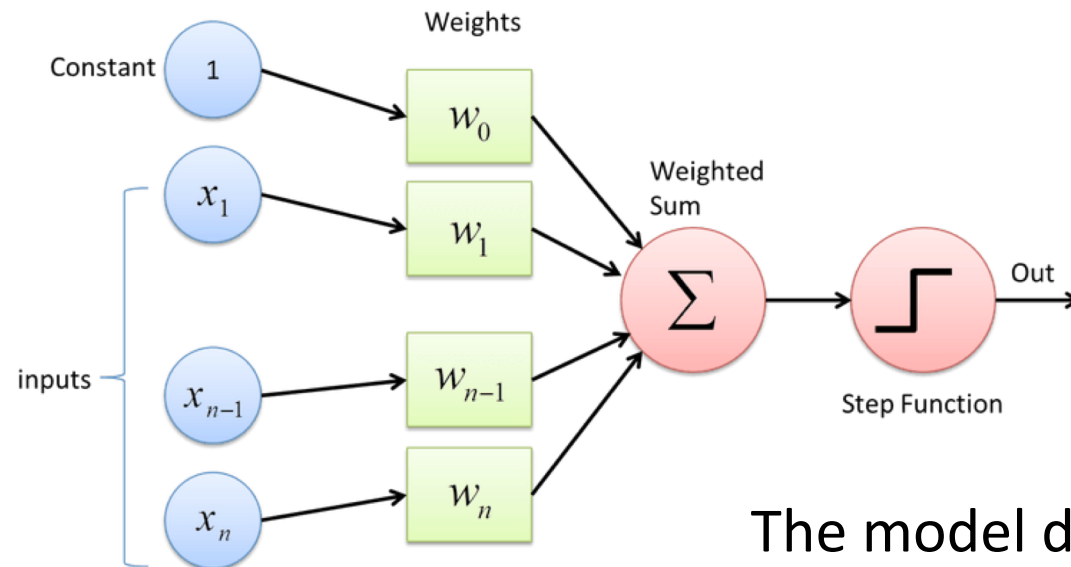
$v_i(t)$: membrane potential (leaky integration of $u_i(t)$)

$$\sigma(t) = \sum_k \delta(t - t_k) \quad u_i(t) = \sum_{j \neq i} w_{i,j} (\alpha_u * \sigma_j)(t) + b_i$$

$\alpha_u = \text{exp. filter}$

$$\dot{v}_i(t) = -\frac{1}{\tau_v} v_i(t) + u_i(t) - \theta_i \sigma_i(t)$$

Simple neuron (+ synapses) model: linearly weighted inputs followed by a nonlinear activation function (W. McCulloch, W. Pitts, 1943)



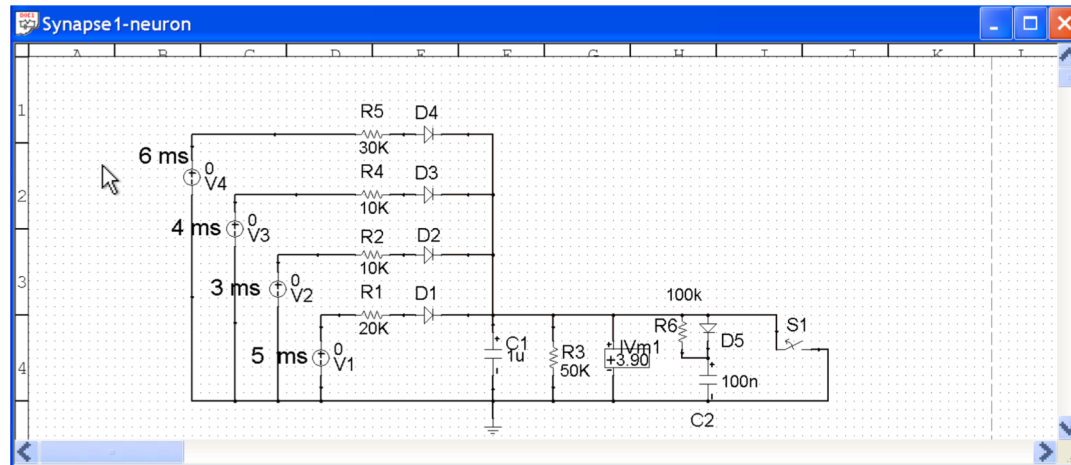
Also known as the "Perceptron"
(F. Rosenblatt 1958)

The model differs from biological neurons:

- static representation of data (no spikes)
- memoryless
- lacks time-dependent plasticity

Implementation of the Leaky Integrate-and-Fire (LIF) neuron model (shown with 4 resistive synapse inputs)

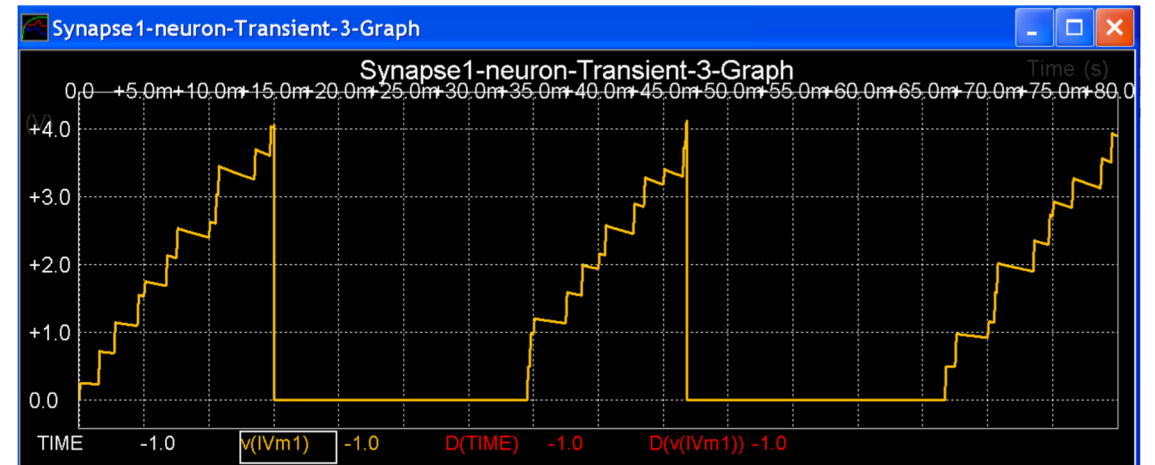
Schematics



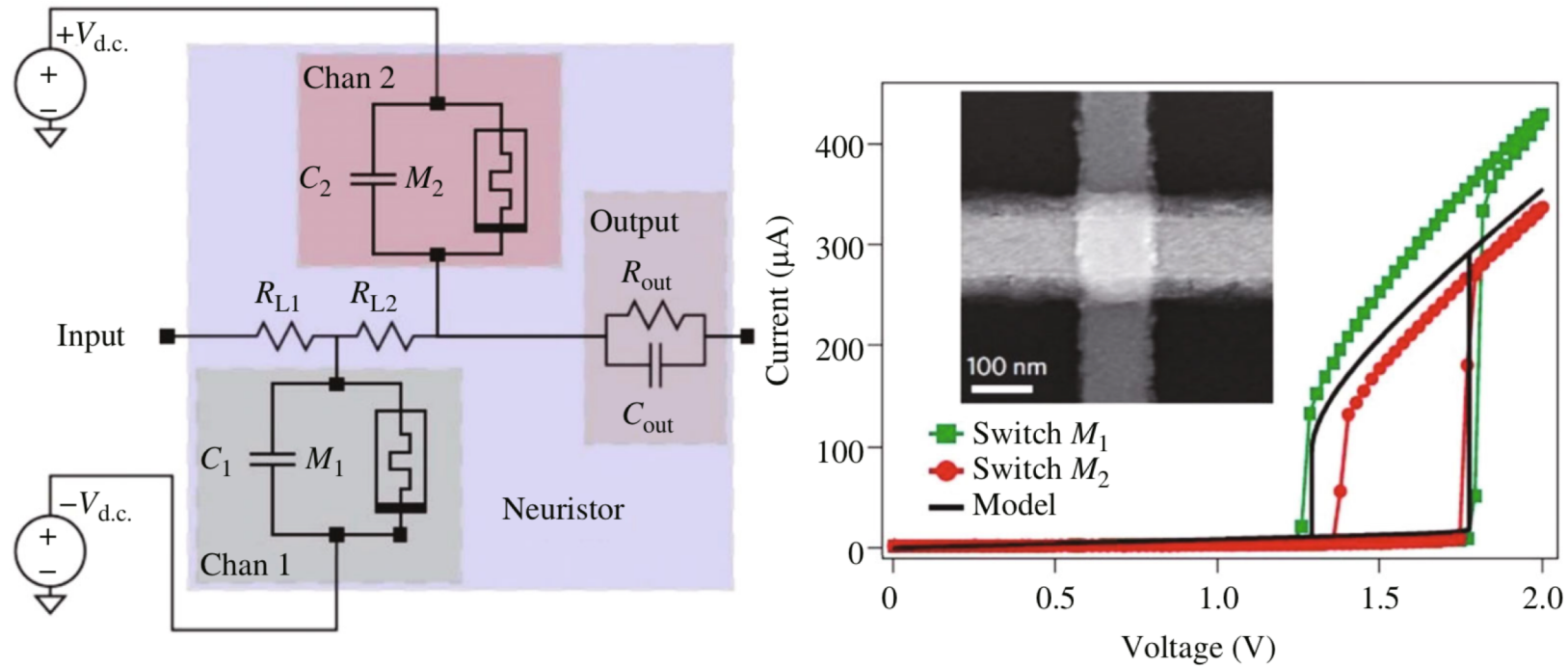
synapses

neuron

Neuron output

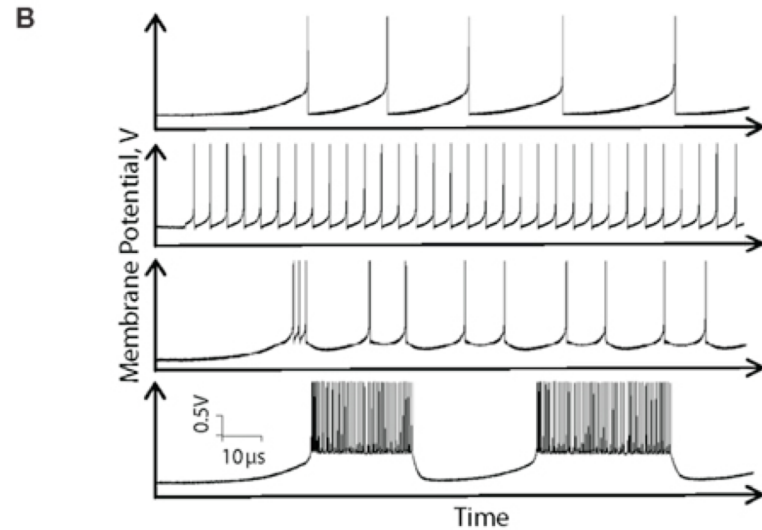
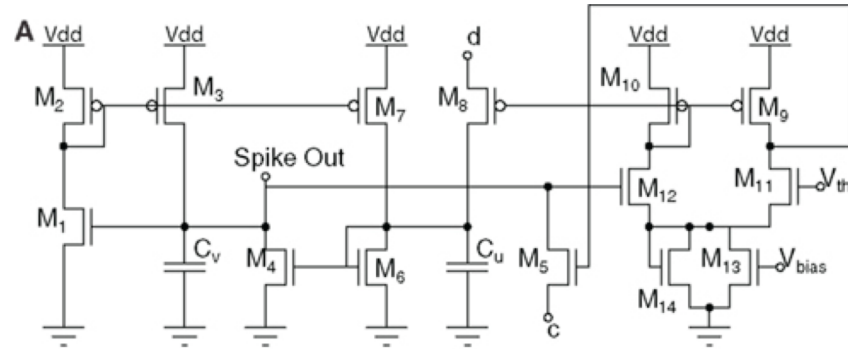


Implementing the LIF neuron with memristors (the "Neuristor")



Reprinted from Picket et al, A scalable neuristor built with Mott memristors. Nat. Material 2013, 12

CMOS neuron implementation



Indiveri et al, Neuromorphic Silicon Neuron Circuits. *Frontiers in neuroscience*. 5. 73. 10.3389/fnins.2011.00073.

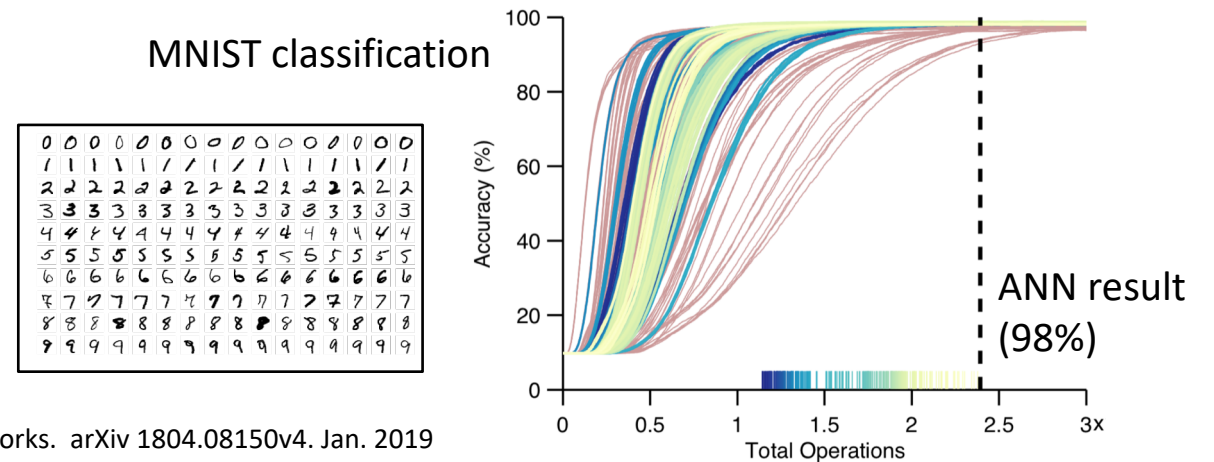
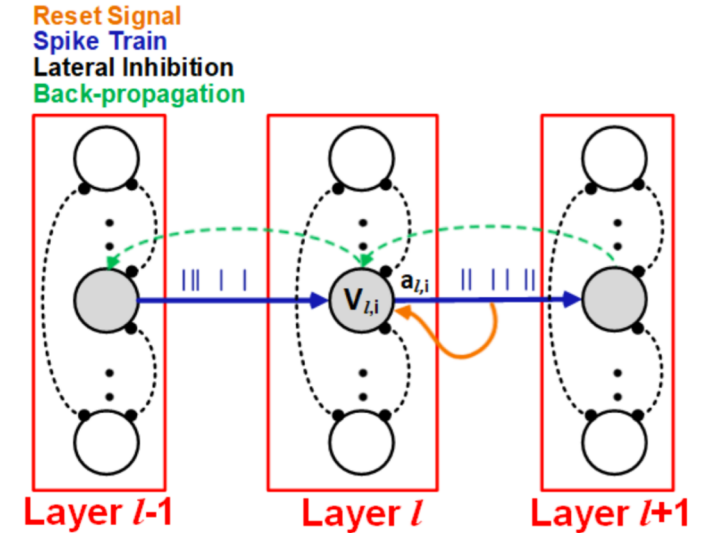
Spiking networks

Architecture

- Fully or convolutionally connected
- Lateral inhibition

Learning

- Unsupervised learning through synaptic plasticity



Reprinted from A. Tavanaei et al, Deep learning in spiking neural networks. arXiv 1804.08150v4. Jan. 2019

MNIST results

Model	Architecture	Learning method	Dataset	Acc
Feedforward, fully connected, multi-layer SNNs				
O'Connor (2016) [137]	Deep SNN	Stochastic gradient descent	MNIST	96.40
O'Connor (2016) [137]	Deep SNN	Fractional stochastic gradient descent	MNIST	97.93
Lee (2016) [57]	Deep SNN	Backpropagation	MNIST	98.88
Lee (2016) [57]	Deep SNN	Backpropagation	N-MNIST	98.74
Neftci (2017) [138]	Deep SNN	Event-driven random backpropagation	MNIST	97.98
Liu (2017) [108]	SNN	Temporal backpropagation (3-layer)	MNIST	99.10
Eliasmith (2012) [129]	SNN	Spaun brain model	MNIST	94.00
Diehl (2015) [130]	SNN	STDP (2-layer)	MNIST	95.00
Tavanaei (2017) [118]	SNN	STDP-based backpropagation (3-layer)	MNIST	97.20
Mostafa (2017) [109]	SNN	Temporal backpropagation (3-layer)	MNIST	97.14
Querlioz (2013) [139]	SNN	STDP, Hardware implementation	MNIST	93.50
Brader (2007) [128]	SNN	Spike-driven synaptic plasticity	MNIST	96.50
Diehl (2015) [140]	Deep SNN	Offline learning, Conversion	MNIST	98.60
Neil (2016) [144]	Deep SNN	Offline learning, Conversion	MNIST	98.00
Hunsberger (2015) [177], [178]	Deep SNN	Offline learning, Conversion	MNIST	98.37
Esser (2015) [141]	Deep SNN	Offline learning, Conversion	MNIST	99.42

Spiking CNNs				
Lee (2016) [57]	Spiking CNN	Backpropagation	MNIST	99.31
Lee (2016) [57]	Spiking CNN	Backpropagation	N-MNIST	98.30
Panda (2016) [173]	Spiking CNN	Convolutional autoencoder	MNIST	99.05
Panda (2016) [173]	Spiking CNN	Convolutional autoencoder	CIFAR-10	75.42
Tavanaei (2017) [171], [172]	Spiking CNN	Layer wise sparse coding and STDP	MNIST	98.36
Tavanaei (2018) [174]	Spiking CNN	Layer-wise and end-to-end STDP rules	MNIST	98.60
Kheradpisheh (2016) [170]	Spiking CNN	Layer wise STDP	MNIST	98.40
Zhao (2015) [169]	Spiking CNN	Tempotron	MNIST	91.29
Cao (2015) [183]	Spiking CNN	Offline learning, Conversion	CIFAR-10	77.43
Neil (2016) [179]	Spiking CNN	Offline learning, Conversion	N-MNIST	95.72
Diehl (2015) [140]	Spiking CNN	Offline learning, Conversion	MNIST	99.10
Rueckauer (2017) [142]	Spiking CNN	Offline learning, Conversion	MNIST	99.44
Rueckauer (2017) [142]	Spiking CNN	Offline learning, Conversion	CIFAR-10	90.85
Hunsberger (2015) [177]	Spiking CNN	Offline learning, Conversion	CIFAR-10	82.95
Garbin (2014) [181]	Spiking CNN	Offline learning, Hardware	MNIST	94.00
Esser (2016) [182]	Spiking CNN	Offline learning, Hardware	CIFAR-10	87.50
Esser (2016) [182]	Spiking CNN	Offline learning, Hardware	CIFAR-100	63.05

Reprinted from A. Tavanaei et al, Deep learning in spiking neural networks. arXiv 1804.08150v4. Jan. 2019

Large-scale neuromorphic designs

Univ. of Manchester: SpiNNaker (research)

IBM: TrueNorth (research)

Intel: Loihi (commercial)

Brainchip: Akida (commercial)

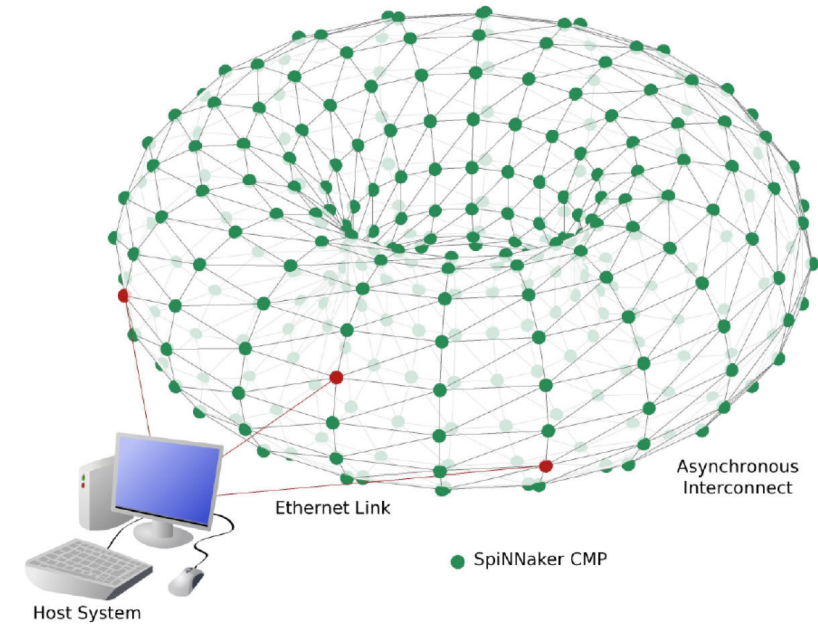
Univ. of Manchester - SpiNNaker (2010)

- Univ. of Manchester: SpiNNaker (Spiking Neural Network Architecture, 2010-2018)
- Uses 57 600 VLSI chips (18-core ARM9)
- Each VLSI chip emulates 18000 neurons
- A “full brain” simulator contains 1 billion neurons and runs in “real-time”.
- Power consumption: 100 kW (100 mW/neuron)
- Funding of 8 MEuro received 2019 to build second generation. Expected to reach 10 billion neurons.



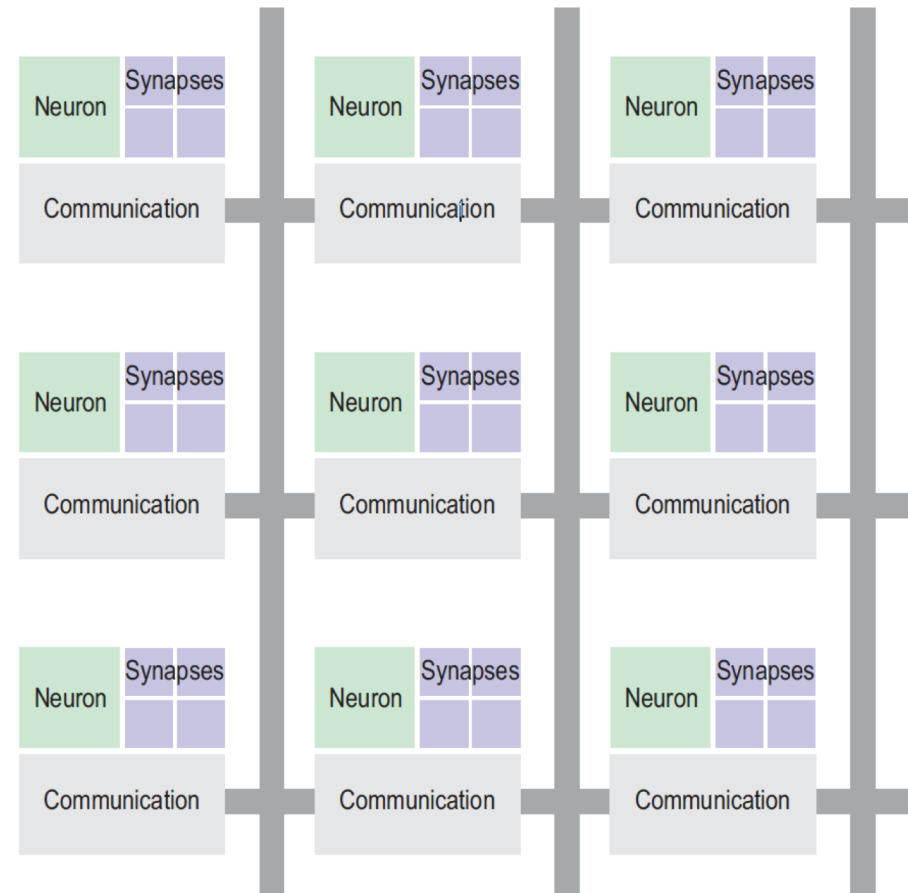
SpiNNaker – System architecture

The VLSI chips are connected in a toroidal network



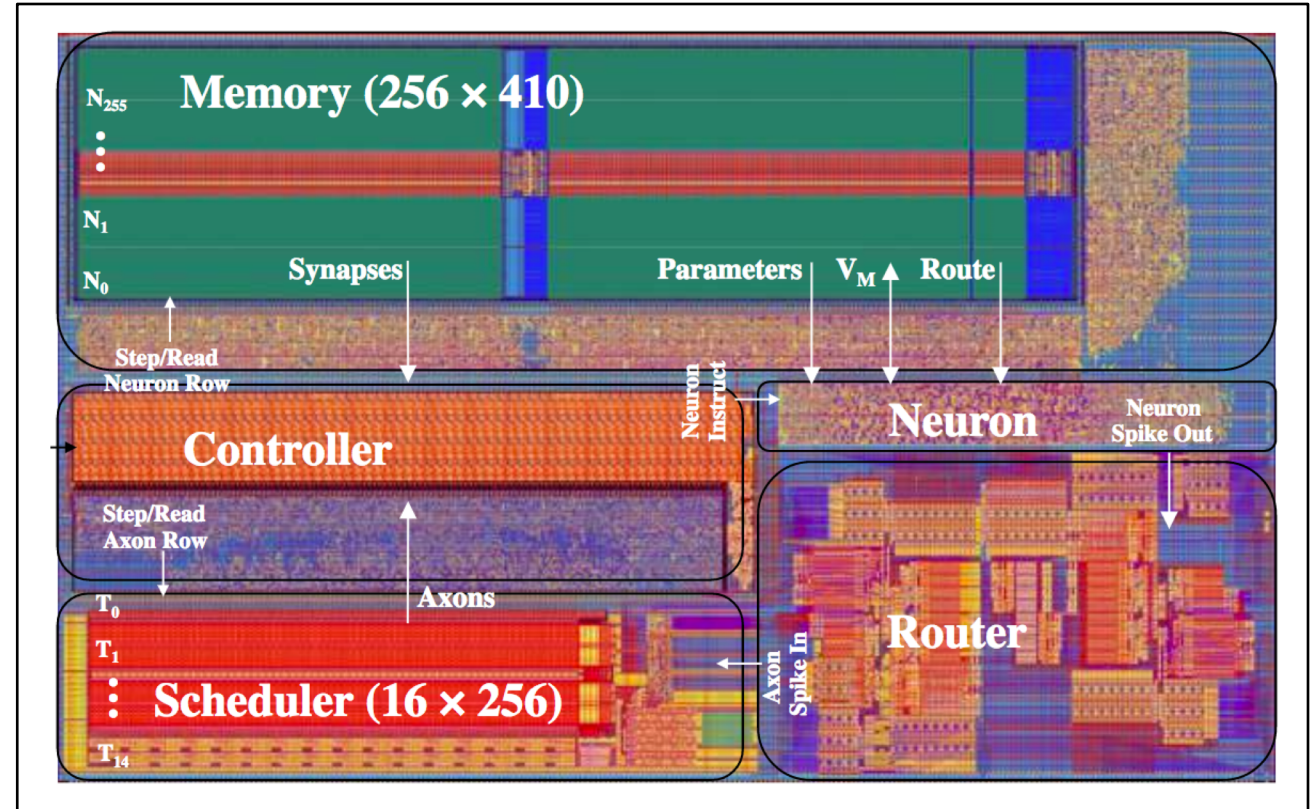
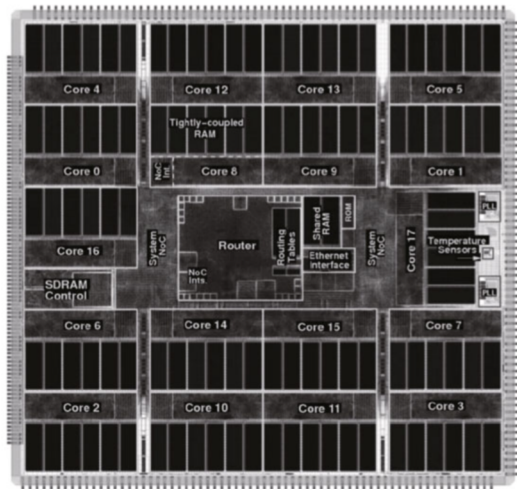
IBM – TrueNorth (2014)

- IBM: TrueNorth (2014)
- Each chip implements 1 million neurons and 256 million synapses
- 5.4 billion transistors
- Power consumption: 73 mW (73 nW/neuron)
- Distributed architecture based on an event-driven network



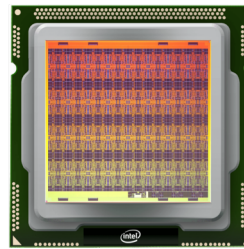
TrueNorth – core element

- Core element 256 neurons, 256*256 synapses
- Size: 240 μm \times 390 μm
- 4 096 cores per die (\approx 400 mm²)

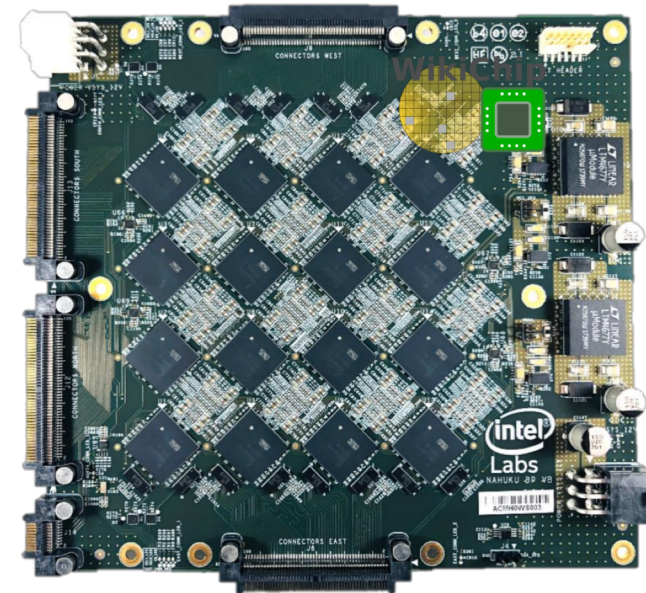


Intel – Loihi (2017)

- Highly flexible neuromorphic chip.
- 131 000 neurons divided into 128 cores
- 130 M synapses
- 14 nm CMOS
- 2.07 billion transistors

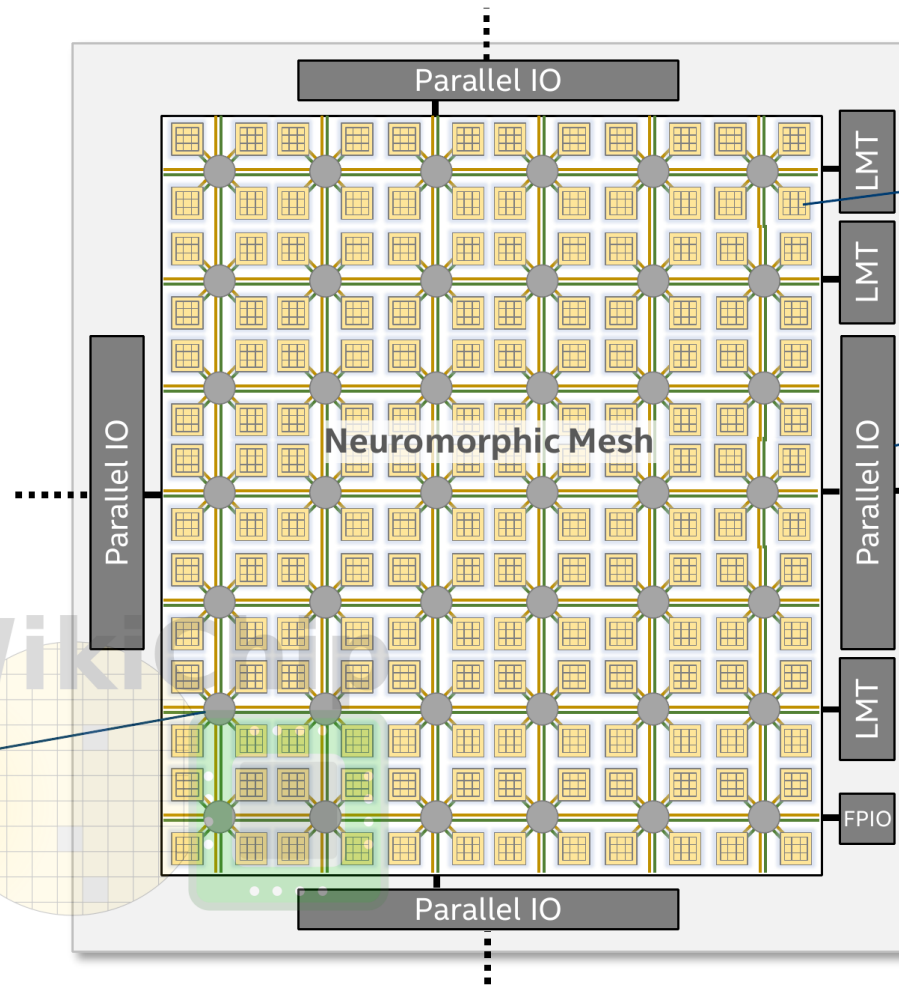


Board with 64 Loihi chips containing 8 M neurons.



Chip Architecture

Technology:	14nm
Die Area:	60 mm ²
Core area:	0.41 mm ²
NmC cores:	128 cores
x86 cores:	3 LMT cores
Max # neurons:	128K neurons
Max # synapses:	128M synapses
Transistors:	2.07 billion



Neuromorphic core

- LIF neuron model
- Programmable learning
- 128 KB synaptic memory
- Up to 1,024 neurons
- Asynchronous design

Parallel off-chip interfaces

- Two-phase asynchronous
- Single-ended signaling
- 100-200 MB/s BW

Embedded x86 processors

- Efficient spike-based communication with neuromorphic cores
- Data encoding/decoding
- Network configuration
- Synchronous design

Low-overhead NoC fabric

- 8x16-core 2D mesh
- Scalable to 1000's cores
- Dimension order routed
- Two physical fabrics
- 8 GB/s per hop

Brainchip – Akida (announced)

- 1.2 M neurons,
- 10 billion synapses
- < 0.5 W
- 28 nm CMOS
- Planned release: 2020
- CNN support

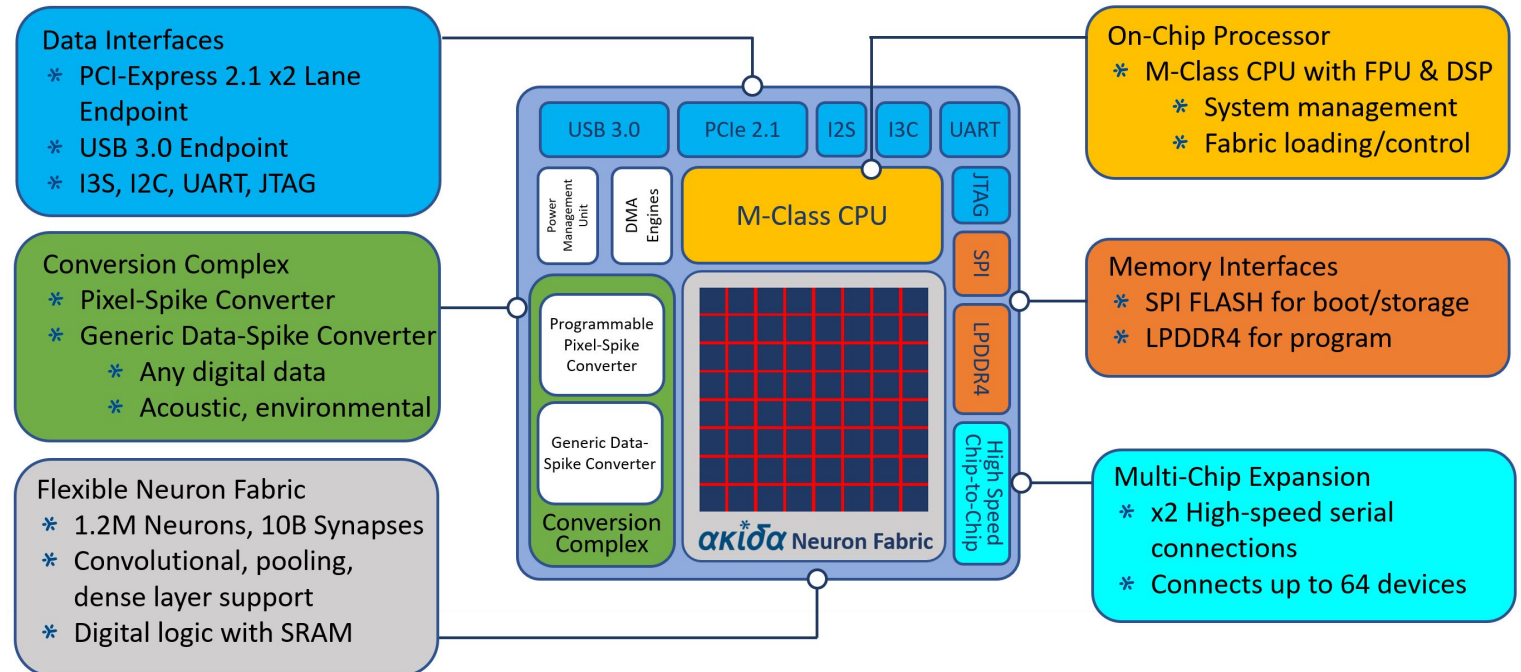


Table 1. A comparison of state-of-the-art neuromorphic chips, along with some performance attributes.

Chip	Technology	Integration Density	Key Functionality/Performance Metrics
SpiNNaker	ARM968, 130-nm CMOS (next-generation prototypes: ARM M4F, 28-nm CMOS)	Up to 1,000 neurpns/core, 1 million cores	Programmable numerical simulations with 72-bit messages, for real-time simulation of spiking networks
TrueNorth	Digital ASIC at 28-nm CMOS	1 million neurons, 256 million synapses; 1-bit synaptic state to represent a connection, with four programmable 9-bit weights per neuron	SNN emulation without on-chip learning; 26 pJ per synaptic operation
Loihi	Digital ASIC at 14-nm CMOS	130,000 neurons, 130 million synapses with variable weight resolution (1–9 bits)	Supports on-chip learning with plasticity rules, such as Hebbian, pairwise, and triplet STDP, 23.6 pJ per synaptic operation (at nominal operating conditions)
BrainScaleS	Mixed-signal wafer-scale system, 180-nm CMOS (next-generation prototype: 65-nm CMOS)	180,000 neurons, 40 million synapses per wafer	10^3 – 10^4 -fold acceleration of spiking network emulations, with hardware-supported synaptic plasticity; next-generation prototype: programmable plasticity
Braindrop	Mixed-signal 28-nm CMOS	4,096 neurons, 64,000 programmable weights (with analog circuits that allow realization of all-to-all connectivity)	0.38 pJ per synaptic update, implements the single core of a planned million-neuron chip
DYNAP-SE	Mixed-signal 180-nm CMOS	1,024 neurons, 64,000 synapses (12-bit content-addressable memory)	Hybrid analog/digital circuits for emulating synapse and neuron dynamics, 17 pJ per synaptic operation
ODIN	Digital ASIC at 28-nm CMOS	256 neurons, 64,000 synapses with 3-bit weight and 1 bit to encode learning	12.7 pJ per synaptic operation, implements on-chip spike-driven plasticity

Spiking neural networks vs Artificial neural networks

	SNN	ANN
Synapse short-term memory	√	—
Synapse long-term memory	√	√
Neuron memory	√	—
Learning mode	unsupervised	supervised
Power consumption	low	high
Hardware efficiency - analog	high	high
Hardware efficiency - digital	high	low
Industrial maturity	low	high

Basic references

- Sergio Davis, *Learning in Spiking Neural Networks*, PhD Thesis, Univ. of Manchester. 2012.
- Schemmel, J., Bruderle, D., Grubl, A., Hock, M., Meier, K., Millner, S., *A wafer-scale neuromorphic hardware system for large-scale neural modeling*. Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium. IEEE, pp. 1947–1950. May 2010.
- Rast, A. D., Jin, X., Galluppi, F., Plana, L. A., Patterson, C., Furber, S., 2010c. *Scalable event-driven native parallel processing: the SpiNNaker neuromimetic system*. In: Proceedings of the 7th ACM international conference on Computing frontiers. CF'10. ACM, New York, NY, USA, pp. 21–30.
- Indiveri, G., Chicca, E., Douglas, R., *A VLSI Array of Low-Power Spiking Neurons and Bistable Synapses With Spike-Timing Dependent Plasticity*. IEEE Transactions on Neural Networks 17 (1), 211–221. Jan. 2006.
- Diorio, C.; Hasler, P.; Minch, A.; Mead, C.A. *A single-transistor silicon synapse*. IEEE Transactions on Electron Devices. **43** (11): 1972–1980. 1995.
- Wijekoon, J. H. B., and Dudek, P. *Compact silicon neuron circuit with spiking and bursting behaviour*. *Neural Netw.* 21, 524–534. 2008.
- Min-Woo Kwon, Hyungjin Kim, Jungjin Park, and Byung-Gook Park., *Integrate-and-Fire Neuron Circuit and Synaptic Device with Floating Body MOSFETs*. Journal of Semiconductor Techn. And Science, Vol.14, No.6, Dec, 2014.
- P. Hasler, C. Diorio, B. Minch, and C. Mead, *Single transistor learning synapses*, in Advances in Neural Information Processing Systems 7, D.S. T. G. Tesauro and T. K. Leen, Eds. Cambridge, MA: MIT Press, 1994, pp. 817–824.
- Merolla P. A., Arthur J. V., Alvarez-Icaza R., Cassidy A. S., Sawada J., Akopyan F., et al. *A million spiking-neuron integrated circuit with a scalable communication network and interface*. Science 345, 668–673. 2014.

Recent publications

- J. Park et al, *A 65-nm Neuromorphic Image Classification Processor With Energy-Efficient Training Through Direct Spike-Only Feedback*, IEEE Journal of Solid-State Physics, Vol. 55, No. 1, Jan 2020.
- H.Cho et al, *An On-Chip Learning Neuromorphic Autoencoder With Current-Mode Transposable Memory Read and Virtual Lookup Table*, IEEE Tr Biomedical Circuits and Systems, Vol 12, No 1, Feb 2018.
- S. Friedmann et al, *Demonstrating Hybrid Learning in a Flexible Neuromorphic Hardware System*, IEEE Tr on Biomedical Circuits and Systems, Vol. 11, No 1, Feb 2017.
- Y. Ahang et al, *A Digital Liquid State Machine With Biologically Inspired Learning and Its Application to Speech Recognition*, IEEE Tr on Neural Networks and Learning Systems, Vol. 26, No. 11, Nov 2015.
- U. Liu et al, *Enabling Non-Hebbian Learning in Recurrent Spiking Neural Processors With Hardware-Friendly On-Chip Intrinsic Plasticity*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol. 9, No. 3, Sept 2019.
- B. Rajendran et al, *Low-Power Neuromorphic Hardware for Signal Processing Applications*, IEEE Signal Processing Magazine, Nov 2019.
- A. R. Young et al, *A Review of Spiking Neuromorphic Hardware Communication Systems*, IEEE Access, Vol 7, 2019.

Home assignment

Describe two examples how the basic learning scheme (Hebbian/STDP) can be augmented to make hardware implementation more efficient.

Your report should include a description of the obtained improvements in terms of reduced hardware or faster computations or more efficient learning.

(Hint: Select the examples from the list of Recent Publications)

Questions?