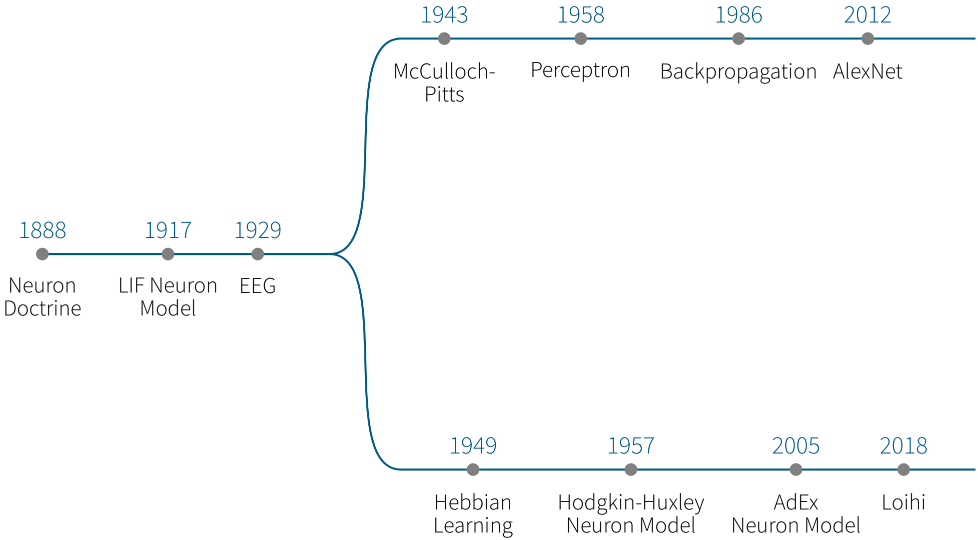
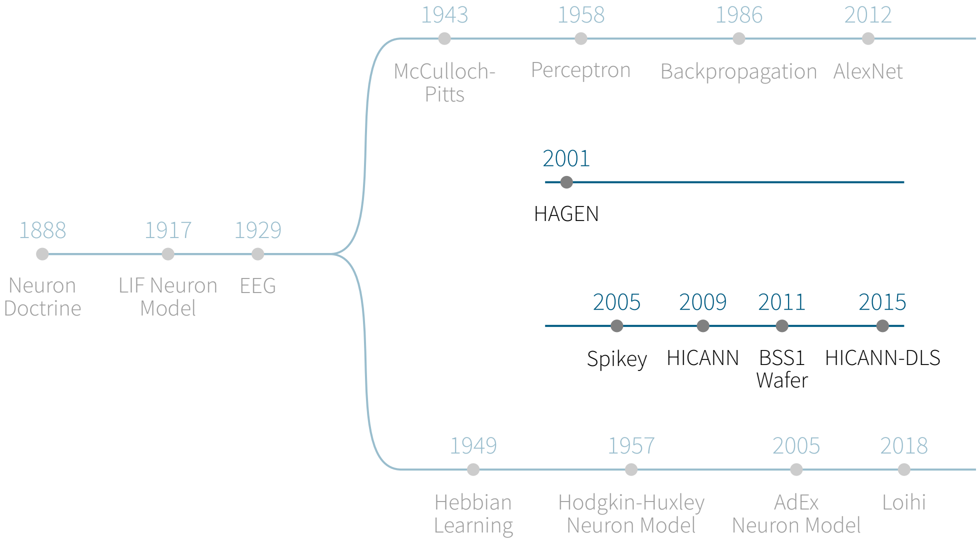
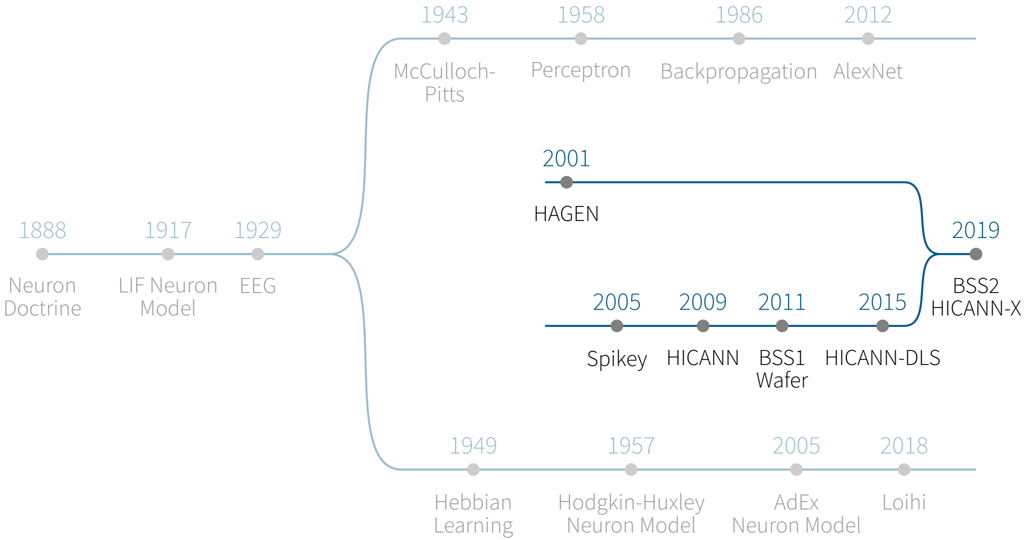
Deep Learning with Analog Neuromorphic Hardware

February 13, 2020 | Yannik Stradmann | Kirchhoff-Institute for Physics, Heidelberg University


1888
Neuron Doctrine
1917
LIF Neuron Model
1929
EEG
1943
McCulloch-Pitts
1958
Perceptron
1986
Backpropagation
2012
AlexNet
1949
Hebbian Learning
1957
Hodgkin-Huxley Neuron Model
2005
AdEx Neuron Model
2018
Loihi

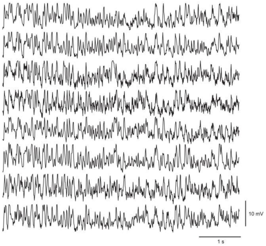

1888
Neuron Doctrine
1917
LIF Neuron Model
1929
EEG
1943
McCulloch-Pitts
1958
Perceptron
1986
Backpropagation
2012
AlexNet
2001
HAGEN
2005
Spikey
2009
HICANN
2011
BSS1 Wafer
2015
HICANN-DLS
1949
Hebbian Learning
1957
Hodgkin-Huxley Neuron Model
2005
AdEx Neuron Model
2018
Loihi


1888
Neuron Doctrine
1917
LIF Neuron Model
1929
EEG
1943
McCulloch-Pitts
1958
Perceptron
1986
Backpropagation
2012
AlexNet
2001
HAGEN
2005
Spikey
2009
HICANN
2011
BSS1 Wafer
2015
HICANN-DLS
2019
BSS2 HICANN-X
1949
Hebbian Learning
1957
Hodgkin-Huxley Neuron Model
2005
AdEx Neuron Model
2018
Loihi

Biophysical Emulation

Biophysical Emulation

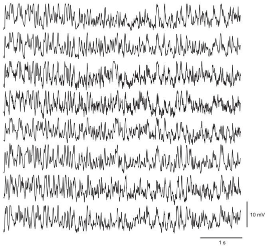
Measure



Measurement by Mohanty, Scholl, and Priebe (2012).

Biophysical Emulation

Measure



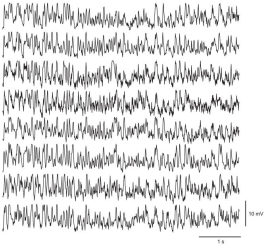
Model

$$C_\mathrm{m} \frac{\mathrm{d}V_\mathrm{m}}{\mathrm{d}t} = -g_\mathrm{leak} \left(V_\mathrm{m} - V_\mathrm{leak}\right) + I_\mathrm{stim}$$

Measurement by Mohanty, Scholl, and Priebe (2012).

Biophysical Emulation

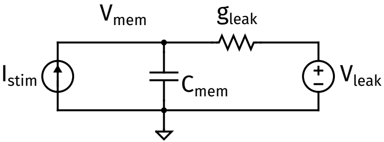
Measure



Model

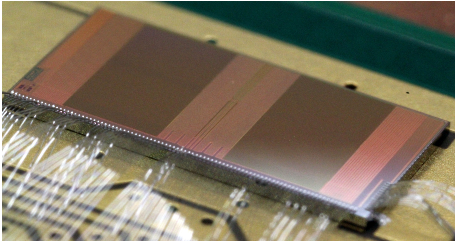
$$C_\mathrm{m} \frac{\mathrm{d}V_\mathrm{m}}{\mathrm{d}t} = -g_\mathrm{leak}\left(V_\mathrm{m} - V_\mathrm{leak}\right) + I_\mathrm{stim}$$

Implement



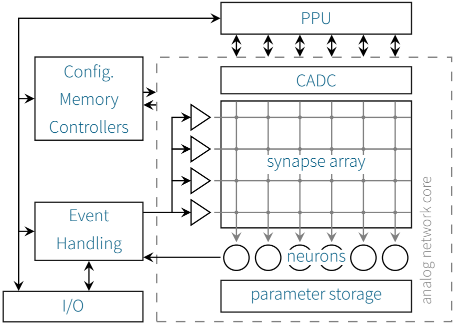
Measurement by Mohanty, Scholl, and Priebe (2012).

BrainScales2 – Overview



- Hybrid neuromorphic system, 65 nm CMOS
- $1000\times$ speedup
- 512 multi-compartment AdEx neurons
- 512×256 synapse circuits
- Two general purpose SIMD processors
- 1024 columnar ADC channels (8 bit)
- $16\,\mathrm{Gbit\,s^{-1}}$ (full duplex) I/O

BrainScales2 – Overview



- Hybrid neuromorphic system, 65 nm CMOS
- $1000\times$ speedup
- 512 multi-compartment AdEx neurons
- 512×256 synapse circuits
- Two general purpose SIMD processors
- 1024 columnar ADC channels (8 bit)
- 16 Gbit s^{-1} (full duplex) I/O

Figure adapted from Billaudelle, Stradmann, et al. (2019).

Simulation vs. Emulation

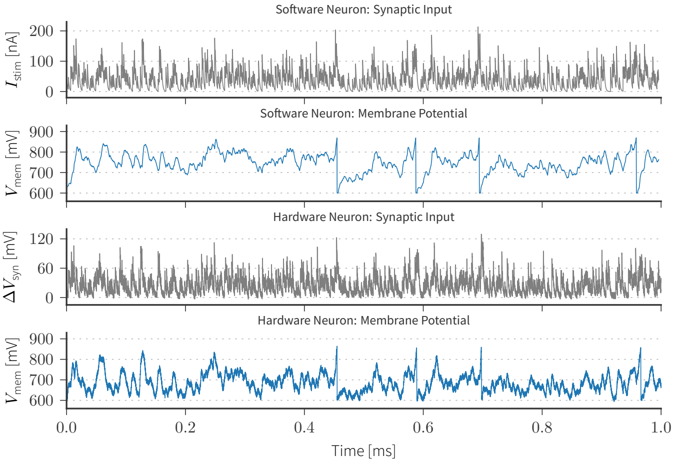
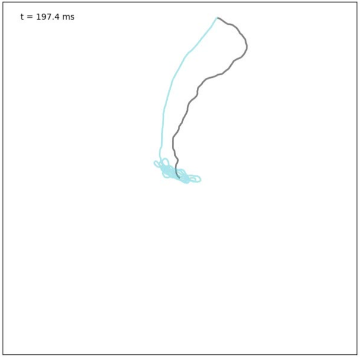
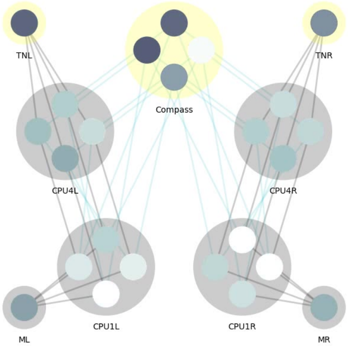


Figure adapted from Aamir et al. (2018).

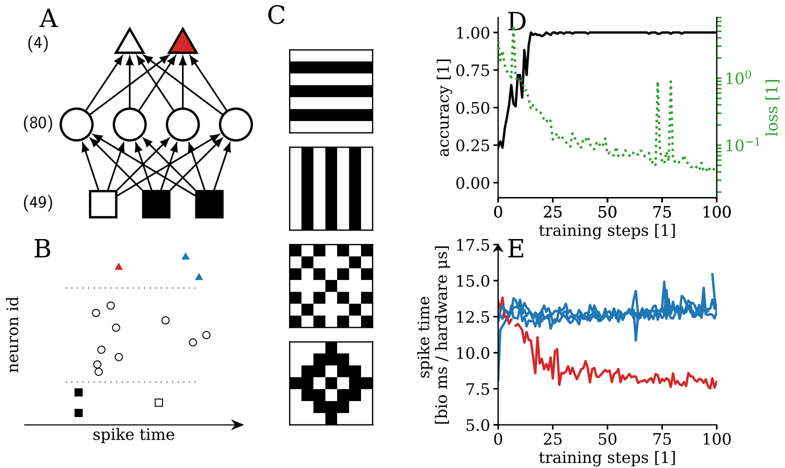
Accelerated Emulation of Spiking Neural Networks





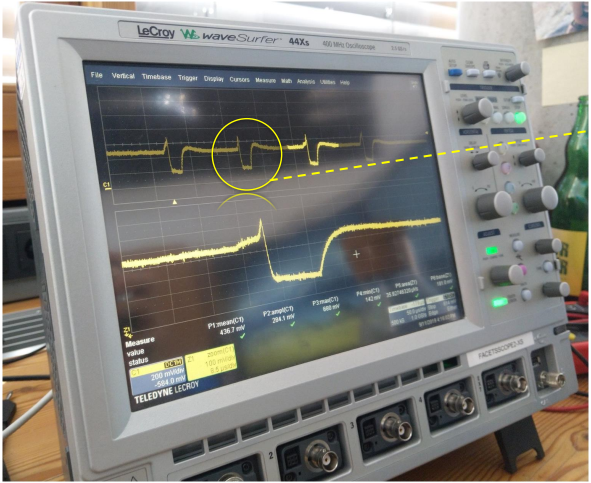
Experiment by Korbinian Schreiber (Billaudelle, Stradmann, et al., 2019).

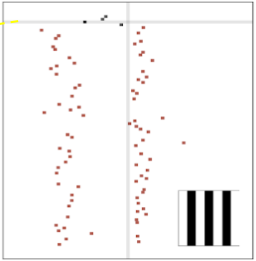
Single Spike Coding – Time to First Spike



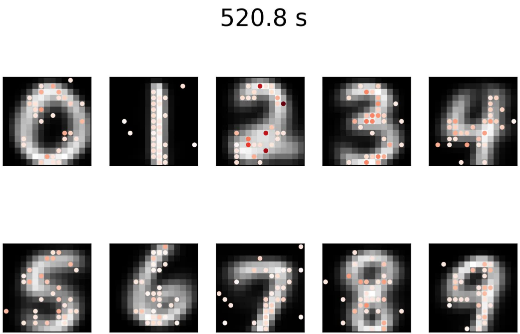
Göltz et al. (2019)

Single Spike Coding – Time to First Spike





On-chip Learning

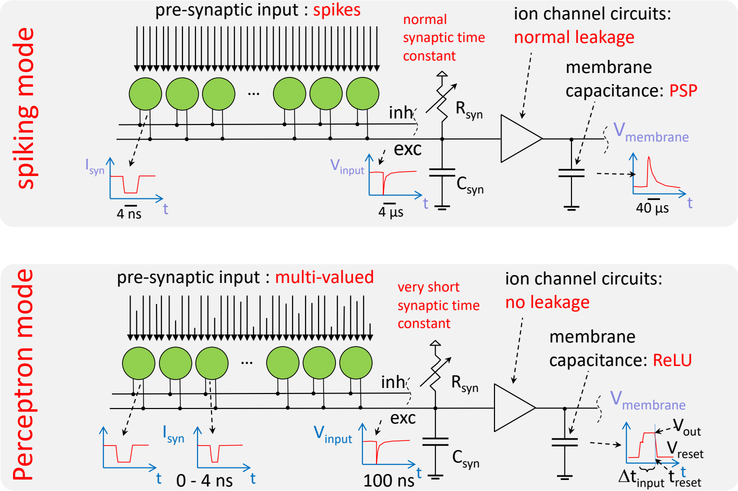


On-chip learning rule:

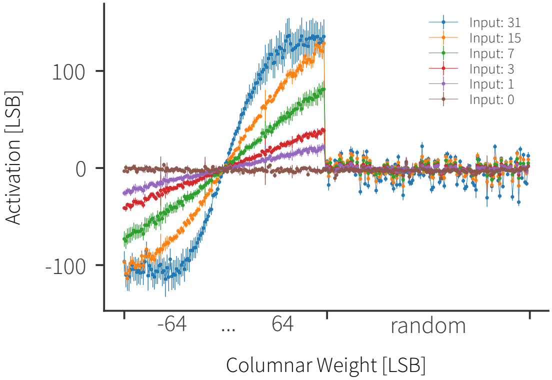
- STDP
- Homeostasis
- Pruning

Experiment by Billaudelle, Cramer, et al. (2019).

BrainScaleS-2 – Spikes and Activations



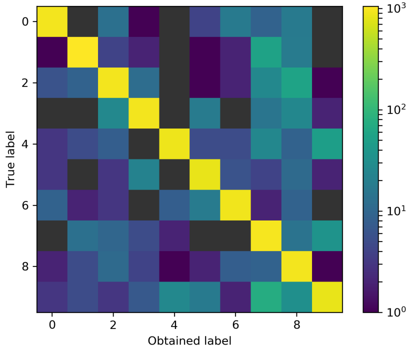
Inference on BSS2 – (Very) Early Results: Analog MAC



input resolution	5 bit
weight resolution	6 bit + sign
activation resolution	8 bit
analog precision	???

Unpublished, measurement designed and executed by Johannes Weis.

Inference on BSS2 – (Very) Early Results: MNIST



- Simple Architecture
  - One convolutional layer (10×10)
  - Two dense layers (128 units, 10 units)
- Achieved accuracy
  - Software: 98.42%
  - Hardware: 91.54% (without re-training)

Unpublished, measurement designed and executed by Johannes Weis.

Model Creation – Hardware in the Loop

CPU/GPU
Cluster

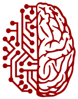
Model
Development

Compute
Graph

Activations

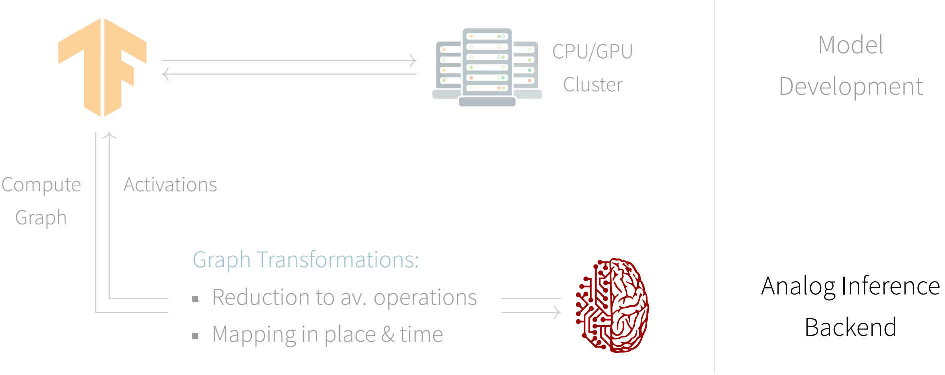
Graph Transformations:

- Reduction to av. operations
- Mapping in place & time



Analog Inference
Backend

Model Application – Analog Inference



In the (near) future…

Software

- “Hardware in the Loop” ANN training
- TensorFlow/PyTorch integration
- SNN abstraction layer
- Compiler support for SIMD operations

Hardware

- Tape-Out in 02/2020
- Inference throughput: up to 131 GOPS
- Spike throughput: up to 250 MEvents s^{-1}
- Power consumption: ≈ 1 W

Summary

BrainScaleS-2 is an analog neural network accelerator

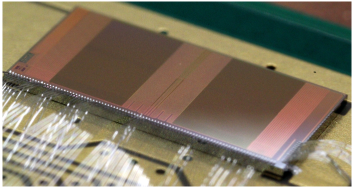
... manufactured in an affordable 65 nm CMOS process

... suitable for artificial neural networks

... suitable for spiking neural networks (1000× speedup)

... optimized for low-power applications

... embedding SIMD microprocessors for on-chip learning



References

Aamir, Syed Ahmed et al. (2018). "An accelerated LIF neuronal network array for a large-scale mixed-signal neuromorphic architecture". In: IEEE Transactions on Circuits and Systems I: Regular Papers 65.12, pp. 4299–4312.

Billaudelle, Sebastian, Benjamin Cramer, et al. (2019). "Structural plasticity on an accelerated analog neuromorphic hardware system". In: arXiv preprint arXiv:1912.12047.

Billaudelle, Sebastian, Yannik Stradmann, et al. (2019). "Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate". In: arXiv preprint arXiv:1912.12980.

Göltz, Julian et al. (2019). "Fast and deep neuromorphic learning with time-to-first-spike coding". In: arXiv preprint arXiv:1912.11443.

Mohanty, Deepankar, Benjamin Scholl, and Nicholas J Priebe (2012). "The accuracy of membrane potential reconstruction based on spiking receptive fields". In: Journal of neurophysiology 107.8, pp. 2143–2153.