Brain-Inspired Computing

An Introduction Into Accelerated Analog Neuromorphic Computing with BrainScaleS

Johannes Schemmel

Electronic Vision(s) Group
Kirchhoff Institute for Physics
Heidelberg University, Germany

Electronic Vision(s)

Kirchhoff Institute of Physics, Heidelberg University

Founded 1995 by Prof. Karlheinz Meier (†2018)

1995 HDR vision sensors

1996 analog image processing

2000 Perceptron based analog neural networks:
EVOOPT and HAGEN

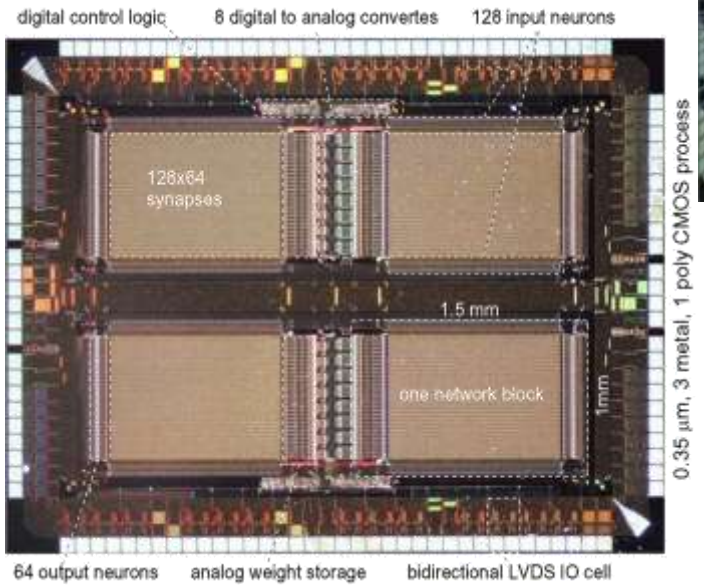
2003 First concepts for spike based analog neural networks

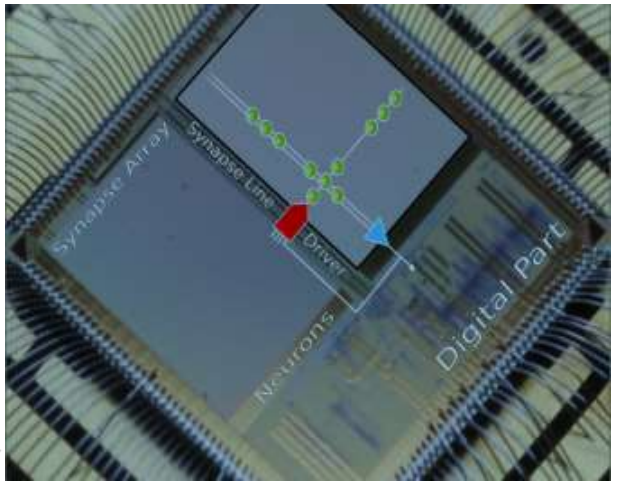
2004 First accelerated analog neural network chip with short and long term plasticity: Spikey

HAGEN: Perceptron-based Neuromorphic chip

introduced:

- accelerated operation
- mixed-signal Kernels





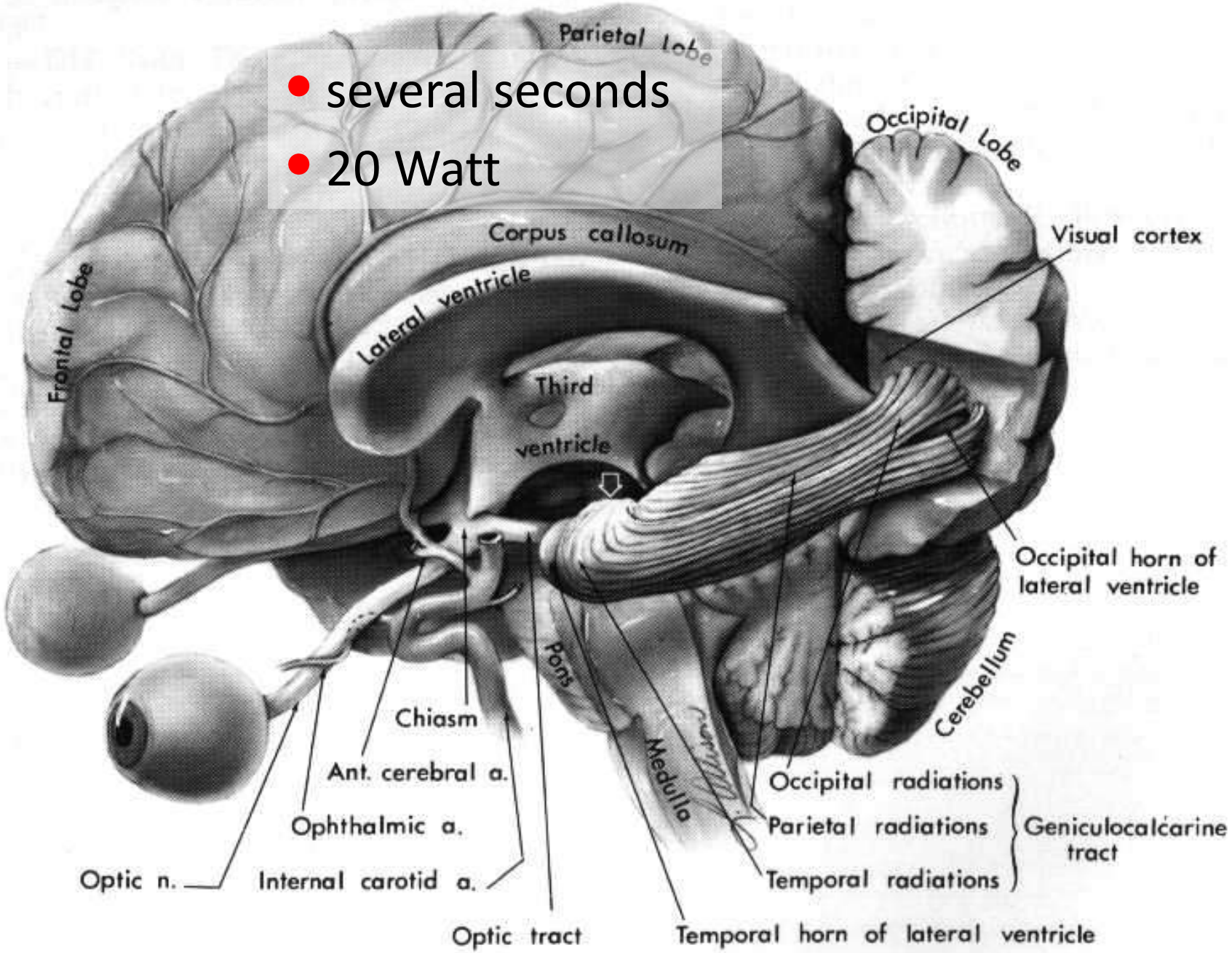
SPIKEY: spike-based Neuromophic chip

introduced:

- fully-parallel Spike-Time-Dependent-Plasticity
- analog parameter storage for calibratable physical model

- 1 Trillion calculations per second
- 5 Watt

- several seconds
- 20 Watt

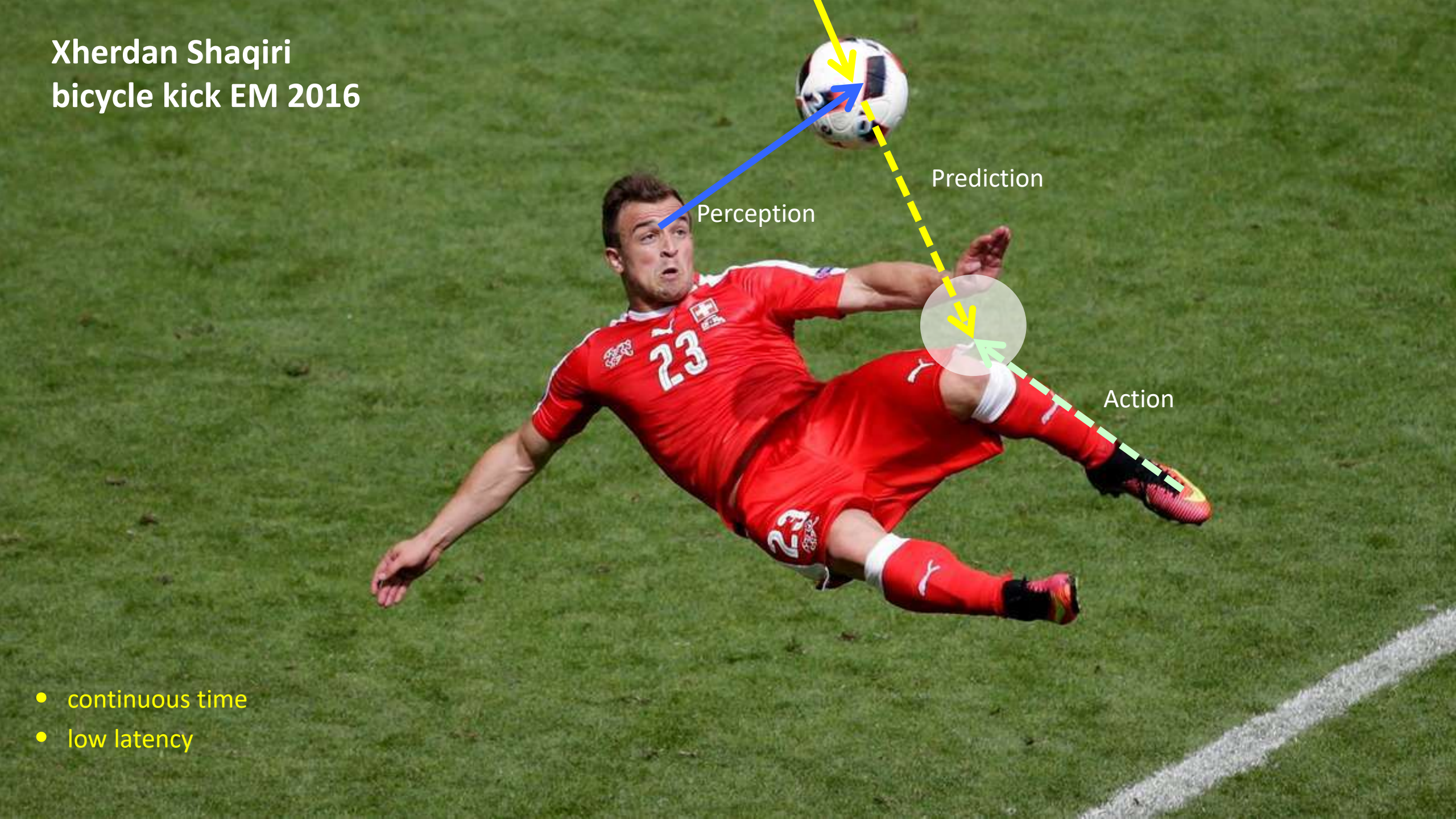


- 0.2 seconds
- 20 Watt

Parietal Lobe
Occipital Lobe
Visual cortex
Corpus callosum
Frontal Lobe
Third ventricle
Occipital horn of lateral ventricle
Cerebellum
Pons
Medulla
Chiasm
Ant. cerebral a.
Ophthalmic a.
Optic n.
Internal carotid a.
Optic tract
Occipital radiations
Parietal radiations
Temporal radiations
Geniculocalcarine tract
Temporal horn of lateral ventricle

Xherdan Shaqiri
bicycle kick EM 2016

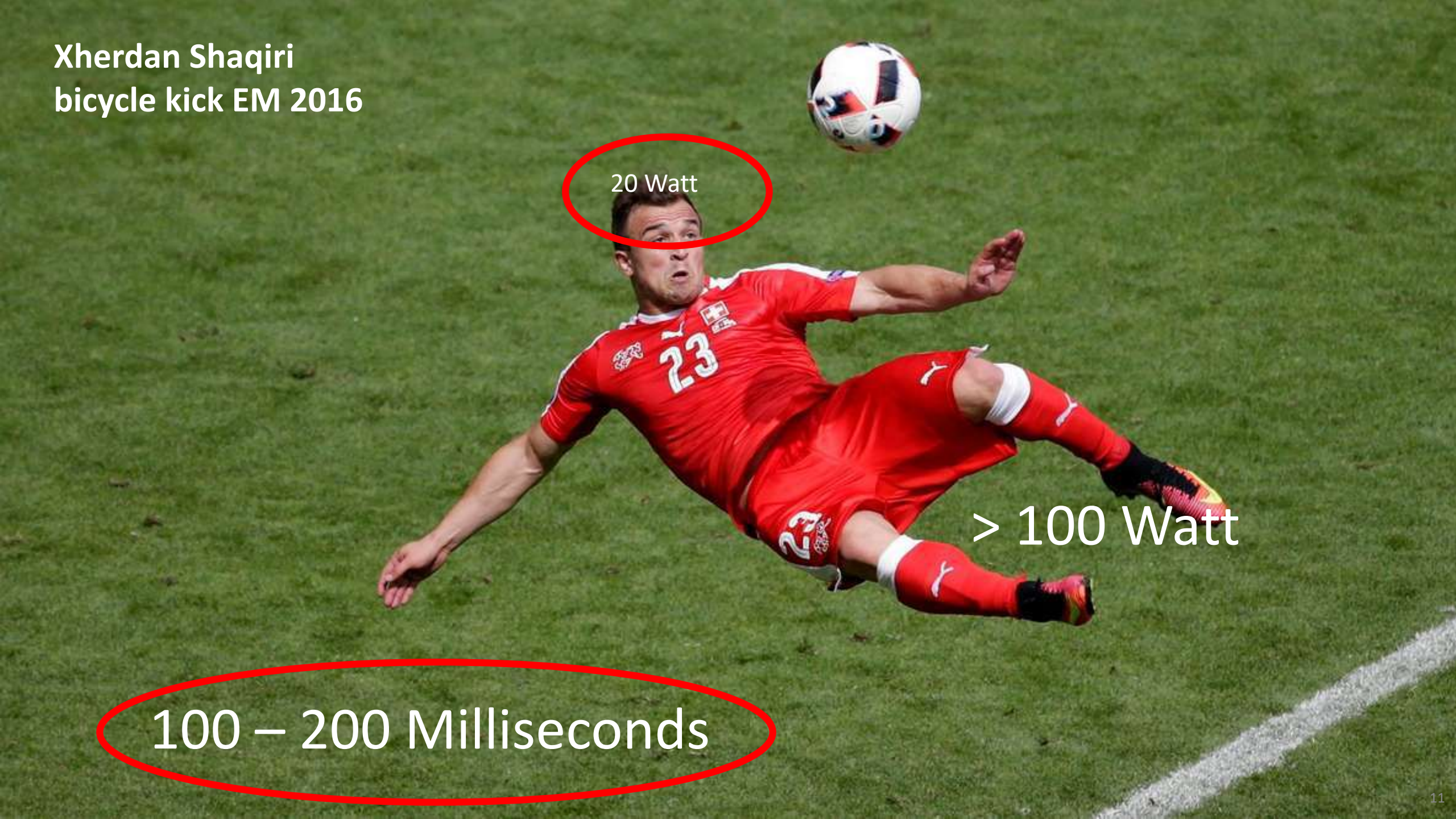
Xherdan Shaqiri
bicycle kick EM 2016



- continuous time
- low latency

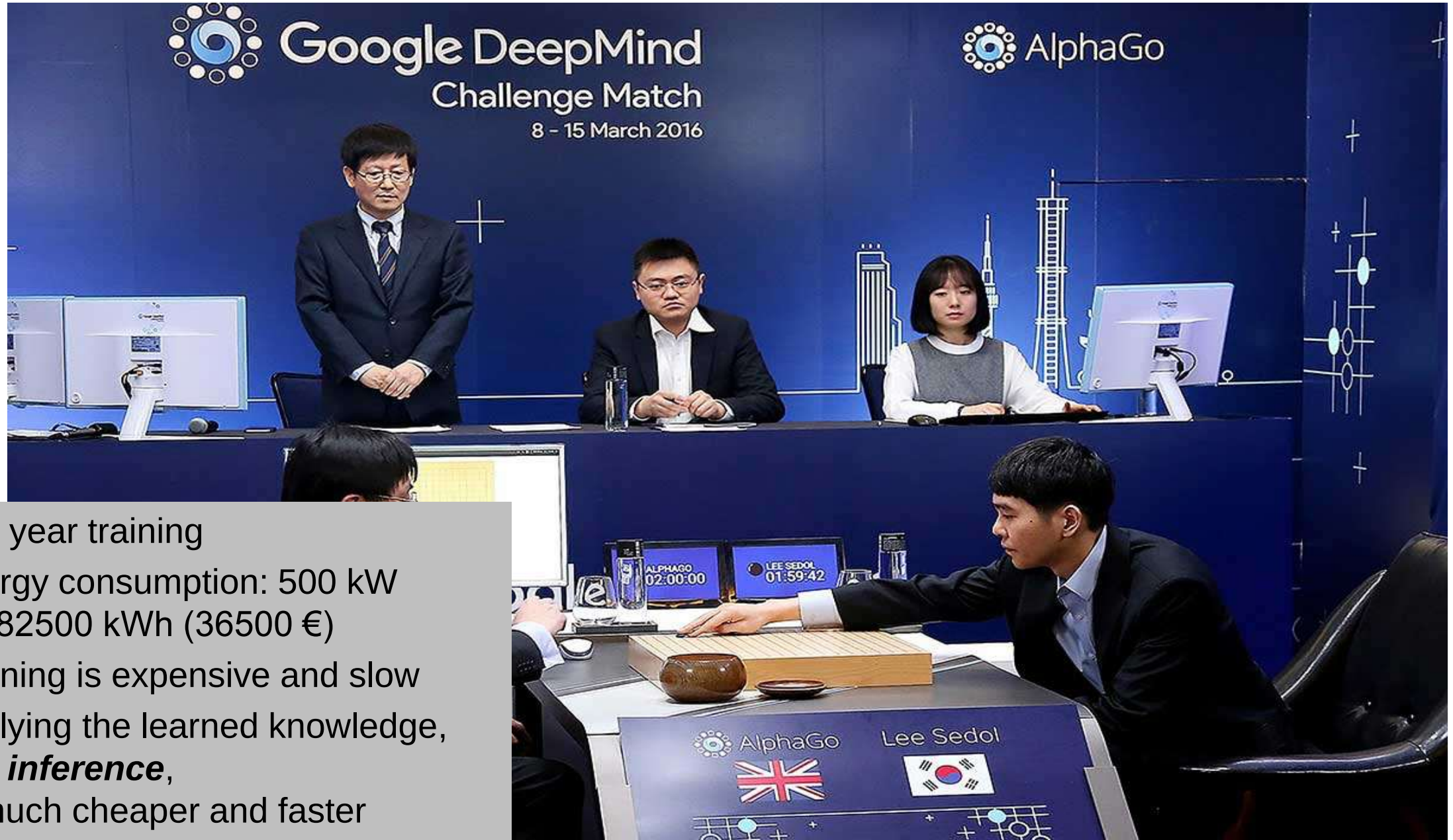

88:07
SUI
1-1
POL
SRF sport

Xherdan Shaqiri
bicycle kick EM 2016




Tots FC

Computers are becoming more brain-like



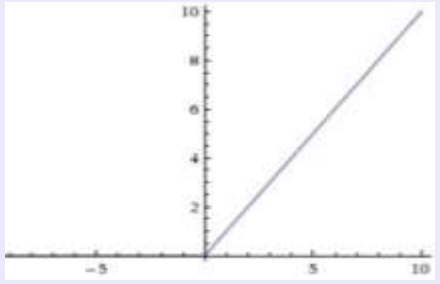
- one year training
- energy consumption: 500 kW
  →182500 kWh (36500 €)
- learning is expensive and slow
- applying the learned knowledge, aka ***inference***,
  is much cheaper and faster

Perceptron model (biology of 1950)

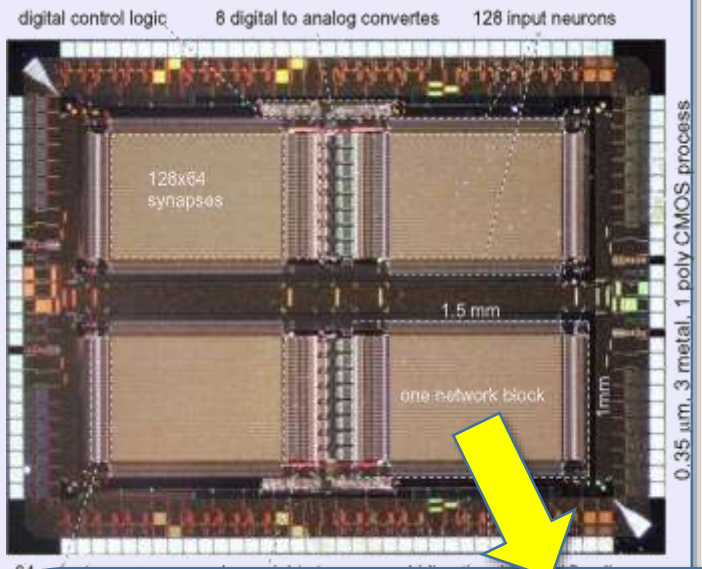
- used in Machine Learning
- vector-matrix multiplication

$$f\left(\sum_i w_i x_i + b\right)$$

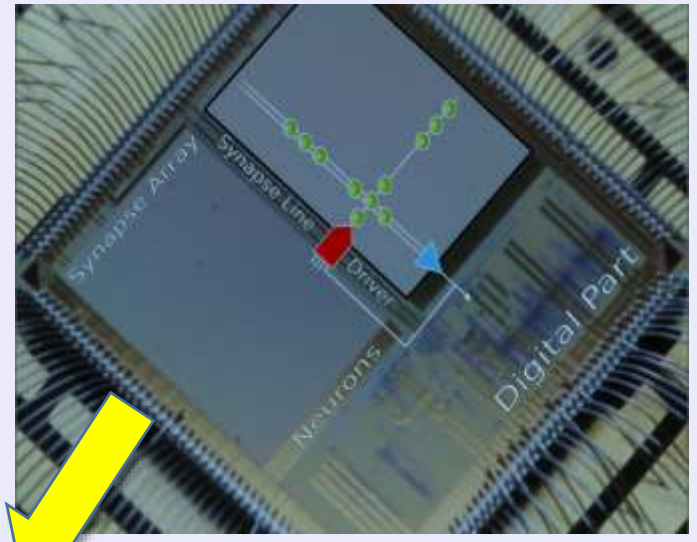
- simple non-linear activation function f (ReLU):



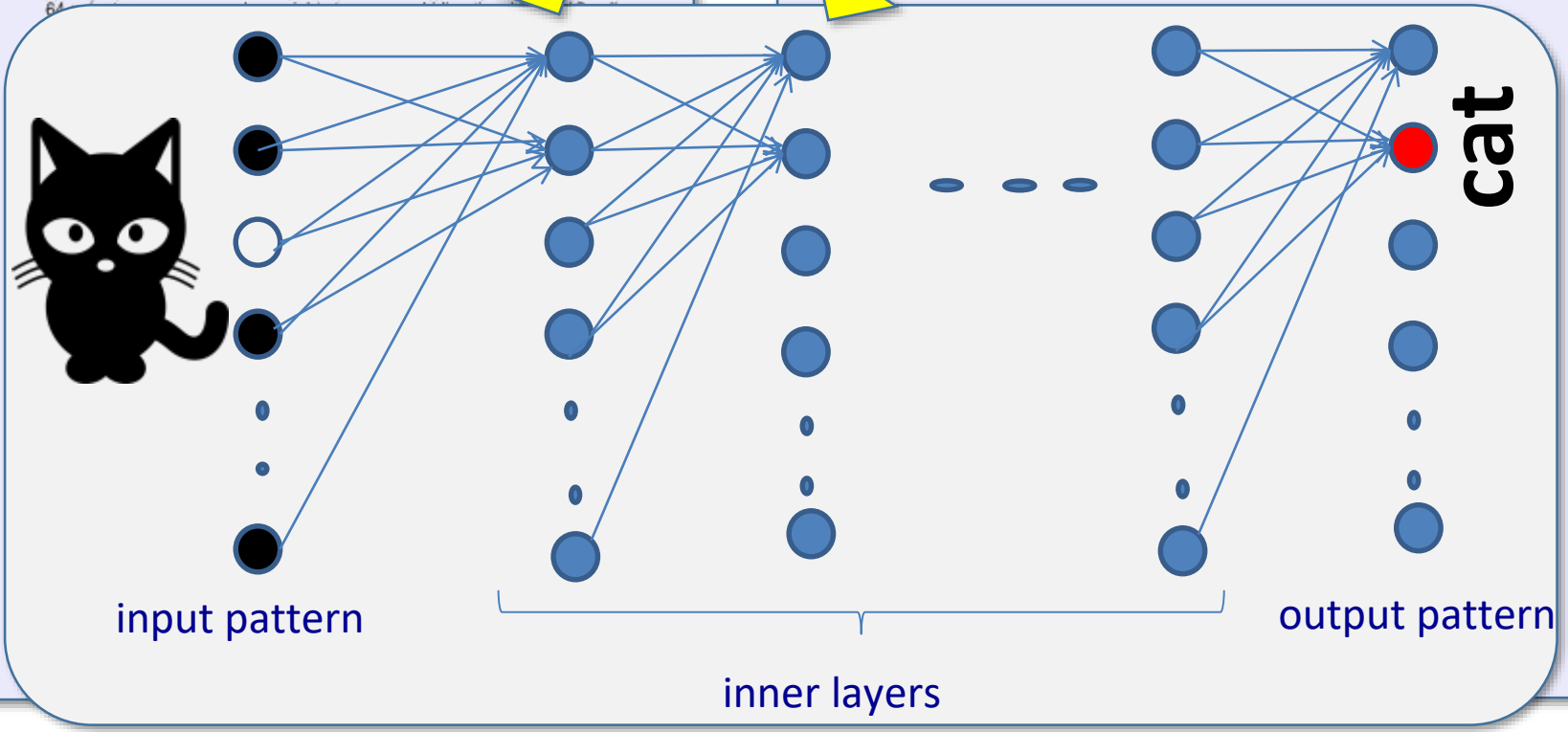
- trained with backpropagation



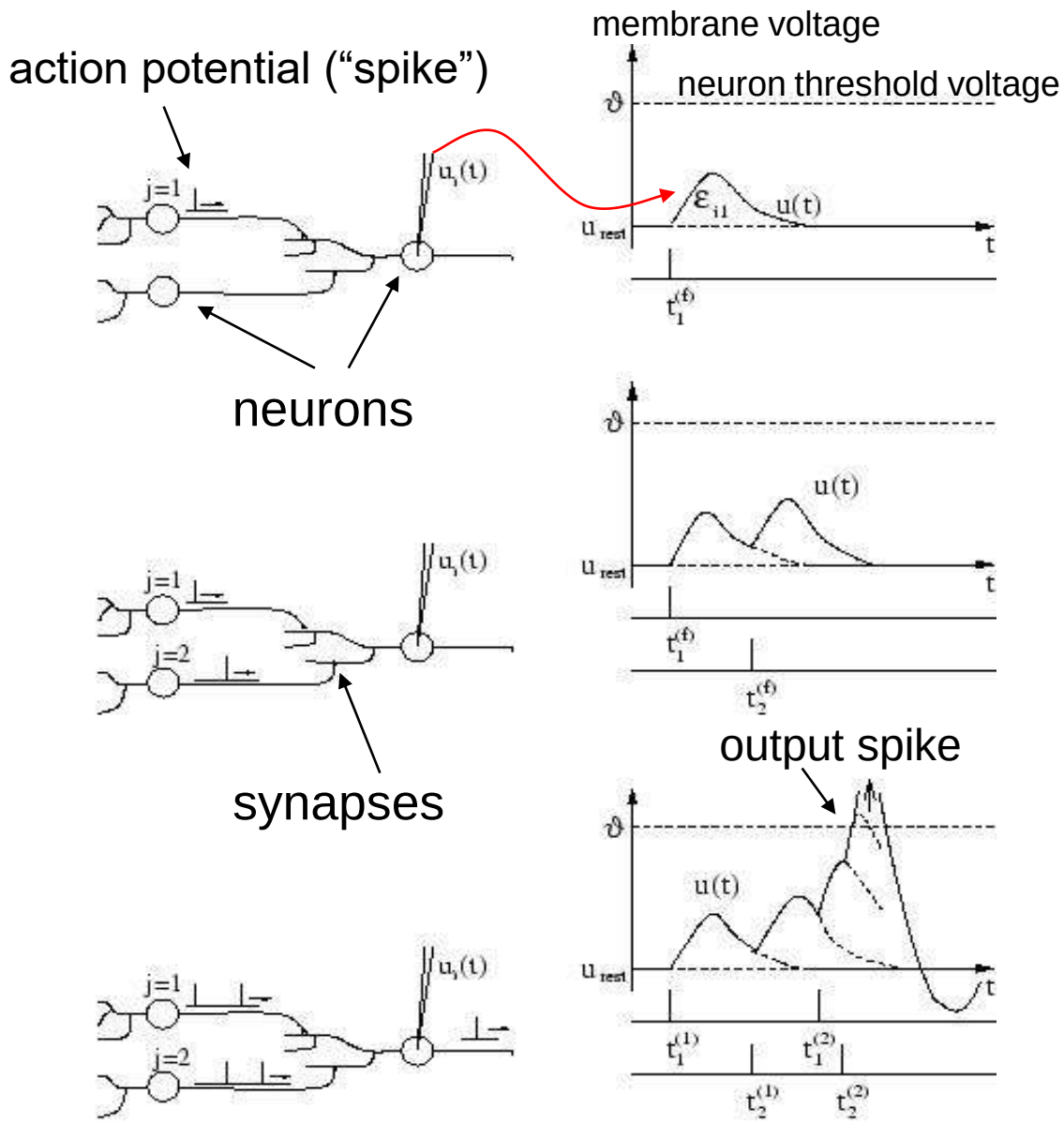
Spike-based model (current biology)



- time-continuous dynamical system
- vector-matrix multiplication
- complex non-linearities
- binary neuron output
- allows to model biological learning mechanisms



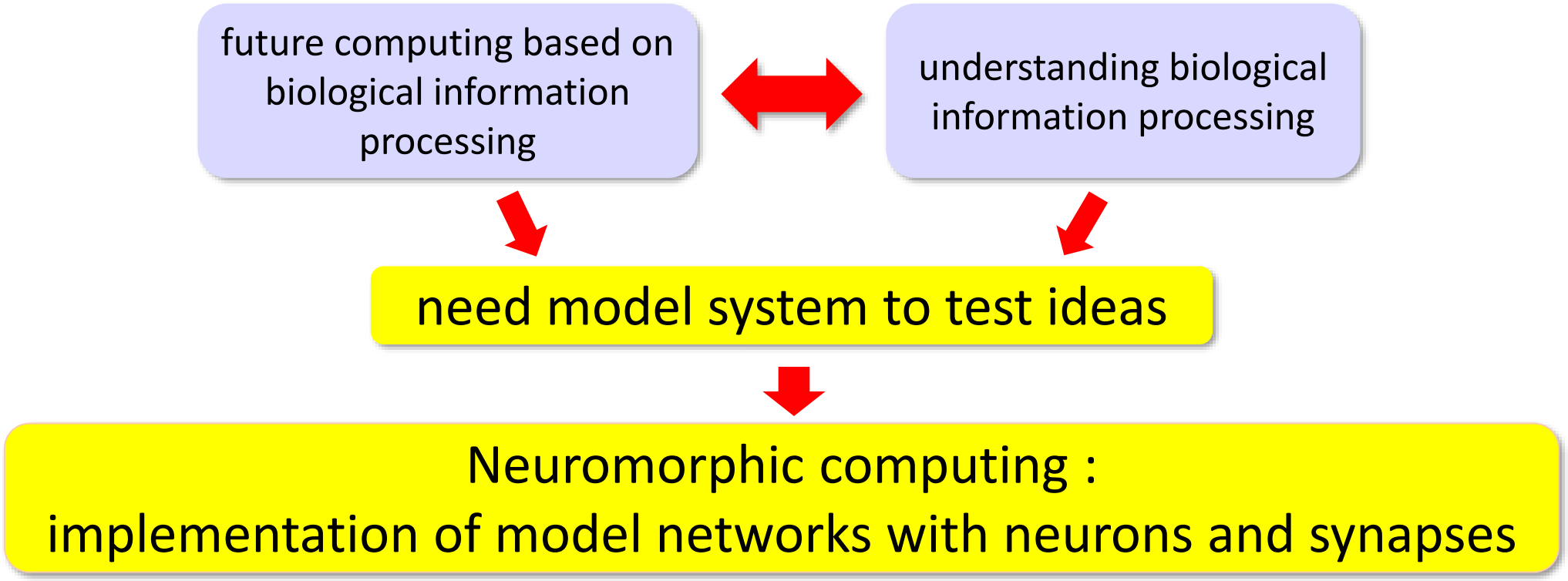
Principles of spike-based neural communication



- neurons integrate over space and time
- temporal correlation is important
- mixed-signal system: action potential (digital) ↔ membrane voltage (analog)
- fault tolerant
- low power consumption → 100 Billion neurons: 20 Watts

Brain-Inspired Computing

Bio-inspired artificial intelligence (Bio-AI)



modeling possibilities:

numerical model : digital simulation

represents model parameters as binary numbers :

→**integer, float, bfloat16**

physical model : analog Neuromorphic Hardware

represents model parameters as physical quantities :

→ **voltage, current, charge**

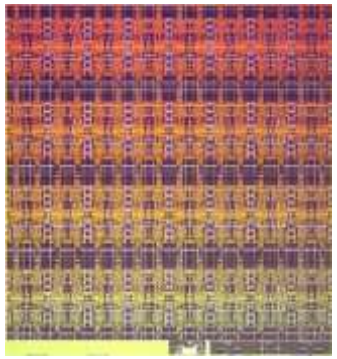
Spike-based Neuromorphic systems worldwide – State-of-the-art and complementarity

SpiNNaker
Biologically Inspired Massively Parallel Architectures

IBM TrueNorth

intel Loihi

BrainScaleS
ScaleS



Biological realism

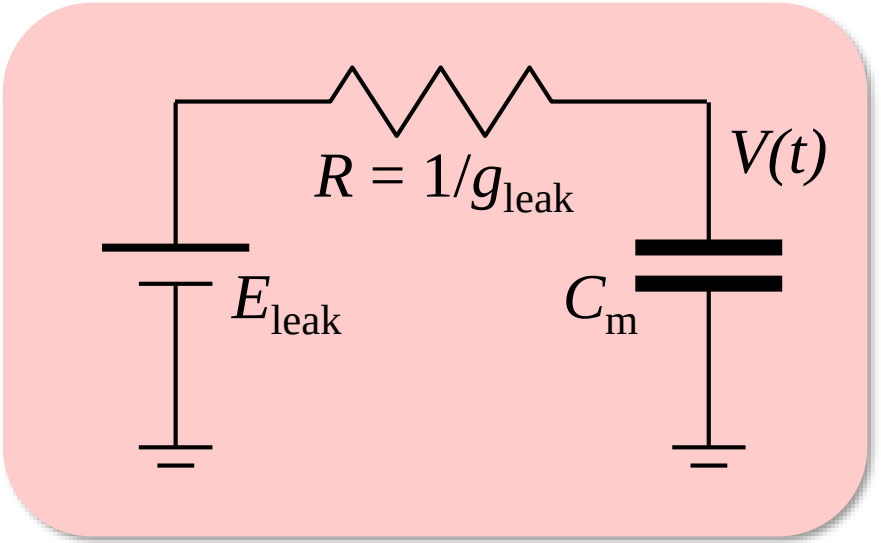
Ease of use

Many-core (ARM) architecture Optimized spike communication network Programmable local learning x0.01 real-time to x10 real-time	Full-custom-digital neural circuits No local learning (TrueNorth) Programmable local learning (Loihi) Exploit economy of scale x0.01 real-time to x100 real-time	Analog neural cores Digital spike communication Biological local learning Programmable local learning x10.000 to x1000 real-time

BrainScaleS : Neuromorphic computing with physical model systems

Consider a simple physical model for the neuron's cell membrane potential *V:*

$$C_{\mathrm{m}} \frac{dV}{dt} = g_{\mathrm{leak}}\left(E_{\mathrm{leak}} - V\right)$$



$$\frac{dV}{dt}_{bio} << \frac{dV}{dt}_{VLSI}$$

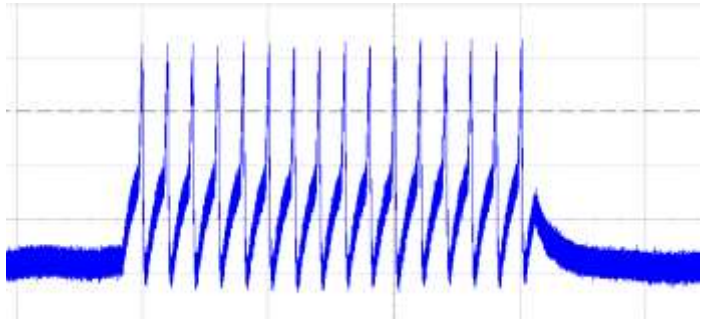
→ accelerated neuron model

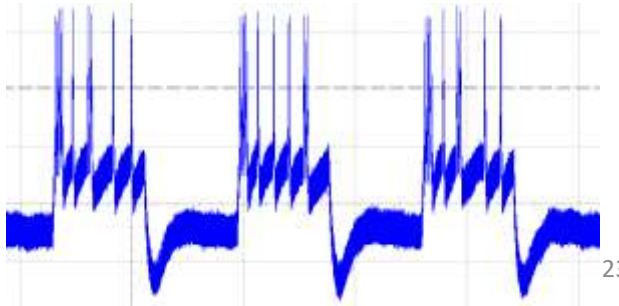
continuous time

- fixed acceleration factor (we use 10^3 to 10^5)

no multiplexing of components storing model variables

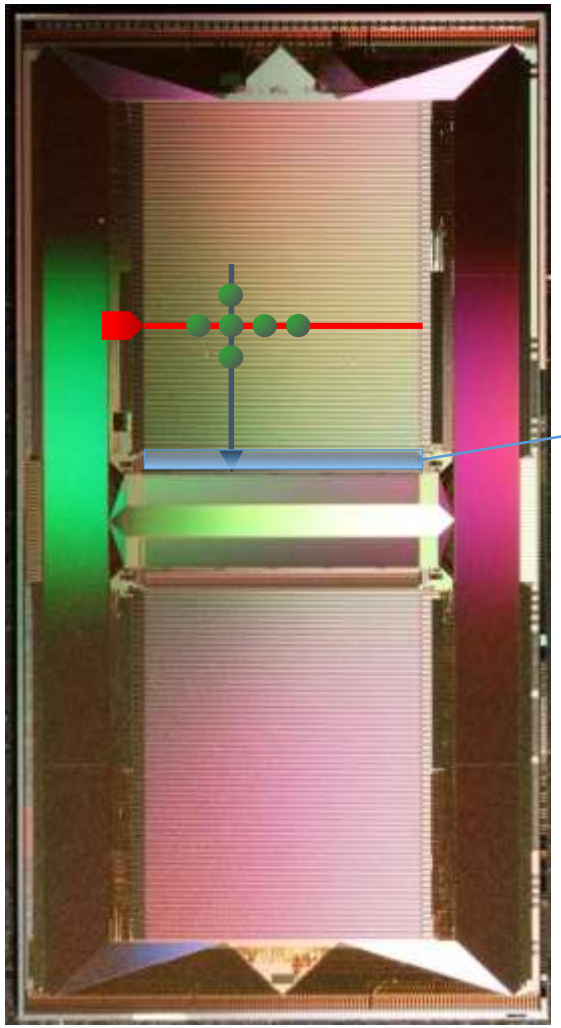
- each neuron has its membrane capacitor
- each synapse has a physical realization

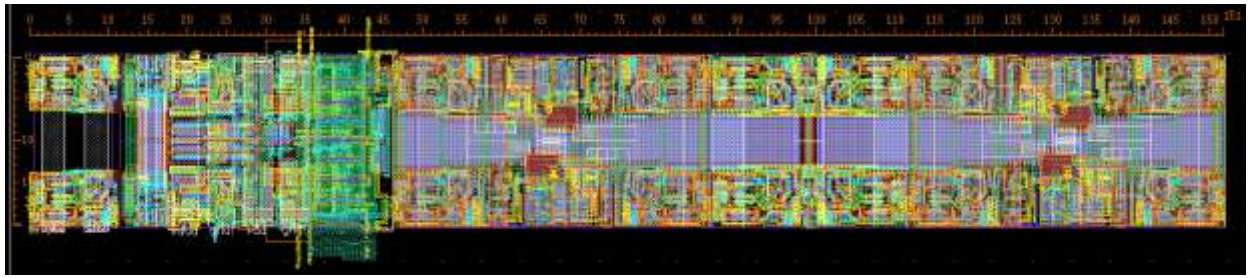


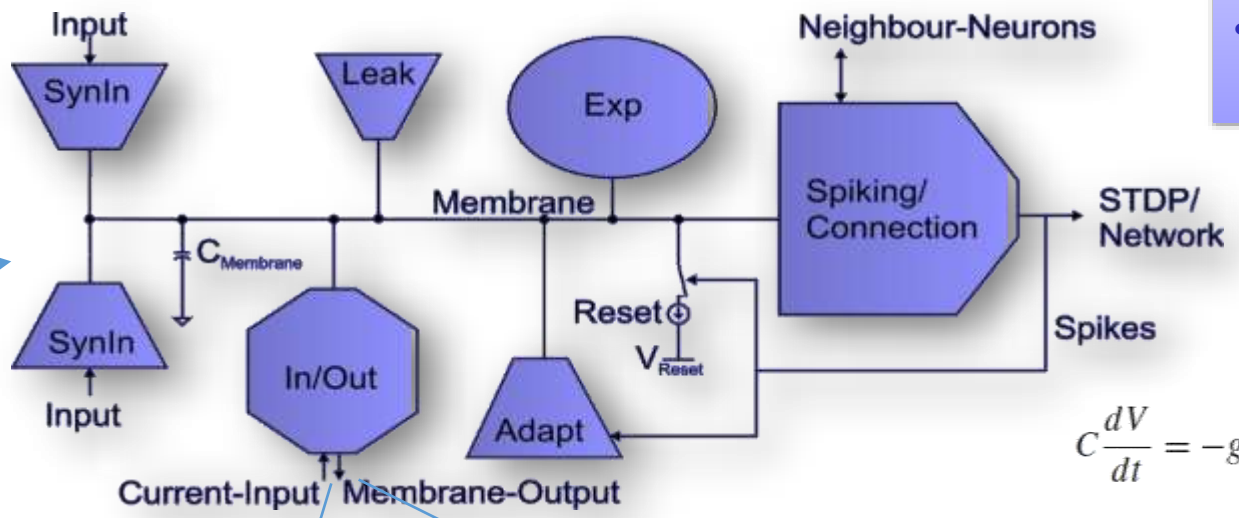


Structure of BrainScaleS neurons: array of parameterized dendrite circuits

photograph of the BrainScaleS 1 neuromorphic chip



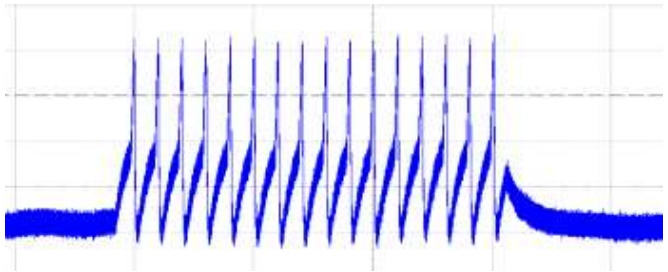


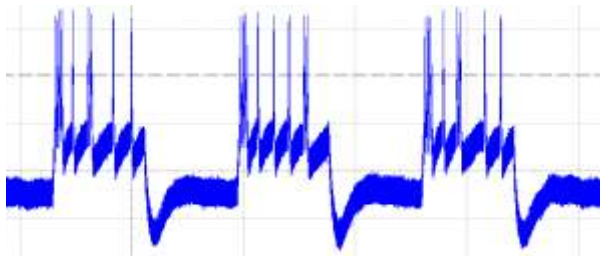


- 180 nm (generation 1) or 65 nm (gen. 2)
- 24 calibration parameters per neuron
- modular structure
- full set of ion-channel circuits for each dendrite

$$C\frac{dV}{dt} = -g_L(V - E_L) + g_L\Delta_T \exp\left(\frac{V - V_T}{\Delta_T}\right) + I - w, \quad (1)$$

$$\tau_w\frac{dw}{dt} = a(V - E_L) - w. \quad (2)$$





Time*Scales*

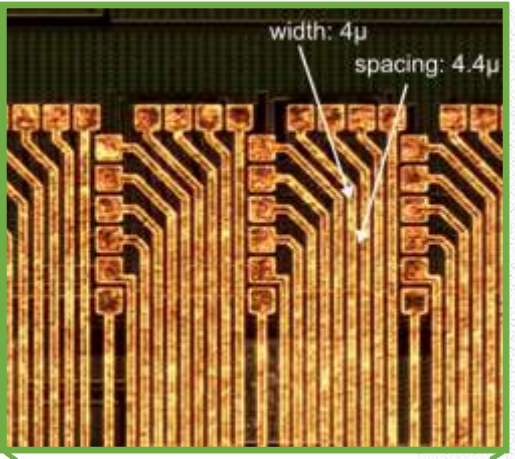
	Nature + Real-time	Simulation	Accelerated Model
Causality Detection	10^{-4} s	0.1 s	10^{-8} s
Synaptic Plasticity	1 s	1000 s	10^{-4} s
Learning	Day	1000 Days	10 s
Development	Year	1000 Years	3000 s

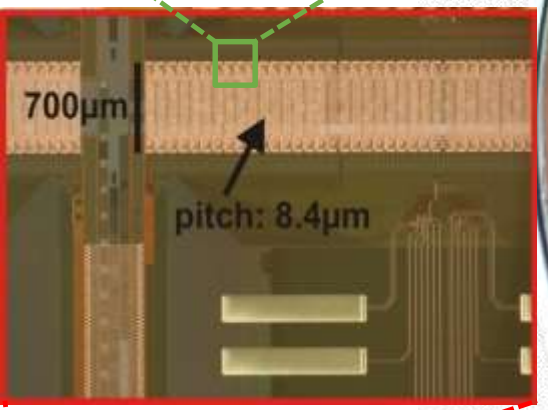
12 Orders of Magnitude

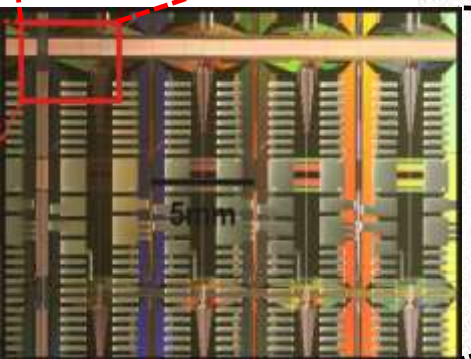
Evolution	> Millenia	> 1000 Millenia	> Months

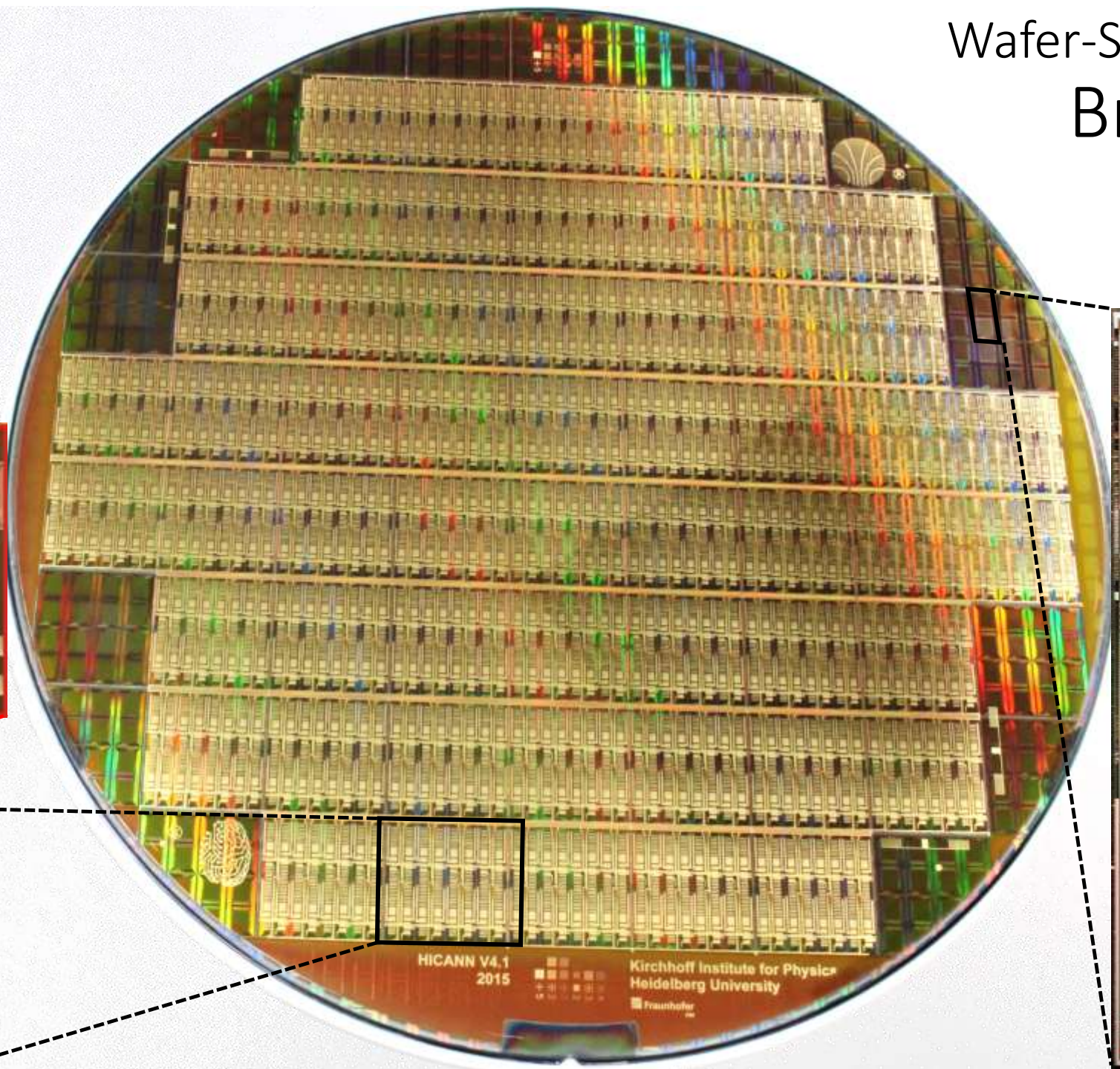
> 15 Orders of Magnitude

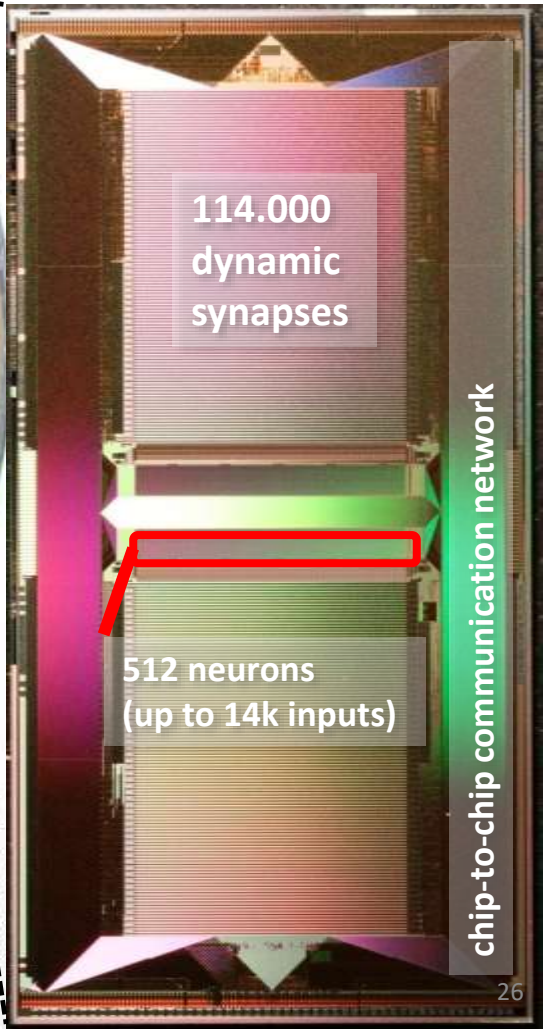
Wafer-Scale Integration : BrainScaleS-1



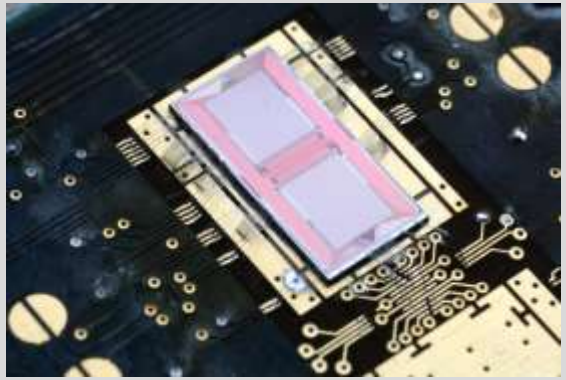








BrainScaleS-1 multi-level architecture



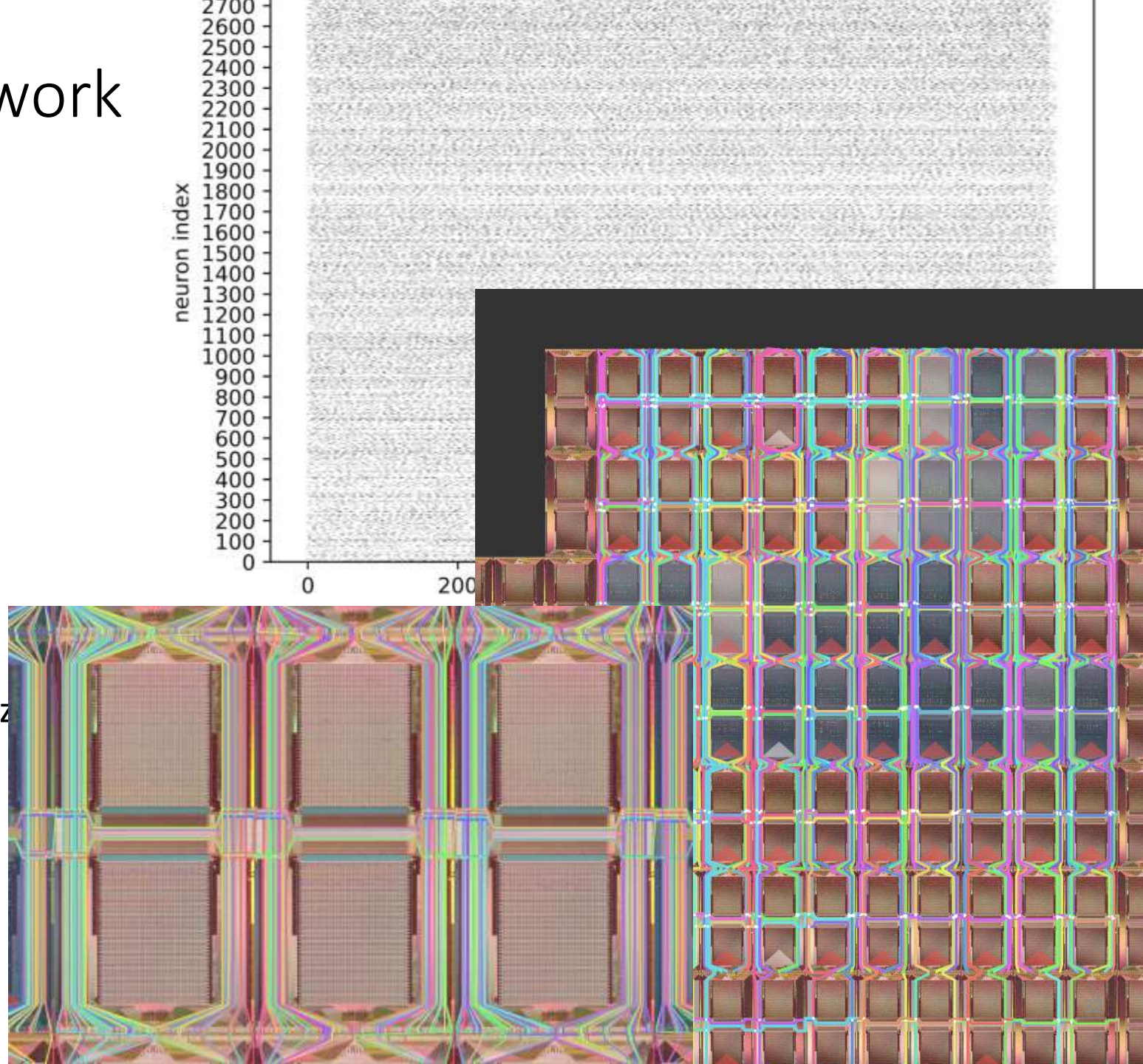
single chip wafer module hybrid system

BrainScales-1 introduced for the first time

- Accelerated (x10.000) mixed-signal implementation of spiking neural networks
- AdEx neurons with very high synaptic imput count (> 10k)
- Wafer-scale event communication

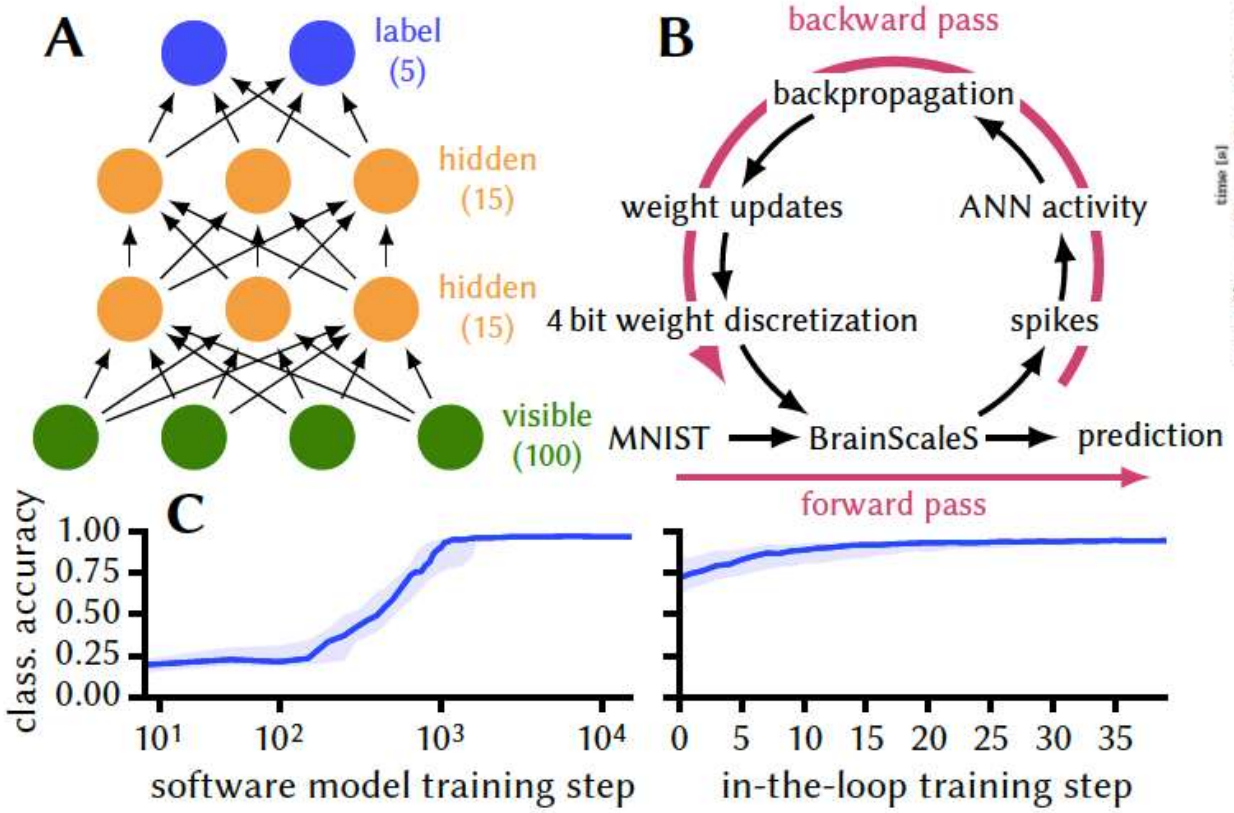
(Balanced) Random Network

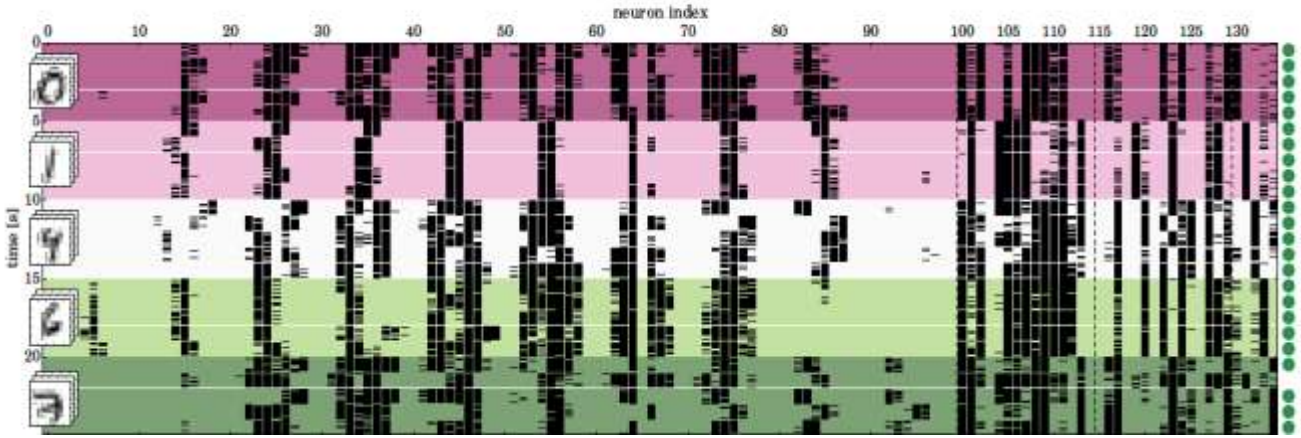
- “Dynamics of Sparsely Connected Networks of Excitatory and Inhibitory Spiking Neurons” (Brunel 2000)
- 3000 neurons (> 1 Gevent/s)
- ~700k synapses (> 0.1 Tconn/s)
- 138 HICANN chips
- 800 individual external poisson sources with 50 Hz each -> 40 kHz (bio) (400 MHz wall clock rate)

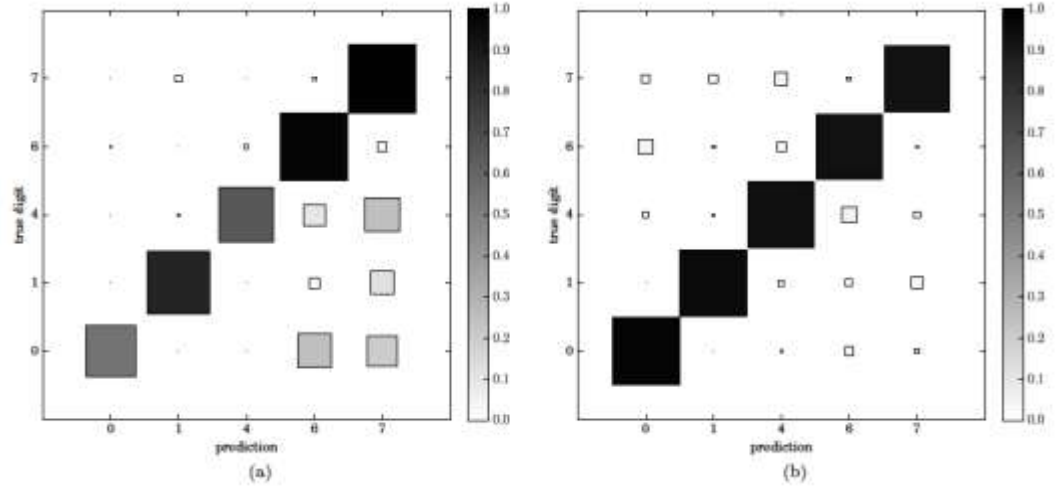


Classification with feed-forward, rate-based network on BrainScaleS-1
Classical machine learning with a physical analog system

Feed-forward, rate-based, spiking network, MNIST classification, hardware in-the-loop learning







Neuromorphic Hardware In The Loop: Training a Deep Spiking Network on the BrainScaleS Wafer-Scale System,
Schmitt, Klaehn, et al., IJCNN, 2017, https://arxiv.org/abs/1703.01909

BrainScaleS-1 :
Observations leading to second-generation BrainScaleS system

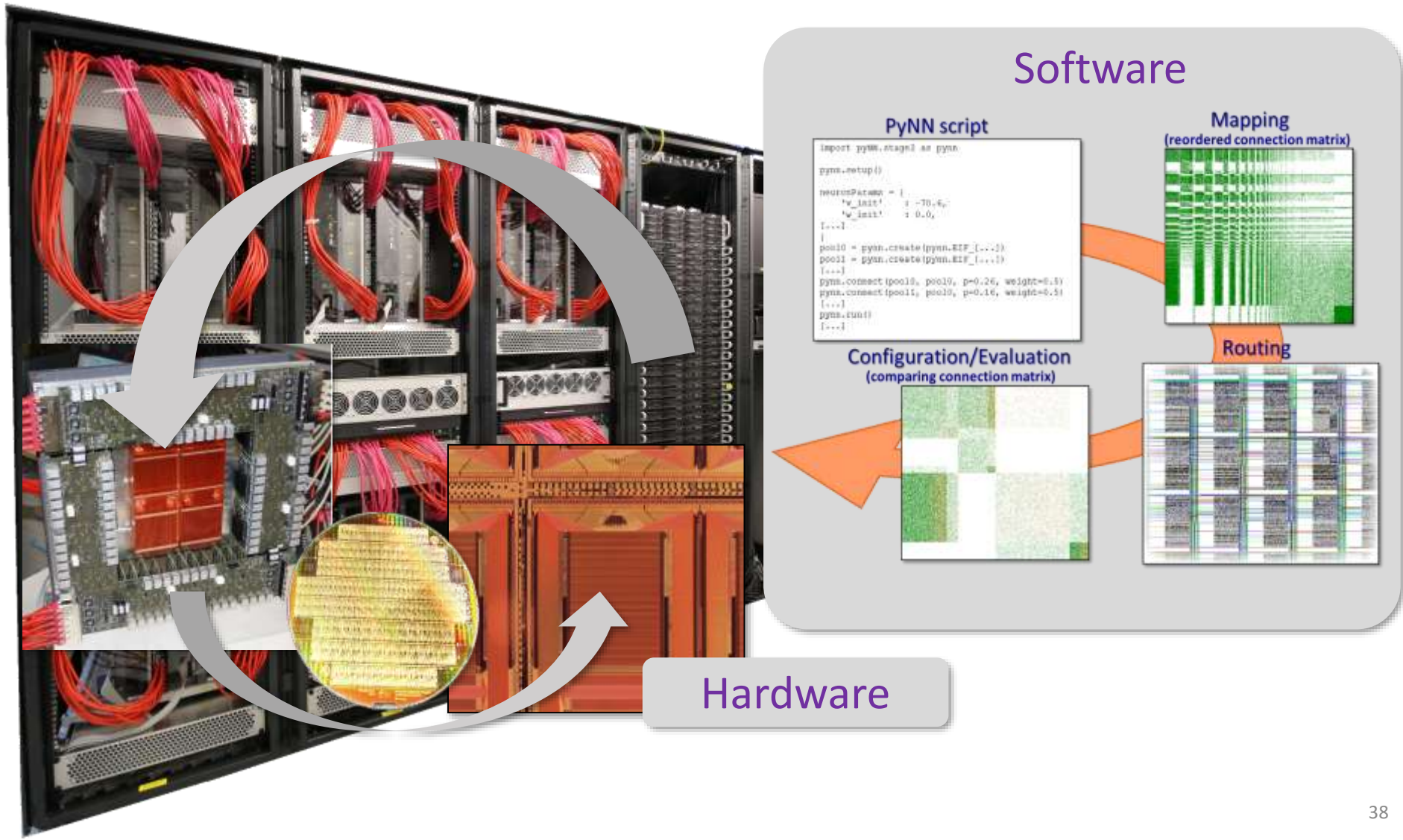
after training:

Non-Turing physical computing system performing autonomously

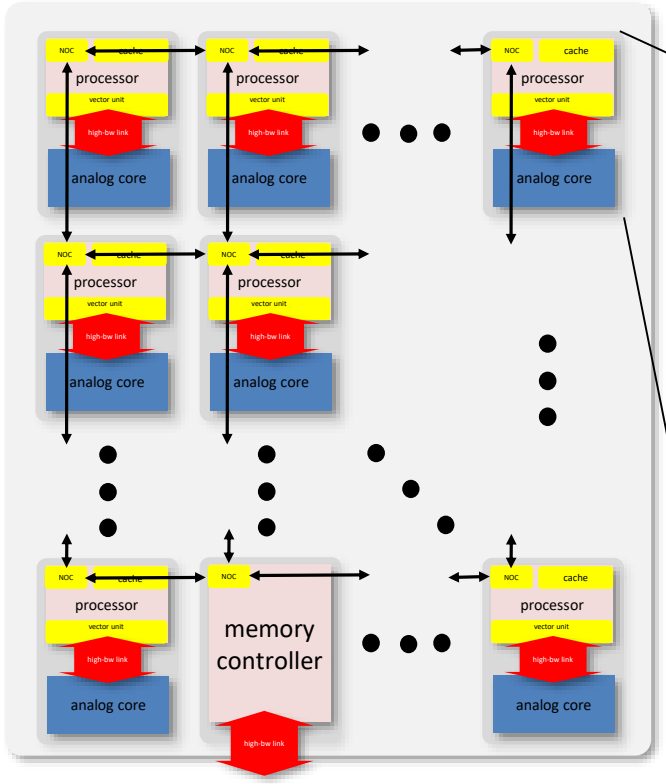
but

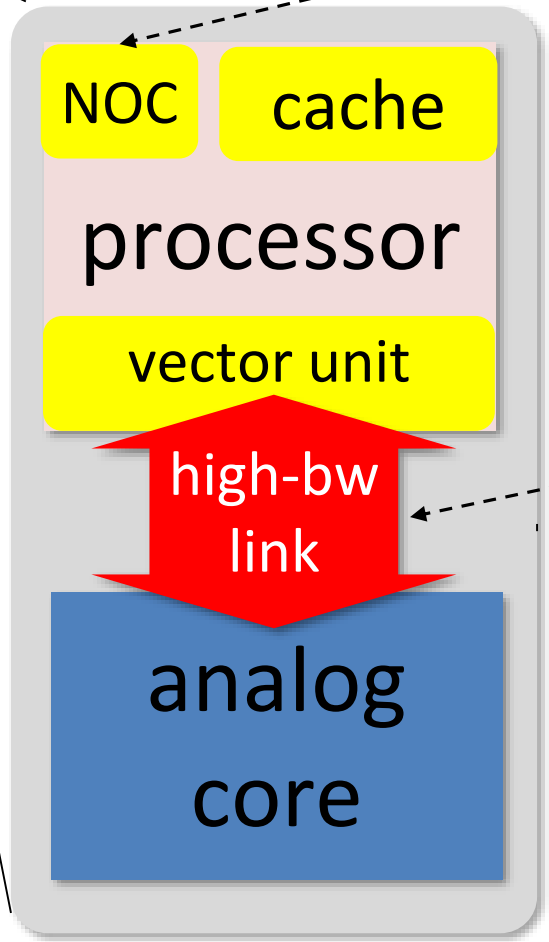
Turing-based computing is used in multiple places:

- training
- system initialization
- hardware calibration
- runtime control
- input/output data handling



Shortening the hardware – software loop : Analog neuromorphic system as coprocessor





Network-on-chip:

- prioritize event data
- unused bw for CPU
- common address space for neurons and CPUs

high-bandwidth link:

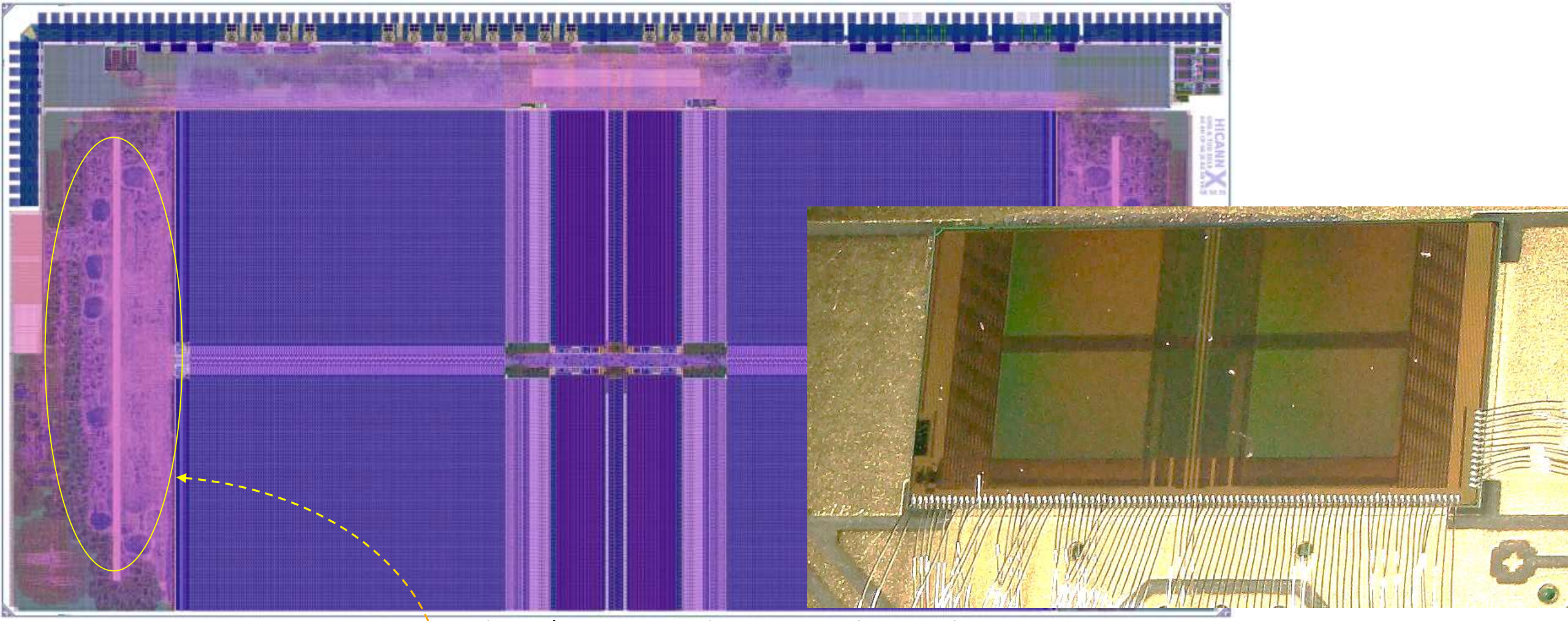
vector unit ←→ NM core

- weights
- correlation data
- routing topology
- event (spikes) IO
- configuration

special function tile:

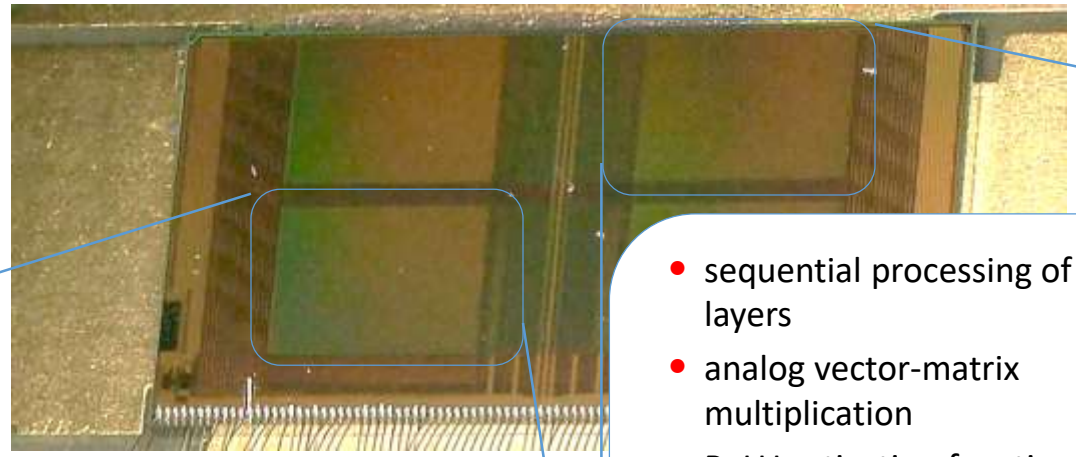
- memory controller
- SERDES IO
- purely digital function unit

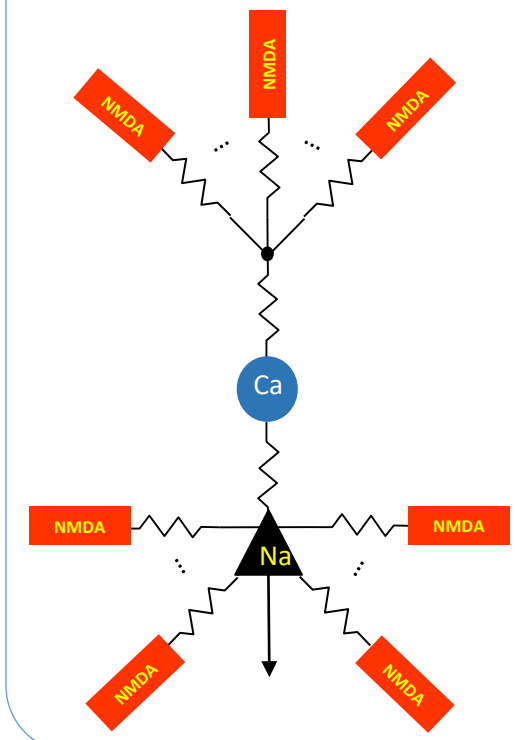
BrainScaleS-2 (BSS-2) ASIC

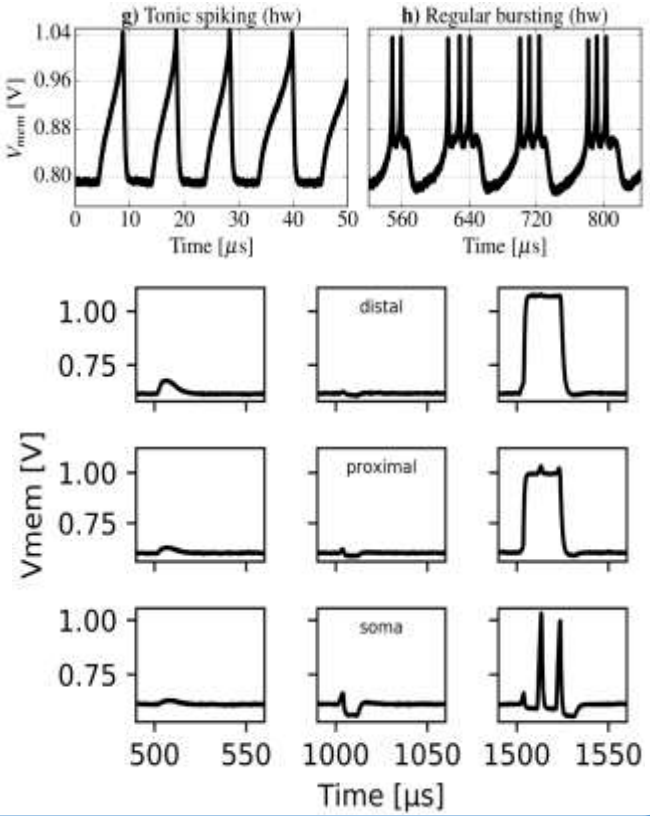


- 65nm LP-CMOS, power consumption O(10 pJ/synaptic event)
- 128k synapses
- 512 neural compartments (Sodium, Calcium and NMDA spikes)
- two SIMD plasticity processing units (PPU)
- PPU internal memory can be extended externally
- fast ADC for membrane voltage monitoring
- 256k correlation sensors with analog storage (> 10 Tcorr/s max)
- 1024 ADC channels for plasticity input variables
- 32 Gb/s neural event IO
- 32 Gb/s local entropy for stochastic neuron operation

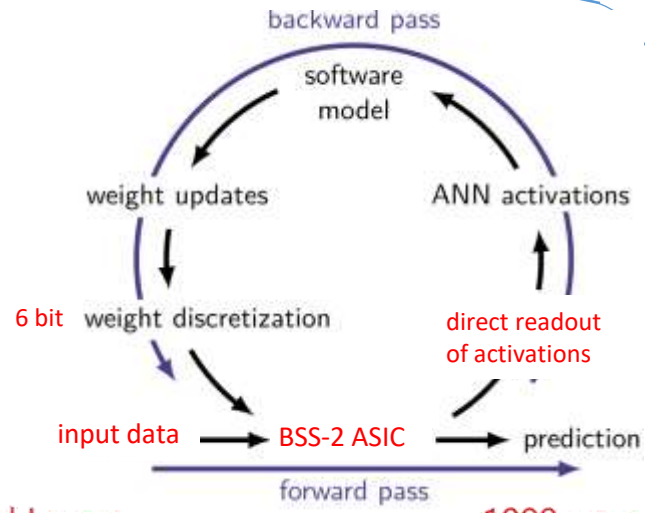
BrainScaleS-2 supports spike-based and Perceptron operation simultaneously

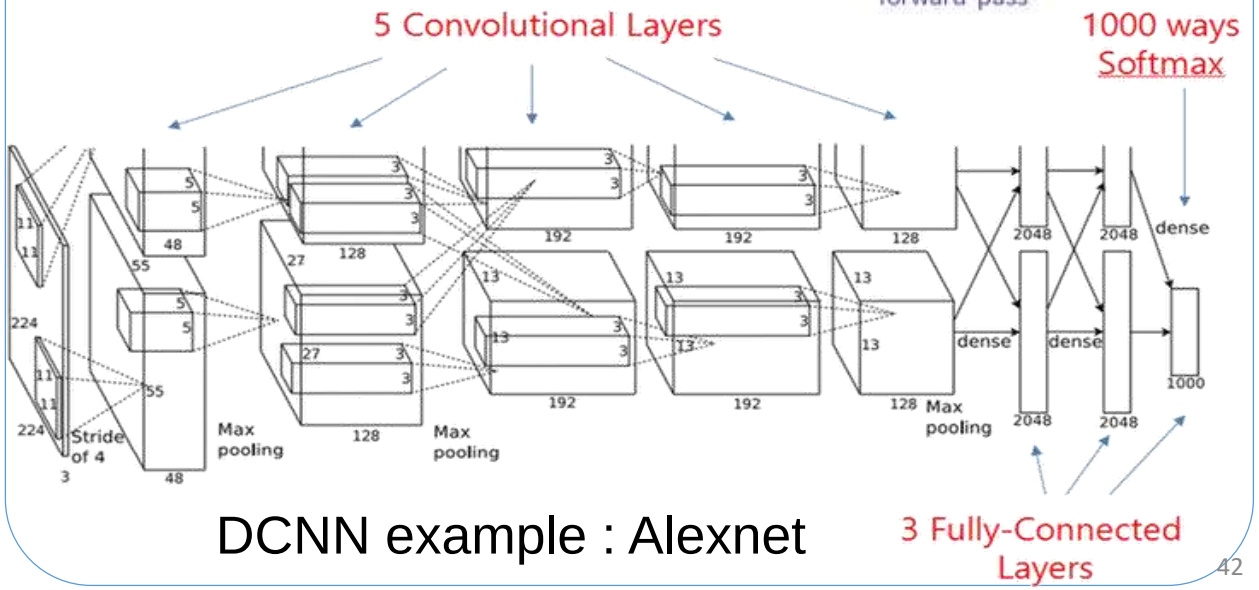






- sequential processing of all layers
- analog vector-matrix multiplication
- ReLU activation function with 4 to 8 bit resolution
- speed mostly limited by external memory





DCNN example : Alexnet

3 Fully-Connected Layers

Learning and plasticity

- ✓ biological relevant neuron model
  → Adaptive Exponential Integrate and Fire (AdExp)
- ✓ biological relevant network topologies
  → more than 10k synapses per neuron
- ✓ high communication bandwidth for scalability
  → wafer-scale integration

Problem:
how to fix millions of parameters

- network topology
- neuron sizes and parameters
- synaptic strengths

Trivial solution: everything is pre-computed on the host-computer

- requires precise calibration of hardware
- takes long time (much longer than running the experiment on the accelerated system)

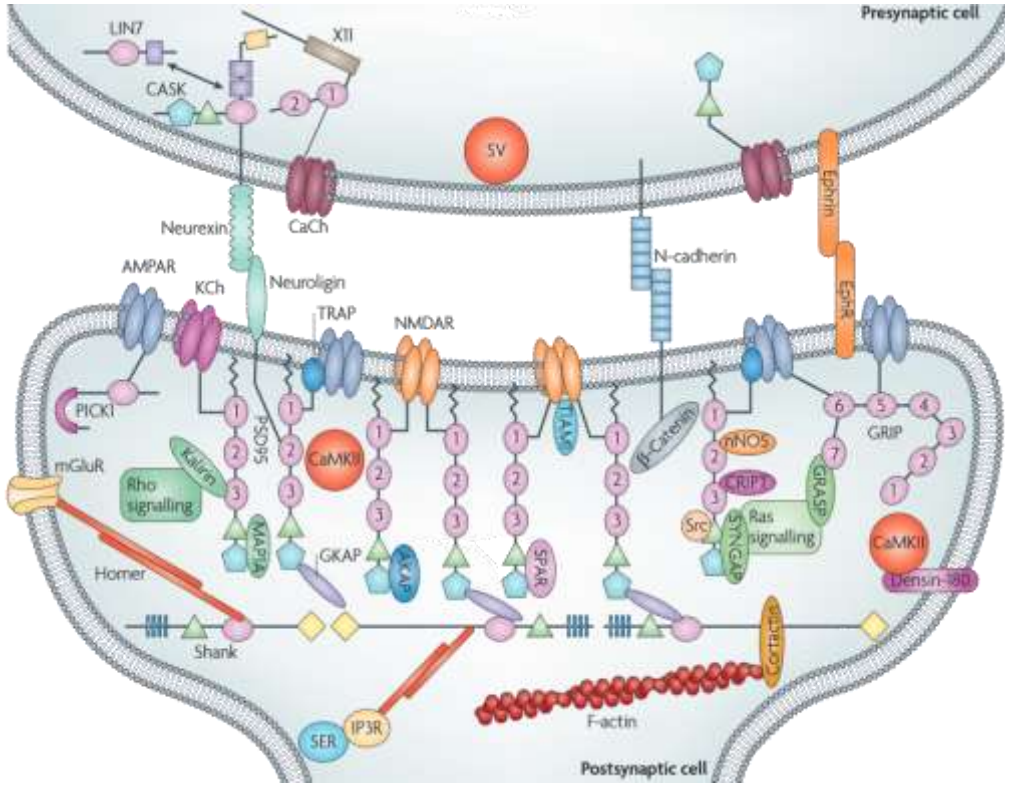
Better approach: hardware in-the-loop training

- makes use of high emulation speed

Biological solution : Integrate some kind of learning or plasticity mechanism

- local feed-back loops, aka *training*, adjust system parameters
- no calibration of synapses necessary → learning replaces calibration
- plastic network topology

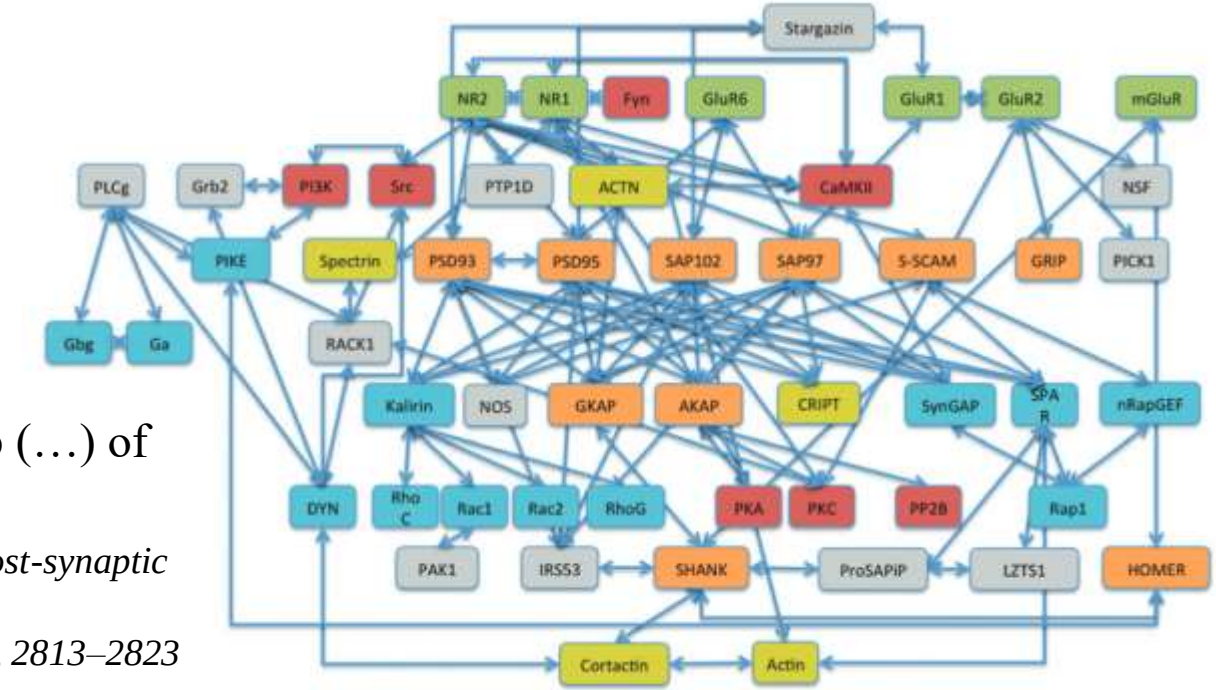
Complexity of synaptic plasticity is key to biological intelligence



Protein complex organization in the postsynaptic density (PSD)

"Organization and dynamics of PDZ-domain-related supramodules in the postsynaptic density"
W. Feng and M. Zhang, Nature Reviews NS, 10/2009

- > 6000 genes primarily active in the brain
- high percentage of regulatory RNA
- evidence for epigenetic effects in plasticity



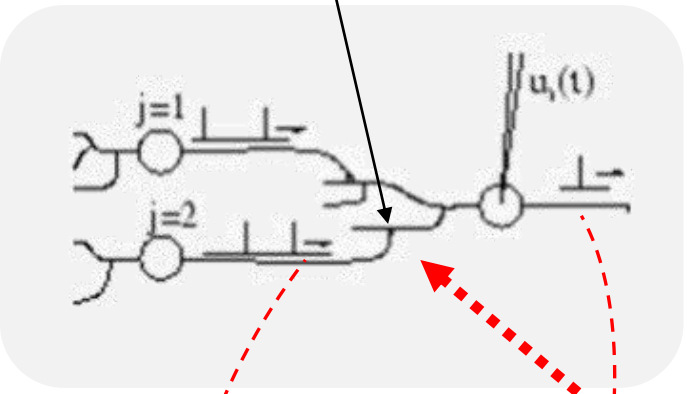
Protein-protein interaction map (…) of post-synaptic density

"*Towards a quantitative model of the post-synaptic proteome"*
O Sorokina et.al., Mol. BioSyst., 2011,7, 2813–2823

BrainScaleS-2: Hybrid Plasticity

- analog correlation measurement in synapses
- A/D conversion by parallel ADC
- digital Plasticity Processing Units can access
  - synaptic weights (ω)
  - configuration data (adr) → structural plasticity
  - neuron voltages and firing rates

plasticity takes place at the synapse

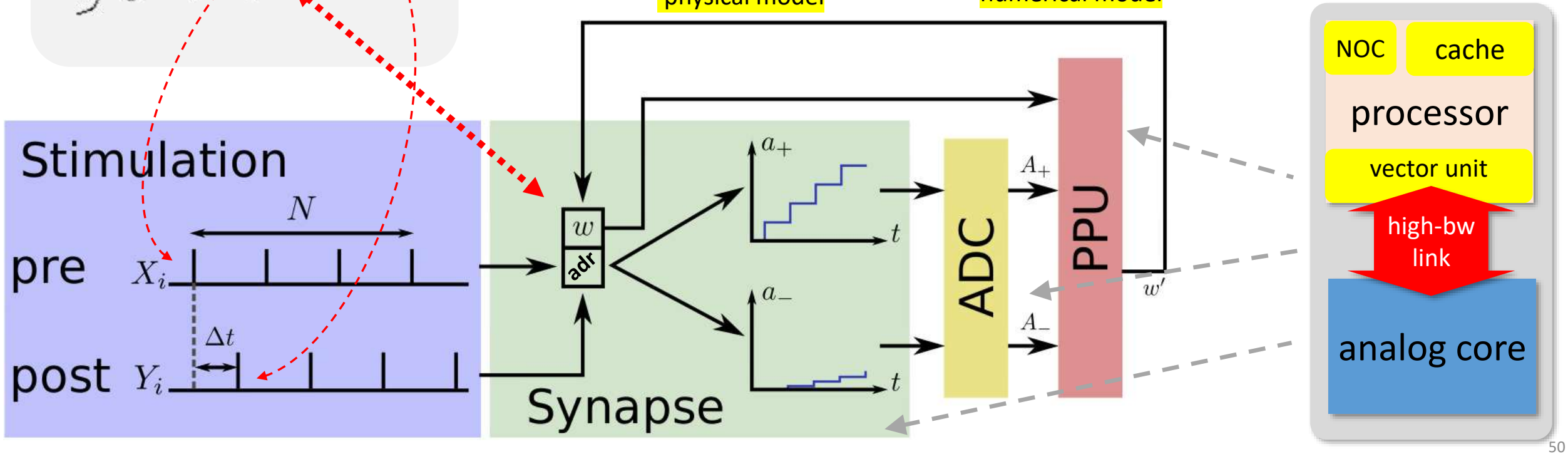


analog

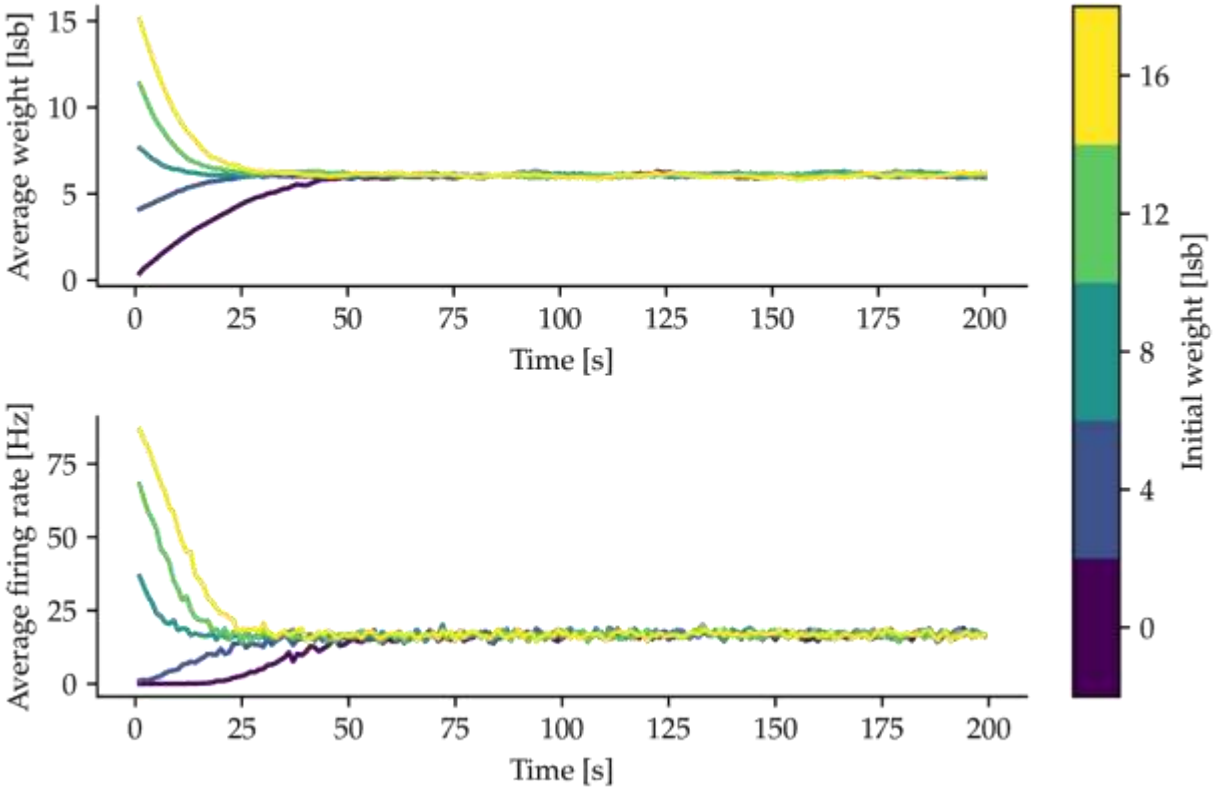
physical model

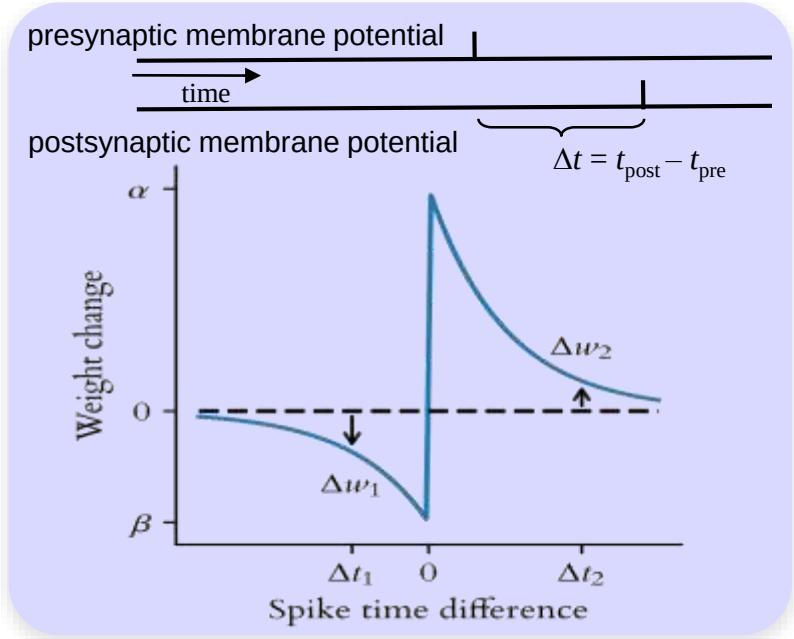
digital

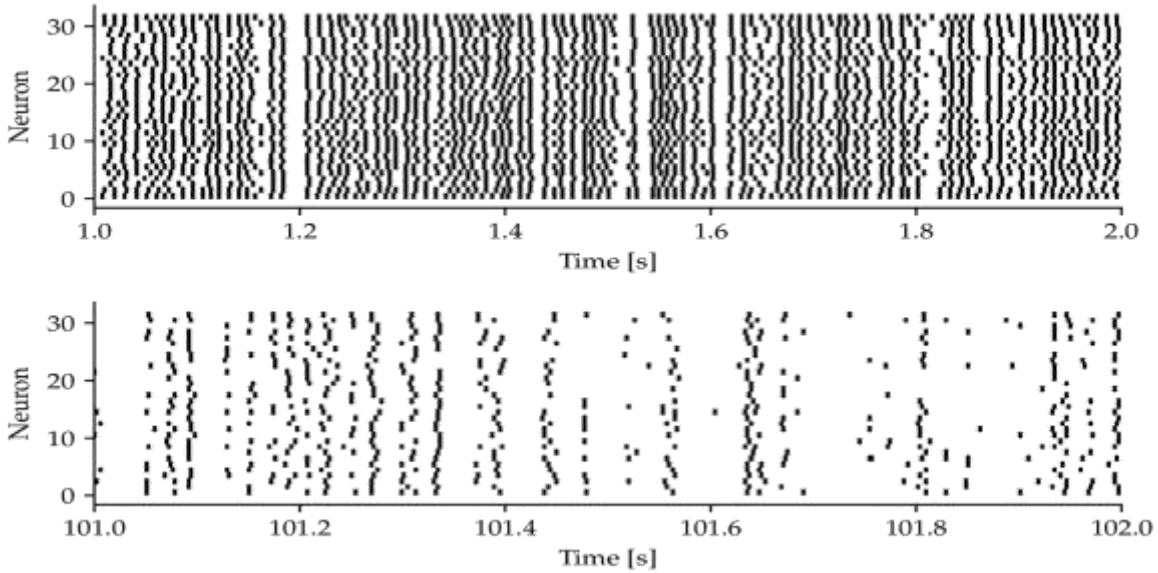
numerical model



Stabilizing firing rates with spike time dependent plasticity





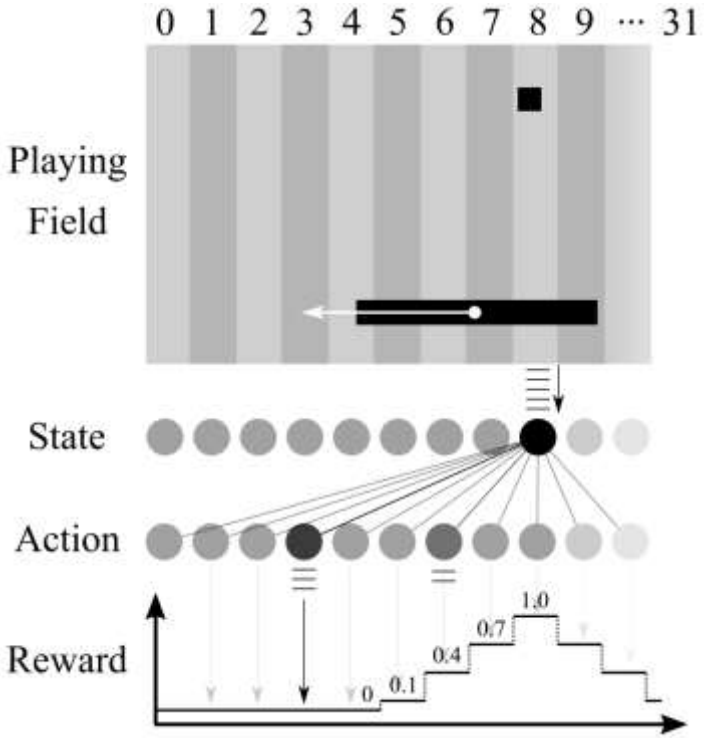


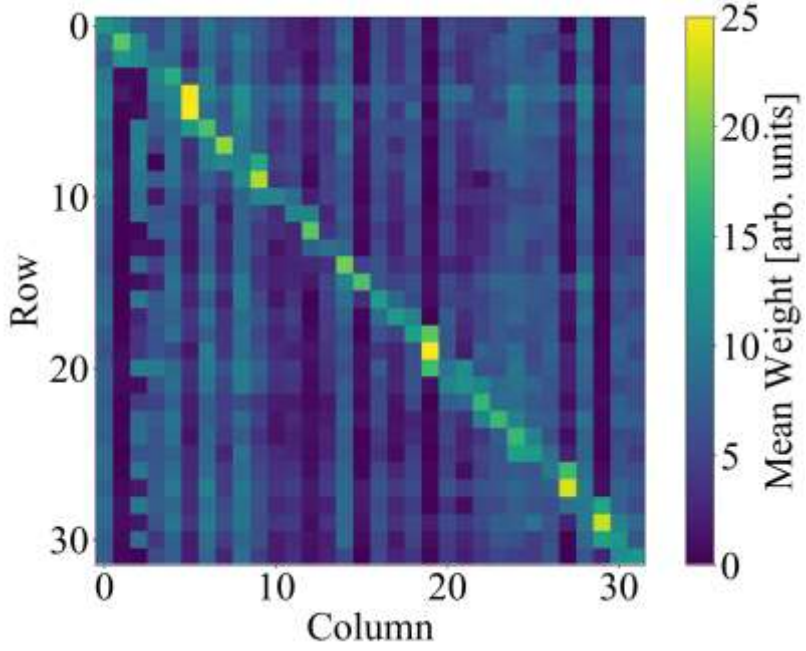
Wall-time per trace: 200ms

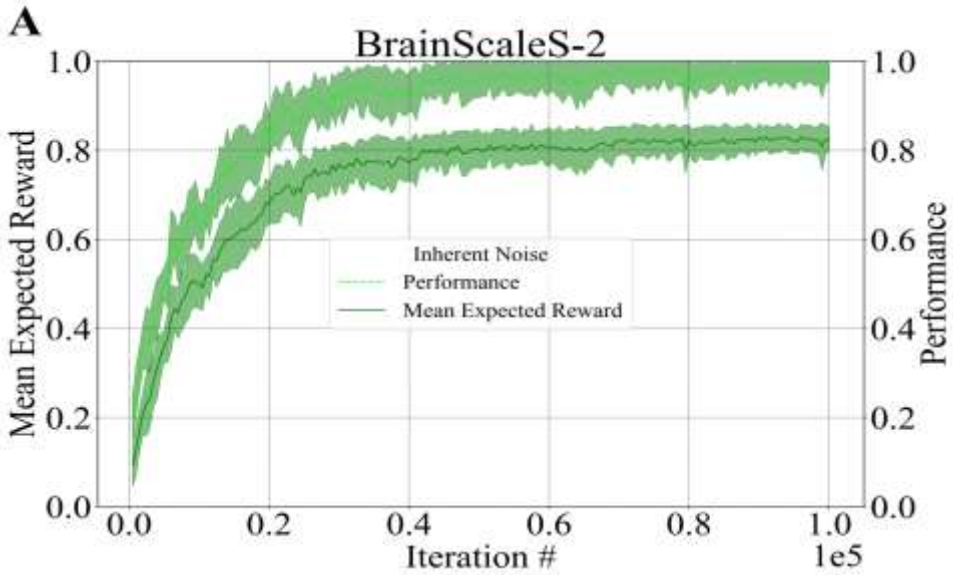
→ acceleration factor of 1000

David Stöckel, Master Thesis, Heidelberg University, 2017

Learning Pong – tech demo using internal PPU only





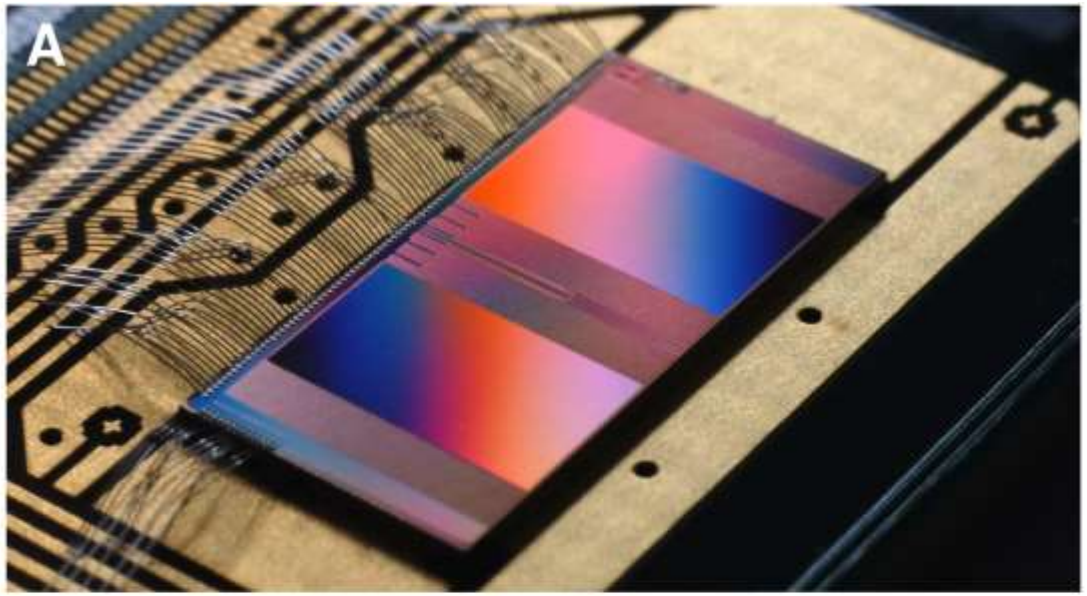


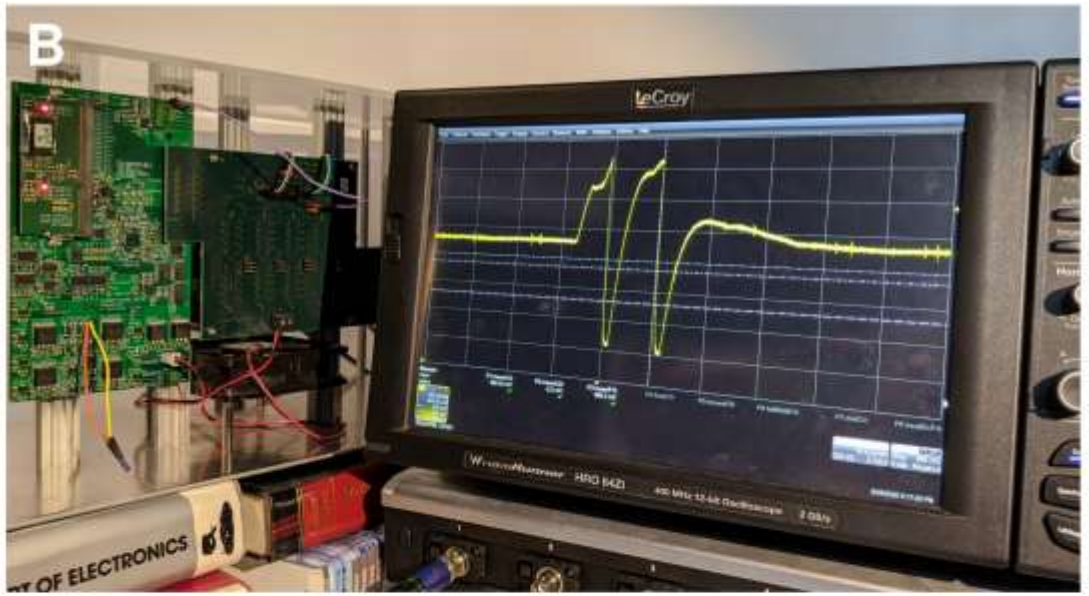
- reinforcement learning rule
- learning is calibration
- experiment runs completely on internal PPU
- 5s for 10k iterations

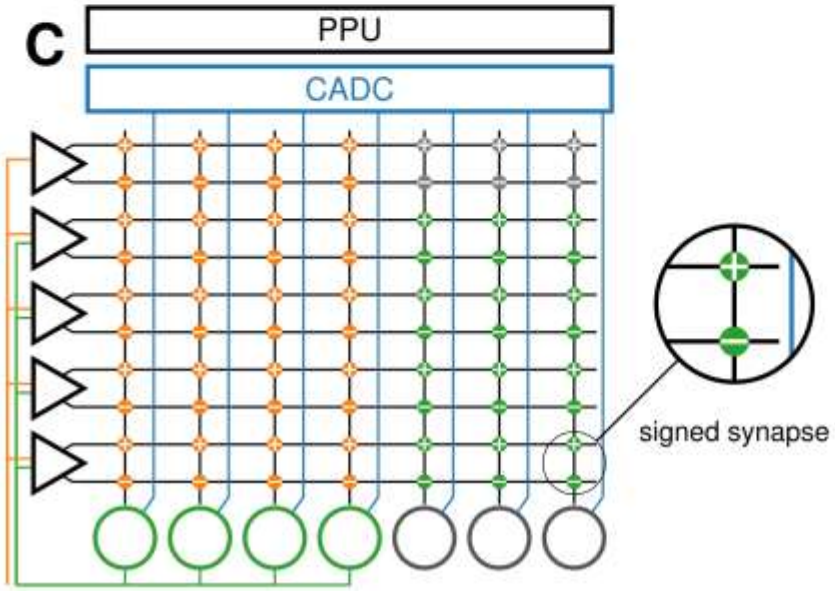
  network time 0.4ms/iteration
  23 µJ total chip energy

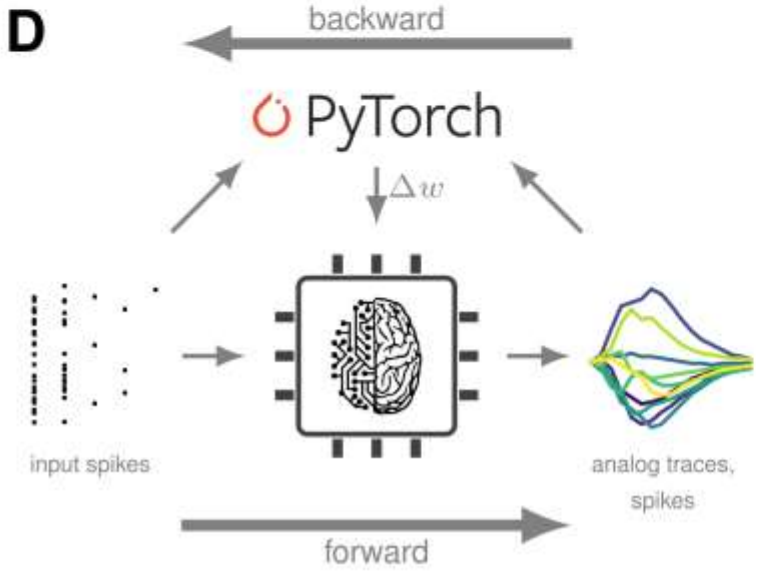
Wunderlich et.al., Demonstrating Advantages ..., Front. Neurosci., 2019

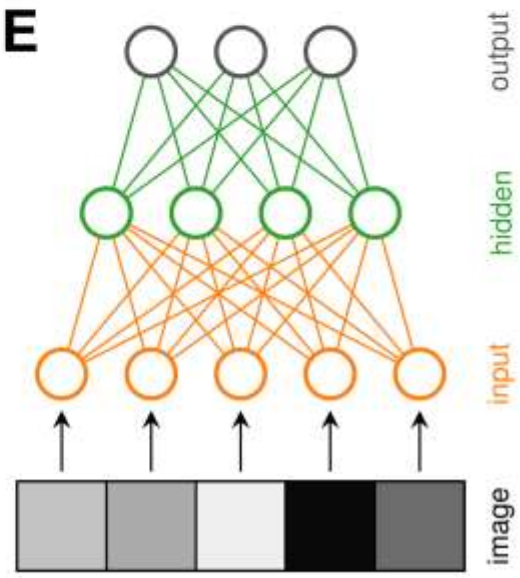
Training multi-layer networks with Surrogate Gradients





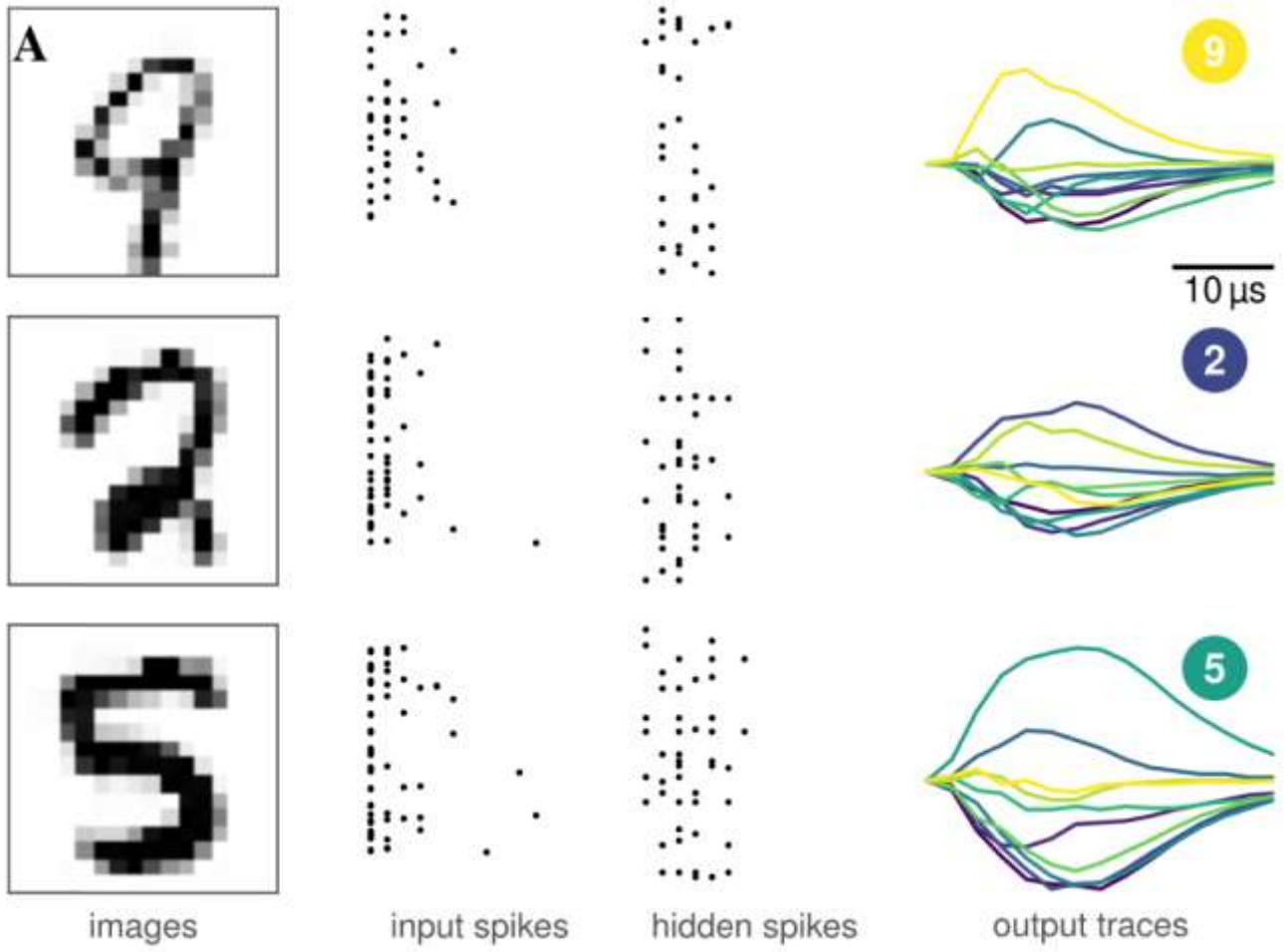






Sebastian Billaudelle, Benjamin Cramer, et.al., publication in preparation, 2020

Spike-latency coding as basis for fast inference



- Classification accuracy on test data: 96,7%
- Possible classification rate: 70k images/s
- Energy per image: 4 µJ
- Energy consumption of ASIC during inference (everything active): 285 mW
- Higher-speed possible
  - interleaving of networks
  - spike-based sensor converts fast serial to slow parallel signal, temporal information becomes partially spatial information (like our ears)
  - could pre-process detector data without digitization
    → higher channel density possible

Sebastian Billaudelle, Benjamin Cramer, et.al., publication in preparation, 2020

What I have learned

- Analog computing is feasible
  - model biology for neuroscience
  - cost and energy efficient inference of DCNNs
  - edge computing (sensor data preprocessing)
- Local learning with closely coupled SIMD CPU
  - Software-based implementation of learning algorithms
    - learning can include calibration
    - supports hyper-parameter learning (L2L)
    - still no solution for deep (i.e. multi-layerd) networks
- Hardware-in-the-loop with backpropagation
  - results comparable to digital systems
  - much better resource efficiency (low cost process)
  - very low latency possible
  - real-time processing of fast sensor data (-> high-energy physics)