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

Co-funded by
the European Union

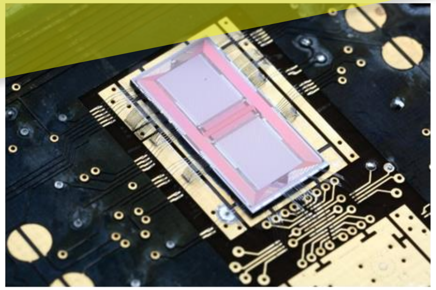
Human Brain Project

Why focus on the brain ? Three Reasons

- **Understanding the brain (Unifying Science Goal)**
  - Underpins what we are,
  - Data & knowledge are fragmented,
  - Integration is needed,
  - Large scale collaborative approach is essential.

- **Understanding brain diseases (Society)**
  - Costs Europe over €800 Billon/year,
  - Affects 1/3 people,
  - Number one cause of loss of economic productivity,
  - No fundamental treatments exist or are in sight
  - Pharma companies pulling out of the challenge.

- **Developing Future Computing (Technology)**
  - Computing underpins modern economies,
  - Traditional computing faces growing hardware, software, & energy barriers,
  - Brain can be the source of energy efficient, robust, self-adapting & compact computing technologies,
  - Knowledge driven process to derive these technologies is missing.



Neuromorphic Computing

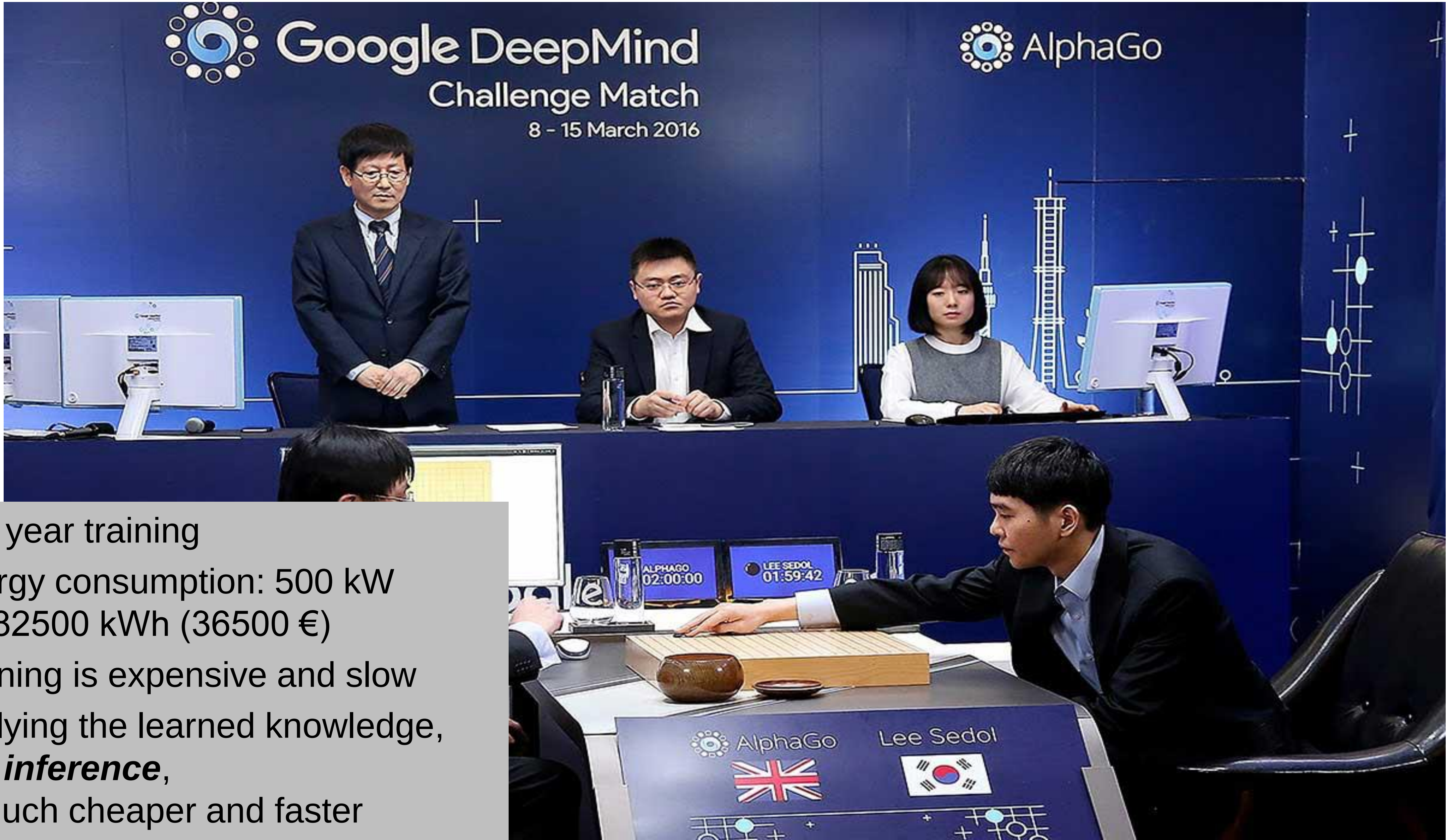
Part of EBRAINS infrastructure

Subproject Leader: Steve Furber

Deputy Leader: Johannes Schemmel

- Neuromorphic Machines
- Algorithms and Architectures for Neuromorphic Computing
  - Theory
  - Applications

Computers are becoming more brain-like



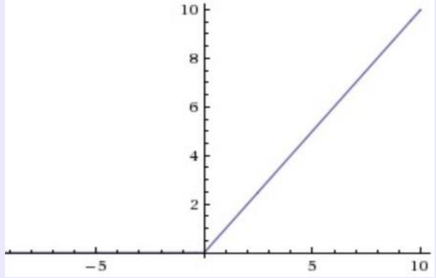
- one year training
- energy consumption: 500 kW
  →182500 kWh (36500 €)
- learning is expensive and slow
- applying the learned knowledge,
  aka ***inference***,
  is much cheaper and faster

Perceptron model (biology of 1950)

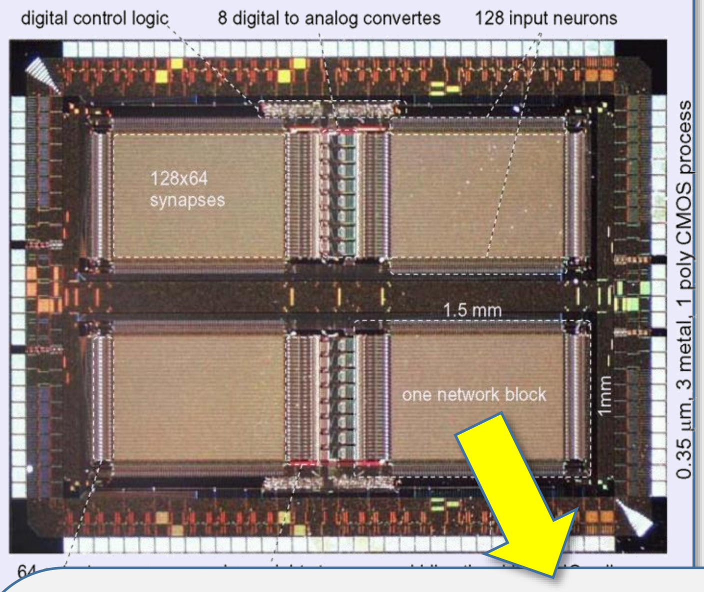
- used in Machine Learning
- vector-matrix multiplication

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

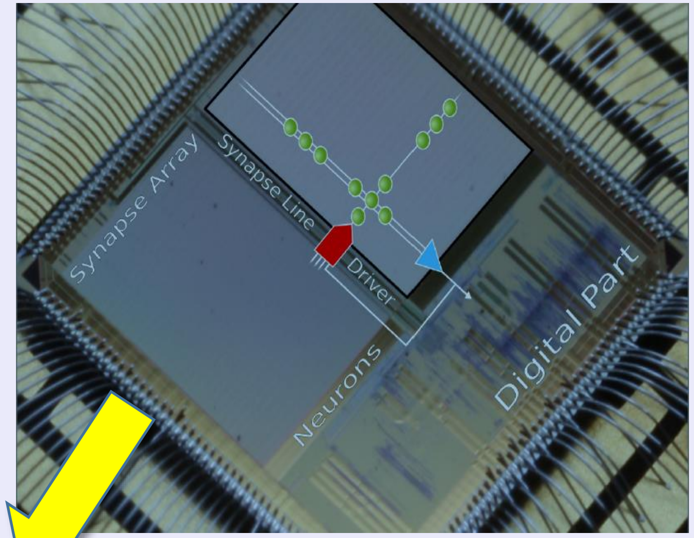
- simple non-linear activation function f (ReLU):



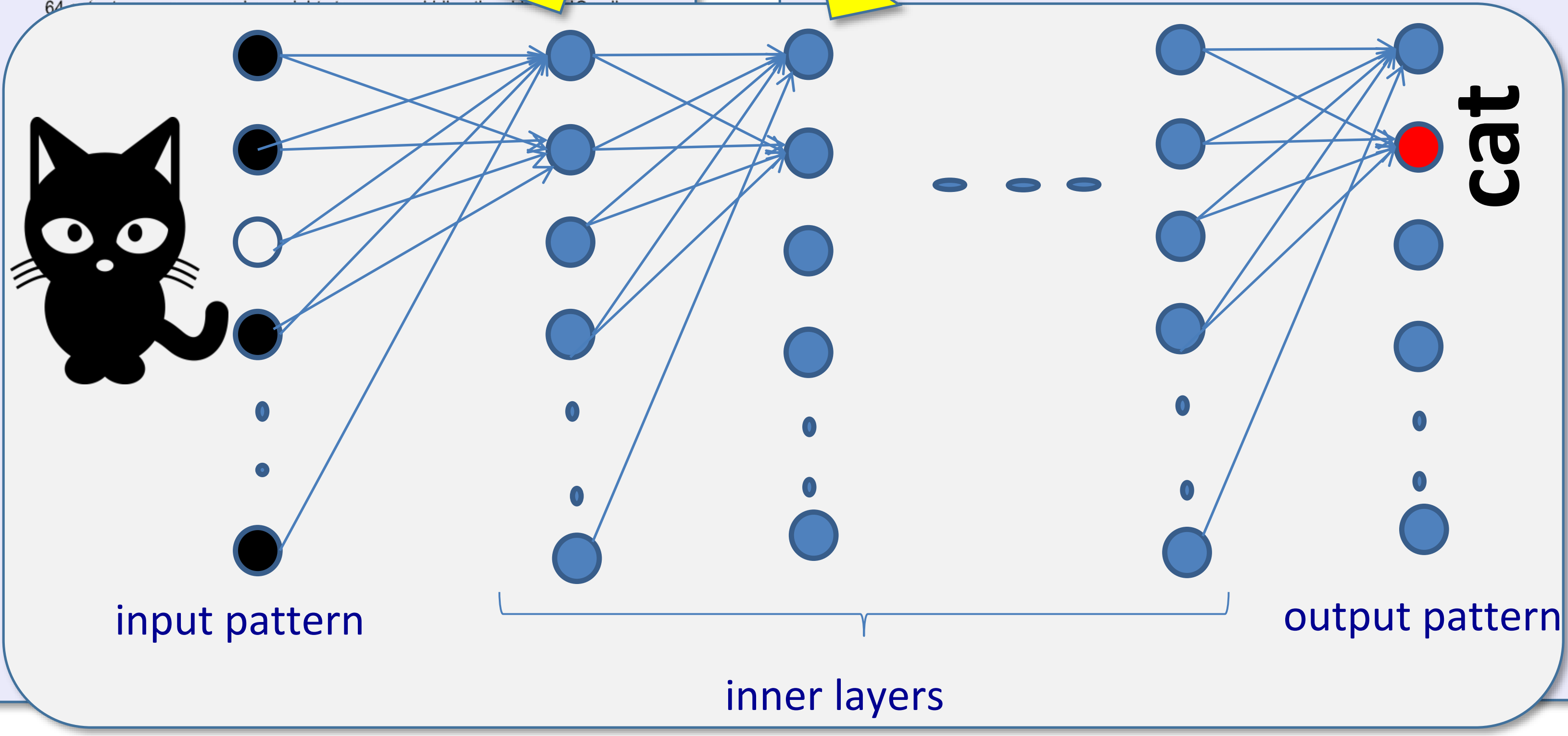
- trained with backpropagation



Spike-based model (current biology)



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



Brain-Inspired Computing

REALIZE future computing based on biological information processing

understanding biological information processing

Neuromorphic Computing :
artificial system of neurons and synapses inspired by neuroscience

hardware realization using dedicated circuits:

→ model embodied in the computing substrate

→ substrate purposely build for a certain class of models

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**

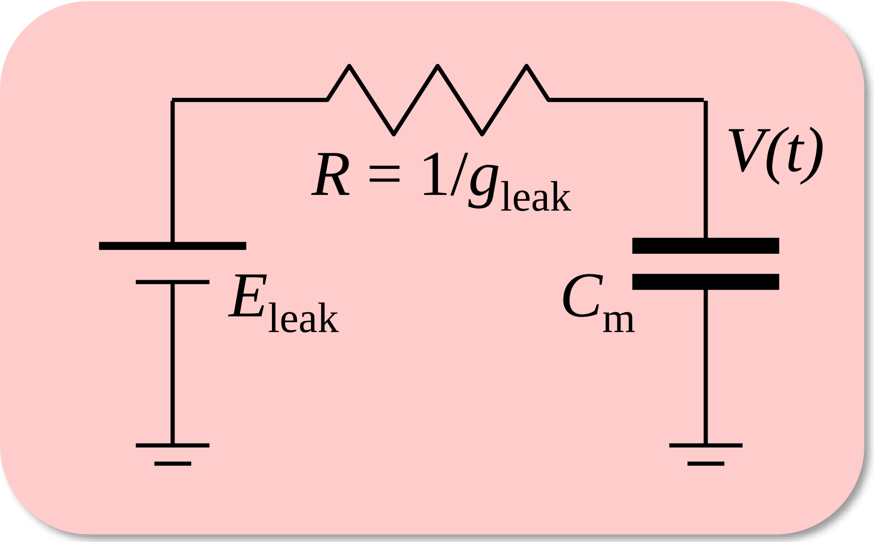
→ BrainScaleS spike-based physical modeling system

- overcoming the power wall of Turing-based computing
- support research of local learning rules
- time-continuous modeling of neuron dynamics
- acceleration of modeling including hierarchical learning schemes

BrainScaleS : Neuromorphic computing with physical model systems

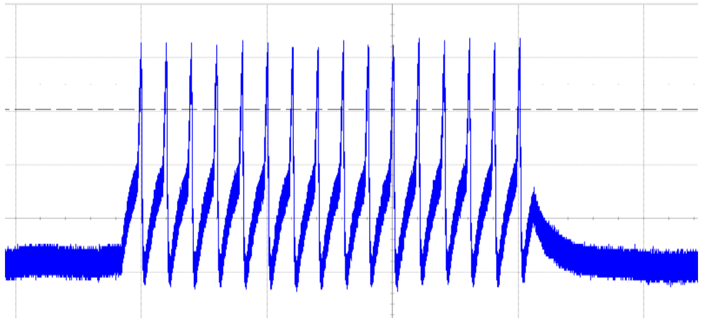
Consider a simple physical model for the neuron's cell membrane potential *V:*

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



$$\left.\frac{dV}{dt}\right|_{bio} << \left.\frac{dV}{dt}\right|_{VLSI}$$

→ accelerated neuron model

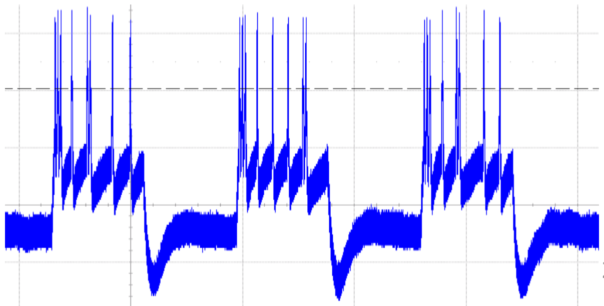


continuous time

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

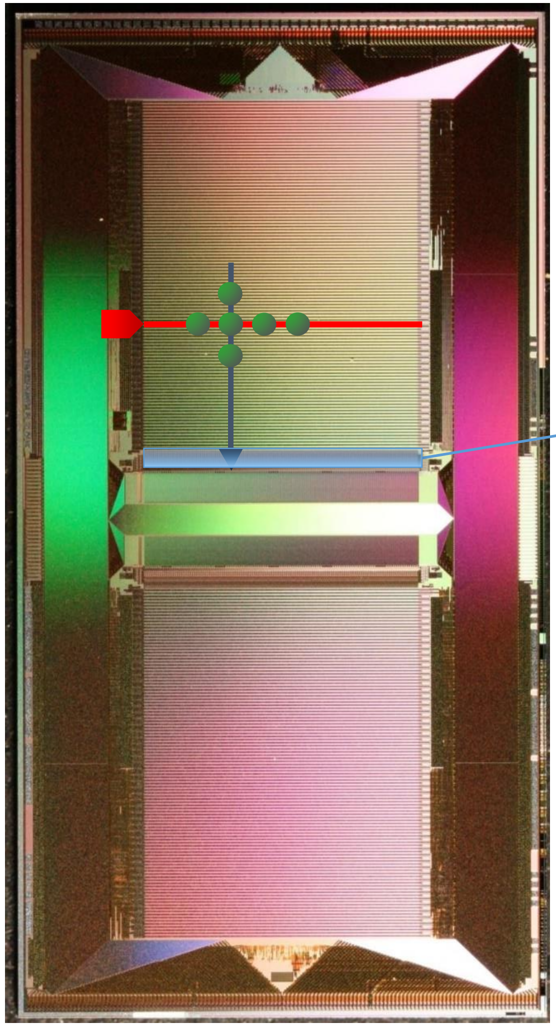
no multiplexing of components storing model variables

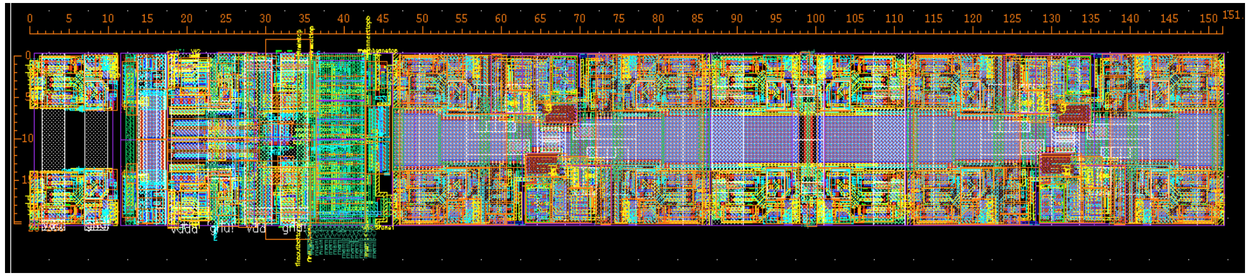
- each neuron has its membrane capacitor
- each synapse has a physical realization

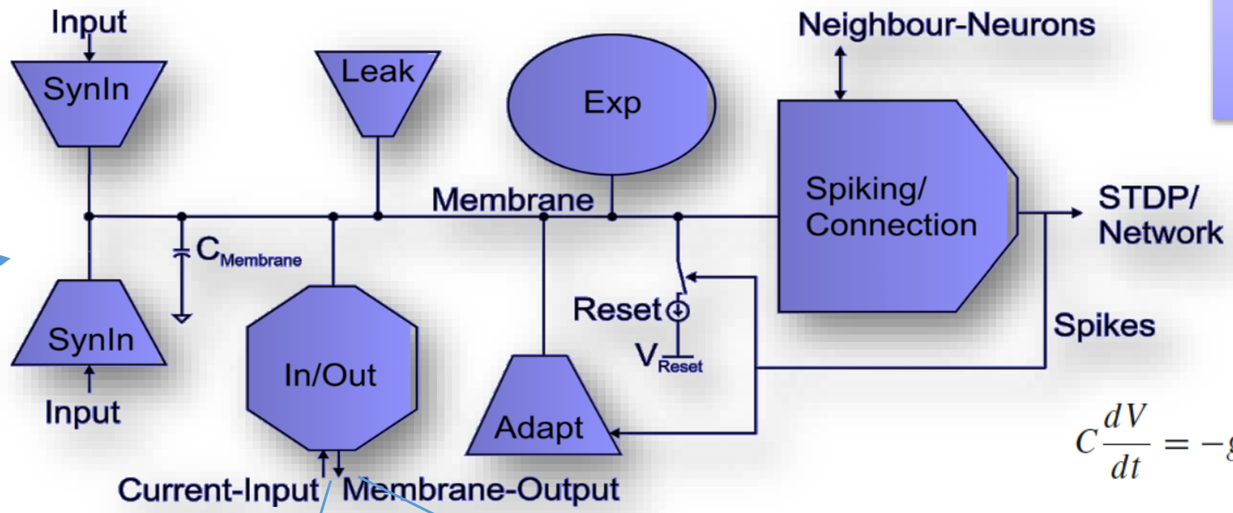


Structure of BrainScaleS neurons: array of parameterized dendrite circuits

photograph of the BrainScaleS 1 neuromorphic chip



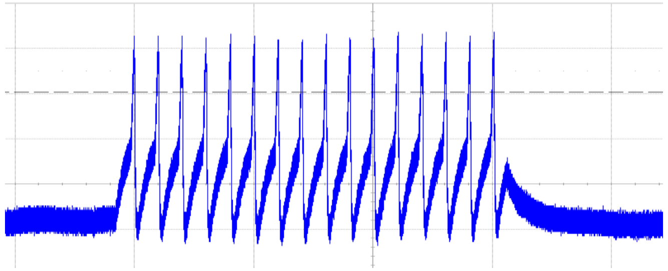


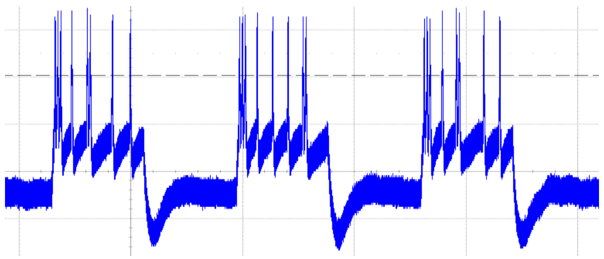


- 180 nm (generation 1) or 65 nm (gen. 2)
- 24 calibration parameters per neuron
- modular structure
- full set of ion-channel circuits for each dendrite

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

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

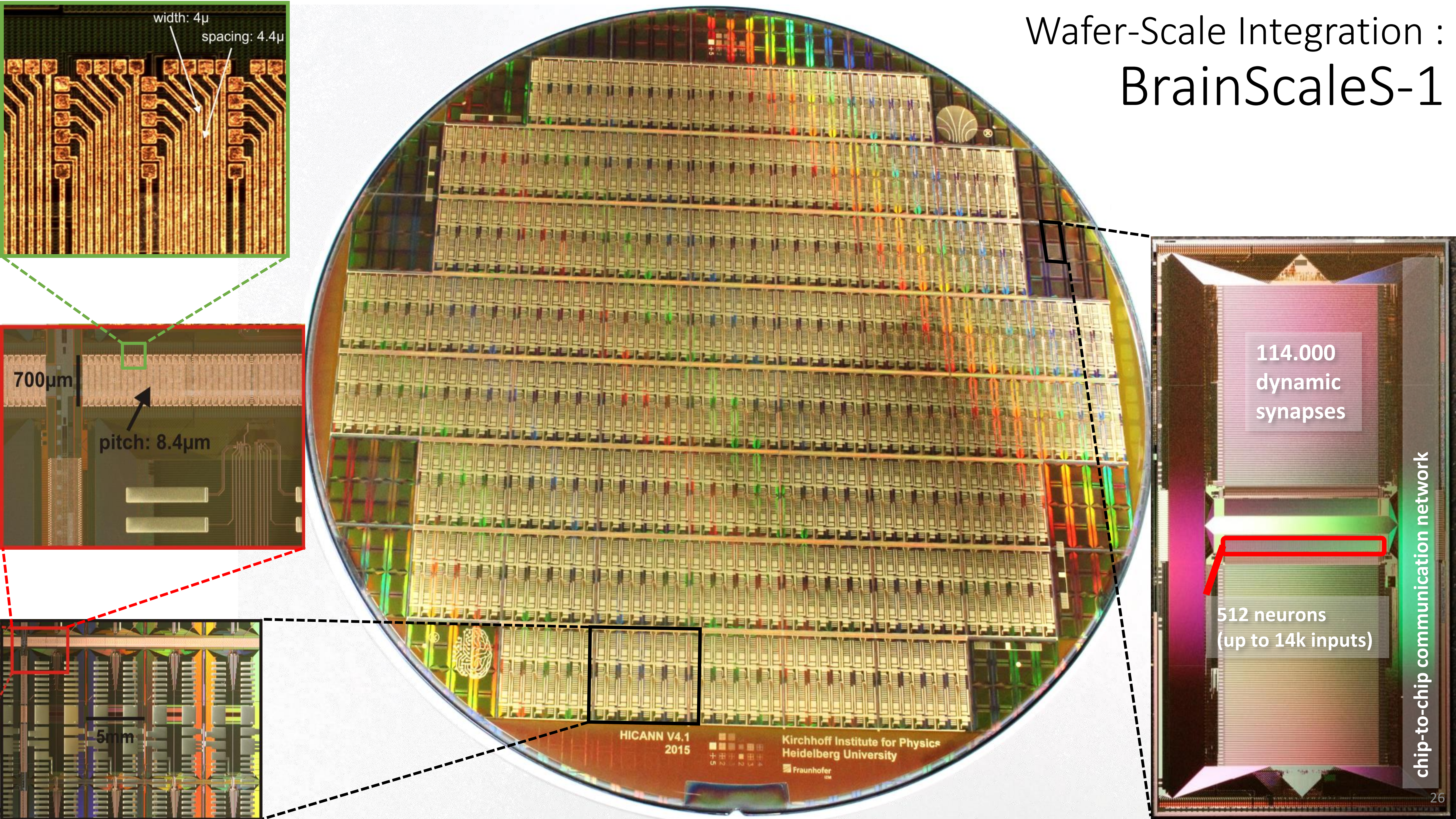




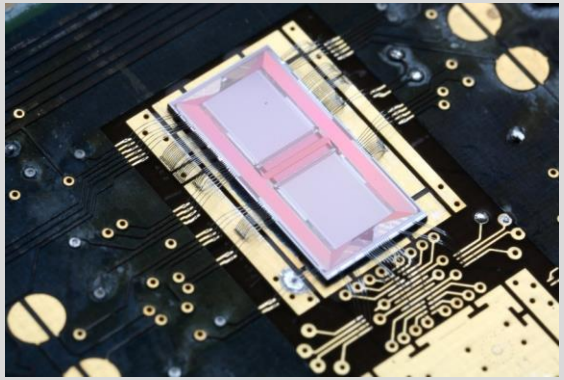
TimeScales

	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



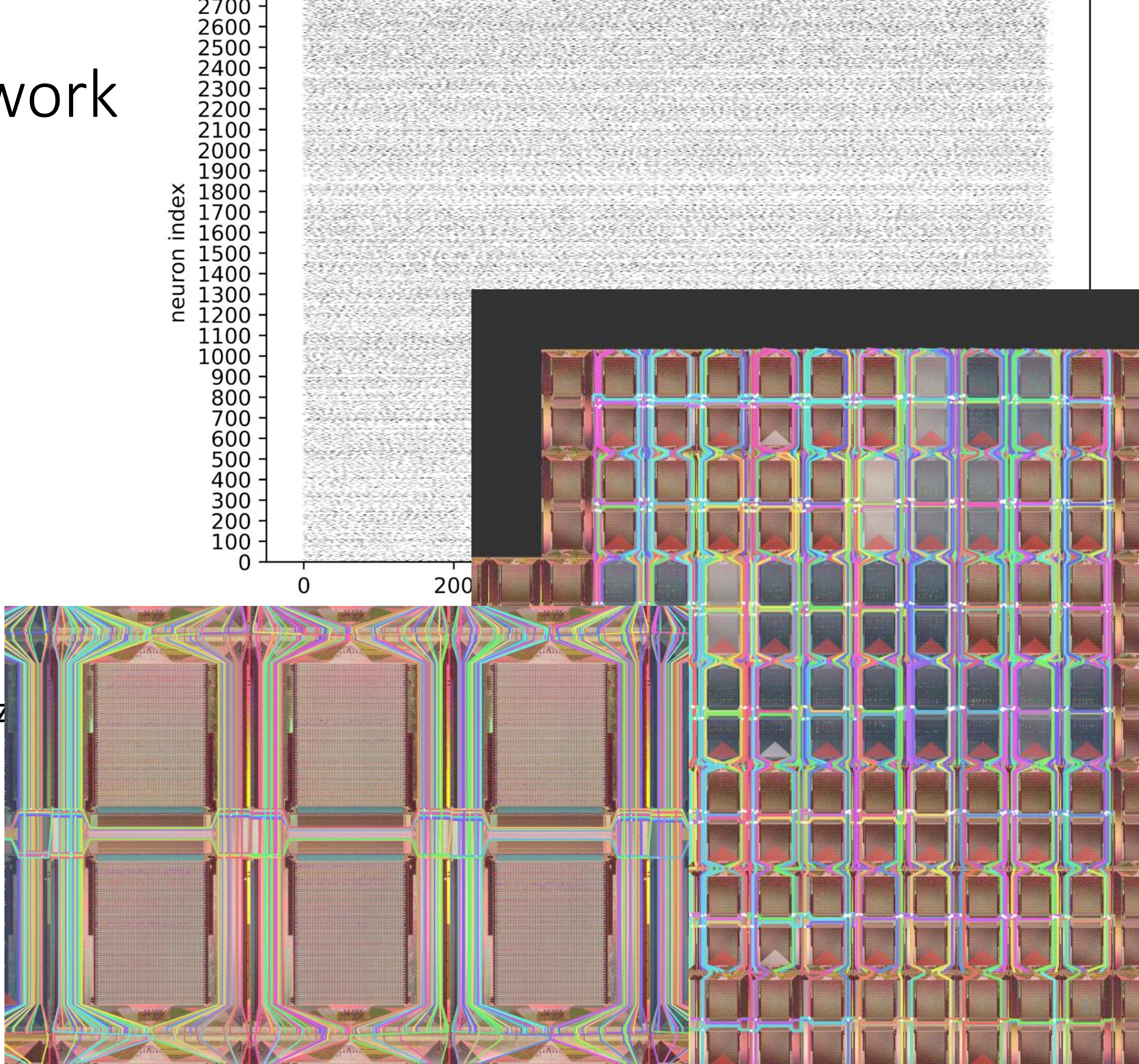
single chip | wafer module | hybrid system

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)



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

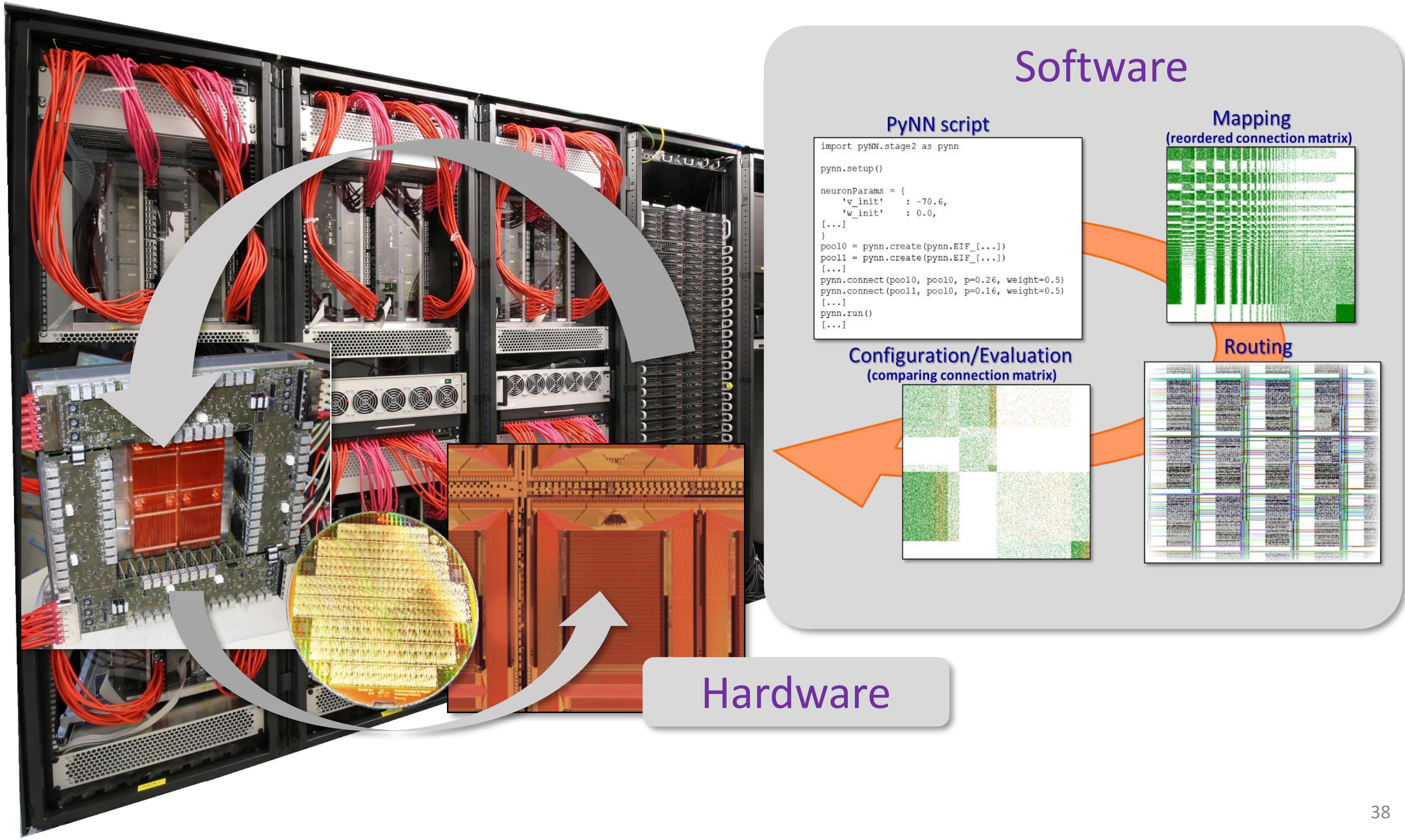
after training:

Non-Turing physical computing system performing autonomously

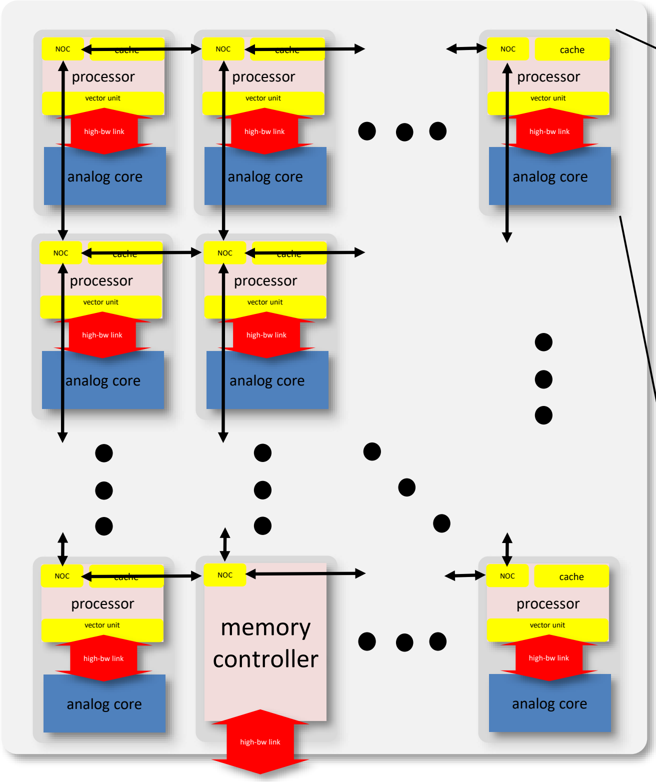
but

Turing-based computing is used in multiple places:

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

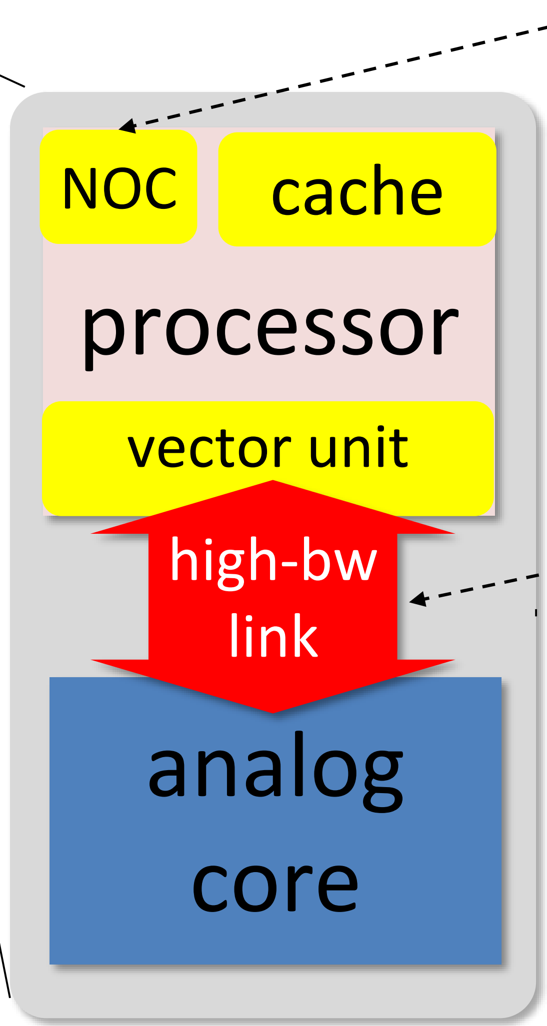


Shortening the hardware – software loop : Analog neuromorphic system as coprocessor



special function tile:

- memory controller
- SERDES IO
- purely digital function unit



Network-on-chip:

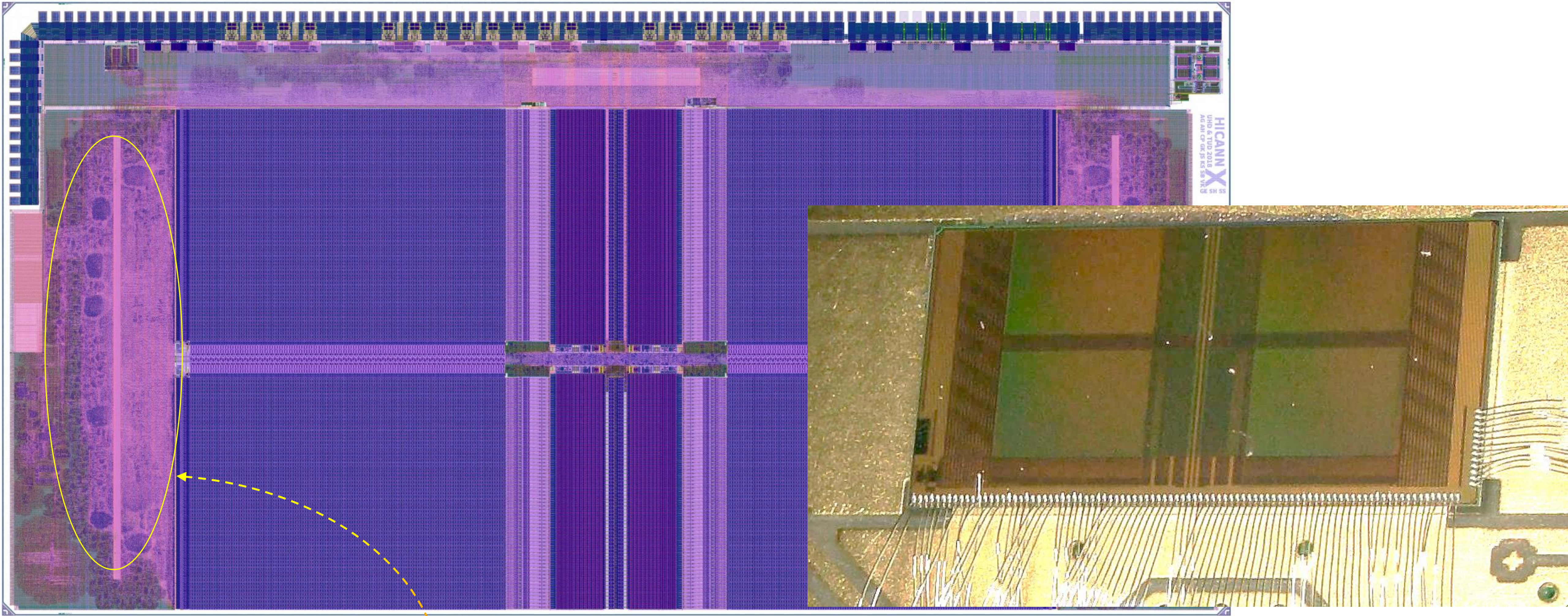
- prioritize event data
- unused bw for CPU
- common address space for neurons and CPUs

high-bandwidth link:

vector unit ←→ NM core

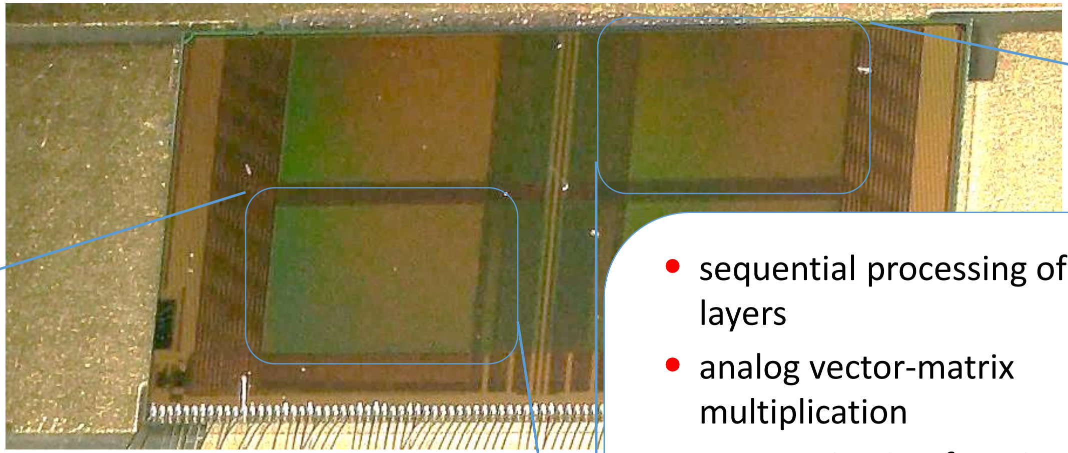
- weights
- correlation data
- routing topology
- event (spikes) IO
- configuration

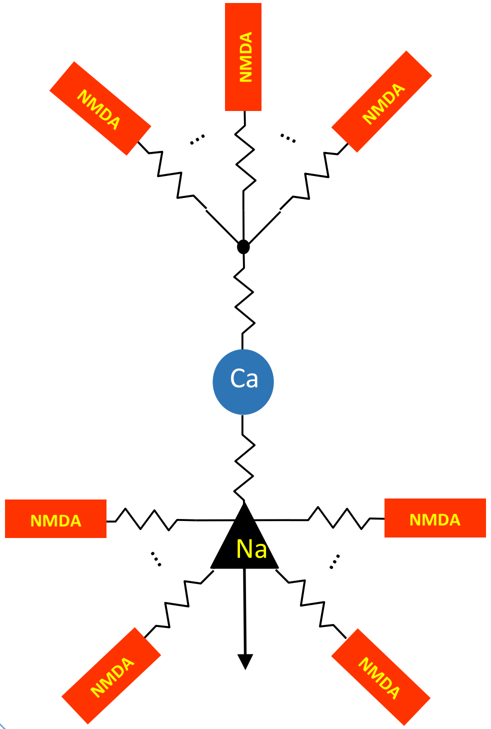
BrainScaleS-2 (BSS-2) ASIC

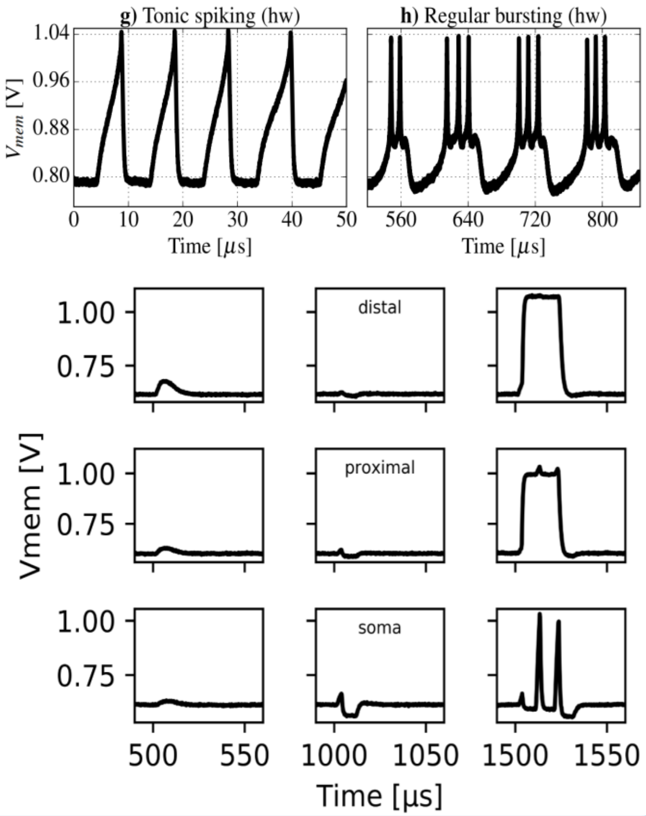


- 65nm LP-CMOS, power consumption O(10 pJ/synaptic event)
- 128k synapses
- 512 neural compartments (Sodium, Calcium and NMDA spikes)
- two SIMD plasticity processing units (PPU)
- PPU internal memory can be extended externally
- fast ADC for membrane voltage monitoring
- 256k correlation sensors with 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

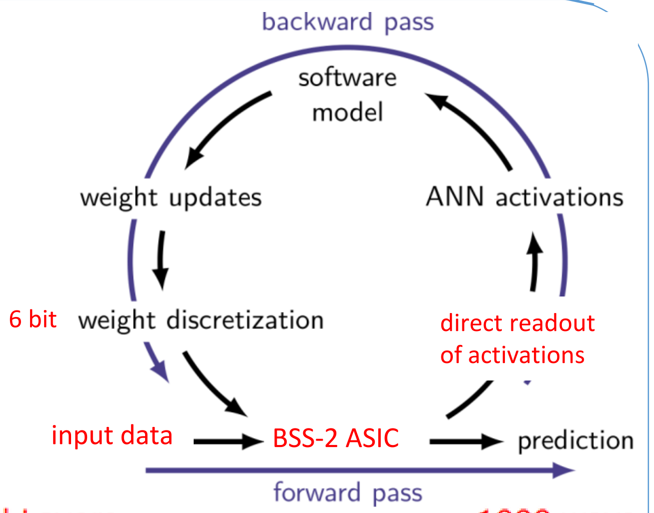
BrainScaleS-2 supports spike-based and Perceptron operation simultaneously

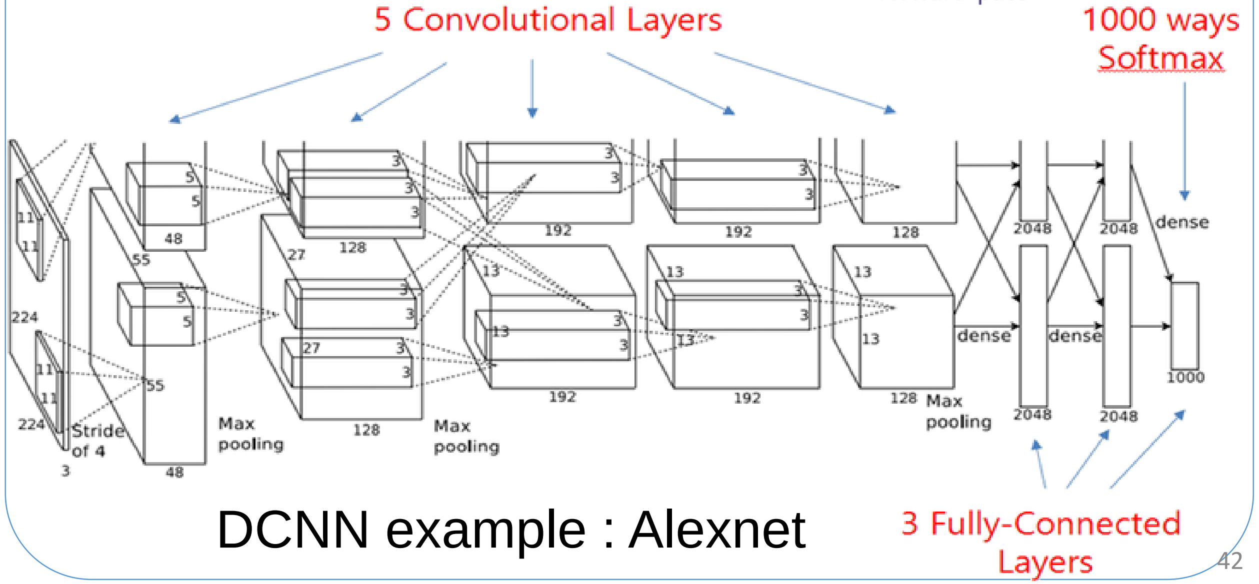






- sequential processing of all layers
- analog vector-matrix multiplication
- ReLU activation function with 4 to 8 bit resolution
- speed mostly limited by external memory





DCNN example : Alexnet

Learning and plasticity

BrainScaleS-2:

✓ biological relevant neuron model
- Adaptive Exponential Integrate and Fire (AdExp)
- NMDA, Ca and Na spikes

✓ biological relevant network topologies
- more than 10k synapses per neuron
- structured neurons with non-linear dendrites

Problem:
how to fix millions of parameters
- network topology
- neuron sizes and parameters
- synaptic strengths

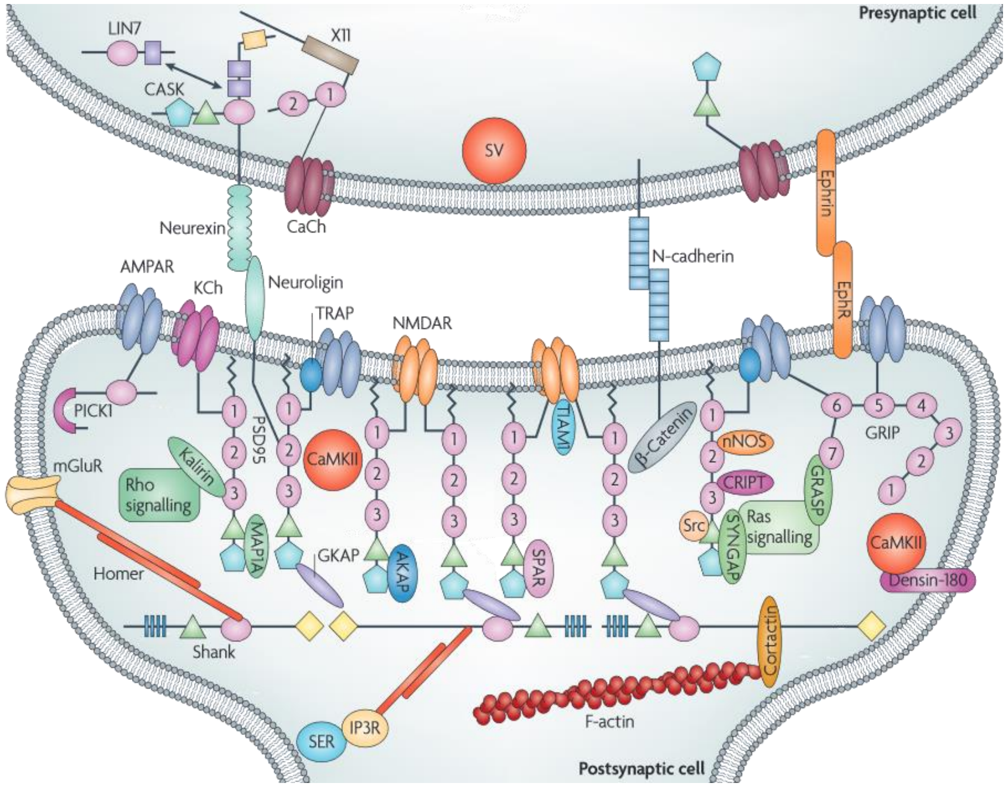
Trivial solution: everything is pre-computed on the host-computer
- requires precise calibration of hardware
- takes long time (much longer than running the experiment on the accelerated system)

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

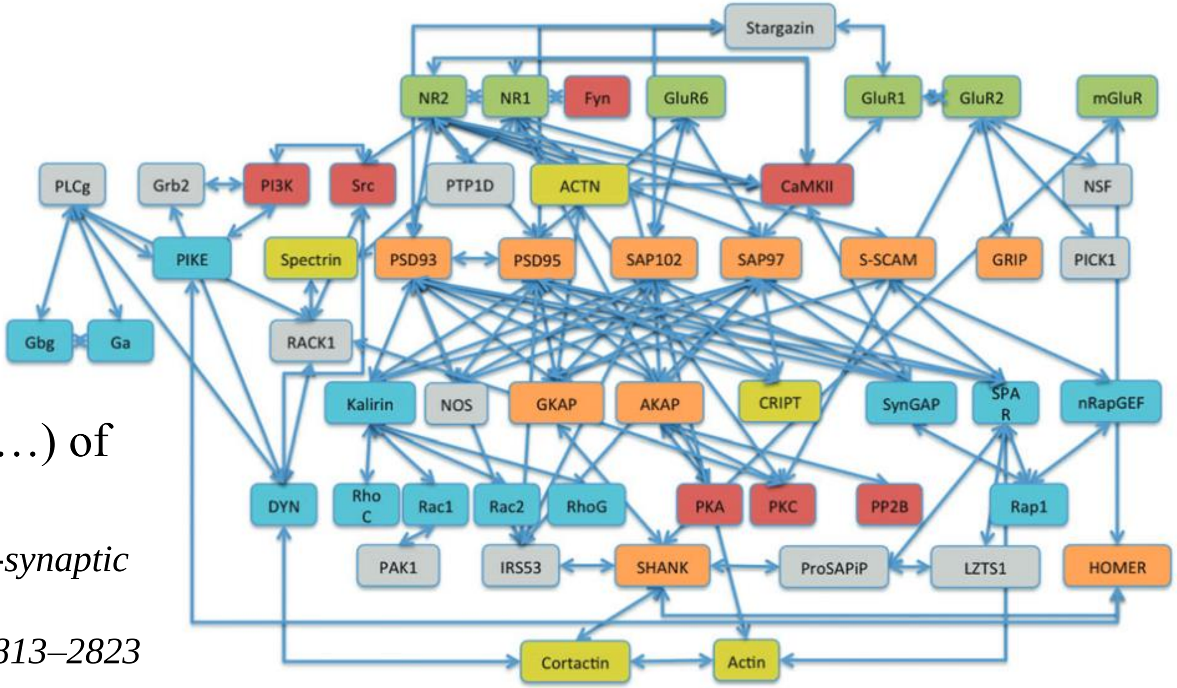
Complexity of synaptic plasticity is key to biological intelligence



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



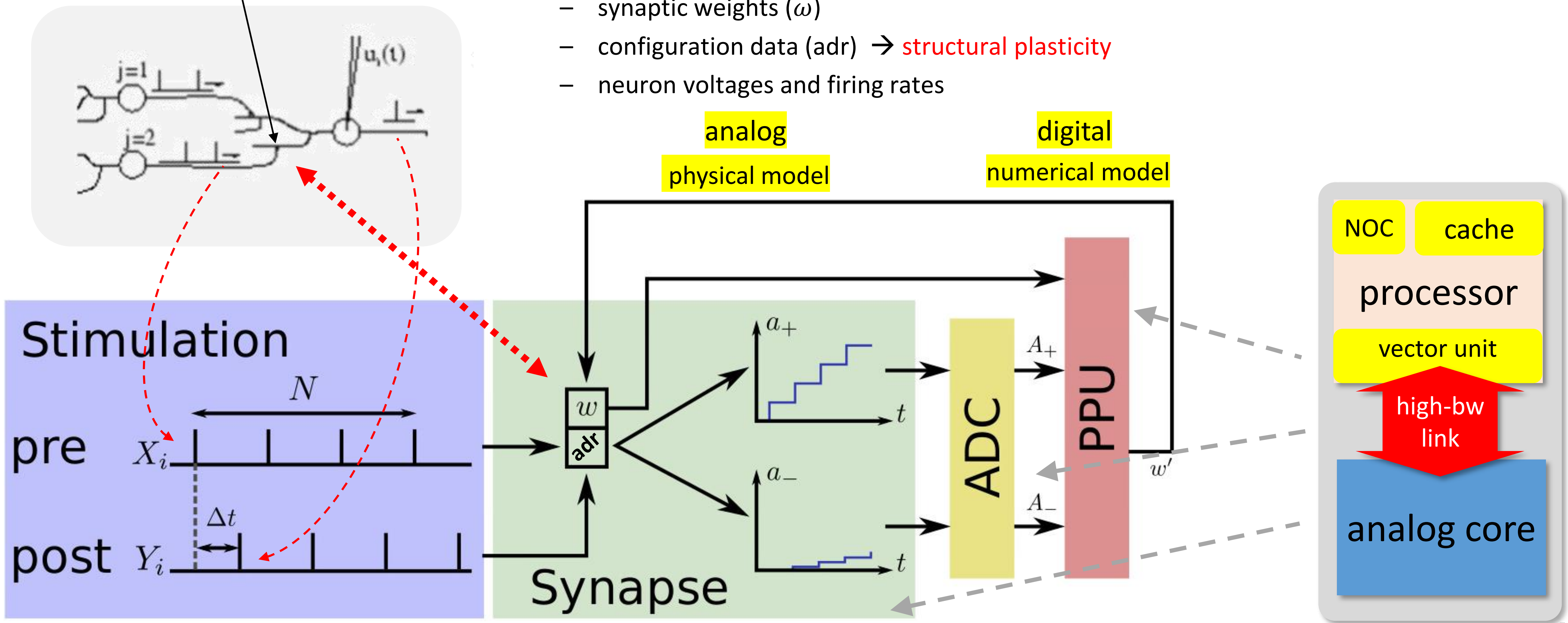
Protein-protein interaction map (…) of post-synaptic density

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

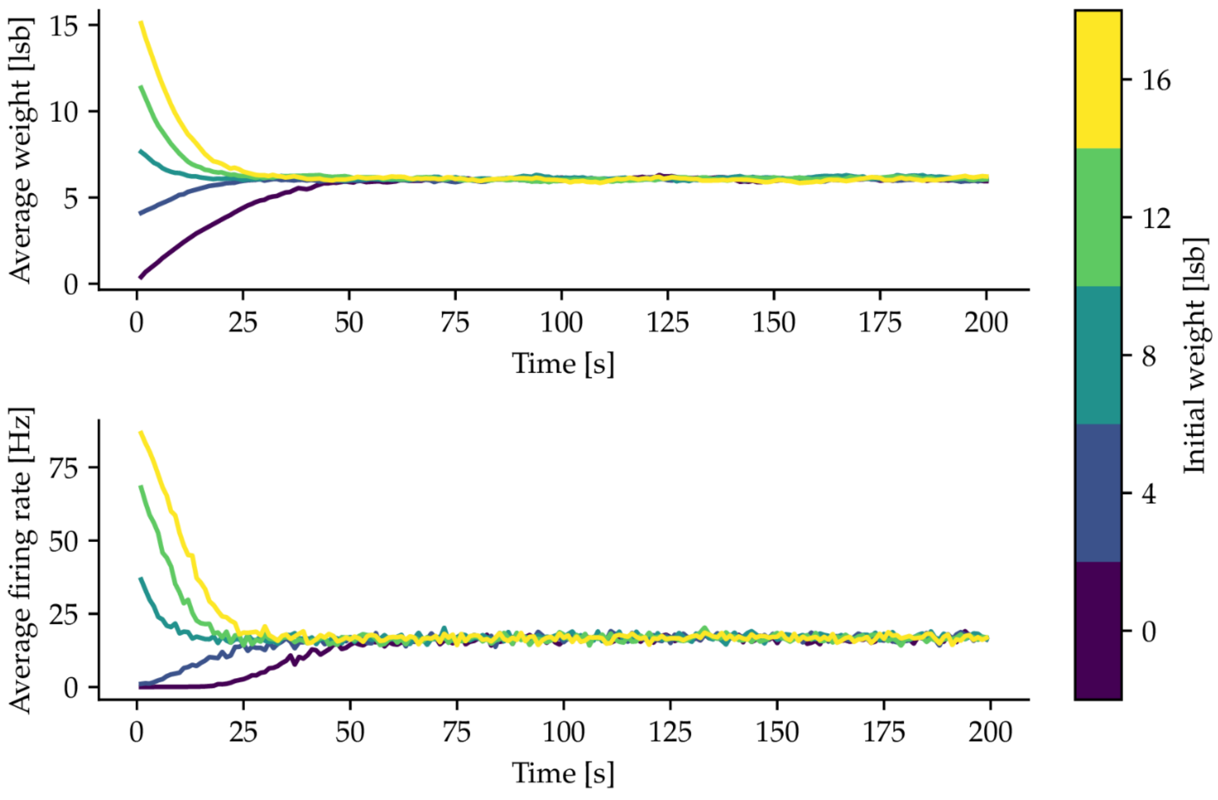
BrainScaleS-2: Hybrid Plasticity

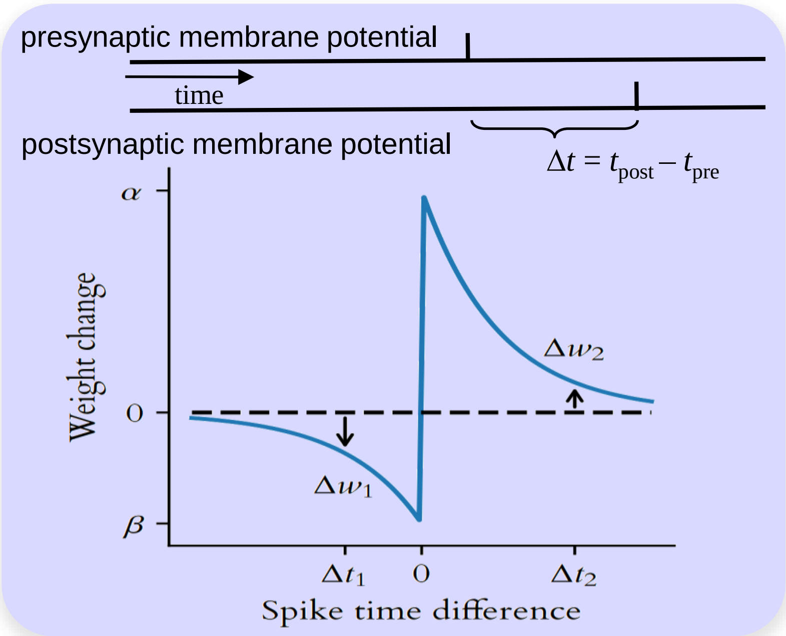
- analog correlation measurement in synapses
- A/D conversion by parallel ADC
- digital Plasticity Processing Units can access
  - synaptic weights (ω)
  - configuration data (adr) → structural plasticity
  - neuron voltages and firing rates

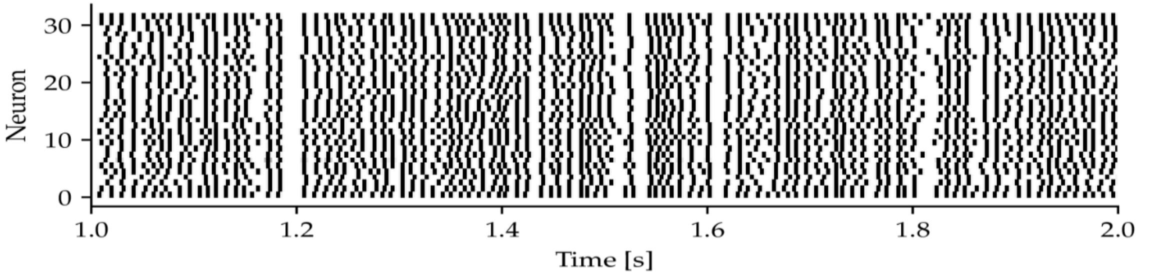
plasticity takes place at the synapse



Stabilizing firing rates with spike time dependent plasticity

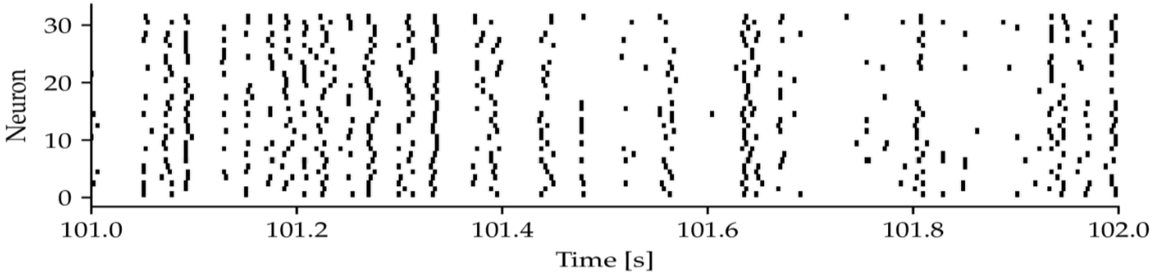






Wall-time per trace: 200ms

→ acceleration factor of 1000

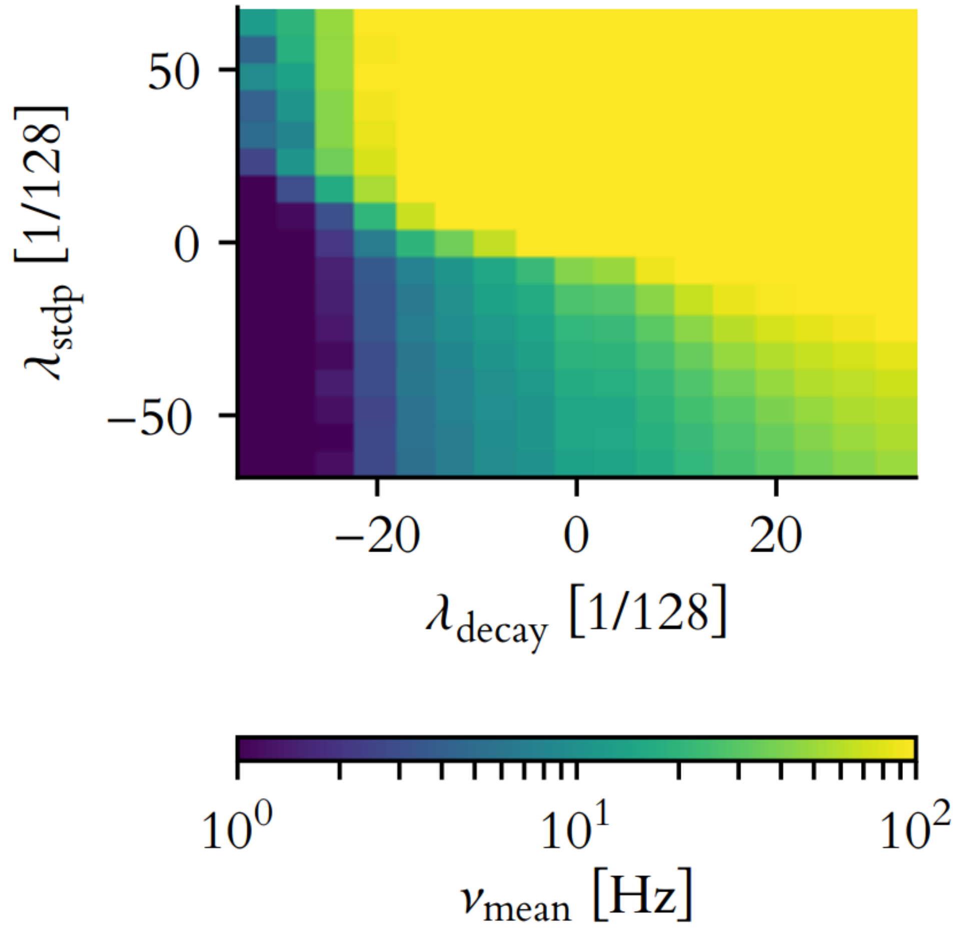


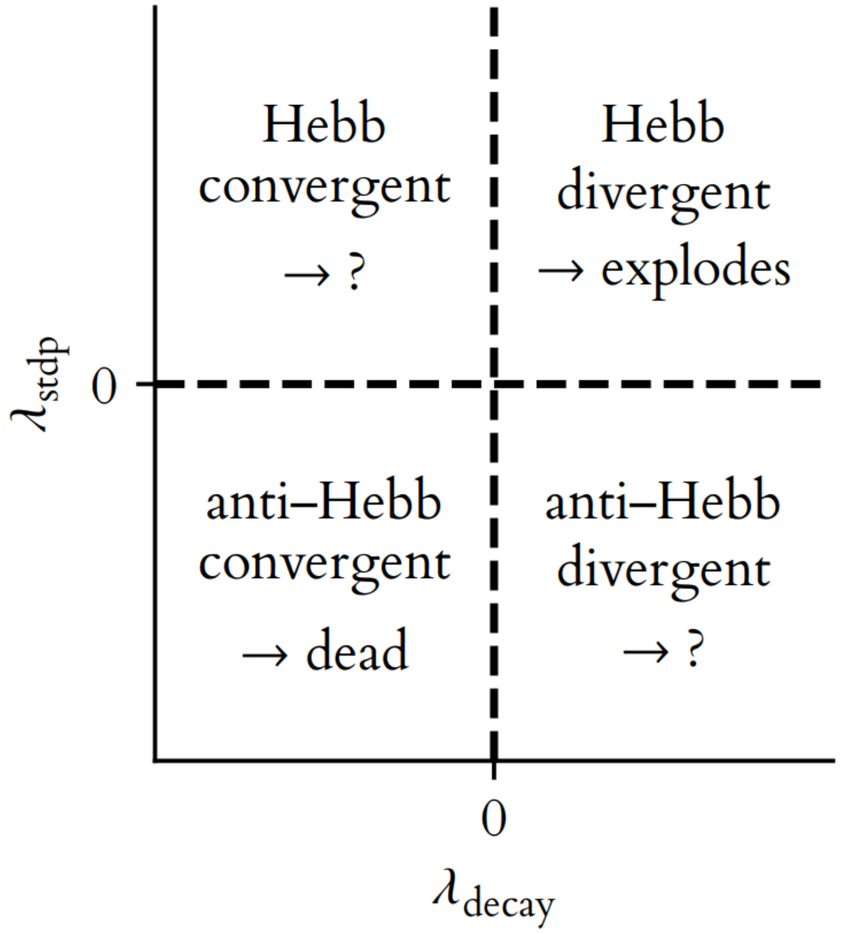
David Stöckel, Master Thesis, Heidelberg University, 2017

Stability analysis for plasticity rules

Measure the plasticity parameter phase space

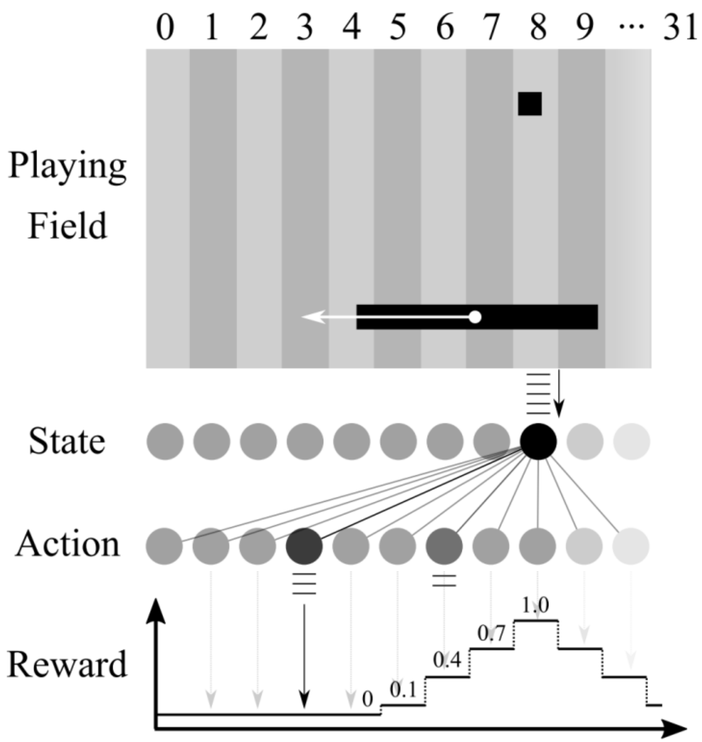
David Stöckel, Master Thesis, Heidelberg University, 2017





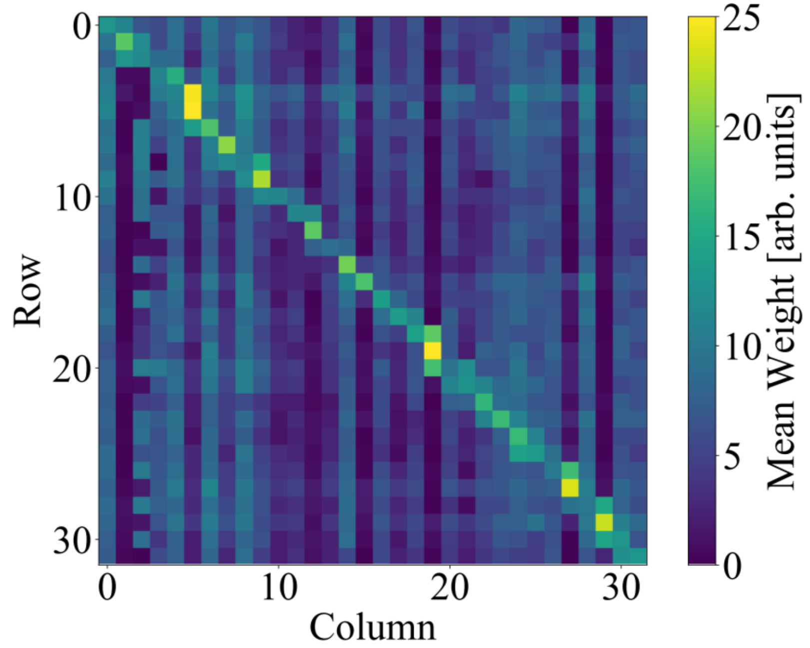
each data point is full plasticity experiment covering 200s biological real time

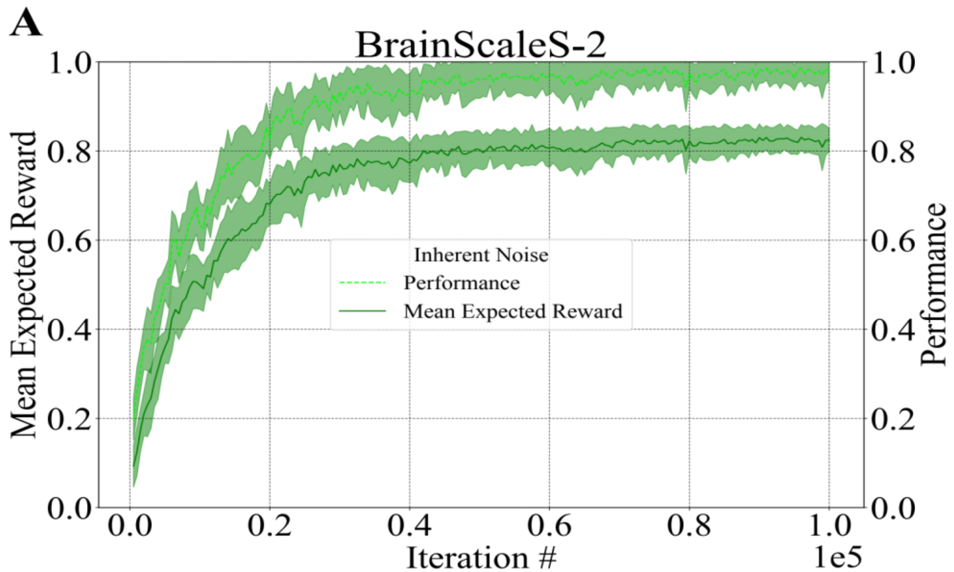
Learning Pong – tech demo using internal PPU only



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

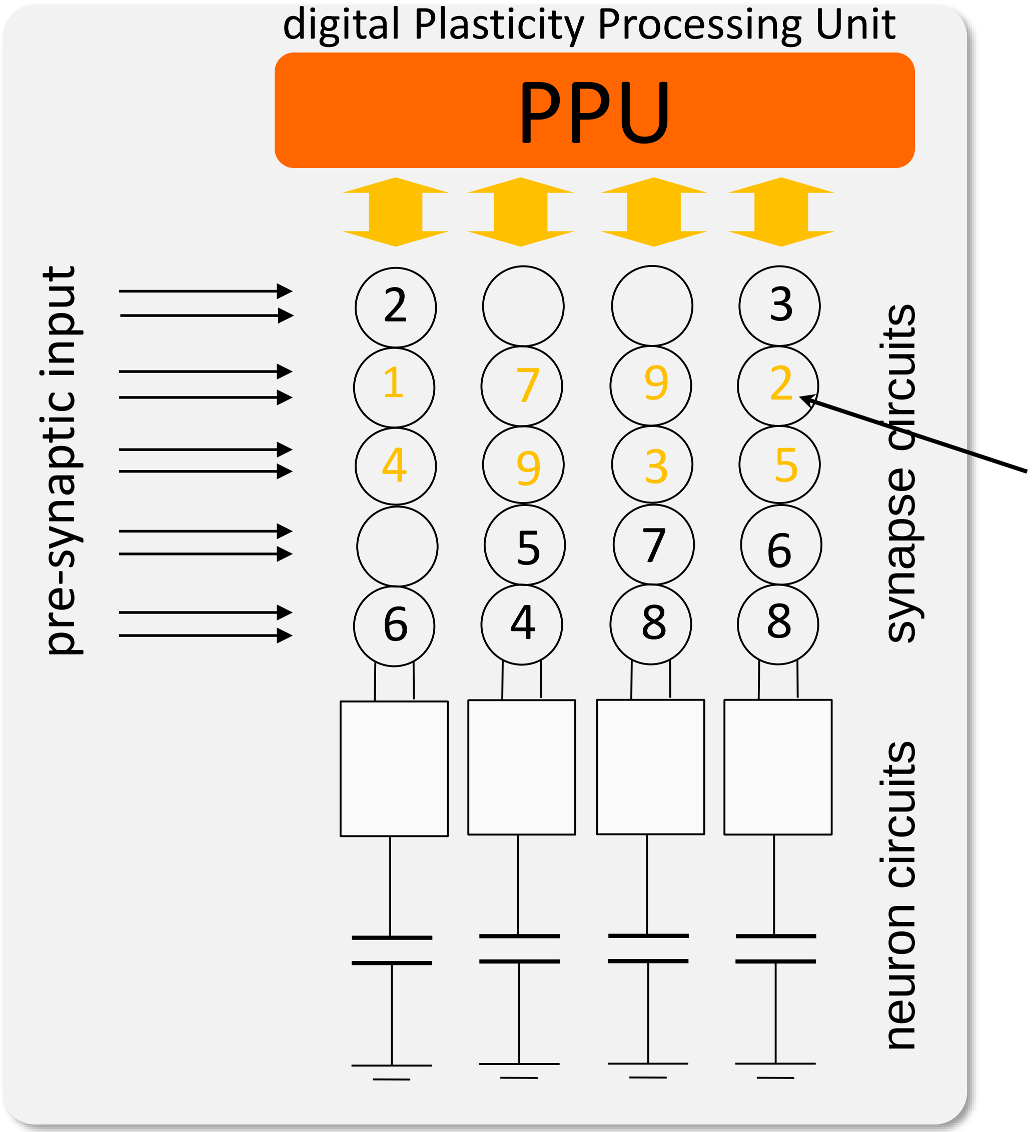
network time 0.4ms/iteration
23 µJ total chip energy





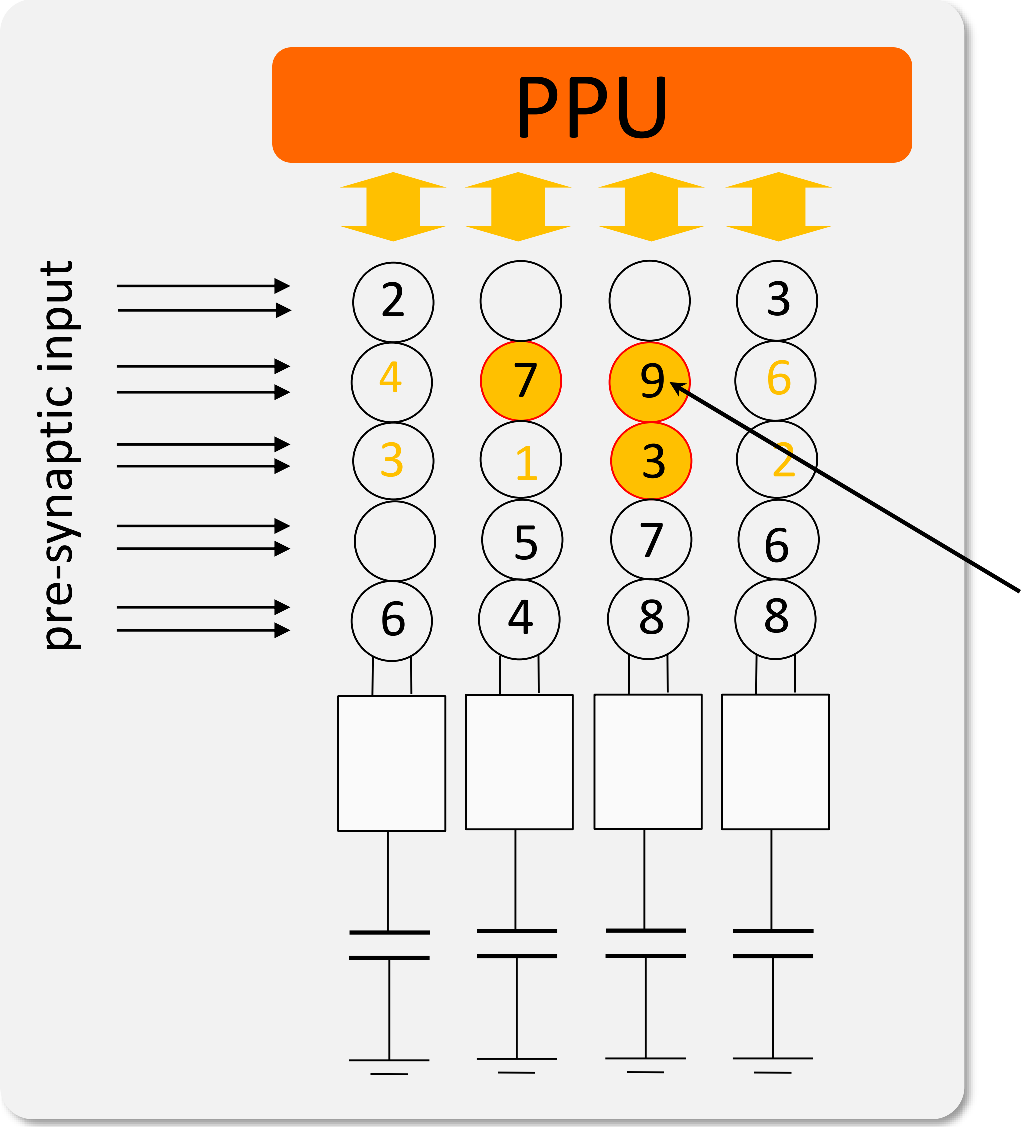
Wunderlich et.al., Demonstrating Advantages ..., Front. Neurosci., 2019

Structural plasticity



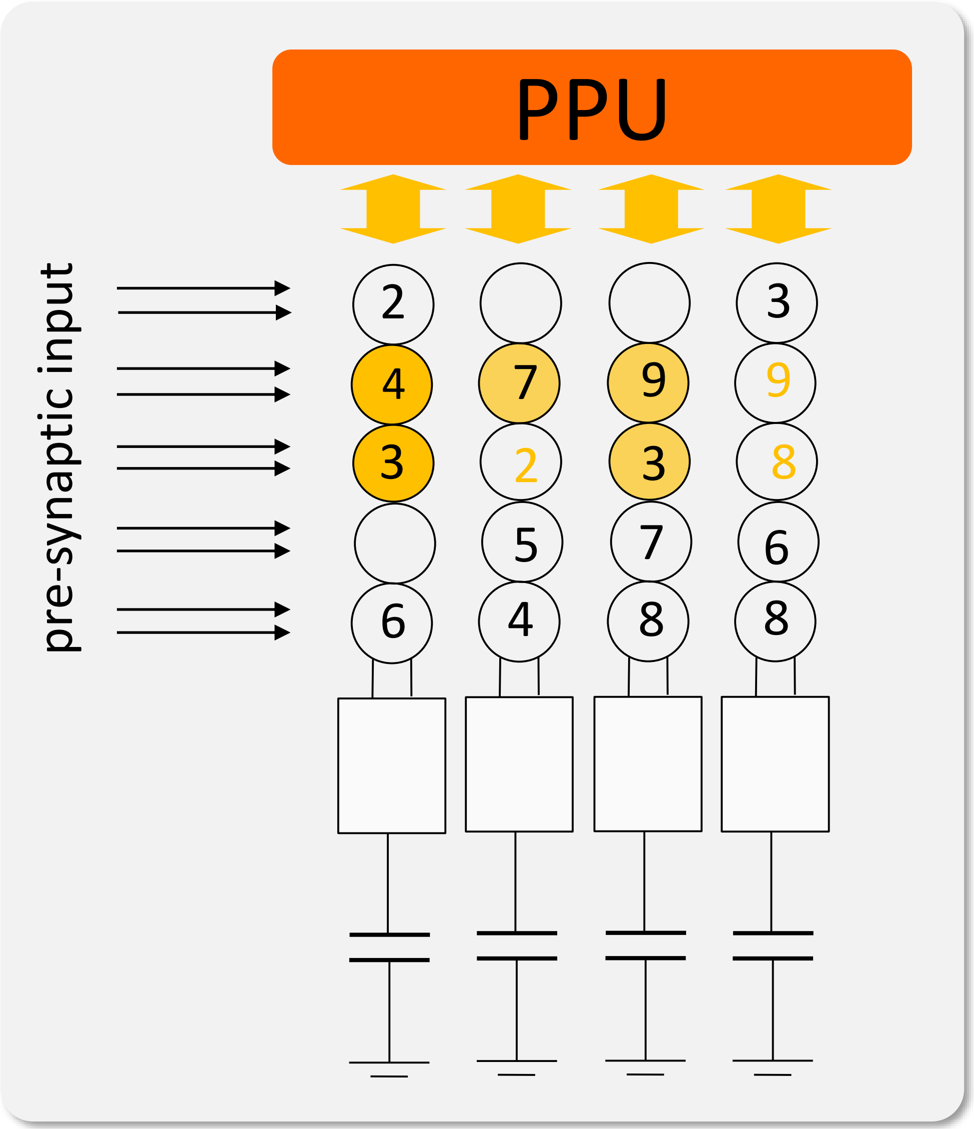
- assign random pre-synaptic neurons
- evaluate correlation

Structural plasticity



- assign random pre-synaptic neurons
- evaluate correlation
- keep the best

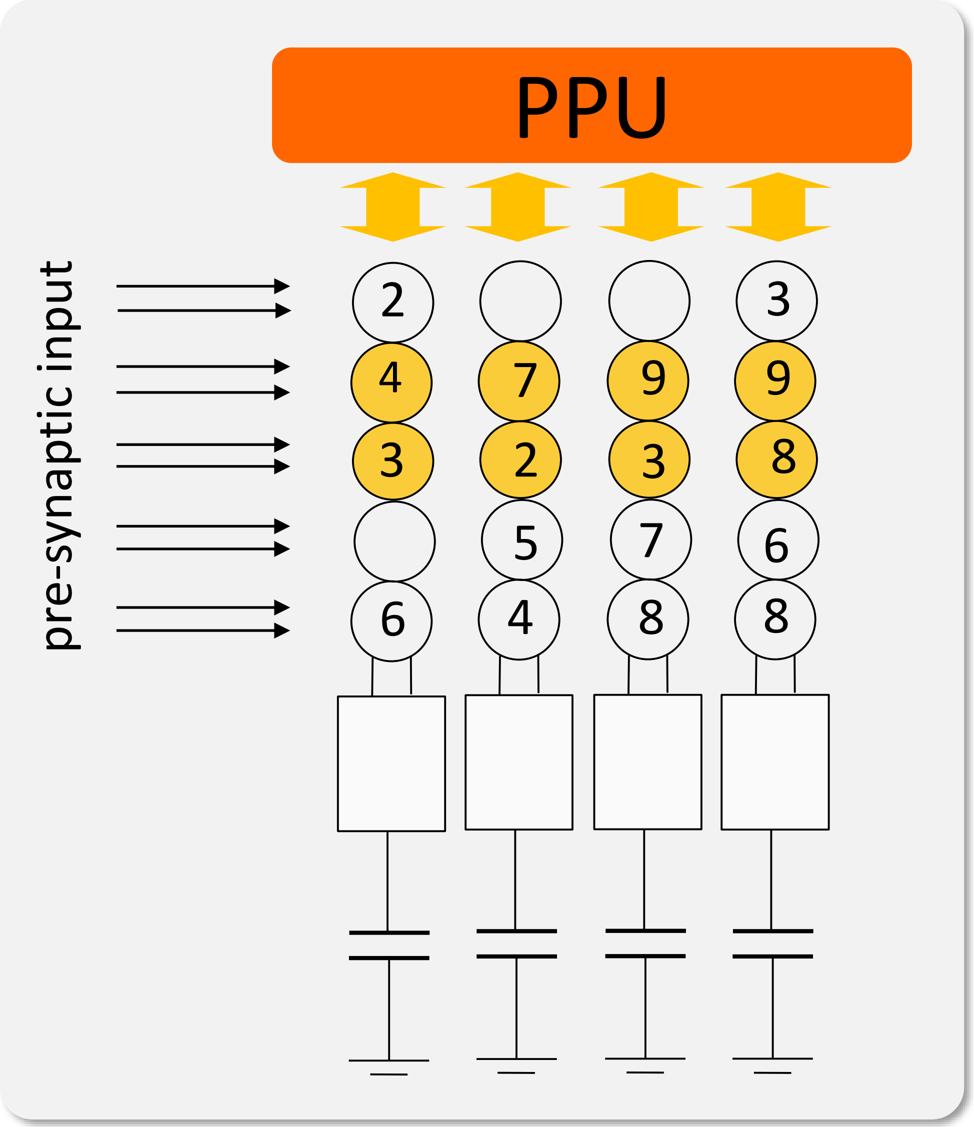
Structural plasticity



- assign random pre-synaptic neurons
- evaluate correlation
- keep the best

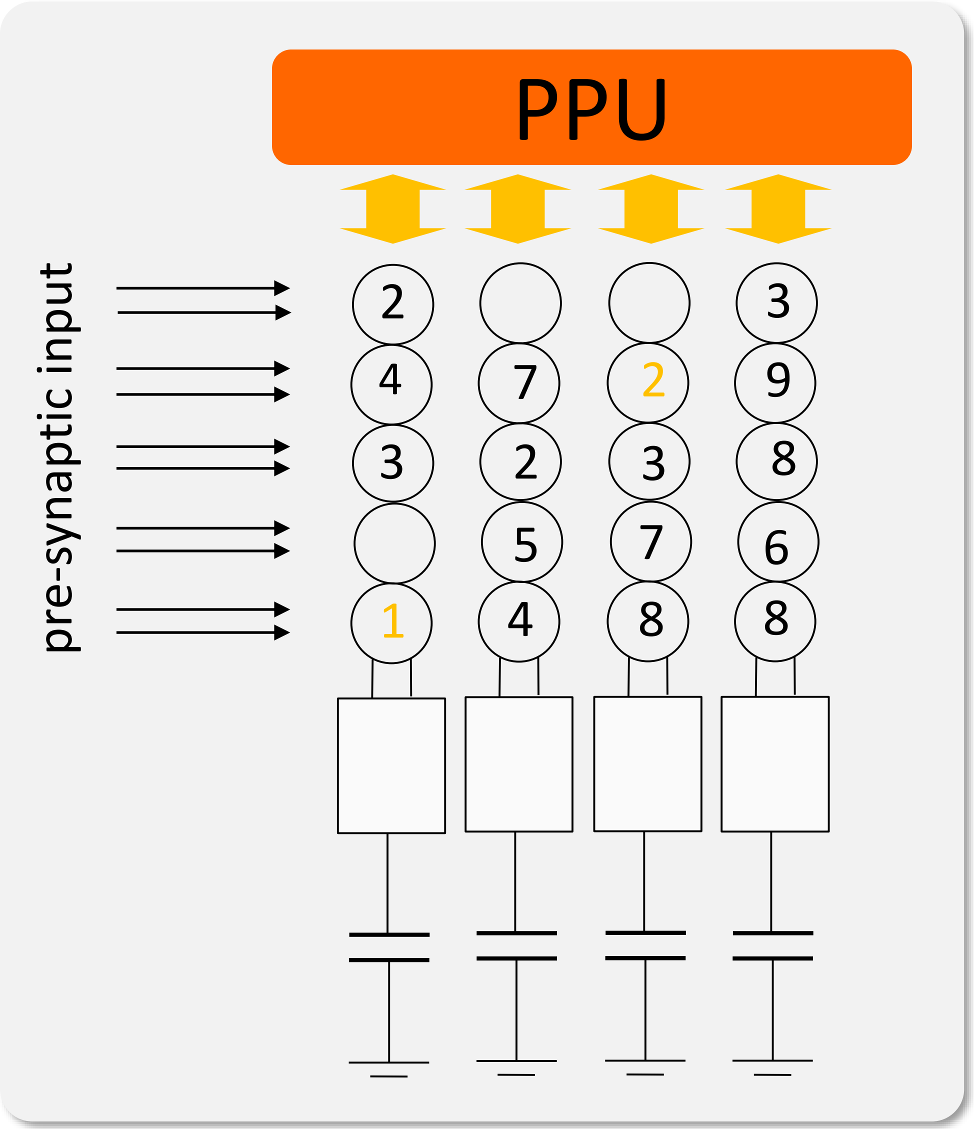
repeat

Structural plasticity



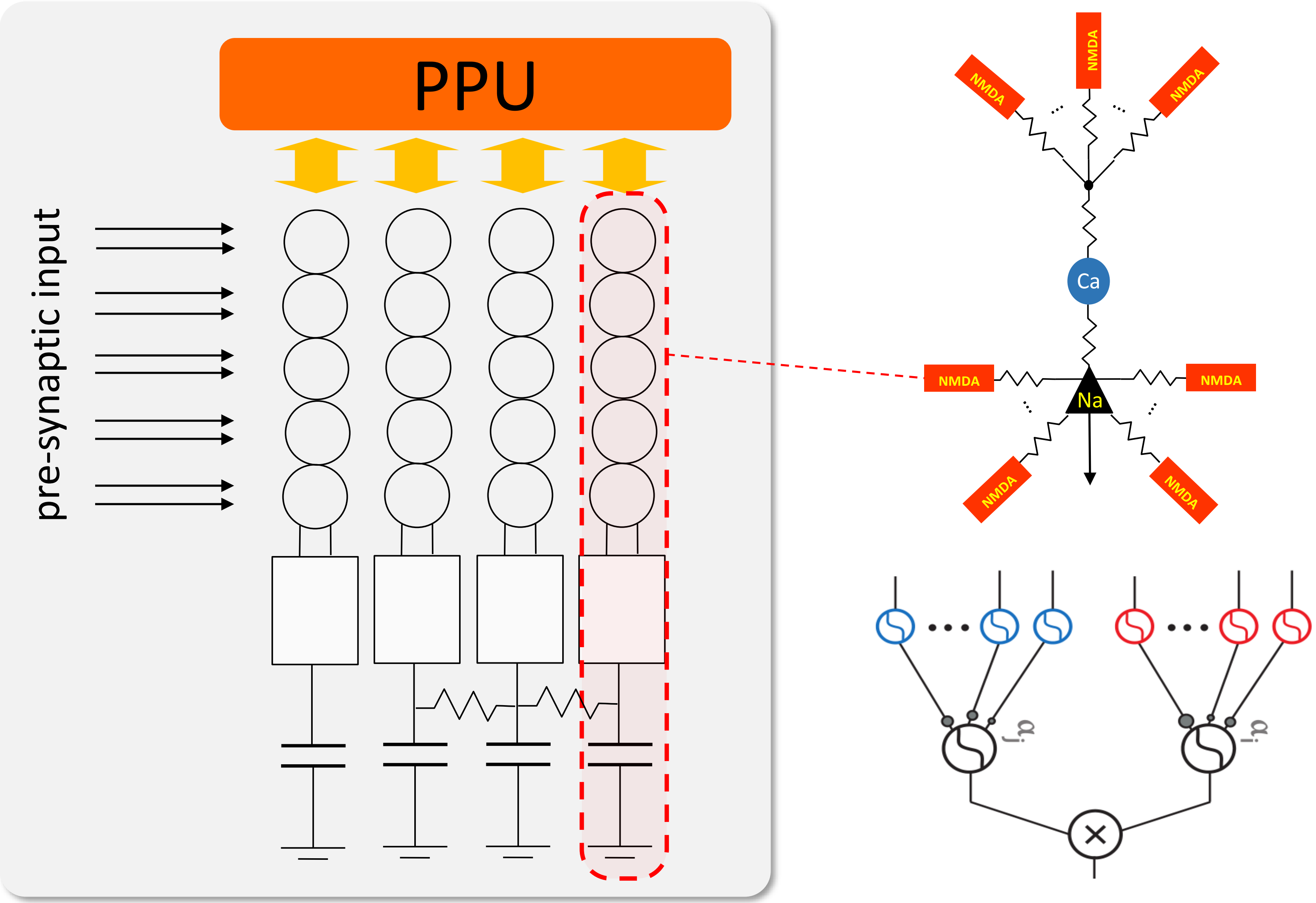
assign random pre-synaptic
neurons
evaluate correlation
keep the best

Structural plasticity

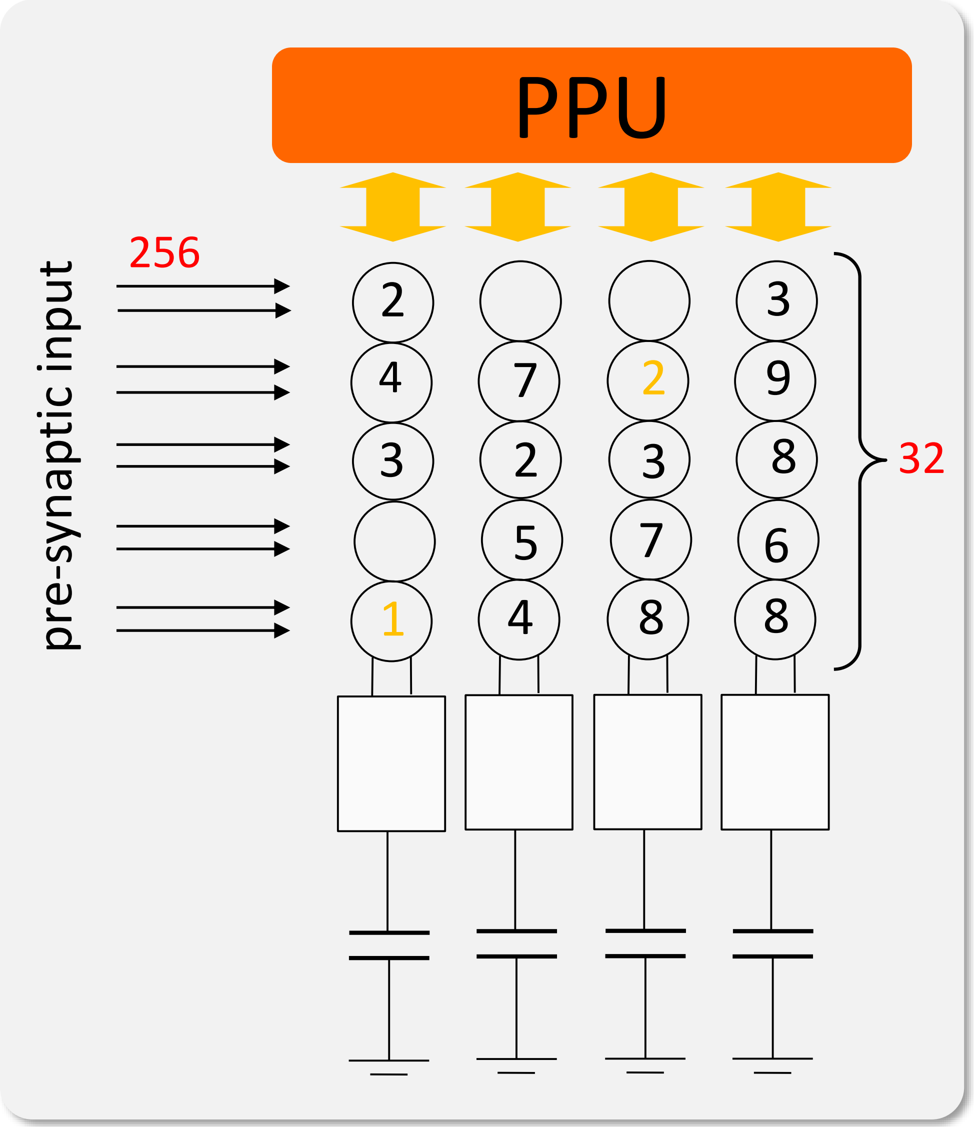


- assign random pre-synaptic neurons
- evaluate correlation
- keep the best
- replace weakly correlating synapses constantly against random new ones

Structural plasticity extends to structured neurons



Experimental example : structural plasticity

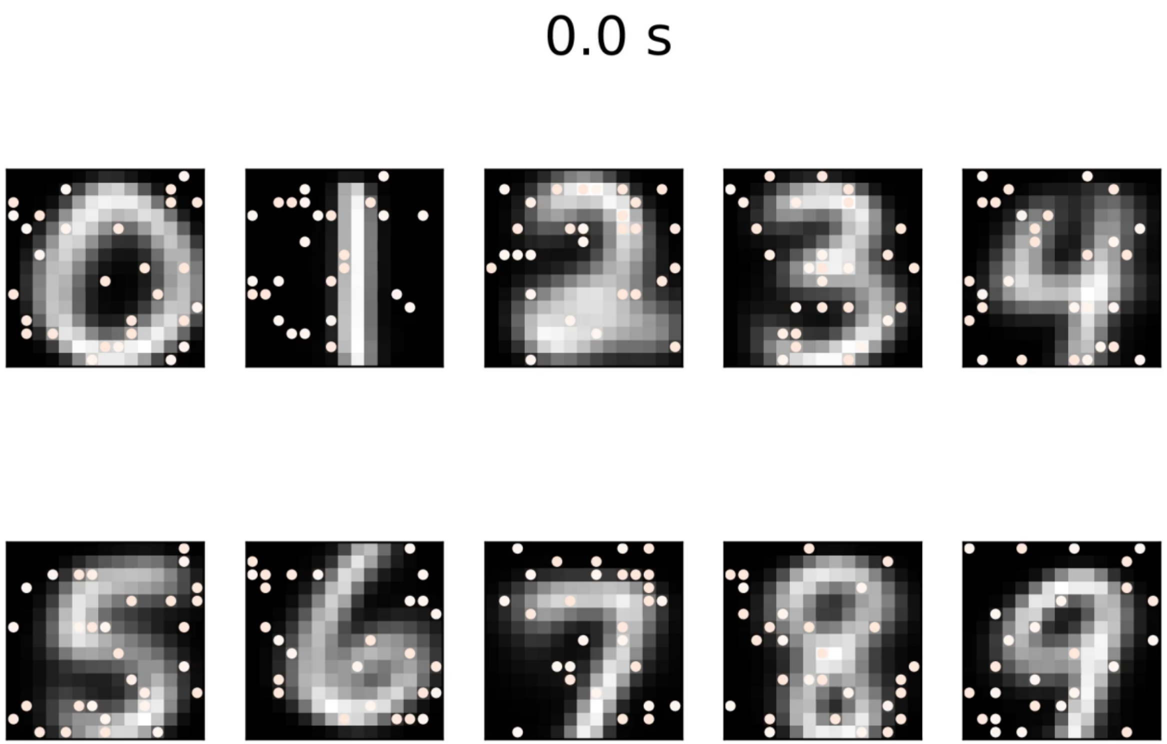


256 pre-synaptic inputs
mapped to single dendrite
with 32 active synapses
plasticity rule combines
structural, STDP and
homeostatic terms:

if $\omega \geq \theta_{\mathrm{rand}}$:
$\omega' \leftarrow \omega$
$+\lambda_{\mathrm{STDP}}(c_+ + c_-)$
$-\lambda_{\mathrm{hom}}\left(\nu + \nu_{\mathrm{target}}\right)$
$a' \leftarrow a$
else:
$\omega' \leftarrow \omega_{\mathrm{init}}$
$a' \leftarrow \mathrm{rand}(0,8)$

B. Cramer and S. Billaudelle, arXiv:1912.12047v1, 2020

Supervised learning using Hybrid Plasticity



dots represent realized (active) synapses
ten target groups (with three dendrites each)
trained simultaneously
1.5 s wall time needed for emulation

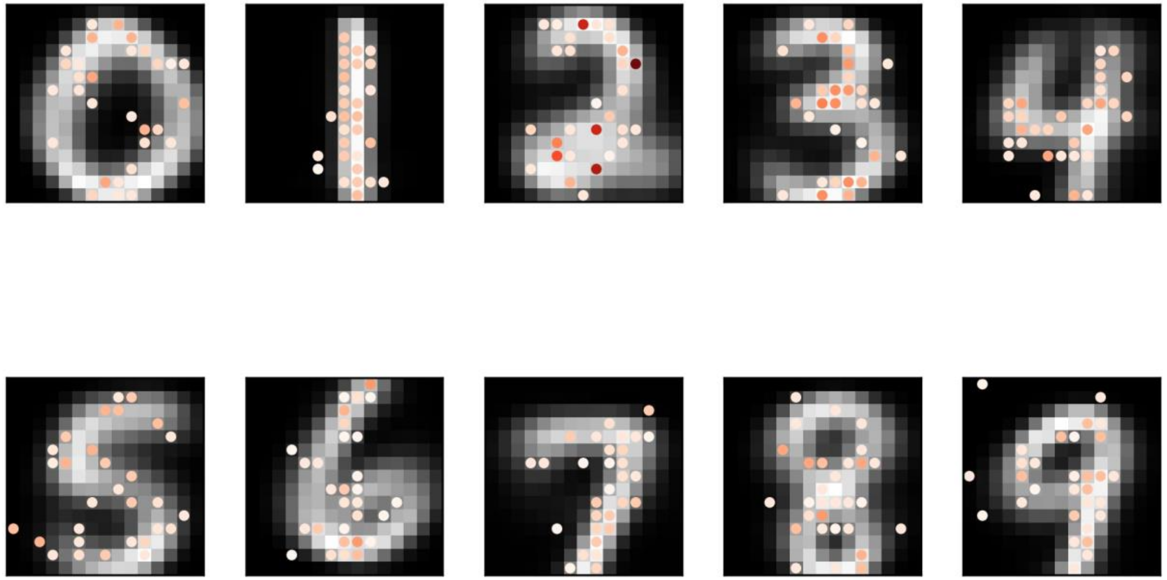
256 pre-synaptic inputs
mapped to single dendrite
with 32 active synapses
plasticity rule combines
structural, STDP and
homeostatic terms:

$$
\begin{aligned}
&\text{if } \omega \geq \theta_{\text{rand}}: \\
&\quad \omega' \leftarrow \omega \\
&\qquad\quad + \lambda_{\text{STDP}}(c_+ + c_-) \\
&\qquad\quad - \lambda_{\text{hom}}\left(\nu + \nu_{\text{target}}\right) \\
&\quad a' \leftarrow a \\
&\text{else:} \\
&\quad \omega' \leftarrow \omega_{\text{init}} \\
&\quad a' \leftarrow \text{rand}(0{,}8)
\end{aligned}
$$

B. Cramer and S. Billaudelle, arXiv:1912.12047v1, 2020

Supervised learning using Hybrid Plasticity

1554.7 s



Hybrid Plasticity
allows simultaneous rules for:

- structutral optimization
- homeostatic balance
- pre-post correlation

and more

$$
\begin{aligned}
&\text{if } \omega \geq \theta_{\text{rand}}: \\
&\quad \omega' \leftarrow \omega \\
&\qquad\qquad + \lambda_{\text{STDP}}(c_+ + c_-) \\
&\qquad\qquad - \lambda_{\text{hom}}\left(\nu + \nu_{\text{target}}\right) \\
&\quad a' \leftarrow a \\
&\text{else:} \\
&\quad \omega' \leftarrow \omega_{\text{init}} \\
&\quad a' \leftarrow \text{rand}(0,8)
\end{aligned}
$$

using software running in parallel to the analog neuron operation

B. Cramer and S. Billaudelle, arXiv:1912.12047v1, 2020

BrainScaleS in EBRAINS

- 2nd generation BrainScaleS with hybrid plasticity support is part of the EBRAIN research infrasturcue for neurosciences
- We are currently developing the high-level user access software, based on PyNN
- Large networks spanning full wafers like 1st generation BrainScaleS are currently not funded
- Small networks of 10 to 50 chips are currently under development


Yannik Stradmann
Björn Kindler
Johannes Schemmel
Eric Müller
Hartmut Schmidt
Christian Mauch
Johannes Weis
Moritz Hornung
Raphael Stock
aronl
Philipp Dauer
Maurice Güttler
Jose Montes
Andreas Baumbach
Timo Gierlich
Sebastian B.
Philipp Spilger
Andreas Grübl
Benjamin Cramer
Milena Cz
Falk Ebert
Tobias Thommes
Akos Kungl
Dan Husmann
julian
Mihai A. Petrovici
Arne Emmel
Vitali Karasenko
Joscha Ilmberger
jakob
Alexander Nock
Oliver Breitwieser
Laura
Markus K
Elias Arnold
Korbi
Carl
Sebastian Schmitt