Brain-Inspired Computing

An Introduction to the Heidelberg Accellerated Analog Neuromorphic Hardware Architecture

BrainScaleS

A Platform for Bio-Inspired AI based on Hybrid Plasticity

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

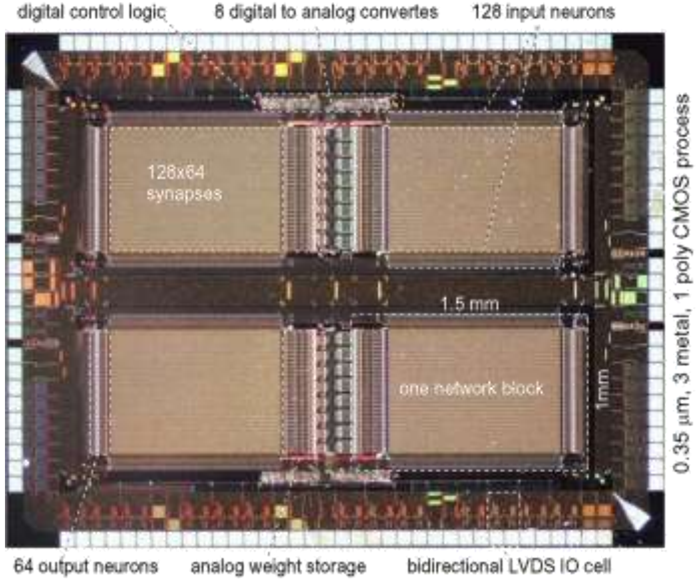
2000 Perceptron based analog neural networks:
EVOOPT and HAGEN

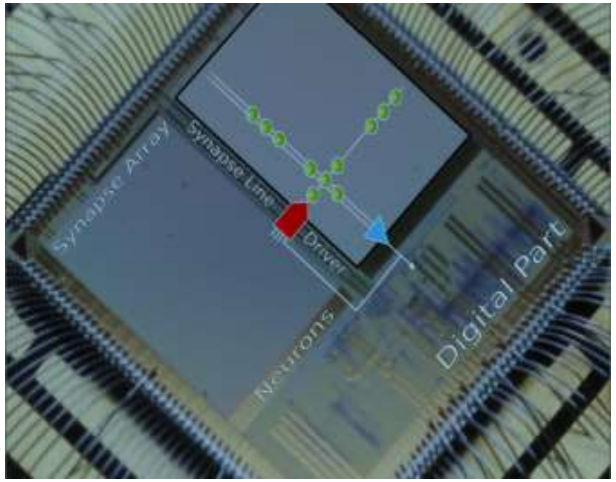
2003 First concepts for spike based analog neural networks

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:

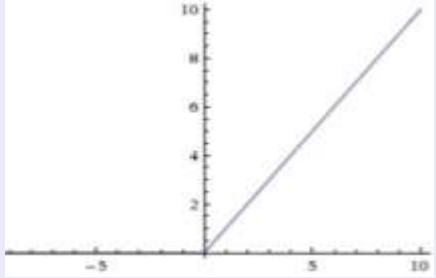
- fully-parallel Spike-Time-Dependent-Plasticity
- analog parameter storage for calibratable physical model

Perceptron model

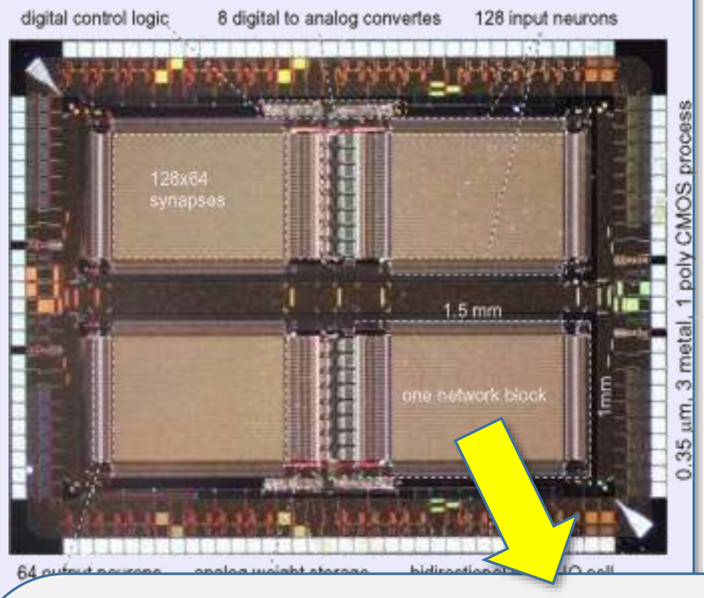
- used in Machine Learning
- vector-matrix multiplication

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

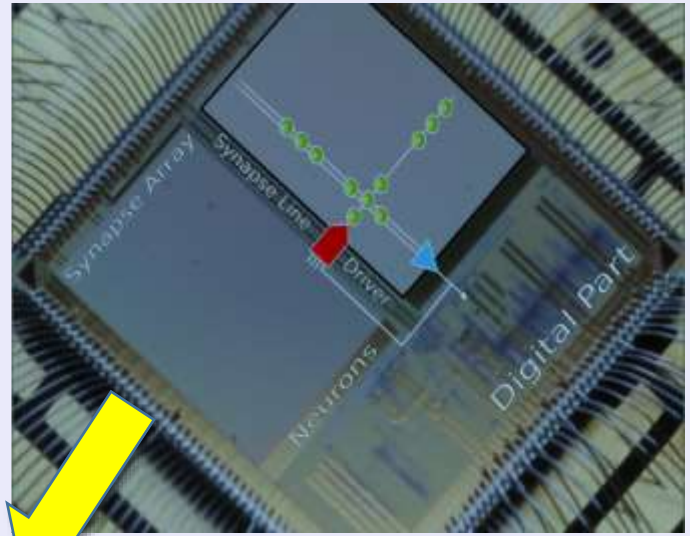
- simple non-linear activation function f (ReLU):



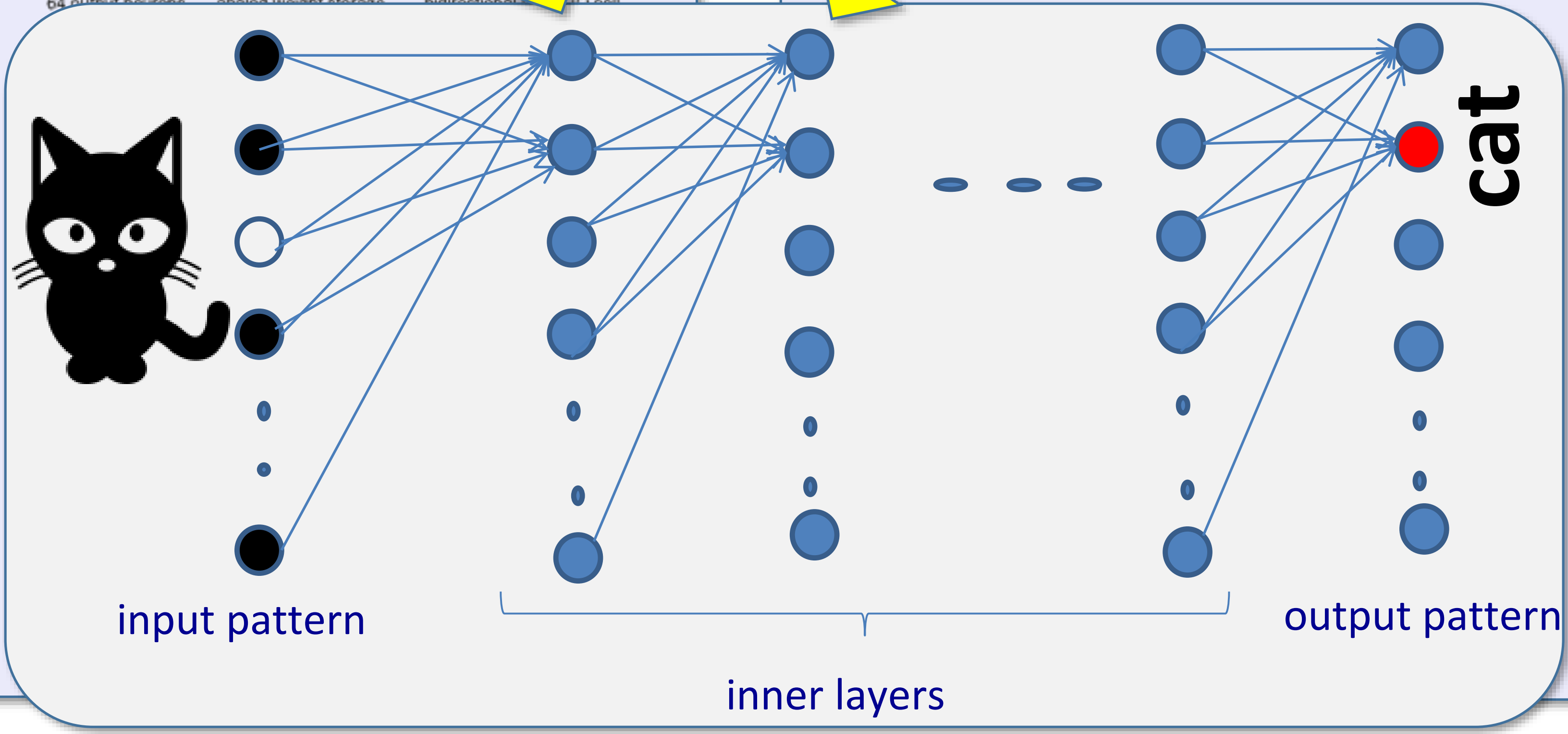
- trained with backpropagation



Spike-based model

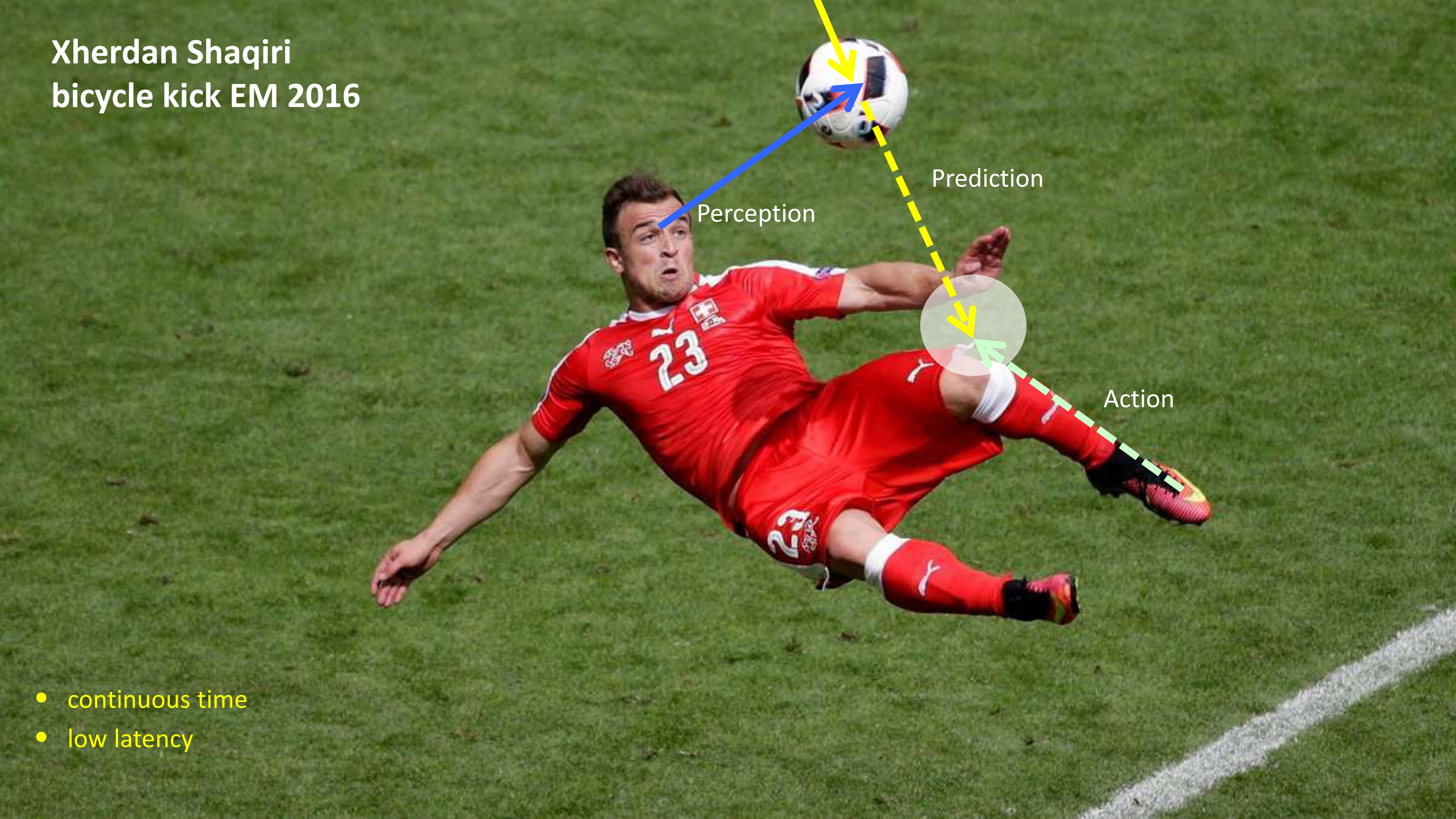


- time-continuous dynamical system
- vector-matrix multiplication
- complex non-linearities
- binary neuron output
- allows to model biological learning mechanisms



Xherdan Shaqiri
bicycle kick EM 2016

Xherdan Shaqiri
bicycle kick EM 2016



- continuous time
- low latency

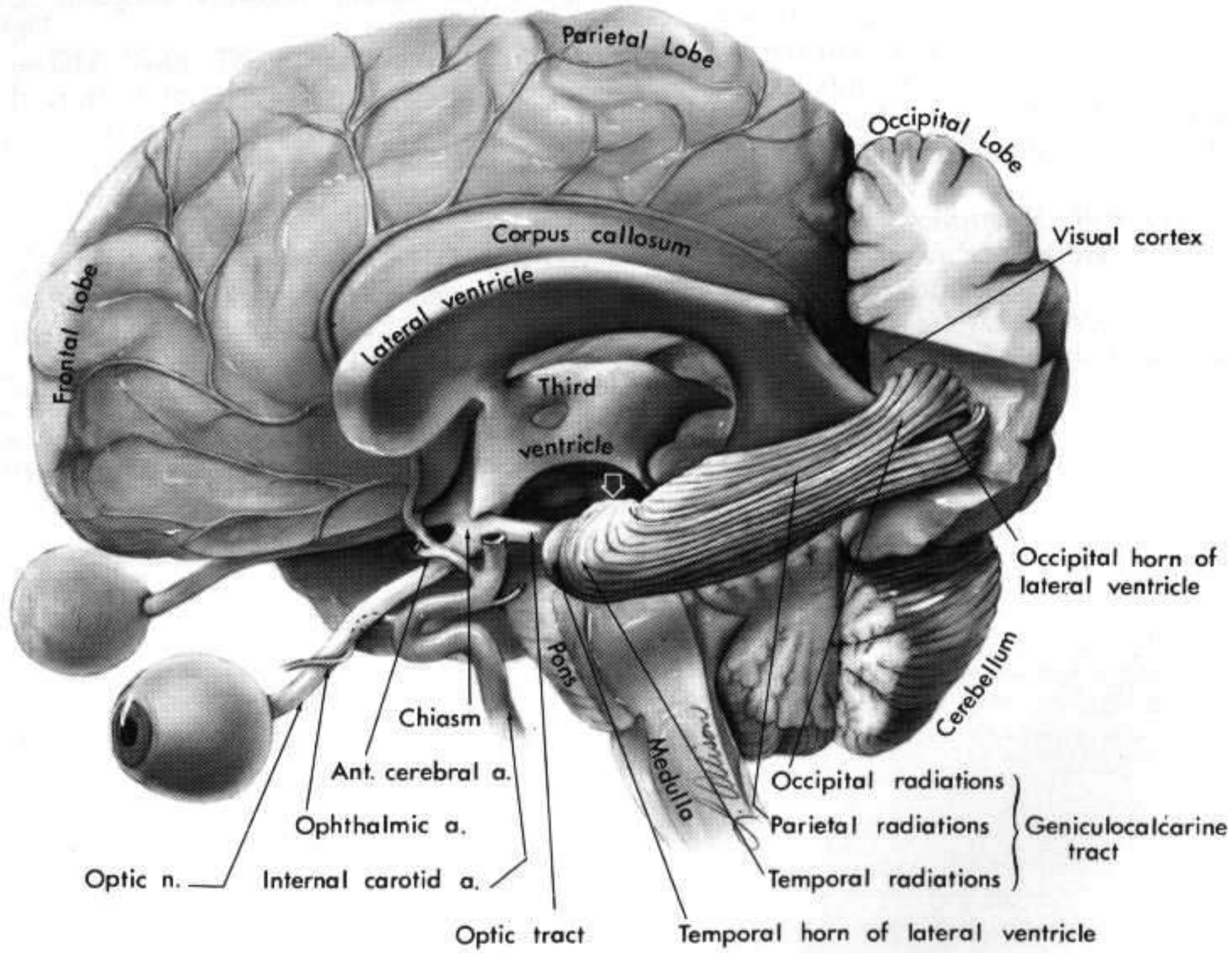

北
North
0
4


88:07
SUI
1-1
POL
SRF sport
容声冰箱


Xherdan Shaqiri
bicycle kick EM 2016
20 Watt
> 100 Watt
100 – 200 Milliseconds


Tots
FC

The human brain is the ultimate cognitive system



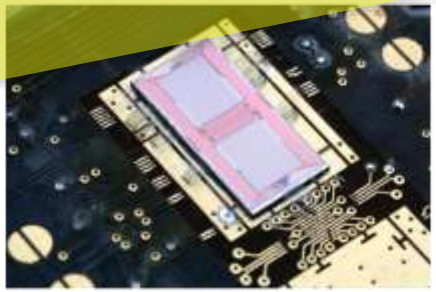
- 100 billion neurons
- 10000 connections per neuron (synapses)
- power consumption of the brain (approx.): 20 Watt

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

Subproject 9 of the HBP

Subproject Leader: Steve Furber

Deputy Leader: Johannes Schemmel

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

What is neuromorphic computing ?

What are
relevant aspects ?
Major research question !
Co-design process

Implement relevant aspects of
structure and function
of biological circuits
as analog or digital images
on electronics substrates

Structure

Cell Cores (Somas) - Networks (Axons and Dendrites) -
Connections (Synapses)

Function

Local Processing - Communication - Learning

Brain-Inspired Computing

Bio-inspired artificial intelligence (Bio-AI)

future computing based on biological information processing

understanding biological information processing

need model system to test ideas

neuromorphic computing :
model networks of neurons and synapses

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**

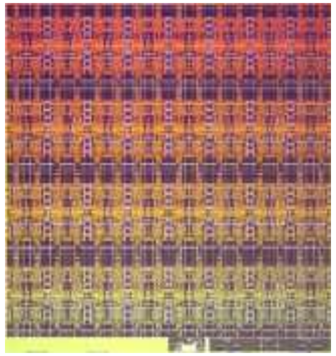
Neuromorphic systems worldwide – State-of-the-art and complementarity

SpiNNaker

TrueNorth

Loihi

BrainScaleS

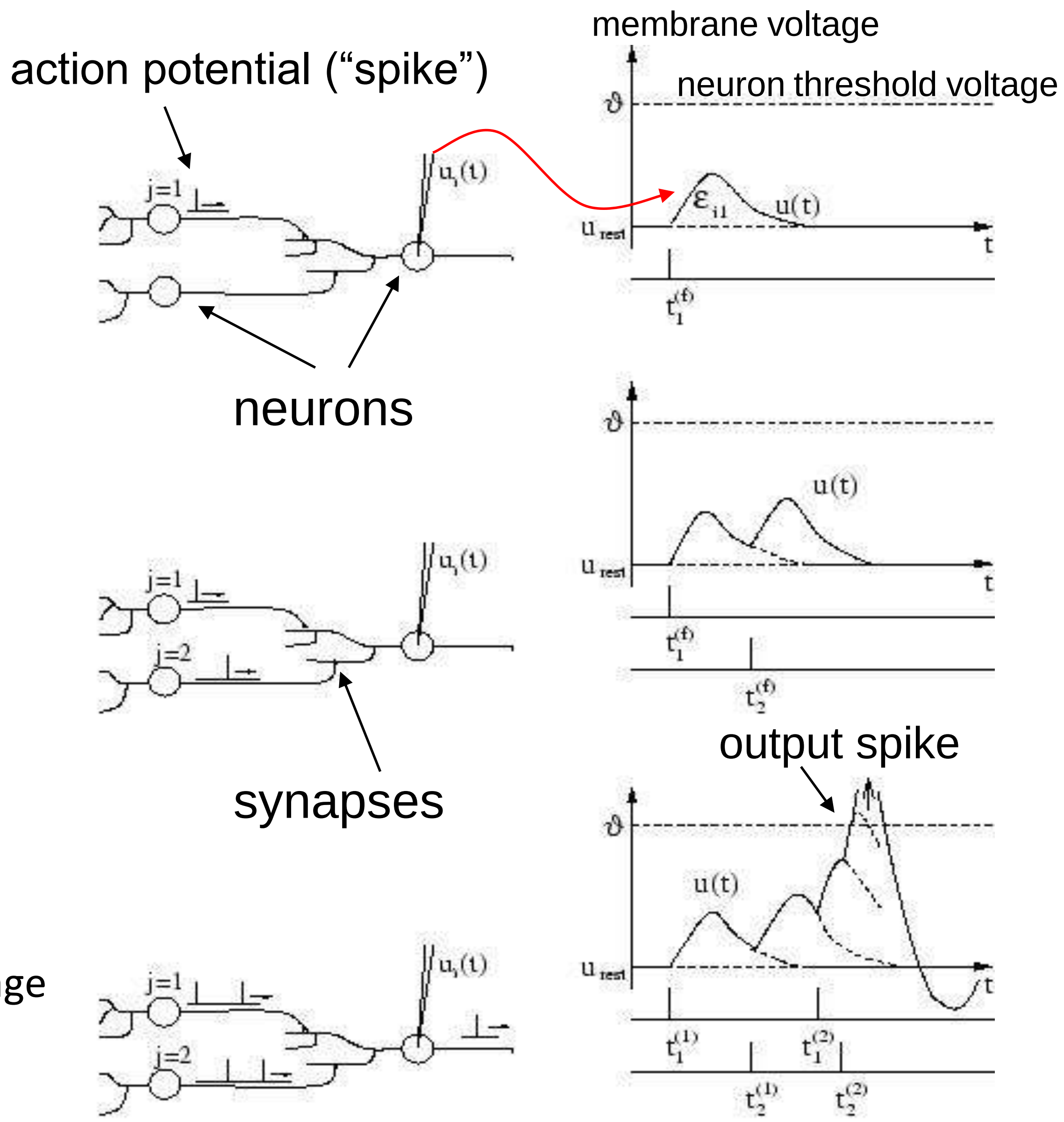


Biological realism

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

Principles of neural communication

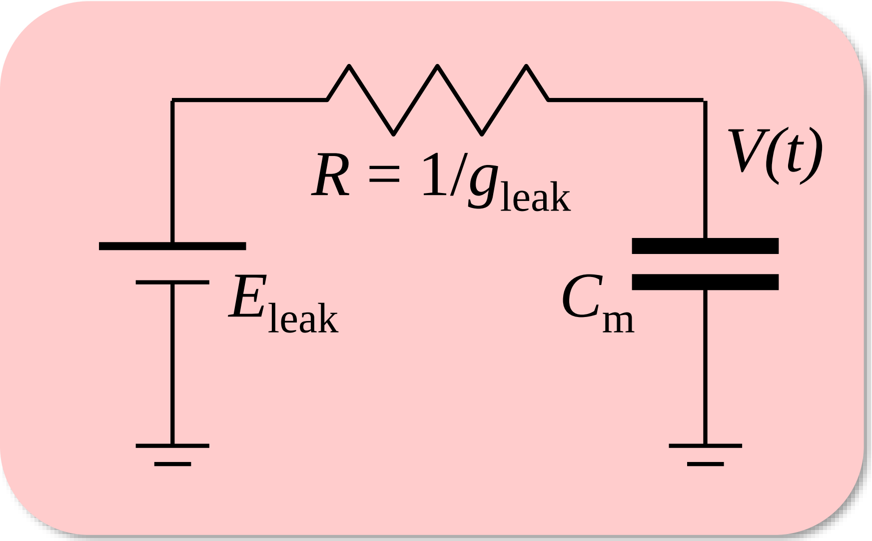


- neurons integrate over space and time
- temporal correlation is important
- kind of mixed-signal system: action potential ↔ membrane voltage
- fault tolerant
- low power consumption → 100 Billion neurons: 20 Watts

BrainScaleS : Neuromorphic computing with physical model systems

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

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

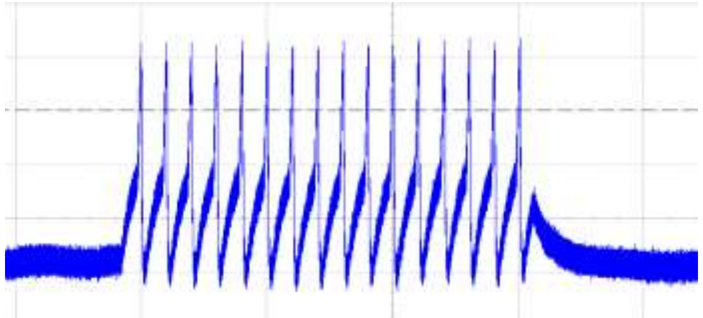


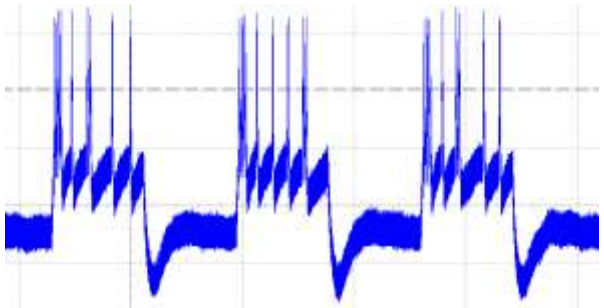
$$\left.\frac{dV}{dt}\right|_{bio} \ll \left.\frac{dV}{dt}\right|_{VLSI}$$

→ accelerated neuron model

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

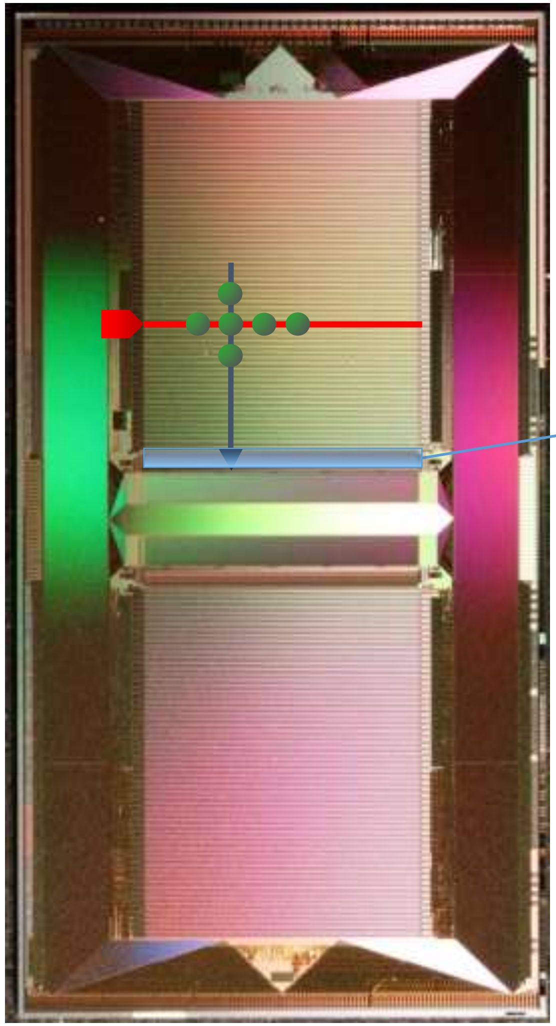
no multiplexing of components storing model variables
- each neuron has its membrane capacitor
- each synapse has a physical realization

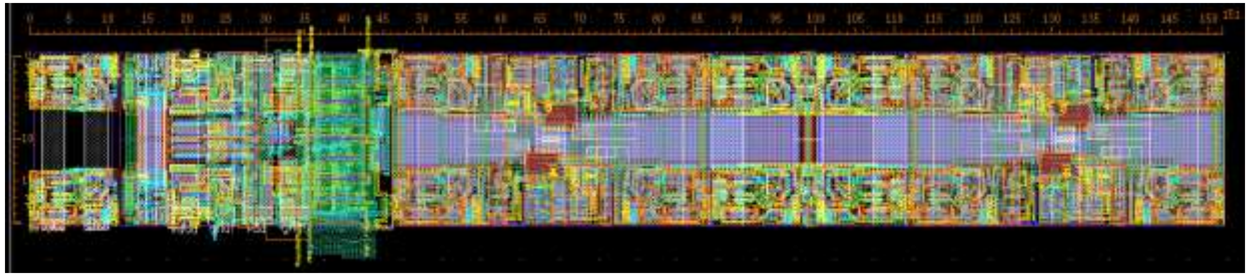


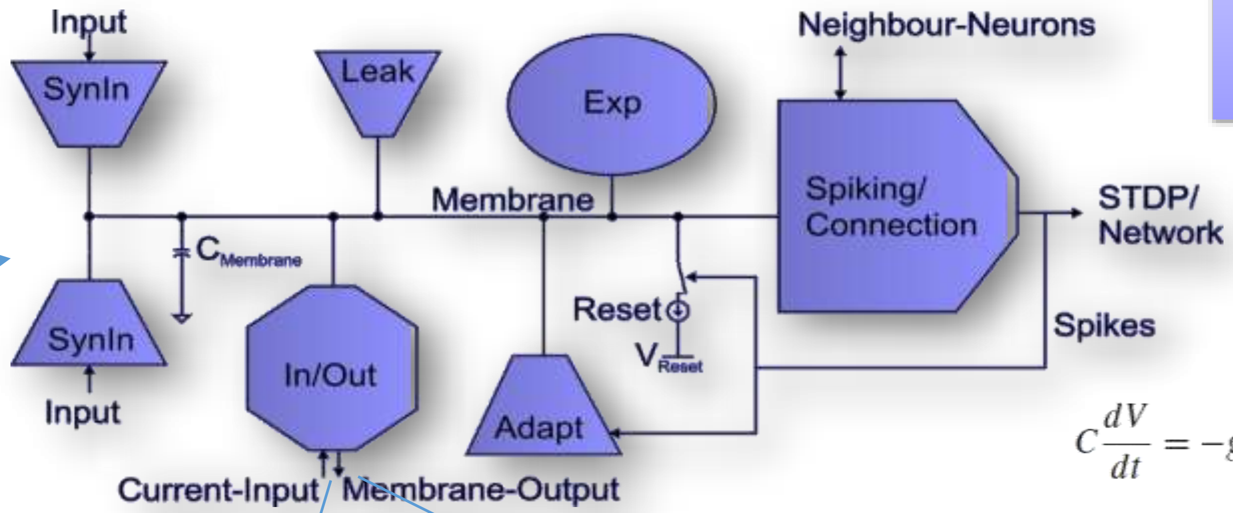


Structure of BrainScaleS neurons: array of parameterized dendrite circuits

photograph of the BrainScaleS 1 neuromorphic chip



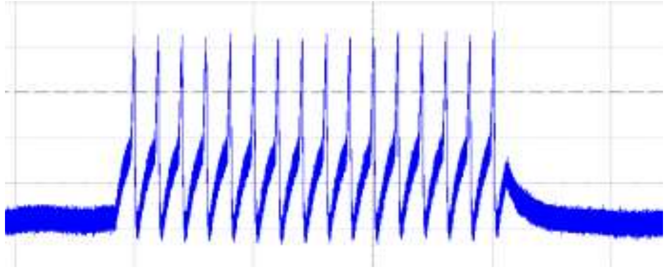


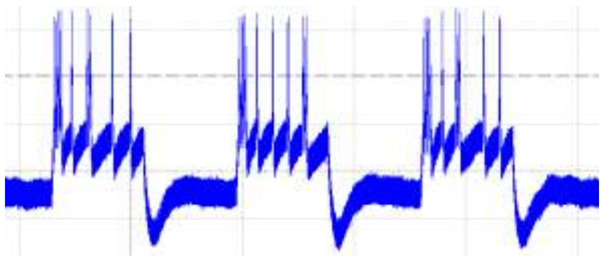


- 180 nm (generation 1) or 65 nm (gen. 2)
- 24 calibration parameters per neuron
- modular structure
- full set of ion-channel circuits for each dendrite

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

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

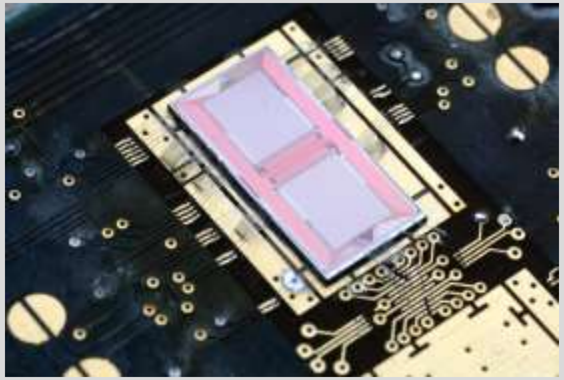




TimeScales

	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			

BrainScaleS-1 multi-level architecture

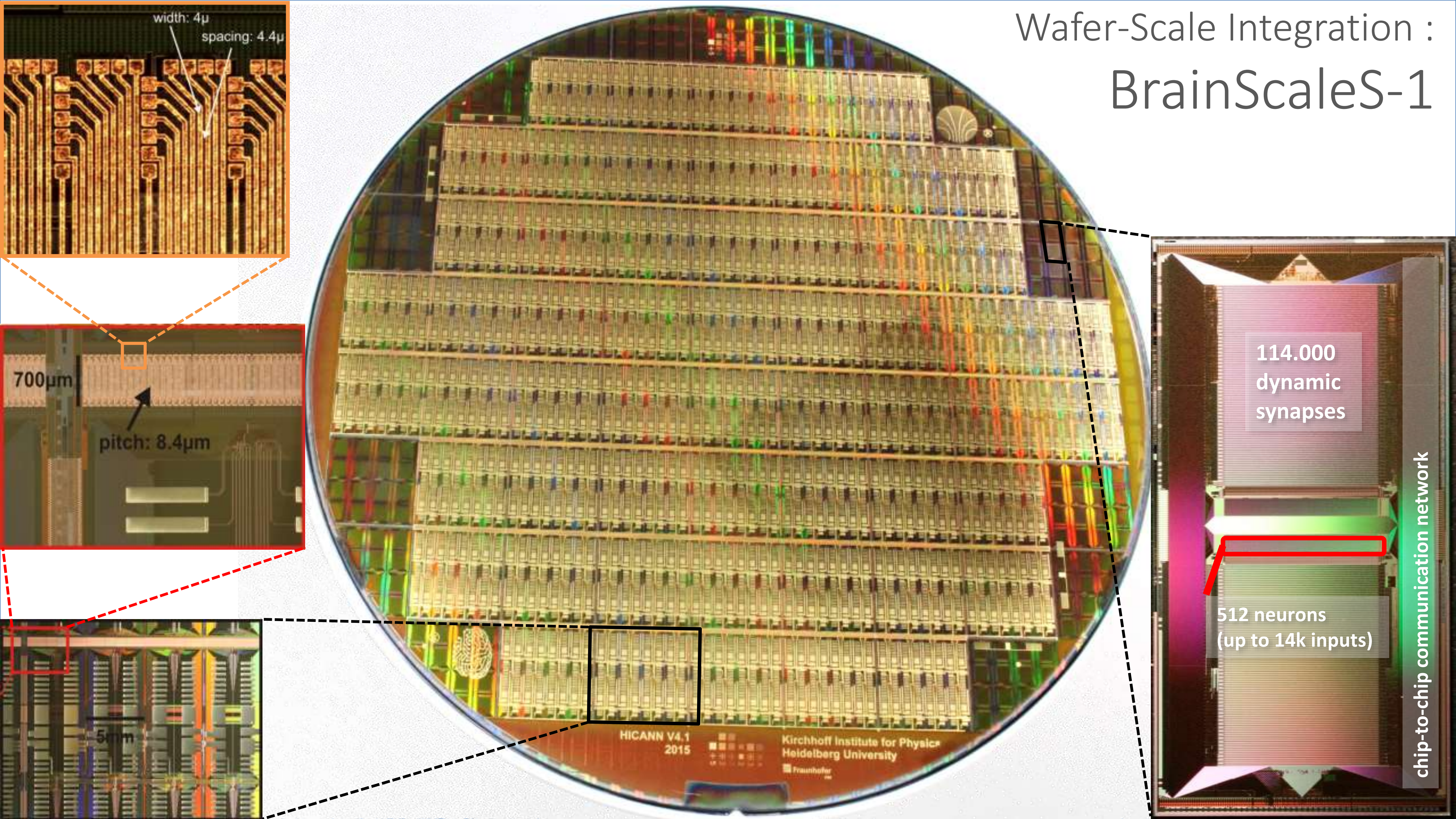


single chip | wafer module | hybrid system

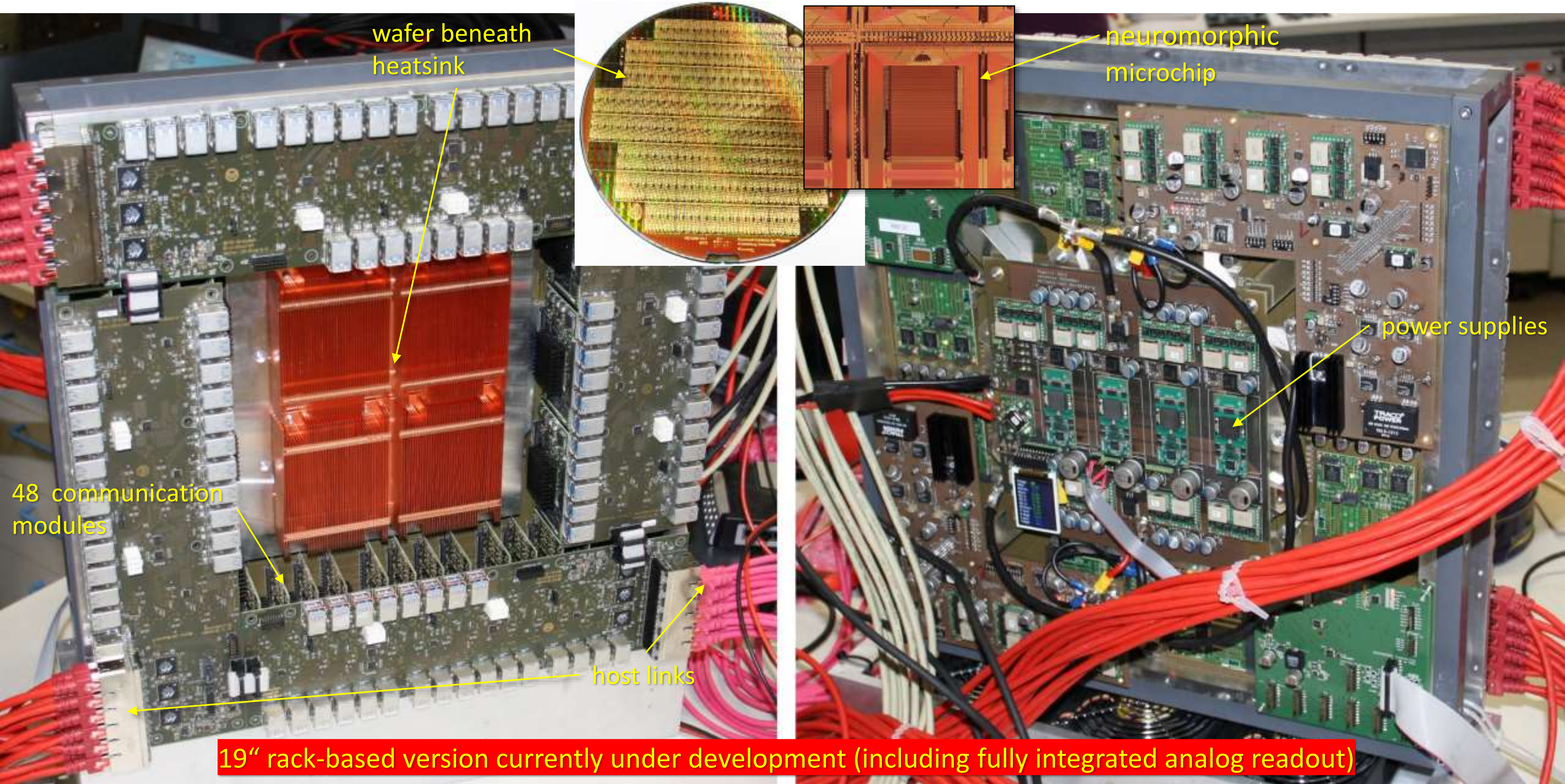
BrainScales-1 introduced for the first time

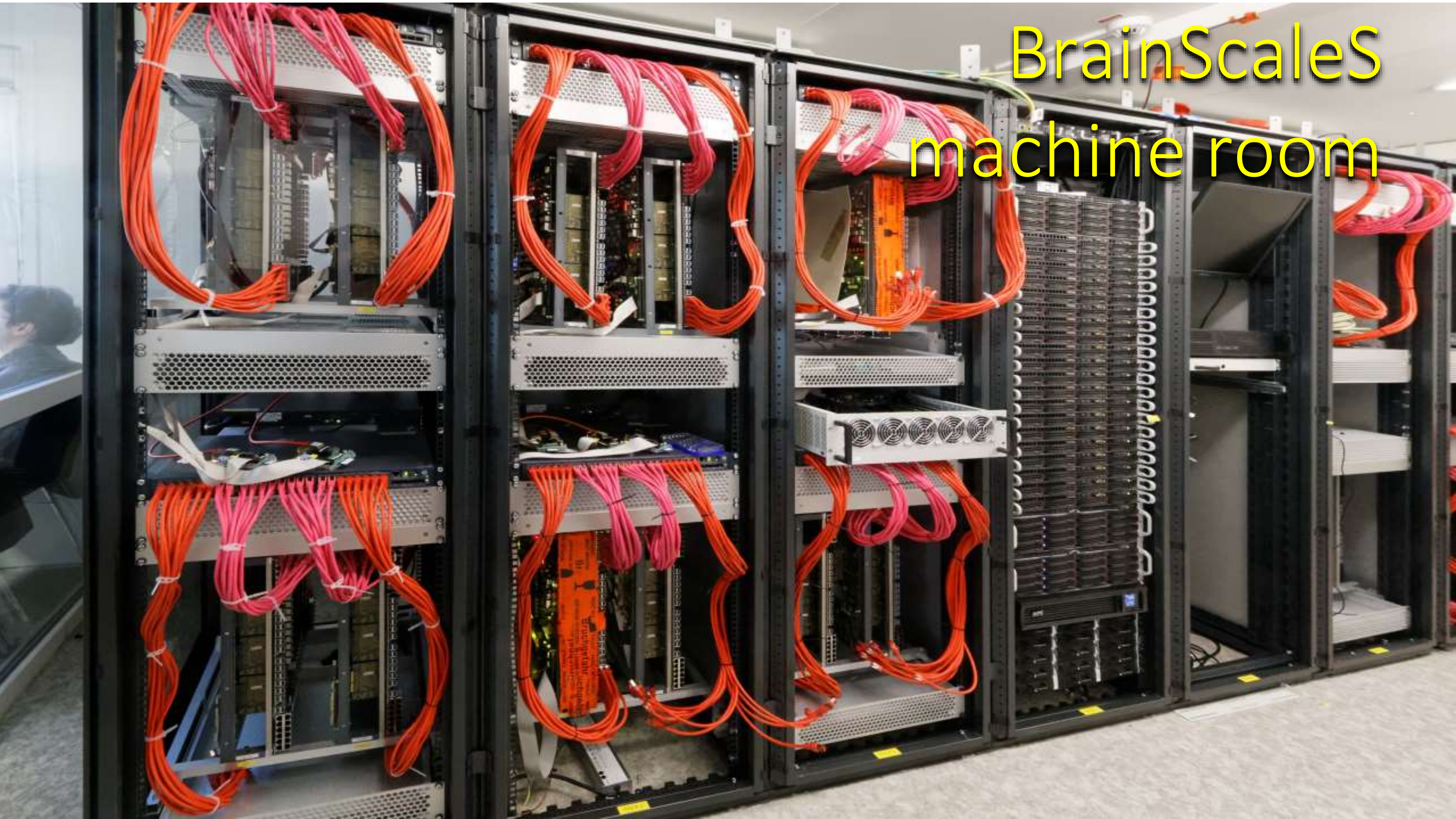
- Accelerated (x10.000) mixed-signal implementation of spiking neural networks
- AdEx neurons with very high synaptic imput count (> 10k)
- Wafer-scale event communication

Wafer-Scale Integration :
BrainScaleS-1



Wafer Module




BrainScaleS
machine room

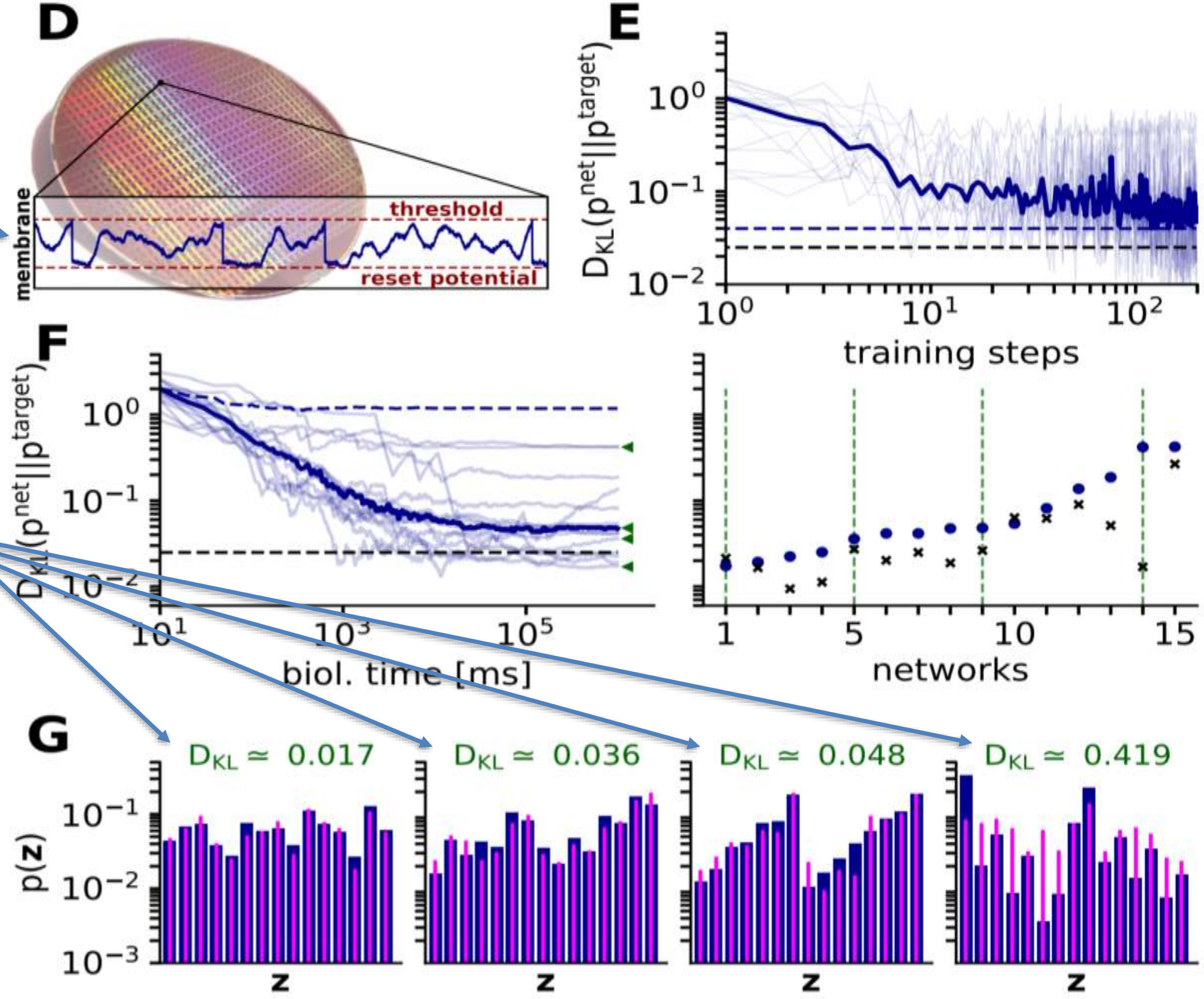
Stochastic model example: sampling from multiple neural Boltzmann machines

analog neurons

autonomously reproduce learned distributions

no software!

?



D. Dold et.al., https://arxiv.org/abs/1809.08045v1

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

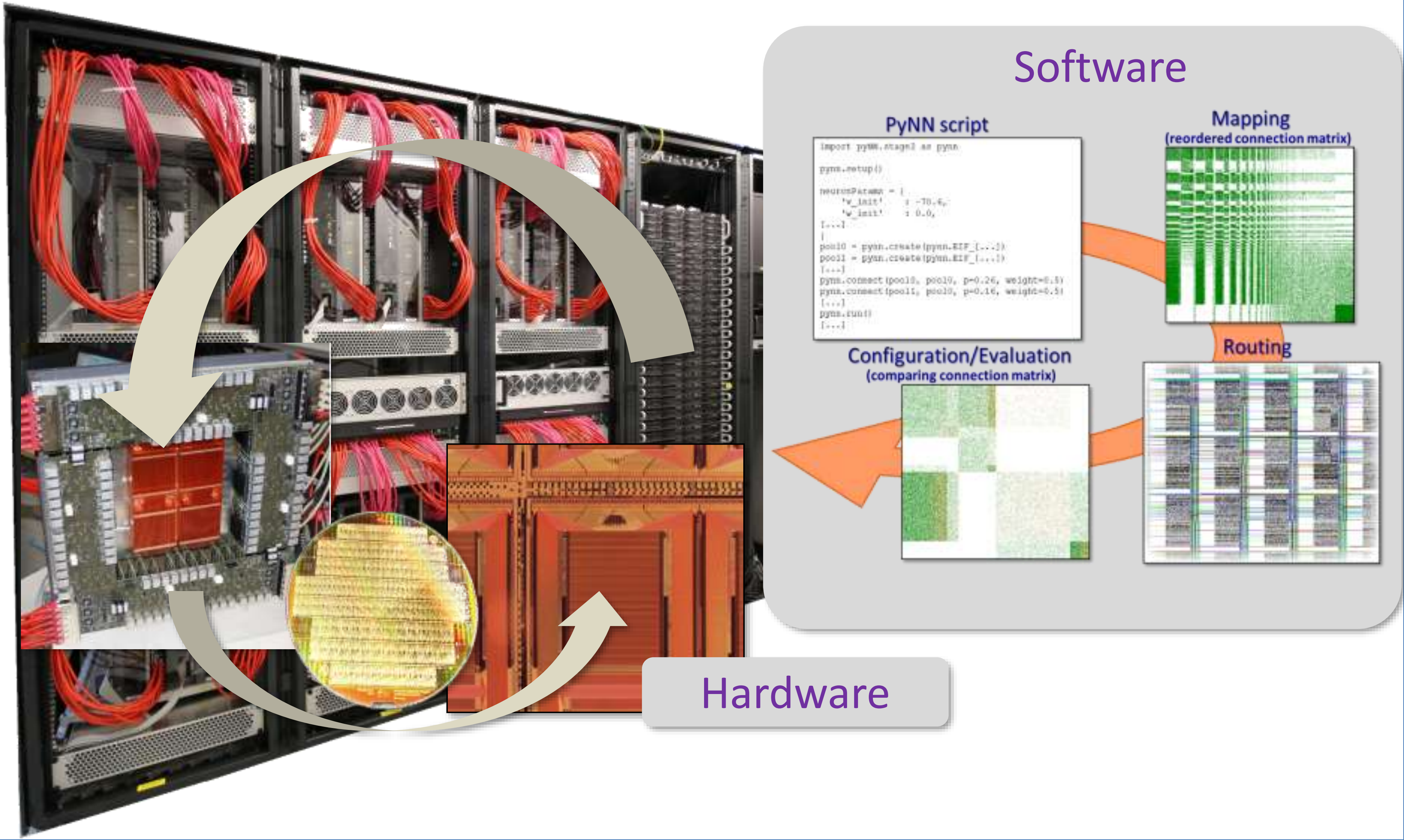
after training:

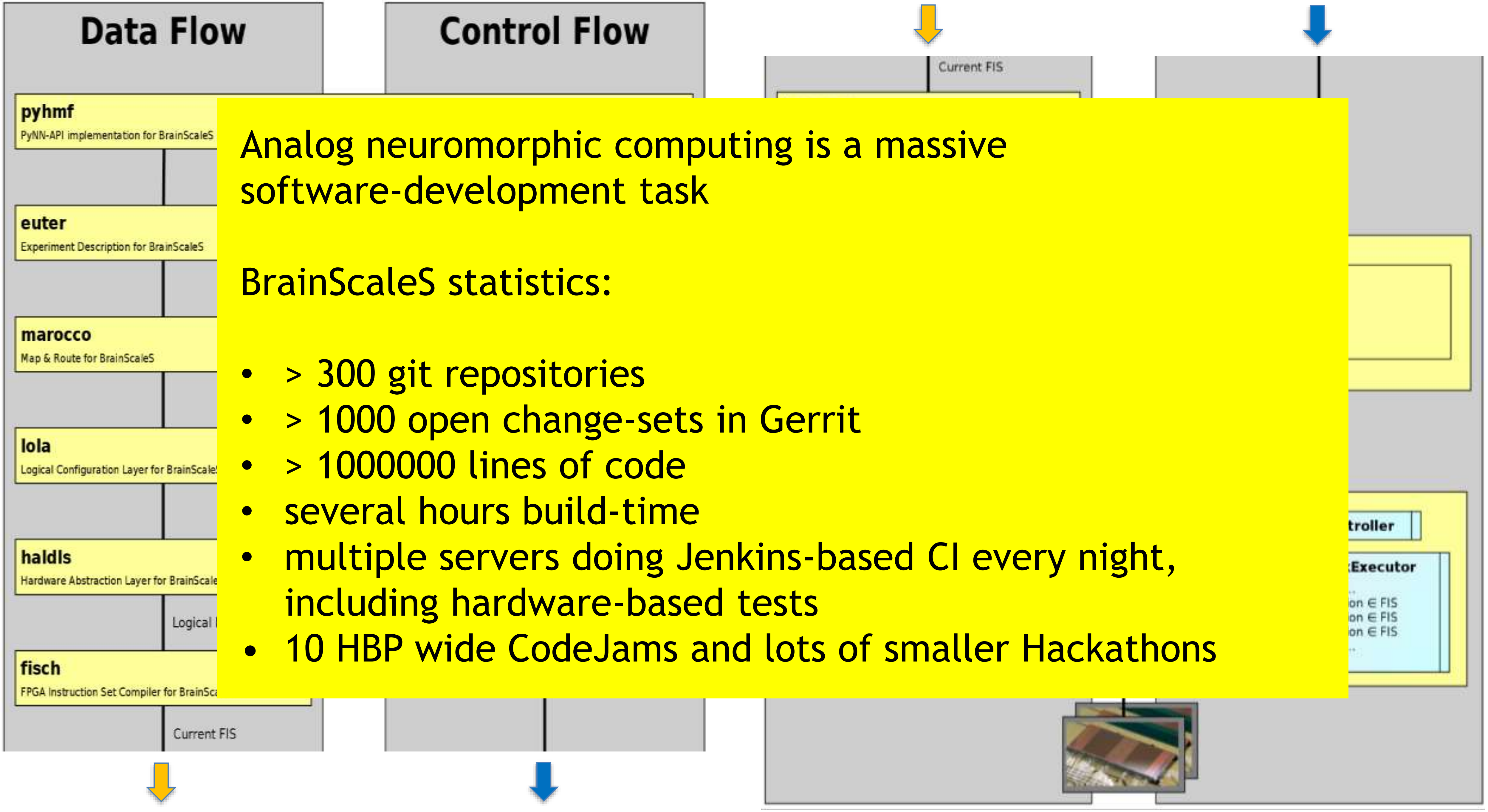
Non-Turing physical computing system performing autonomously

but

Turing-based computing is used in multiple places:

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



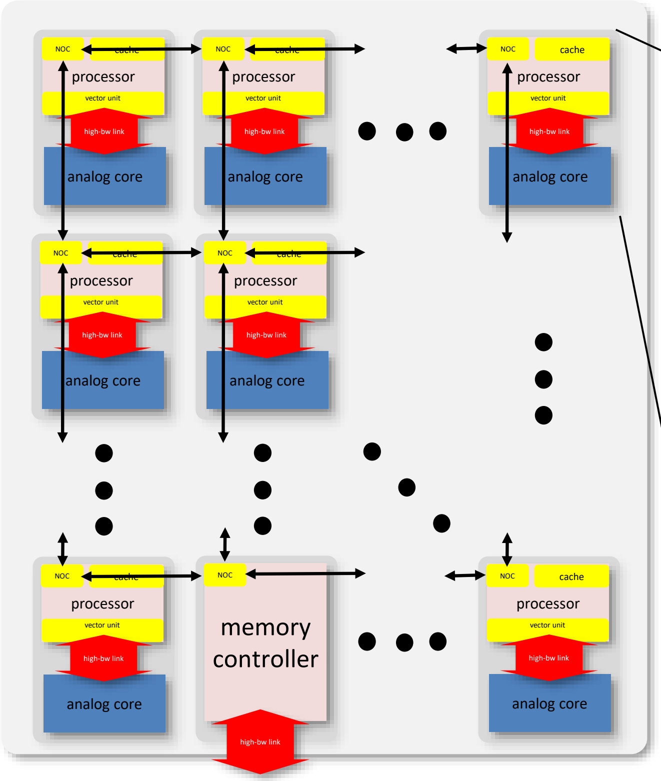


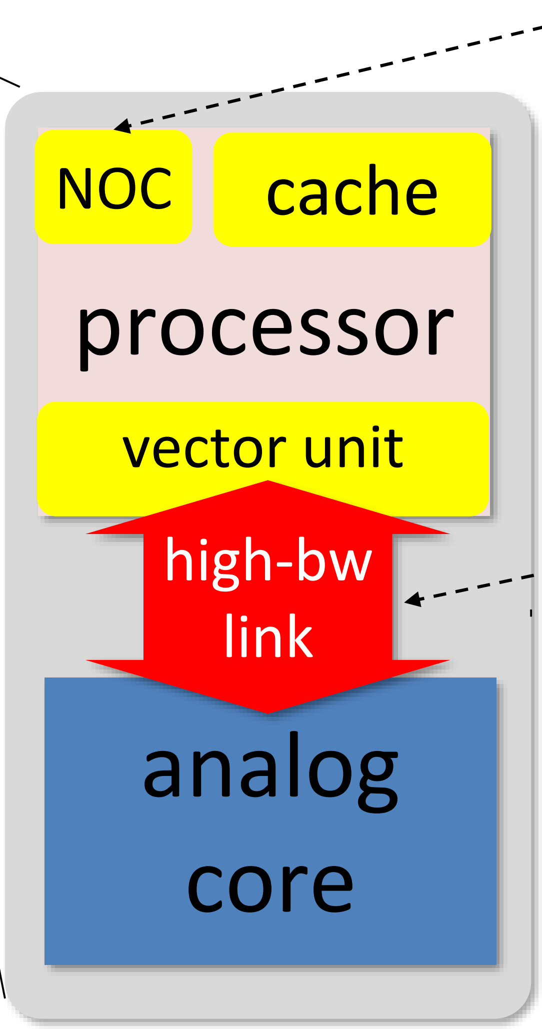
Analog neuromorphic computing is a massive software-development task

BrainScaleS statistics:

- > 300 git repositories
- > 1000 open change-sets in Gerrit
- > 1000000 lines of code
- several hours build-time
- multiple servers doing Jenkins-based CI every night, including hardware-based tests
- 10 HBP wide CodeJams and lots of smaller Hackathons

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:

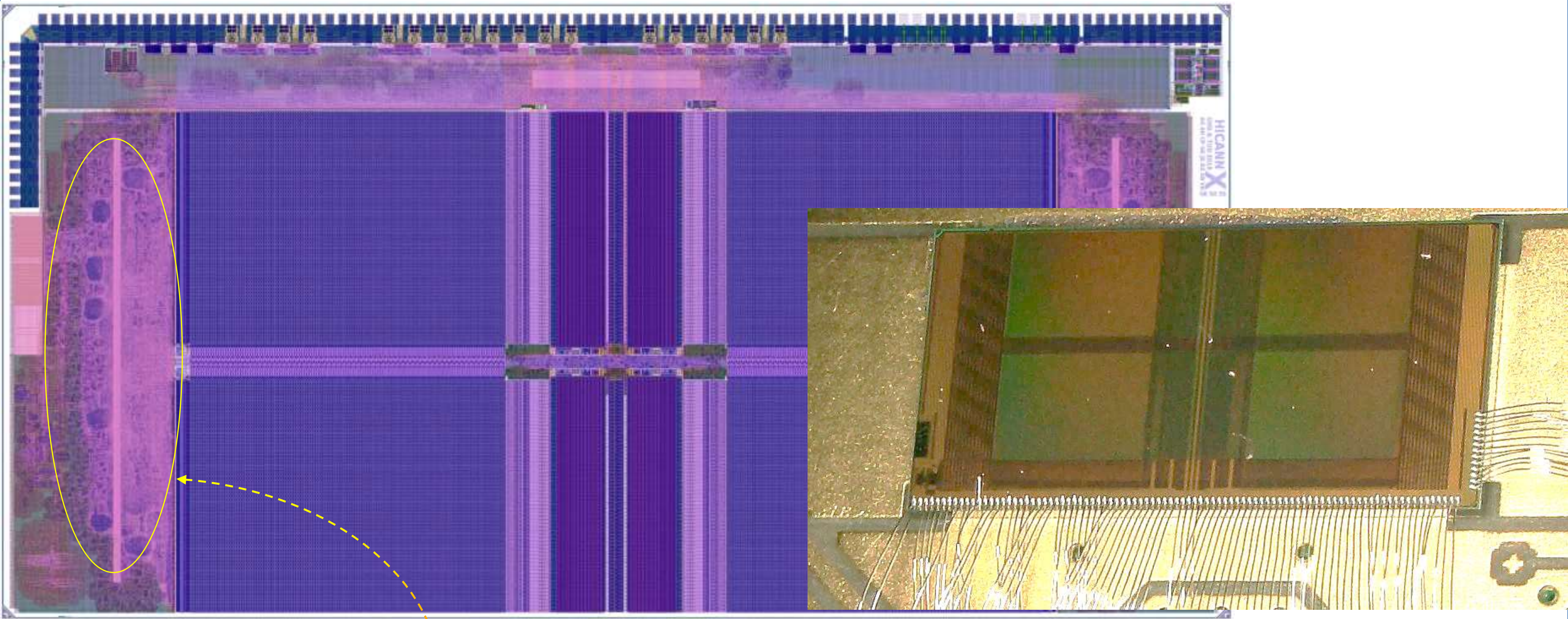
vector unit ←→ NM core

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

special function tile:

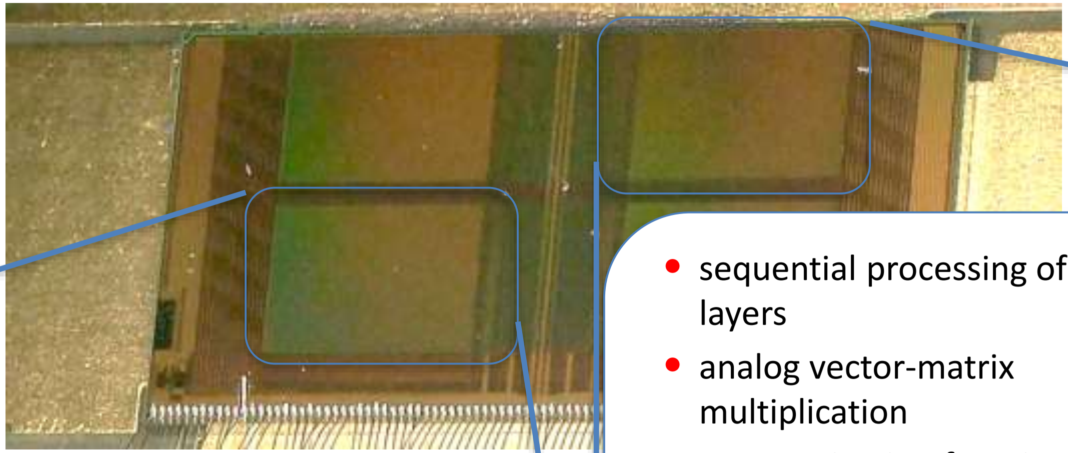
- memory controller
- SERDES IO
- purely digital function unit

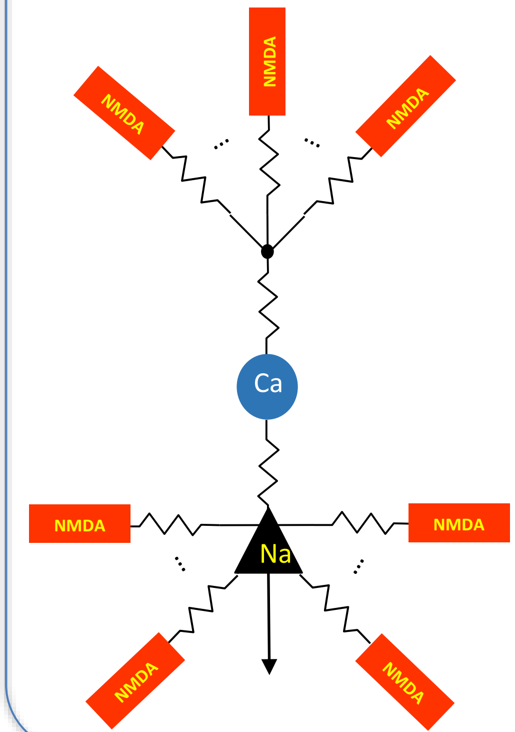
BrainScaleS-2 (BSS-2) ASIC

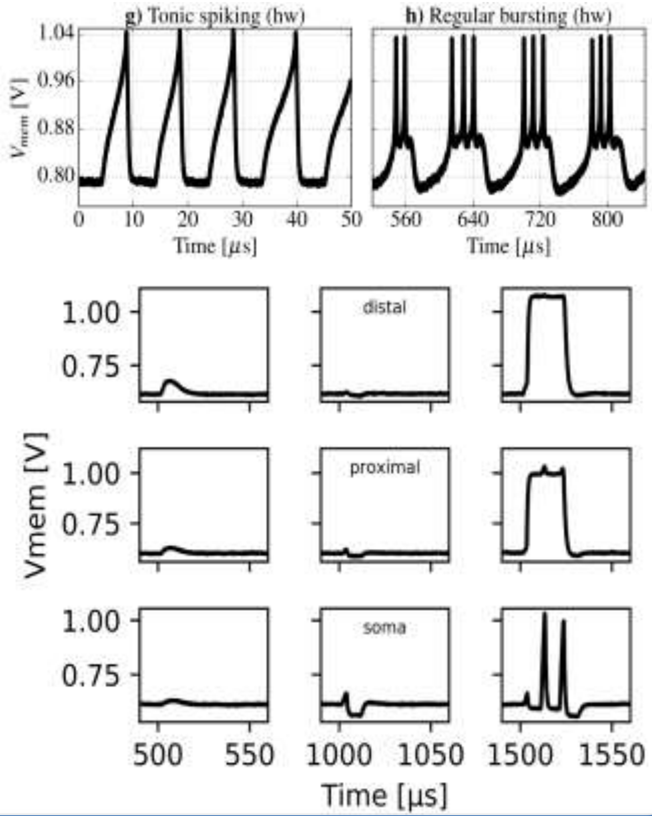


- 65nm LP-CMOS, power consumption O(10 pJ/synaptic event)
- 128k synapses
- 512 neural compartments (Sodium, Calcium and NMDA spikes)
- two SIMD plasticity processing units (PPU)
- PPU internal memory can be extended externally
- fast ADC for membrane voltage monitoring
- 256k correlation sensors with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

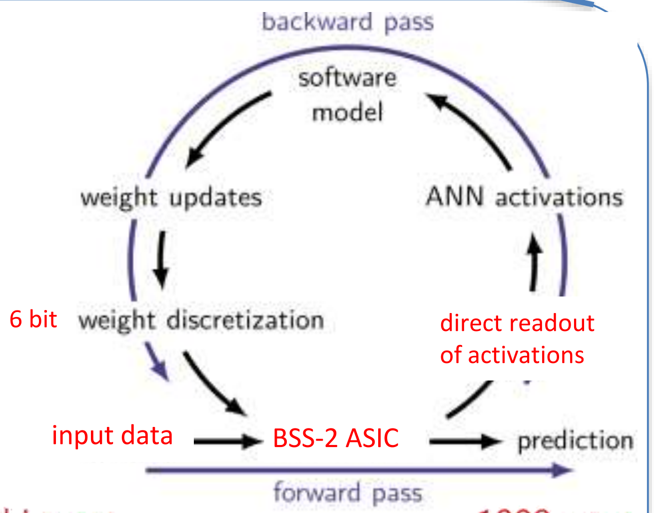
BrainScaleS-2 supports spike-based and Perceptron operation simultaneously

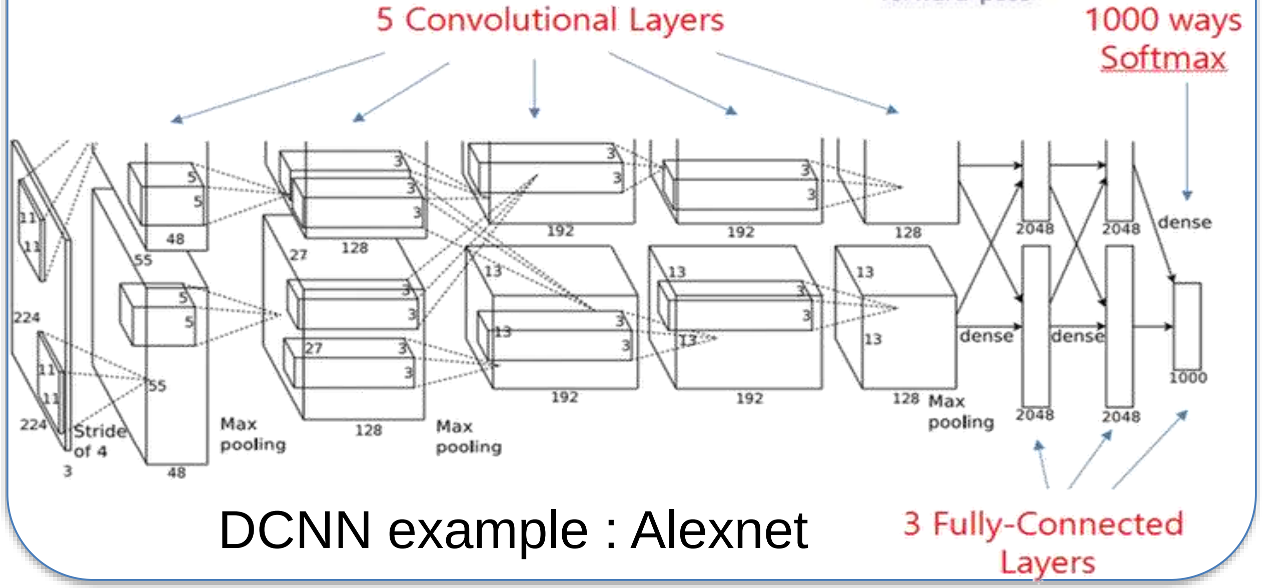






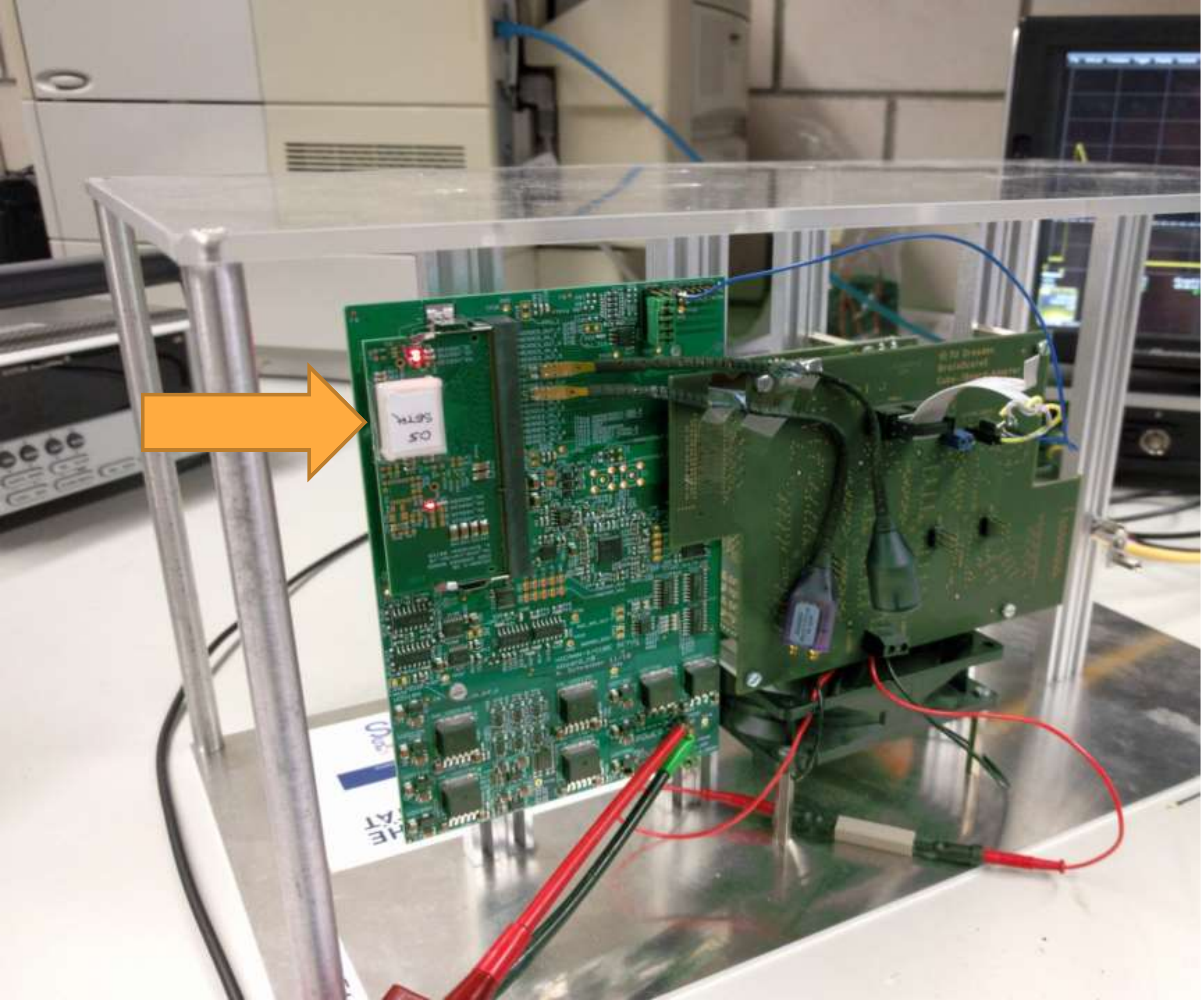
- sequential processing of all layers
- analog vector-matrix multiplication
- ReLU activation function with 4 to 8 bit resolution
- speed mostly limited by external memory





DCNN example : Alexnet

BrainScaleS-2

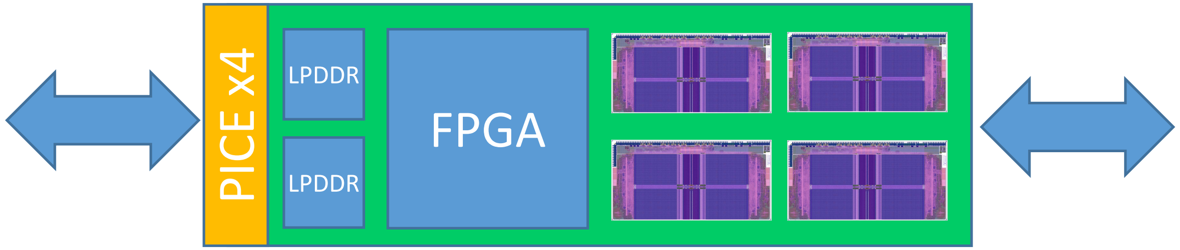


- 8Gbit/s raw bandwidth between BSS ASIC and host
- Latency < 300ns
- Event rates up to 250MHz real-time (250kHz bio) full duplex

Outlook : Edge-computing with BrainScaleS

M2 standard with PCIE x4
- backplanes and server
- pre-processed sensor data
- NM accelerators

M2 formfactor



BrainScaleS ASICs

event-based direct IO
- neuromorphic detectors
- neuromorphic sensors
  - event-based cameras
  - bio-sensors
  - etc

network structure :

feature extraction with pretrained DCNNs

classification : - spike based with online learning
- activation based (pretrained)

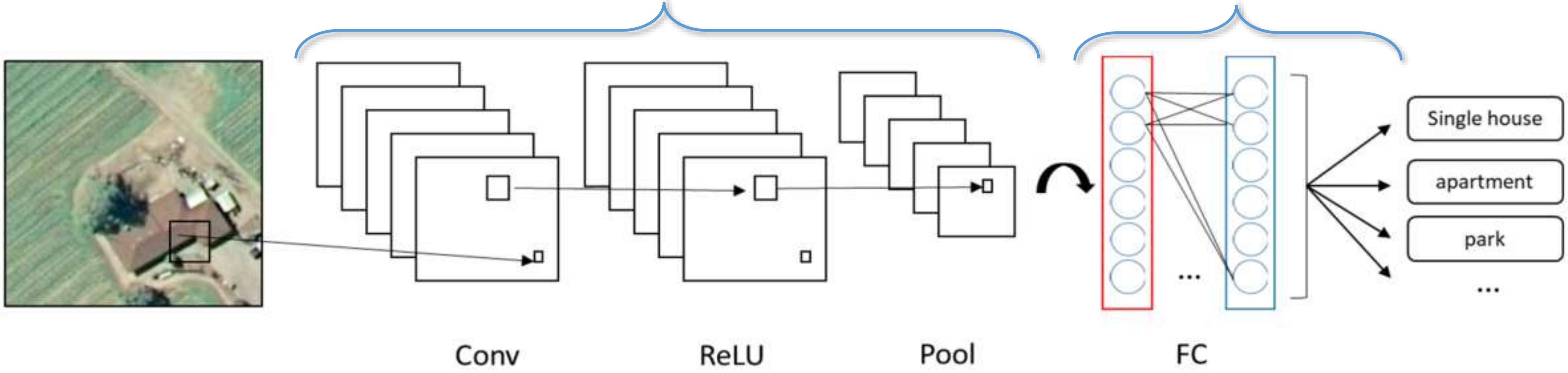
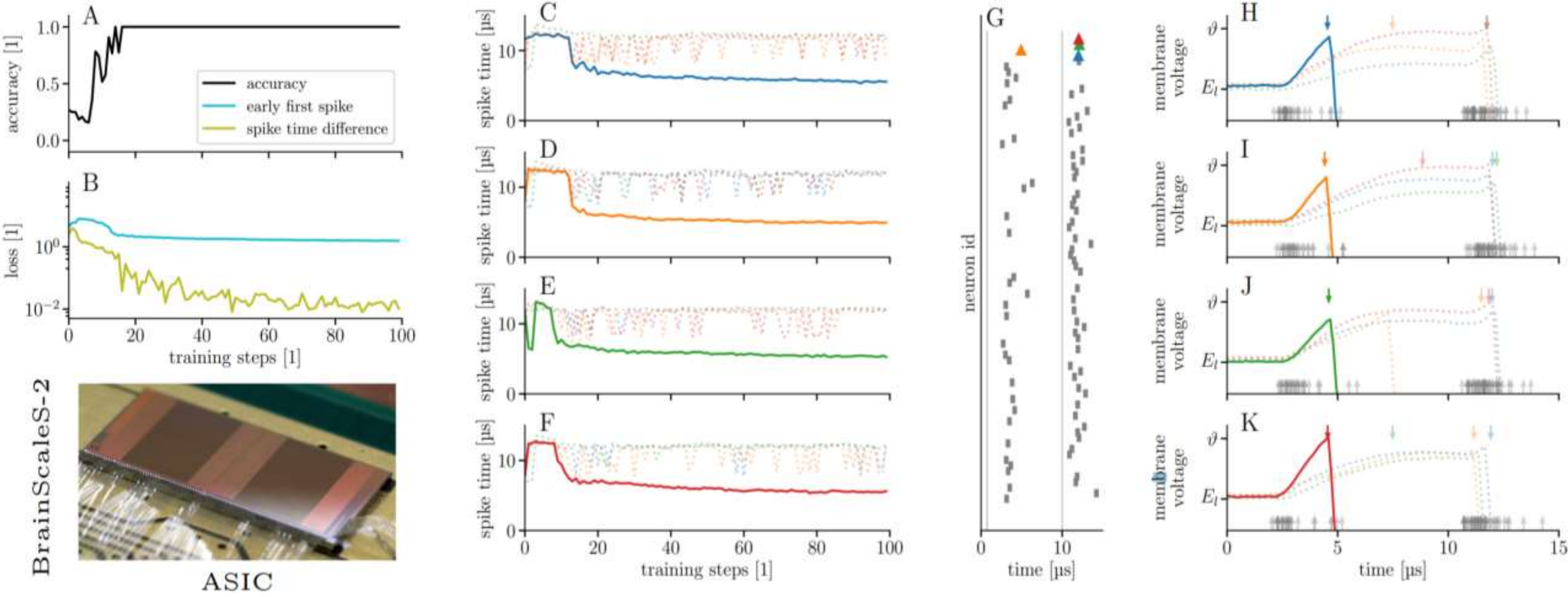
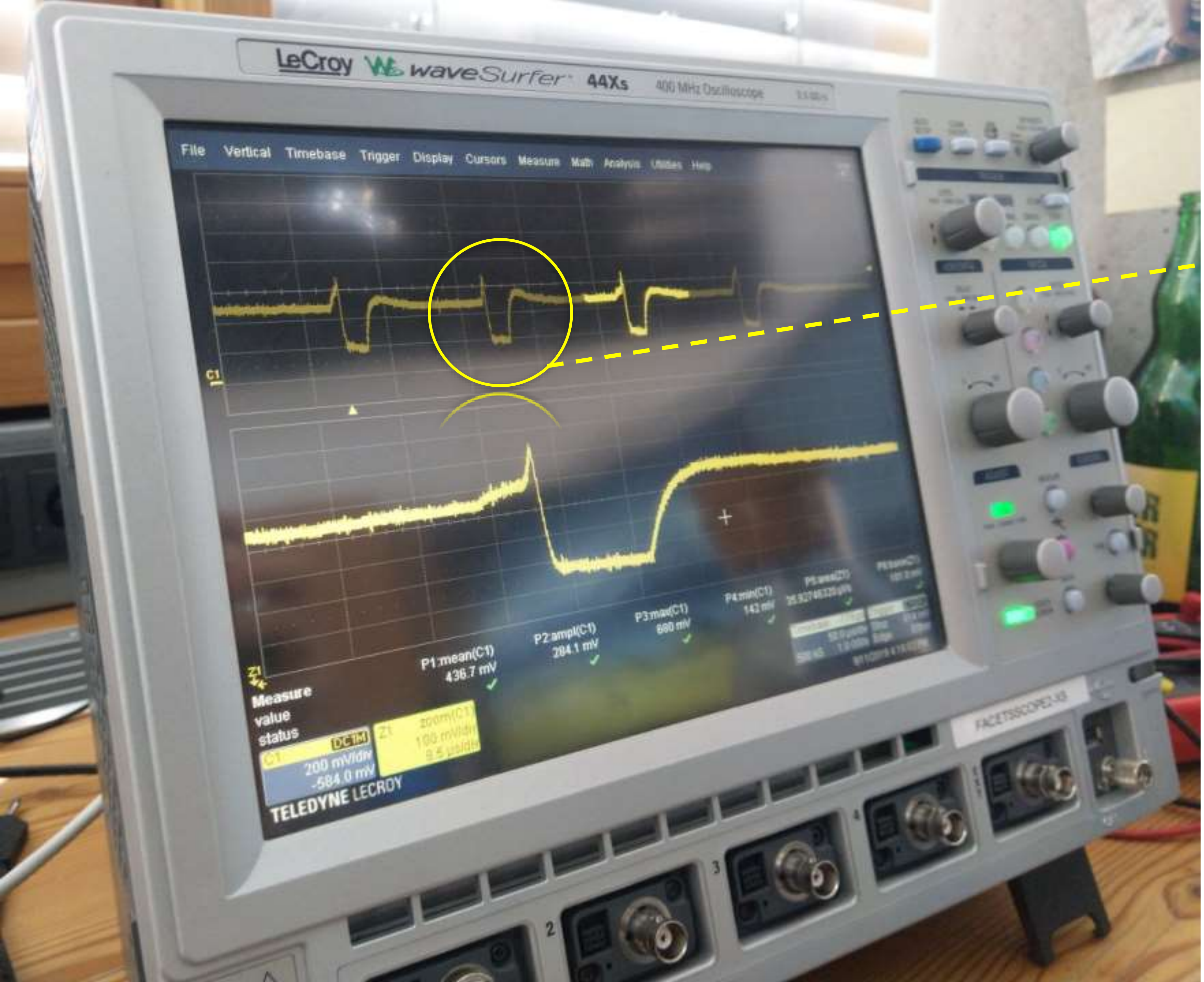
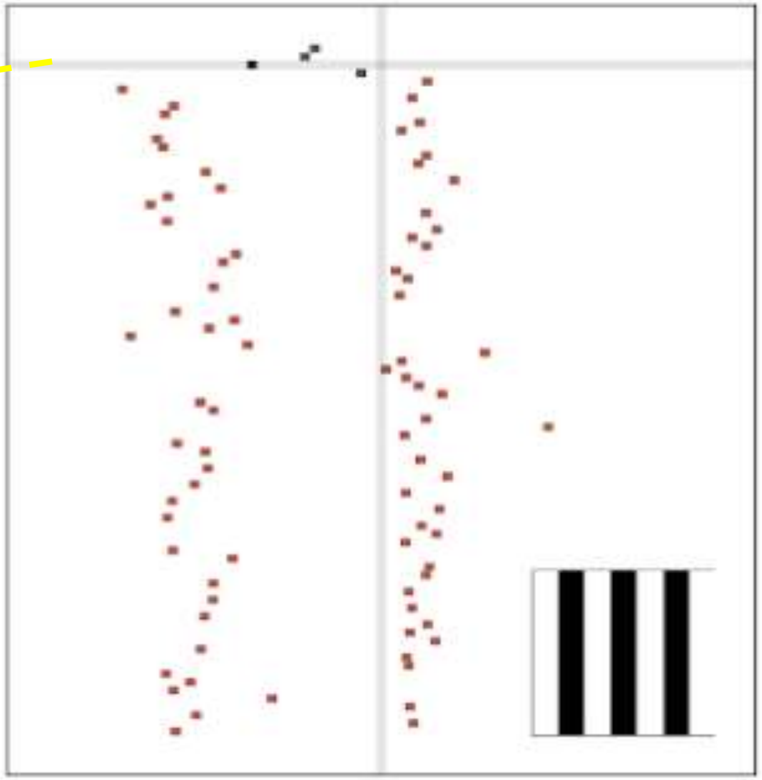


image from C. Cao, https://doi.org/10.3390/environments6020025

Training deep networks with time-to-first-spike coding



J. Goeltz et. al, *„Fast and deep neuromorphic learning with time-to-first-spike coding“*, arXiv:1912.11443


LeCroy
waveSurfer 44Xs
400 MHz Oscilloscope
File
Vertical
Timebase
Trigger
Display
Cursors
Measure
Math
Analysis
Utilities
Help
C1
Measure
value
status
P1:mean(C1)
436.7 mV
P2:ampl(C1)
284.1 mV
P3:max(C1)
580 mV
P4:min(C1)
143 mV
C1
DC1M
200 mV/div
-584.0 mV
Z1
zoom(C1)
100 mV/div
TELEDYNE LECROY


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

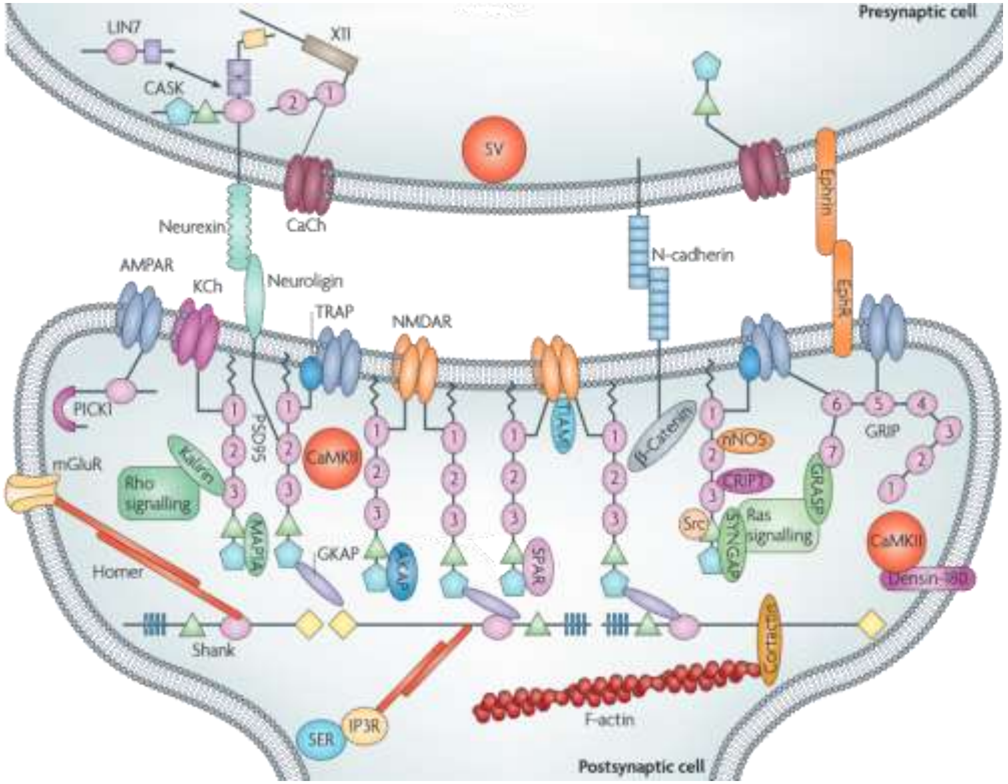
Problem:
how to fix millions of parameters
- network topology
- neuron sizes and parameters
- synaptic strengths

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

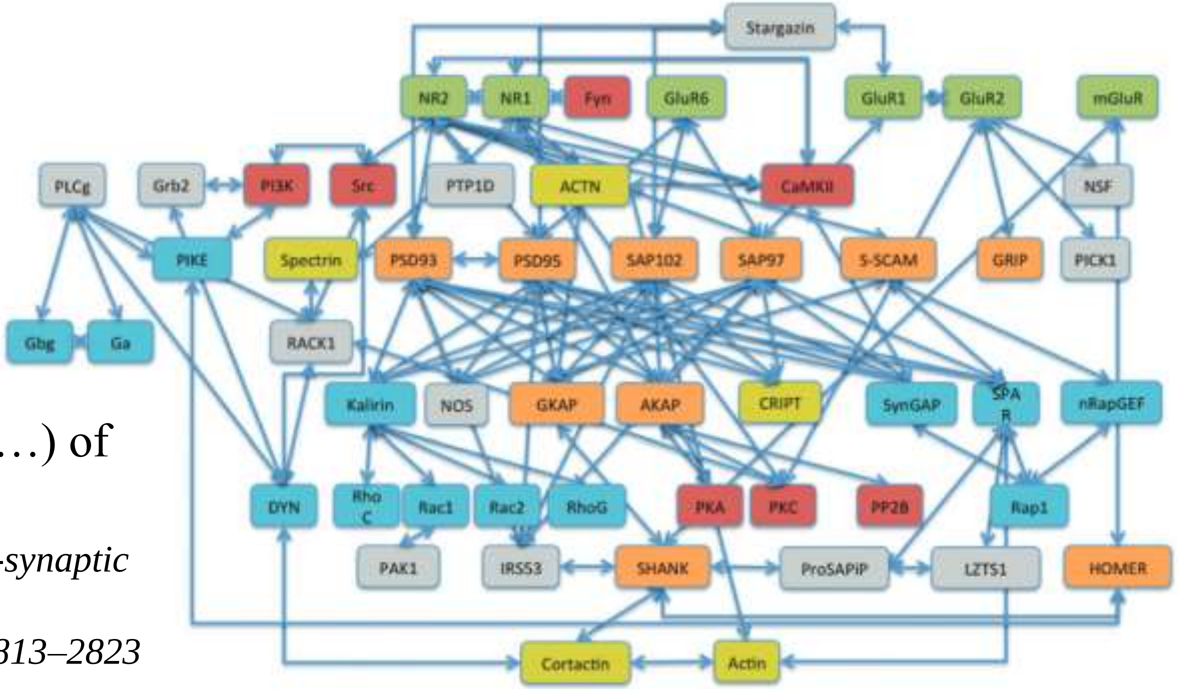
Complexity of synaptic plasticity is key to biological intelligence



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

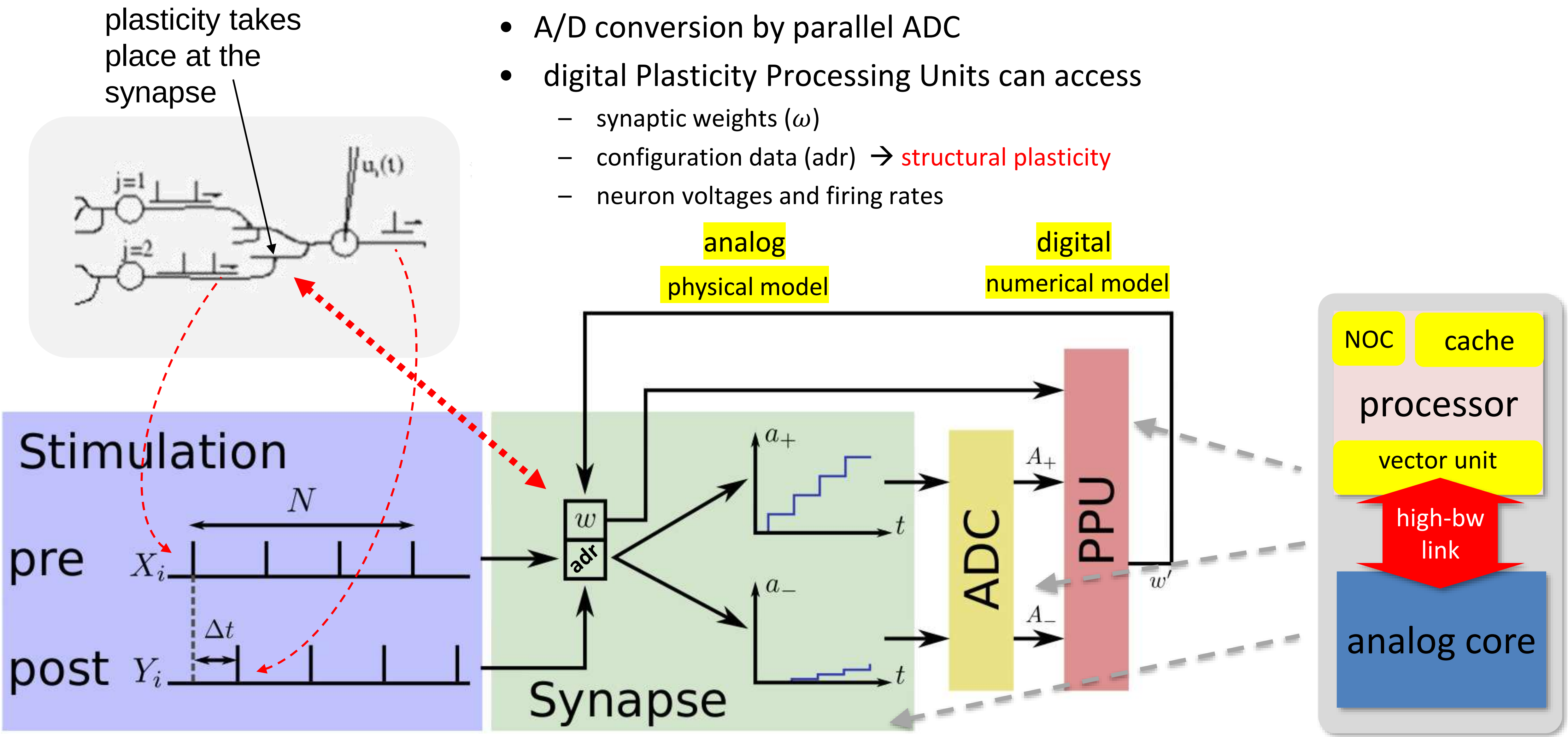


Protein-protein interaction map (…) of post-synaptic density

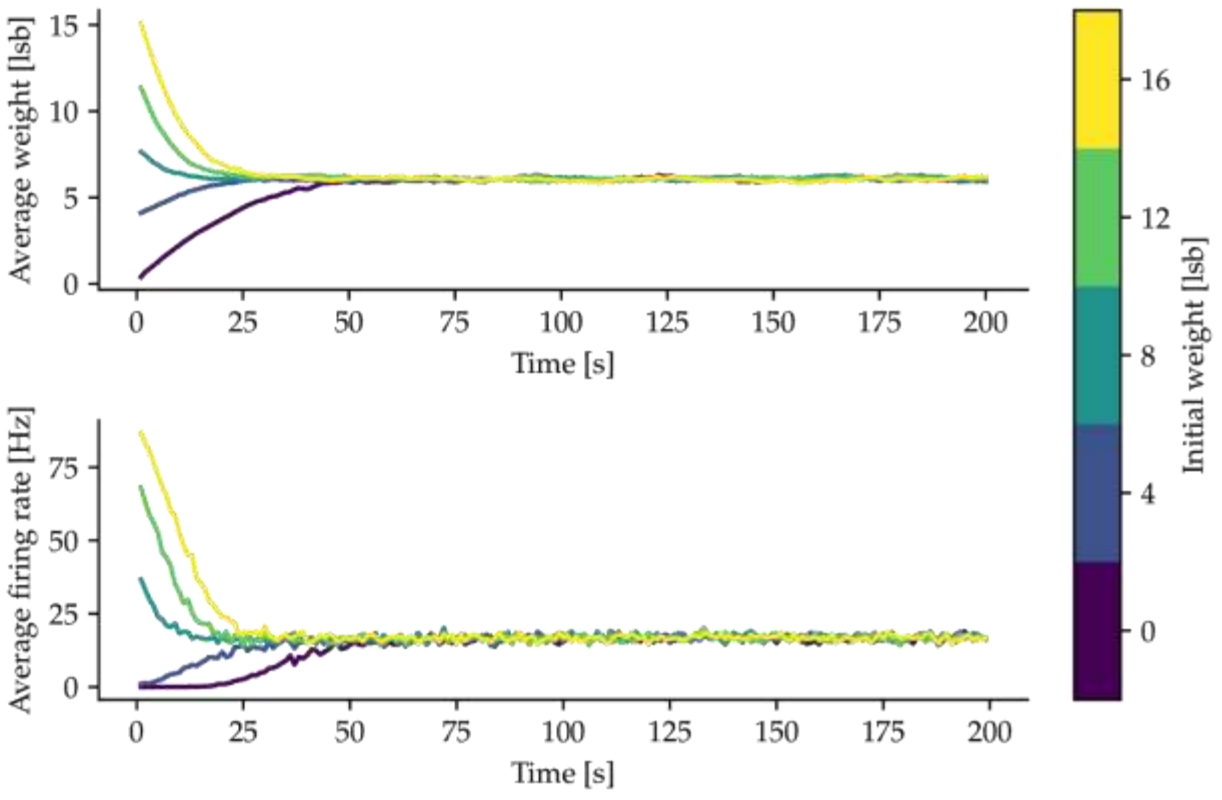
"*Towards a quantitative model of the post-synaptic proteome"*
O Sorokina et.al., Mol. BioSyst., 2011,7, 2813–2823

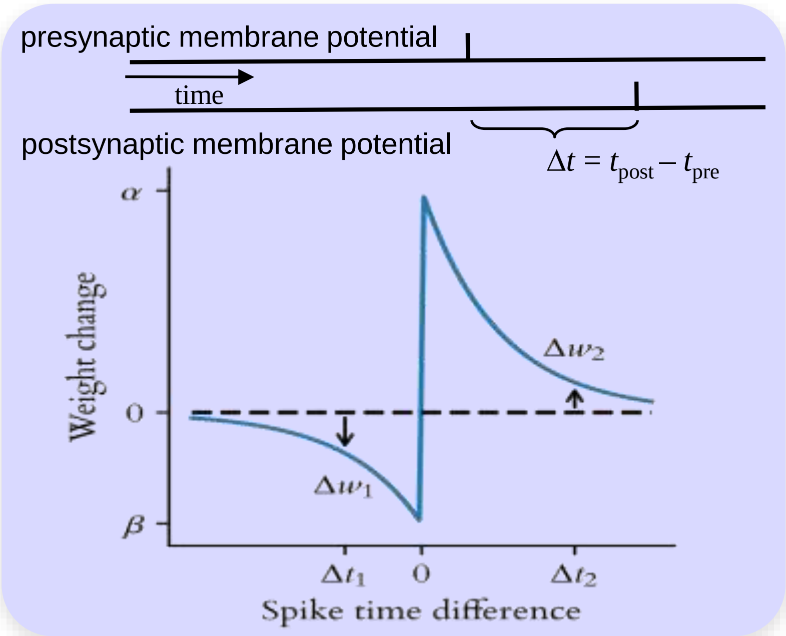
BrainScaleS-2: Hybrid Plasticity

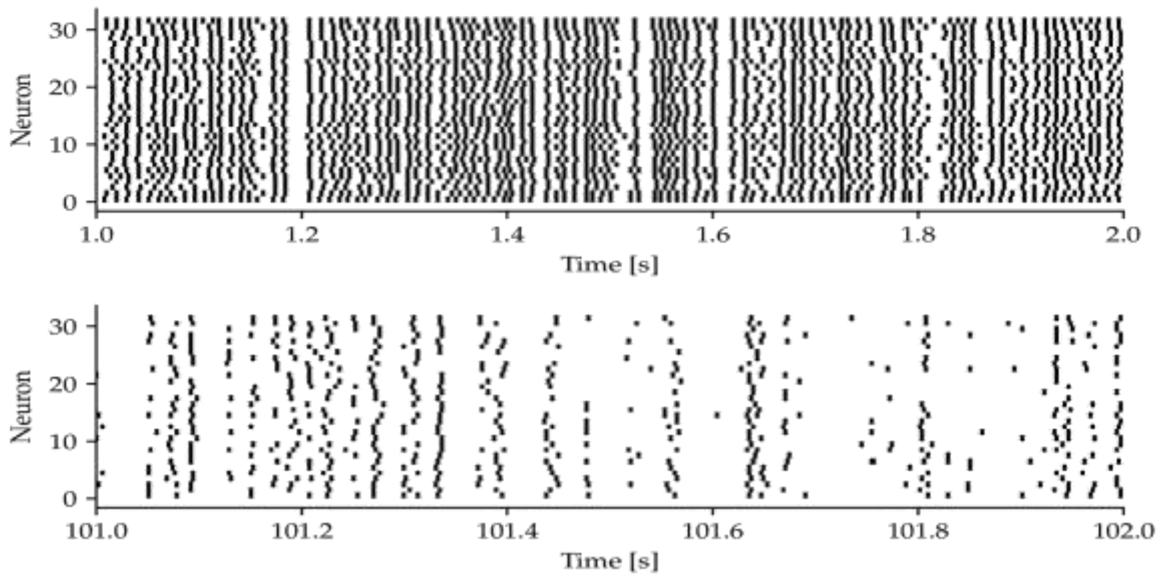
- analog correlation measurement in synapses
- A/D conversion by parallel ADC
- digital Plasticity Processing Units can access
  - synaptic weights (ω)
  - configuration data (adr) → structural plasticity
  - neuron voltages and firing rates



Stabilizing firing rates with spike time dependent plasticity





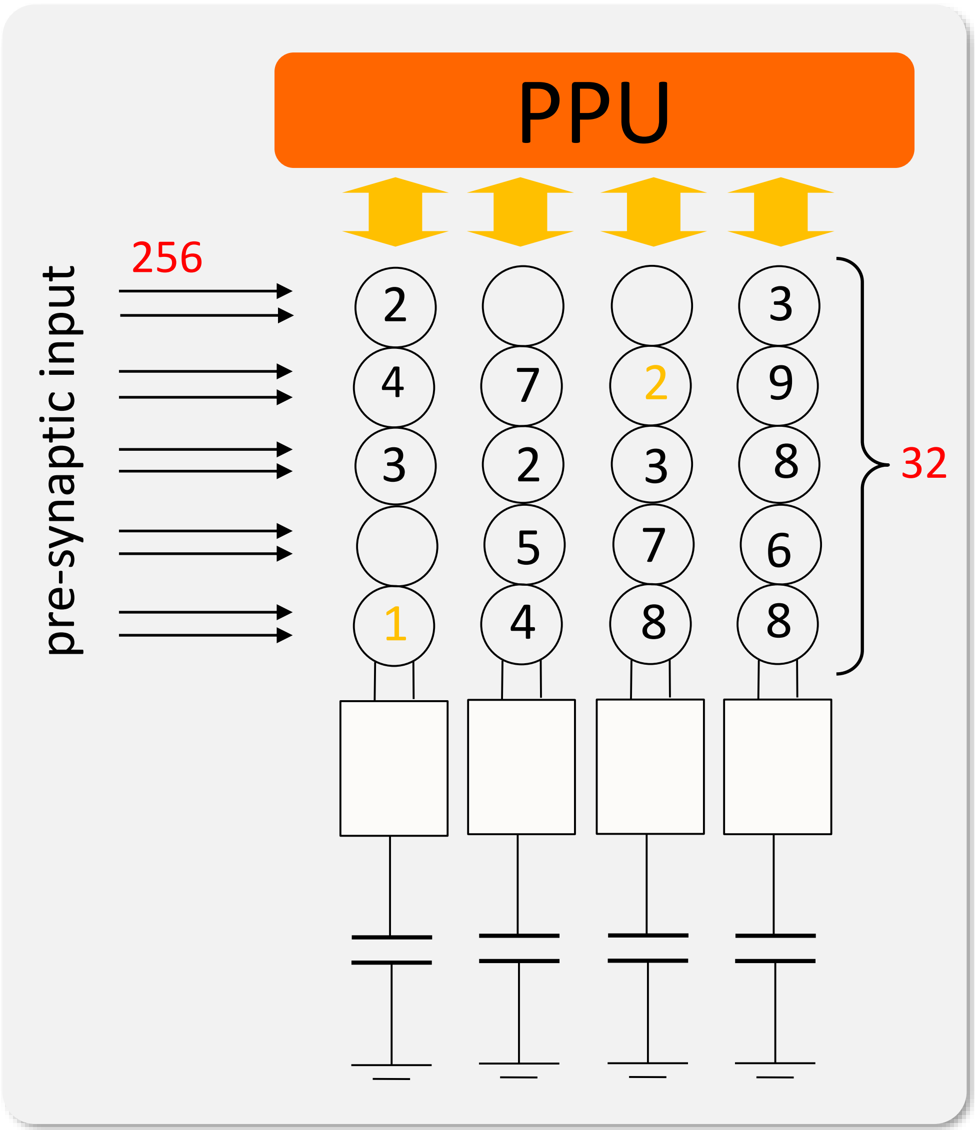


Wall-time per trace: 200ms

→ acceleration factor of 1000

David Stöckel, Master Thesis, Heidelberg University, 2017

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 $\omega \geq \theta_{\text{rand}}$:

$\quad \omega' \leftarrow \omega$

$\qquad + \lambda_{\text{STDP}}(c_+ + c_-)$

$\qquad - \lambda_{\text{hom}}\left(\nu + \nu_{\text{target}}\right)$

$\quad a' \leftarrow a$

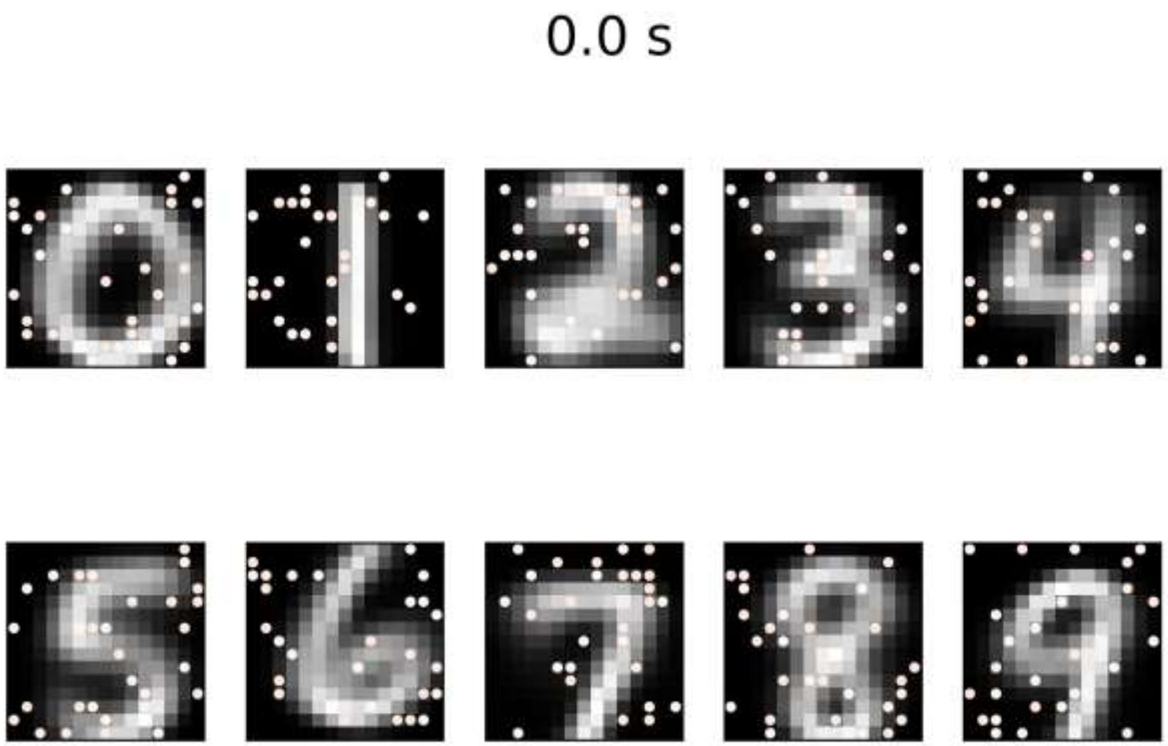
else:

$\quad \omega' \leftarrow \omega_{\text{init}}$

$\quad a' \leftarrow \text{rand}(0,8)$

B. Cramer and S. Billaudelle, unpublished work, 2018

Supervised learning using Hybrid Plasticity



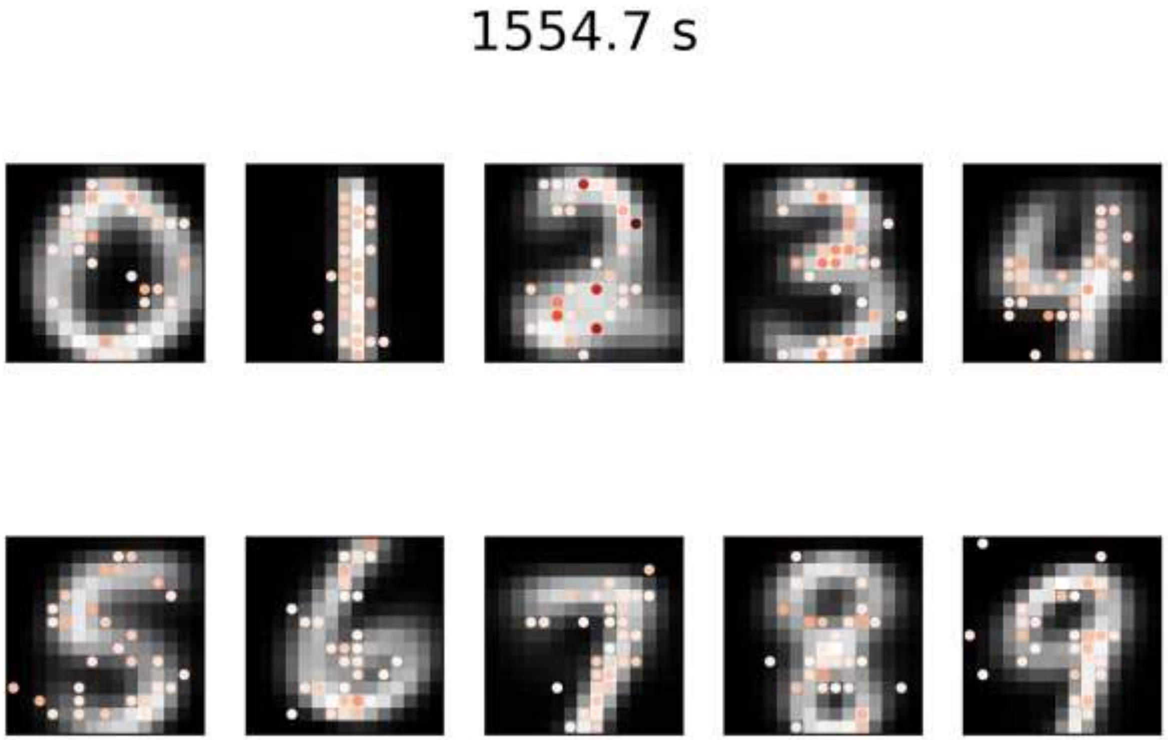
256 pre-synaptic inputs
    mapped to single dendrite
    with 32 active synapses
plasticity rule combines
    structural, STDP and
    homeostatic terms:

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

B. Cramer and S. Billaudelle, unpublished work, 2018

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

Supervised learning using Hybrid Plasticity



Hybrid Plasticity
allows simultaneous rules for:

- strucutral optimization
- homeostatic balance
- pre-post correlation

and more

if $\omega \geq \theta_{\text{rand}}$:

$$\omega' \leftarrow \omega$$
$$+\lambda_{\text{STDP}}(c_+ + c_-)$$
$$-\lambda_{\text{hom}}\left(\nu + \nu_{\text{target}}\right)$$
$$a' \leftarrow a$$

else:

$$\omega' \leftarrow \omega_{\text{init}}$$
$$a' \leftarrow \text{rand}(0,8)$$

using software running in parallel to the analog neuron operation

B. Cramer and S. Billaudelle, unpublished work, 2018

What I have learned

- analog computing is feasible
  - model biology for neuroscience
  - cost and energy efficient inference of DCNNs
  - edge computing (sensor data preprocessing)
- works best if closely coupled to SIMD CPU
  - Software-based implementation of learning algorithms
    - learning can include calibration
    - supports hyper-parameter learning (L2L)
  - initialization
  - configuration
  - debugging
  - calibration
- future considerations
  - find the optimum hybrid (digital vs. analog) system for a given technology
  - replacing CMOS will be very difficult (>20 years from now)
  - CMOS is good enough, but cost might be prohibitive
    → efficient in-memory computing needs large amounts of silicon

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

NICE 2020

March 17 - 20th 2020

Neuro-Inspired
Computational Elements
Workshop

Im Neuenheimer Feld 227
D-69120 Heidelberg
Germany

Workshop:	March 17-20th 2020
Tutorials:	March 20th 2020

Heidelberg - Germany


Picture: fotolia.com / Sergey Borisov

Kirchhoff Institute for Physics


Thank you!