Exploring the possibilities of 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

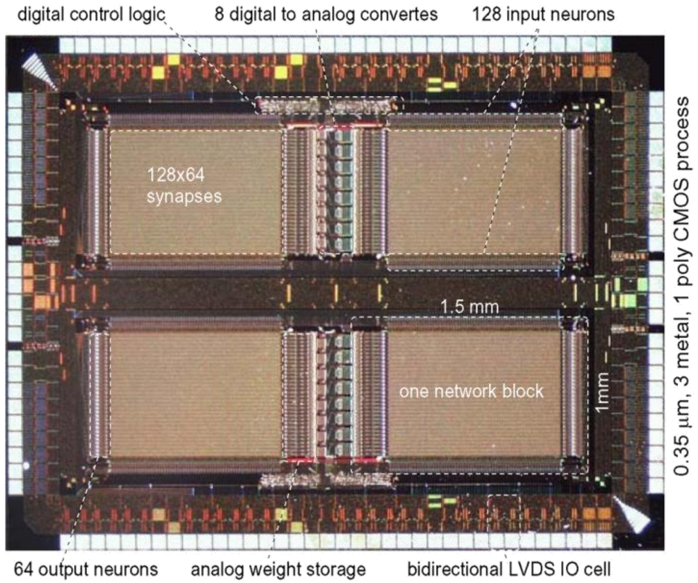
2003 First concepts for spike based analog neural networks

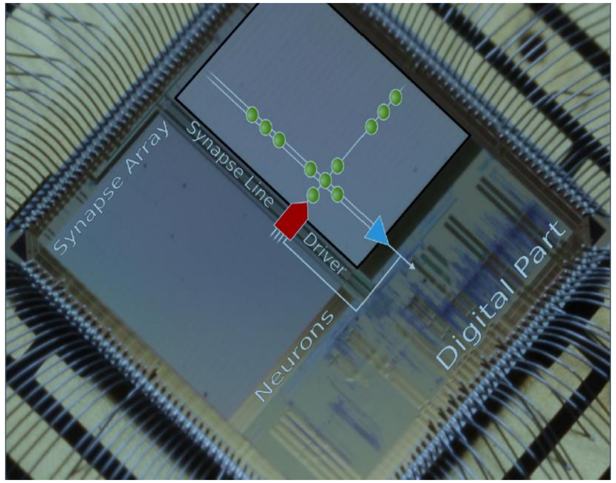
2004 First accelerated analog neural network chip with short and long term plasticity: Spikey

HAGEN (2000):

Perceptron-based Neuromorphic chip introduced:

- accelerated operation
- mixed-signal Kernels





SPIKEY (2004):

spike-based Neuromorphic chip introduced:

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

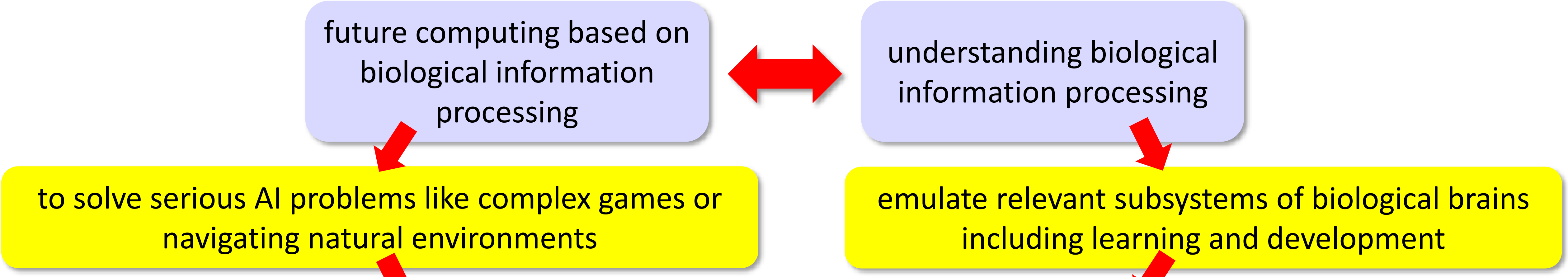
since the year 2000: Computers became more brain-like

Neuromorphic computing (analog or digital):

- at least similar performance
- faster learning
- lower energy consumption
- closer to biological concepts

Neuromorphic Computing

”Learning from nature to advance future computing.”



future computing based on biological information processing ↔ understanding biological information processing

to solve serious AI problems like complex games or navigating natural environments

emulate relevant subsystems of biological brains including learning and development

→ requires circuits and architectures for scalable neuromorphic computing systems:

☺

1. in size – problem size, dimensionality of sensor date
2. in speed – to find a solution, lots of trial runs needed
3. in model complexity – to model biology appropriately
4. in learning capabilities – problem complexity
5. while keeping energy consumption reasonable

☹

Perceptron-based digital machine learning

Cerebras CS-1

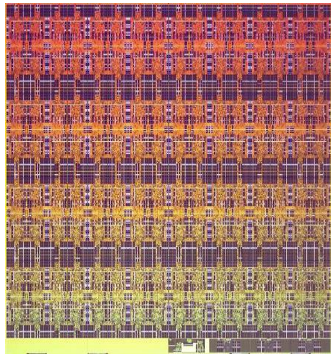
NVIDIA DGX A100

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

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

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

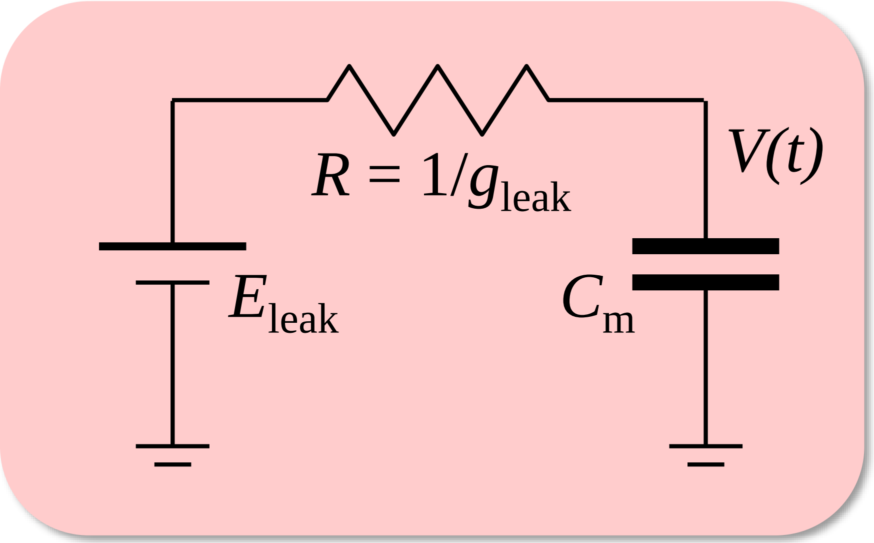
Analog neural cores
Digital spike communication
Biological local learning
Programmable local learning
x10.000 to x1000 real-time

Analog computing helps for scaling-up speed : Neuromorphic Computing with physical model systems

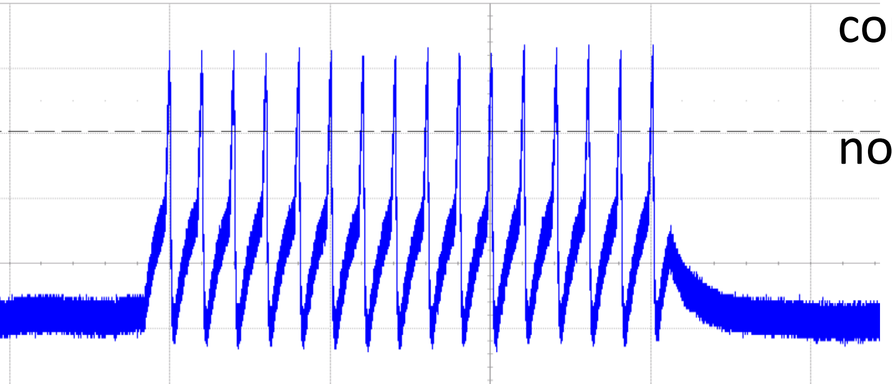
- Consider a simple physical model for the neuron's cell membrane potential V:

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

- representing model parameters as physical quantities : **voltage, current, charge**



$$\left.\frac{dV}{dt}\right._{bio} << \left.\frac{dV}{dt}\right._{VLSI} \rightarrow \text{accelerated neuron model}$$

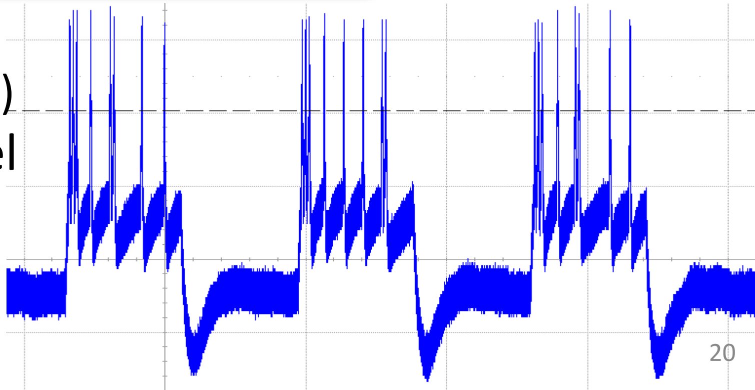


continuous time

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

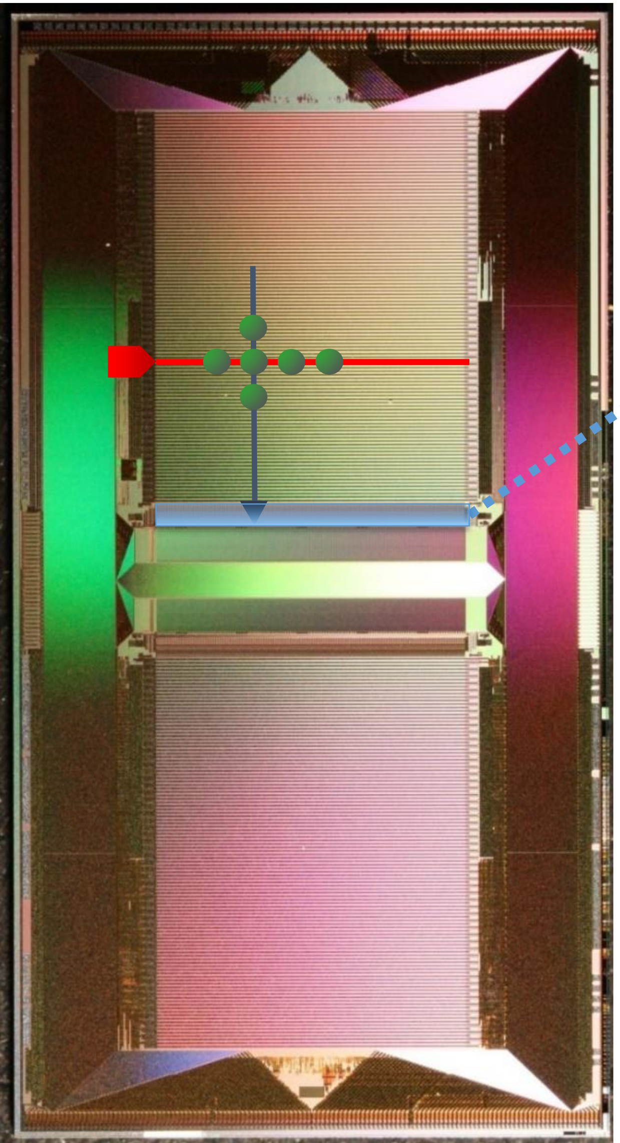
no multiplexing of components storing model variables

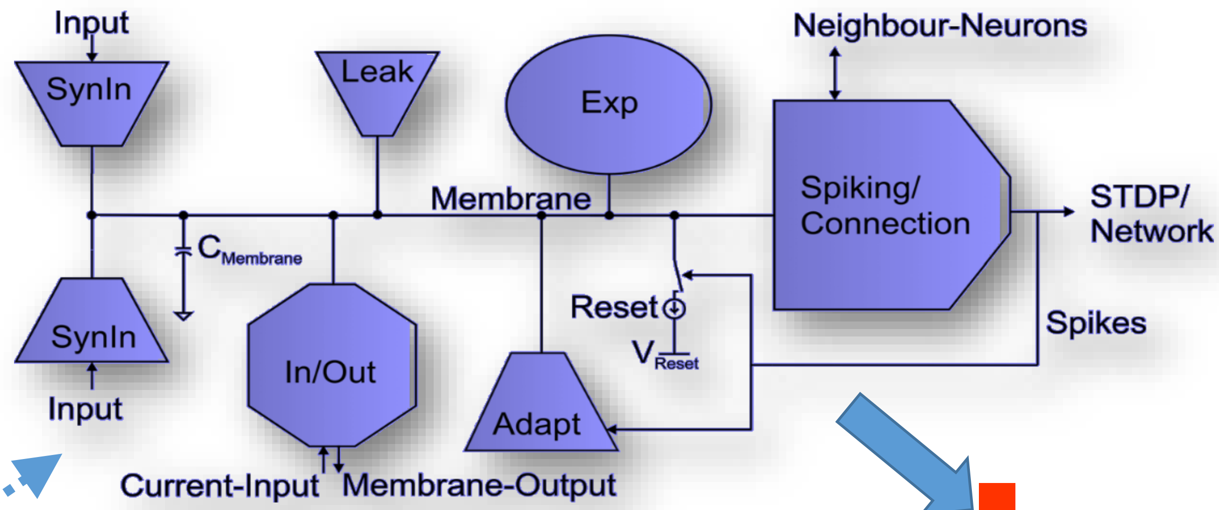
- each neuron has its membrane capacitor
- each synapse has a physical realization



Analog helps for energy-efficient scaling-up of model complexity : neurons built from parameterized dendritic compartments

photograph of the BrainScaleS 1 neuromorphic chip



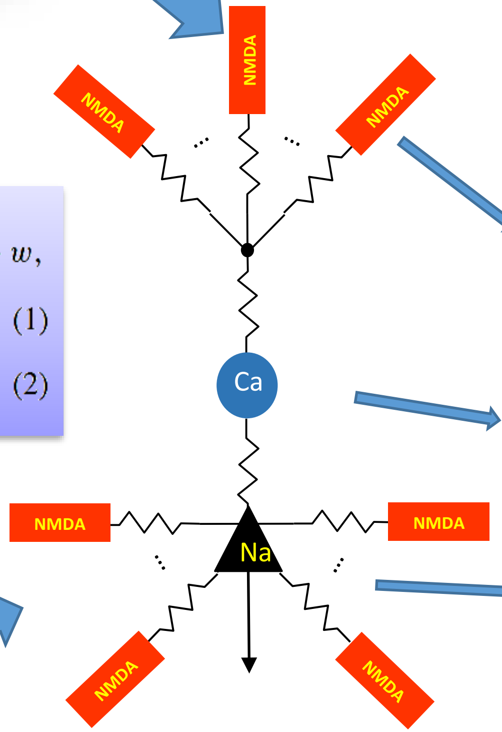


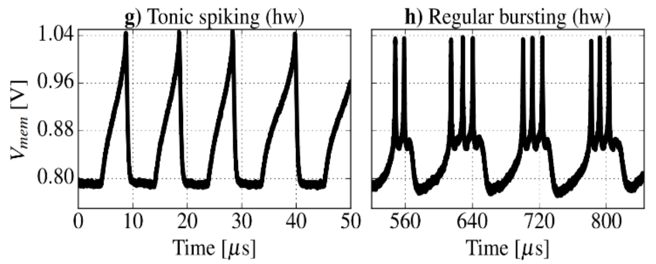
- modular structure
- Adaptive Exponential I&F model
- full set of ion-channel circuits for each compartment
- 24 calibration parameters per compartment

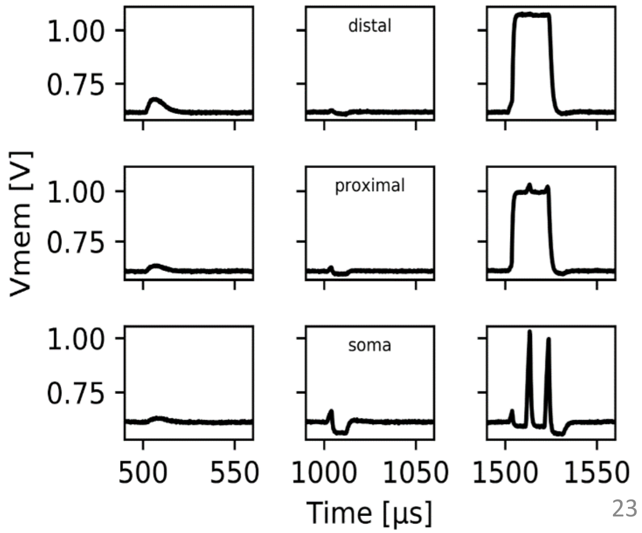
$$C\frac{dV}{dt} = -g_L(V-E_L) + g_L\Delta_T \exp\left(\frac{V-V_T}{\Delta_T}\right) + I - w, \qquad (1)$$

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

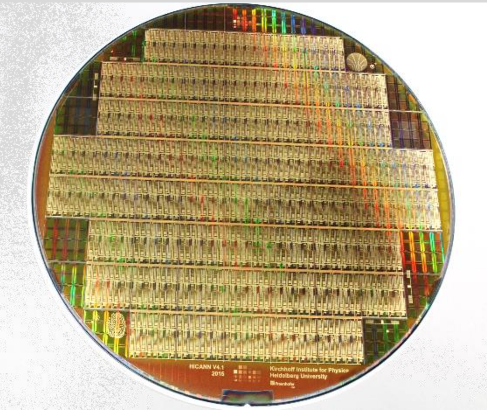
complex neurons can be build by connecting individual compartments







BrainScaleS-1 : large-scale analog



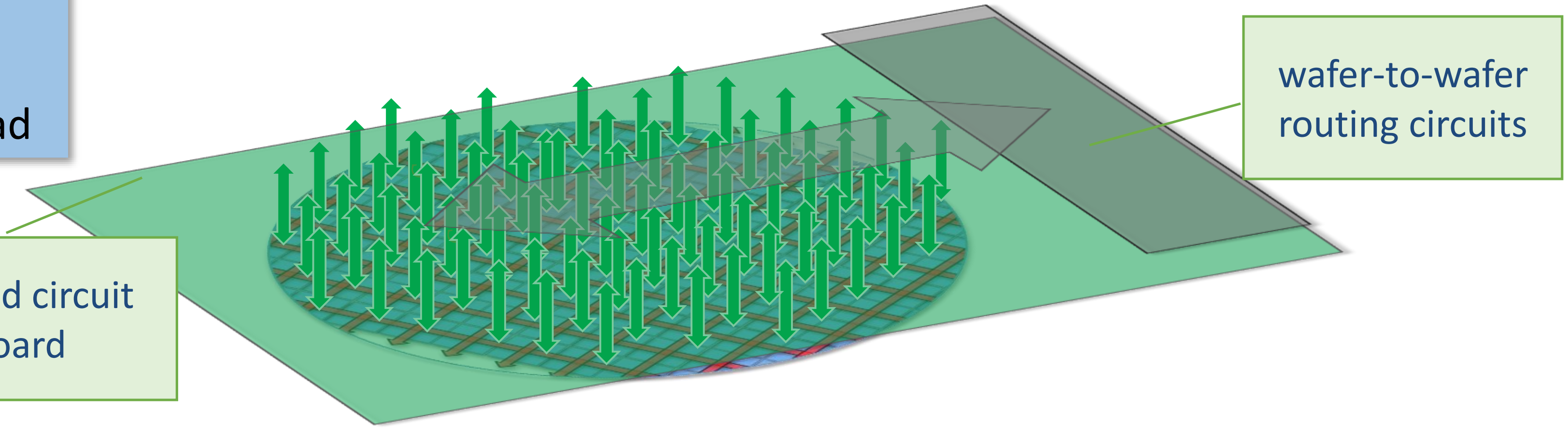
silicon wafer with BSS ASICs

wafer module

hybrid system

custom circuits locally interconnect the chips on the wafer

- creates wafer-wide network
- distributes communication load



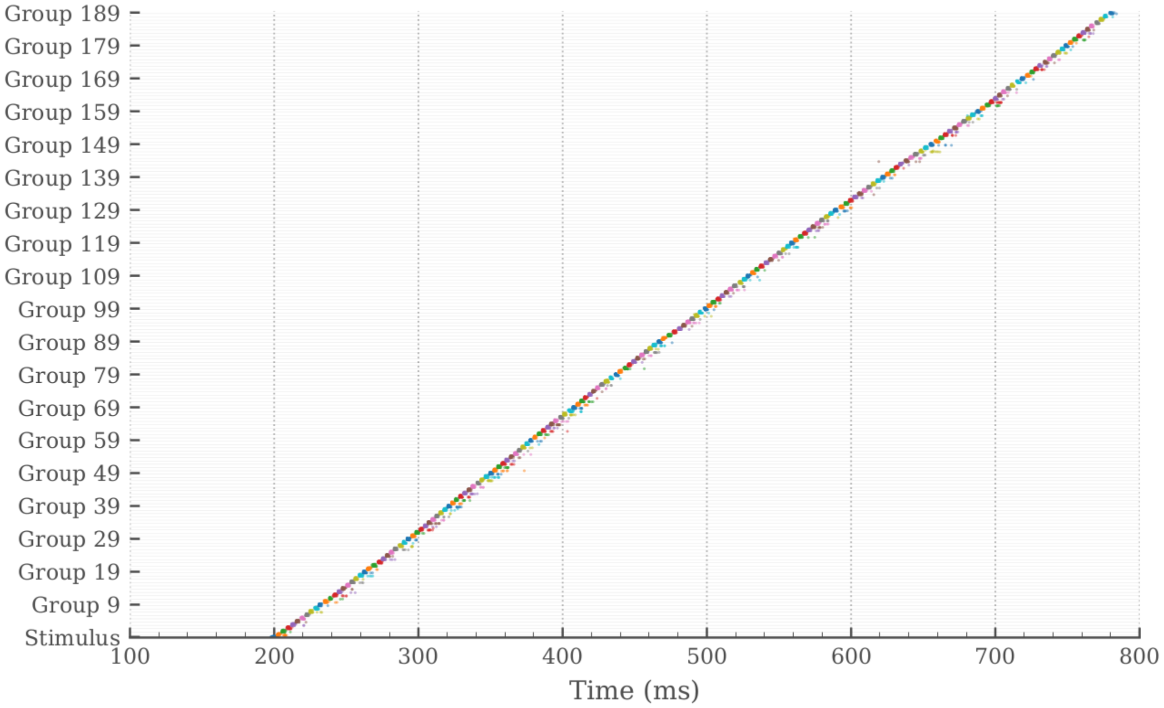
wafer-to-wafer routing circuits

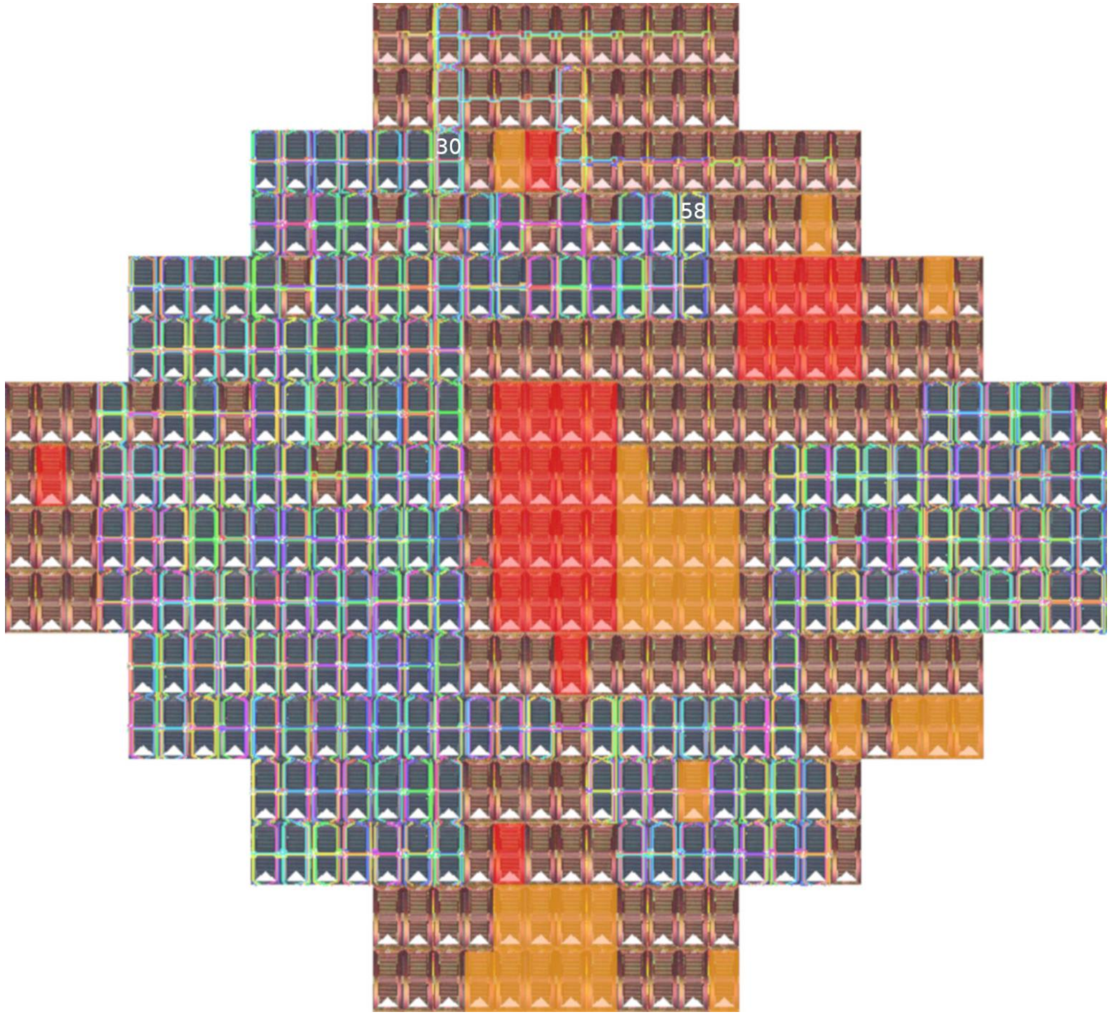
printed circuit board

BrainScaleS-1 - large scale analog network model

- Synfire chain with feedforward inhibition
- 19000 neurons (190 chain links)
- Over 1.4 million synapses
- Acceleration: 10,000

→ compute performance equivalent to 14 billion synapses and 19 million neurons in real time





next step: cortical column model currently under development

Talk:

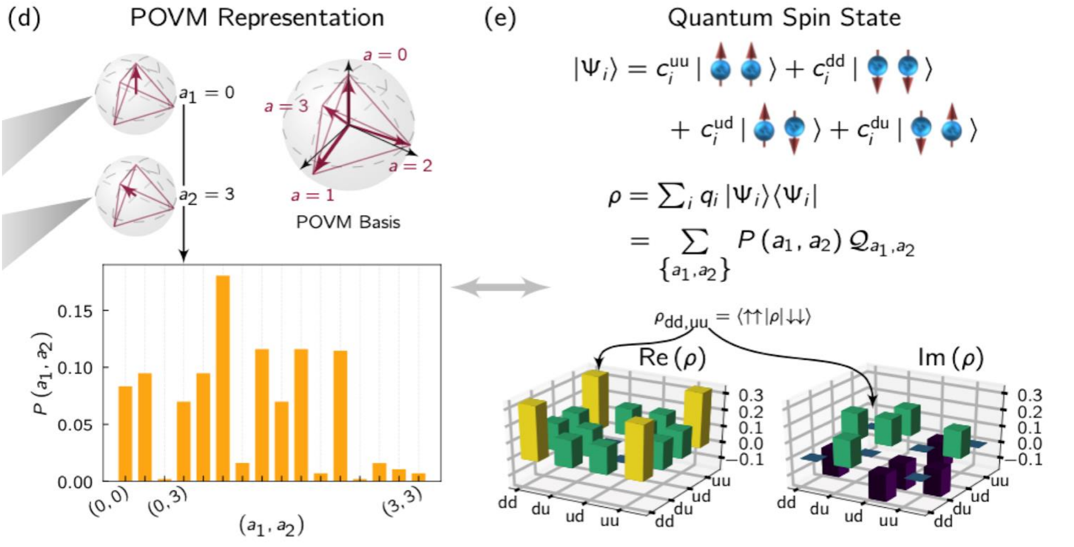
From clean room to machine room: towards accelerated cortical simulations on the BrainScaleS wafer-scale system (S. Schmitt)

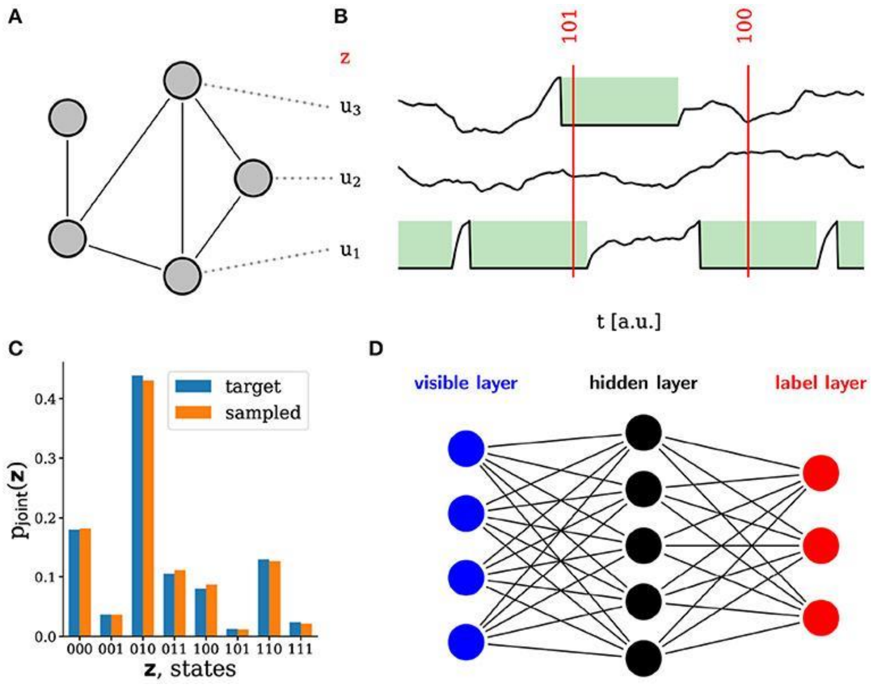
Today 17:20 (CET)

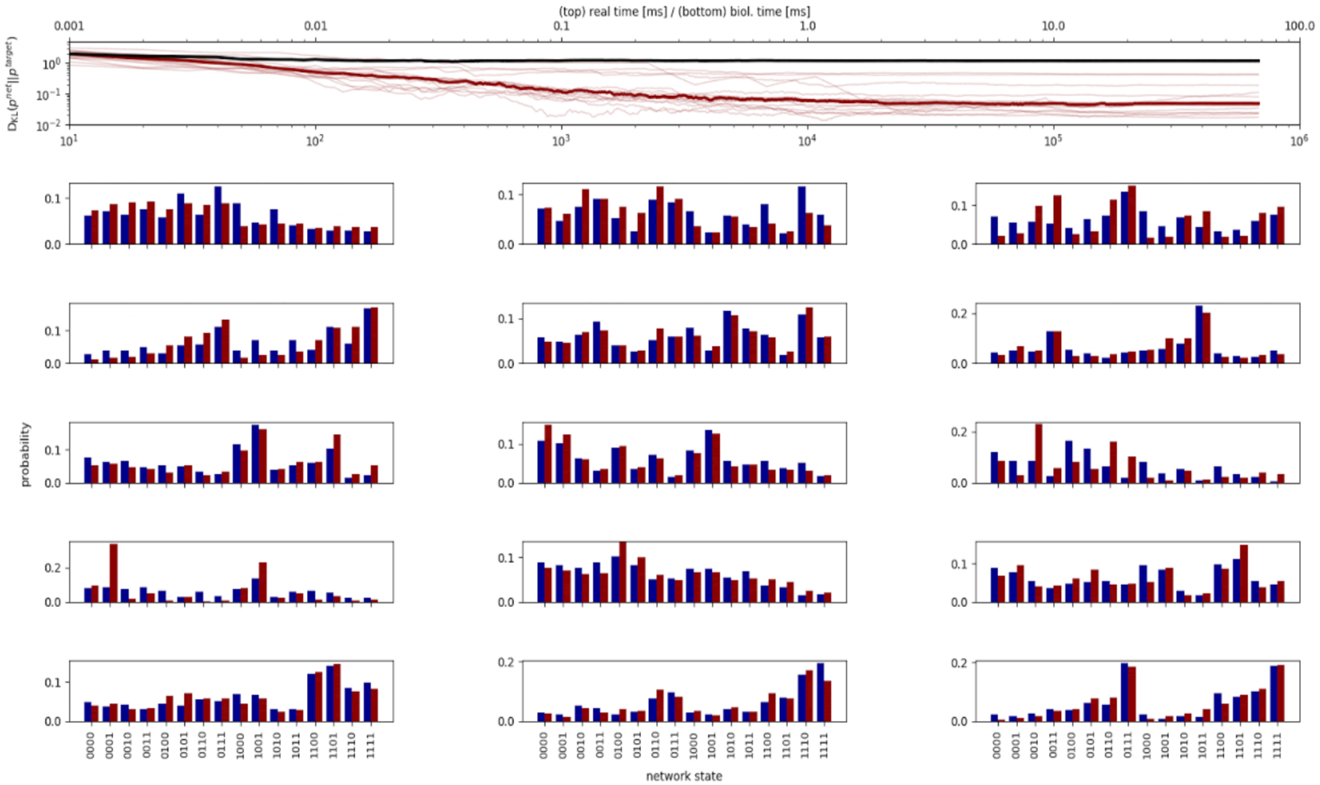
work by Jakob Kaiser [Master Thesis 2020], Sebastian Schmitt

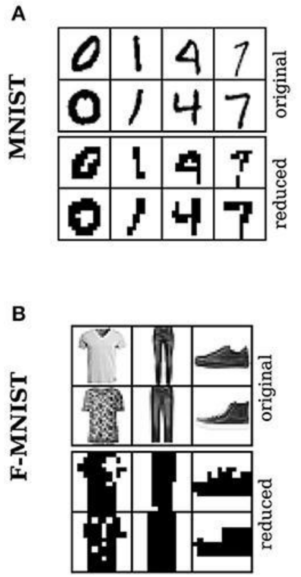
Analog spike-based Bayesian Inference

- implements *generative* models
- applications to image datasets and quantum states
- analog neurons autonomously sample from learned distributions
- no numerical calculations









Work by: A. Baumbach, S. Czischek, D. Dold, T. Gierlich, A. Kungl, M. Petrovici [D. Dold et al. 2019], [A. Kungl et al. 2019], [S. Czischek et al. 2020]

Analog neuromorphic hardware ≠ no software

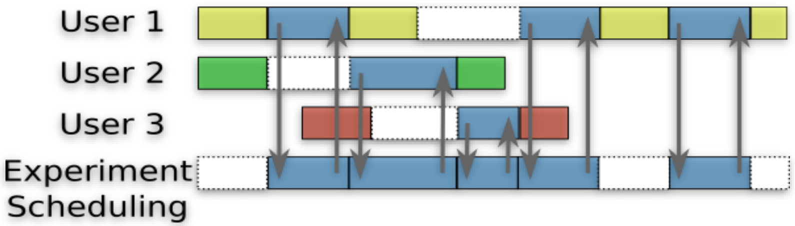
after training:

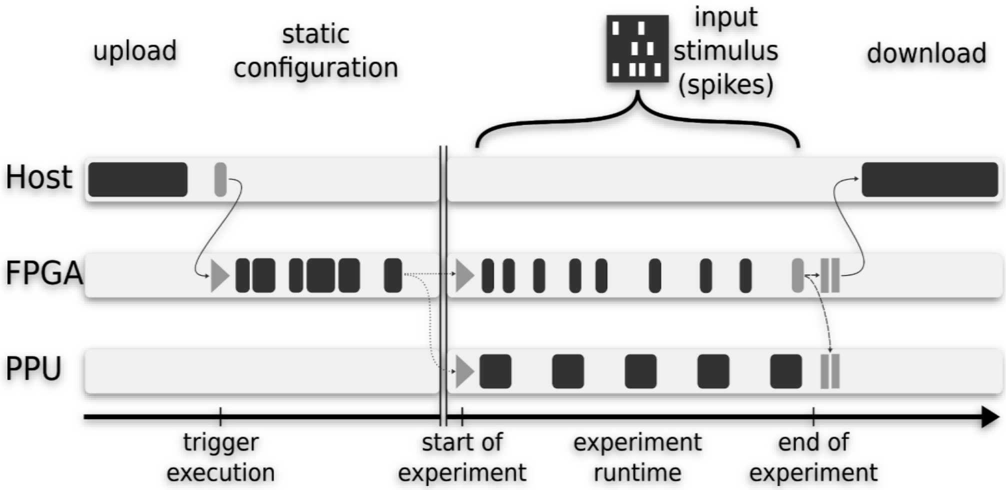
Non-Turing analog computing system performs autonomously

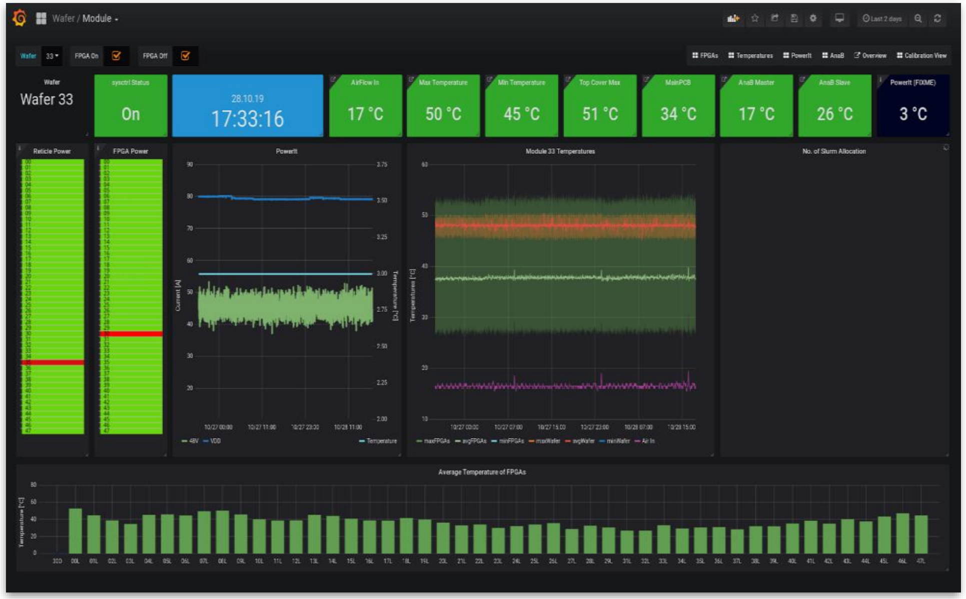
but

Turing-based digital computing is used in multiple places:

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






Wafer / Module

Wafer 33 | FPGA On ☑ | FPGA Off ☑

FPGAs | Temperatures | PowerIt | AnaB | Overview | Calibration View

Wafer: Wafer 33

sysctrl Status: On

28.10.19 17:33:16

AirFlow In: 17 °C

Max Temperature: 50 °C

Min Temperature: 45 °C

Top Cover Max: 51 °C

MainPCB: 34 °C

AnaB Master: 17 °C

AnaB Slave: 26 °C

PowerIt (FIXME): 3 °C

BraScaleS-2 User high-level APIs:

- hxtorch for PyTorch
- PyNN.brainscales2

Components:

- Hardware Testing & Calibration
- Hardware Configuration
- Experiment Encapsulation & Scheduling

Operation:

- Monitoring
- Resource Management
- Software Environment

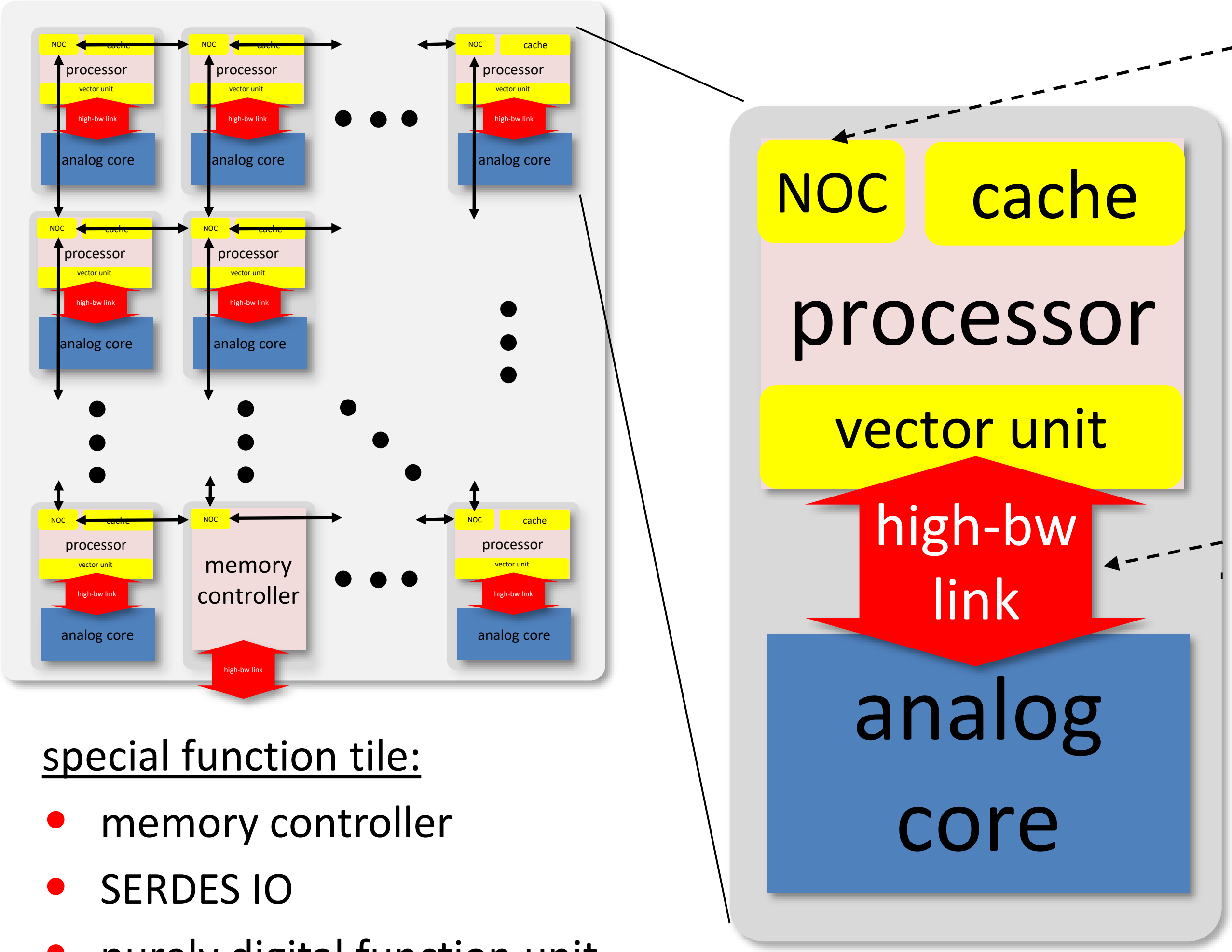
Talk on Friday

Work by:
E. Müller, O. Breitwieser, S. Schmitt, C. Mauch, P. Spilger, Y. Stradmann, H. Schmidt, J. Montes, A. Emmel, M. Czierlinski, J. Kaiser, S. Billaudelle,F. Ebert, M. Güttler, J. Ilmberger, A. Leibfried, J. Weis.

Optimum combination of analog and digital processing allows scaling-up of learning capabilites

BrainScaleS-2 : analog coprocessor with digital learning support

- on-chip training with complex learning rules
- learning capabilities scale with system size
- can cope with scaled-up speed of accelerated physical model



special function tile:

- memory controller
- SERDES IO
- purely digital function unit

Network-on-chip:

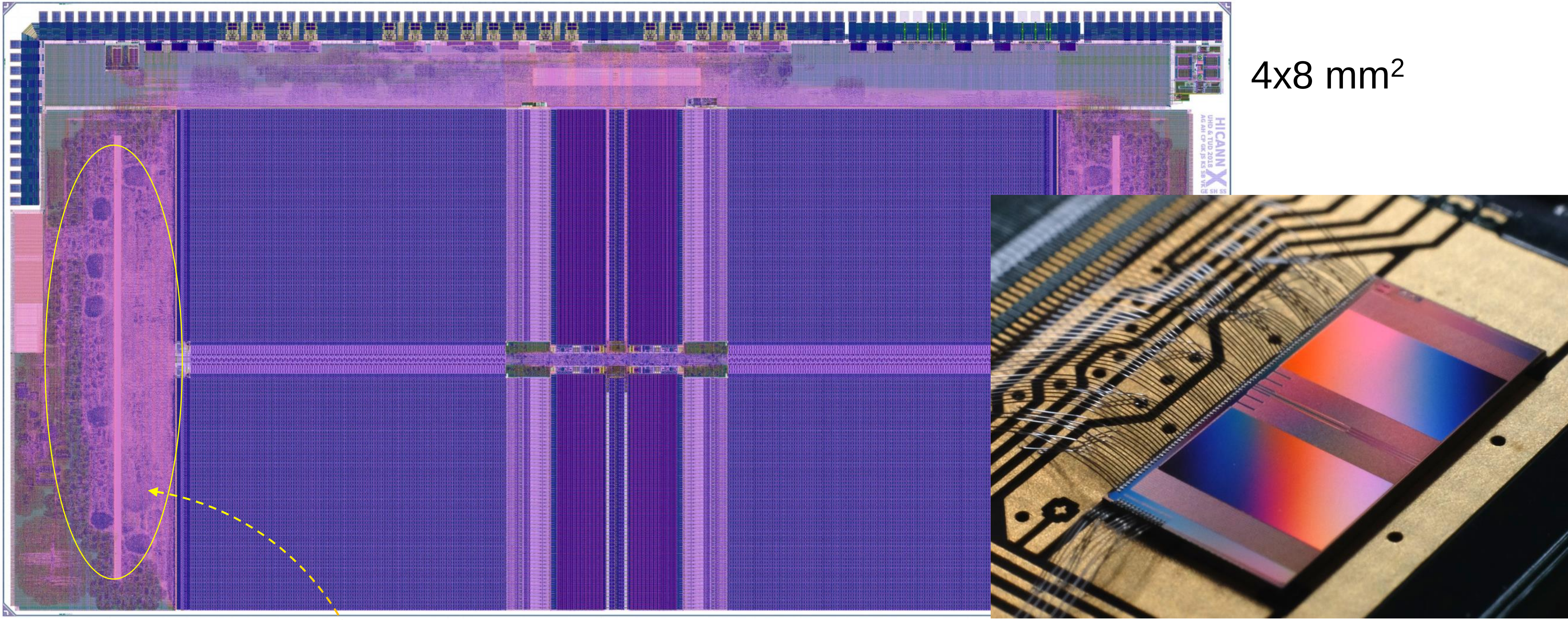
- prioritize event data
- unused bw for CPU
- common address space for neurons and CPUs

high-bandwidth link:

vector unit ←→ NM core

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

BrainScaleS-2: two cores with neuromorphic coprocessors



4x8 mm^2

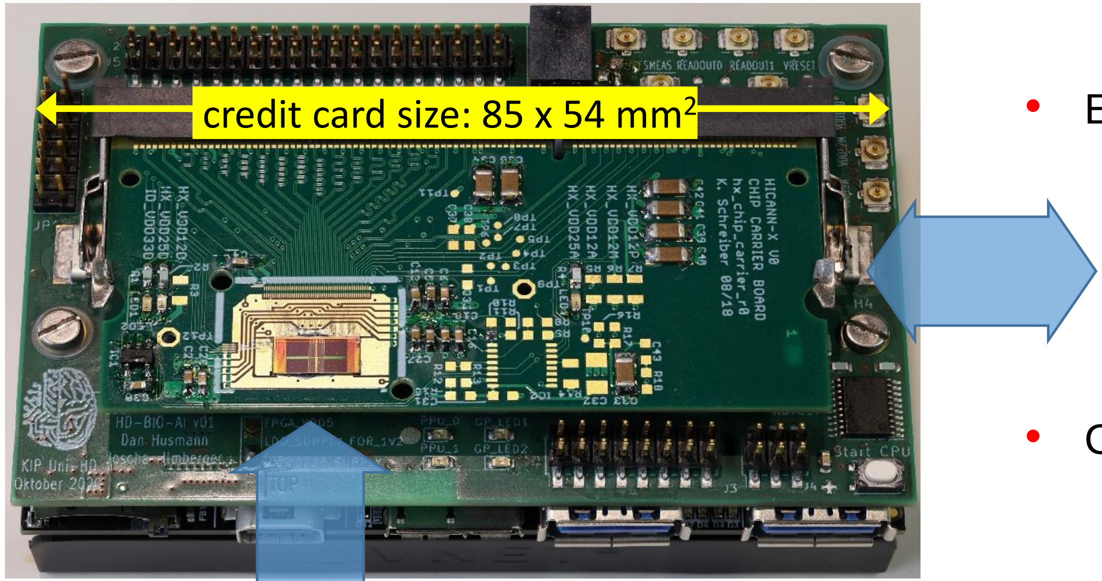
- 65nm LP-CMOS, power consumption O(10 pJ/synaptic event)
- 128k synapses
- 512 neural compartments (Sodium, Calcium and NMDA spikes)
- two CPU cores for learning (PPU)
- PPU internal memory can be extended externally

- 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

Application: edge-AI with BrainScaleS

BrainScaleS mobile system

- Small, cost-efficient system
  - low-power FPGA base board
  - interface board for BrainScaleS ASIC
  - ASIC carrier board
- Multi-chip operation with different ASIC carrier boards
- Direct applicable for applications



credit card size: 85 x 54 mm^2

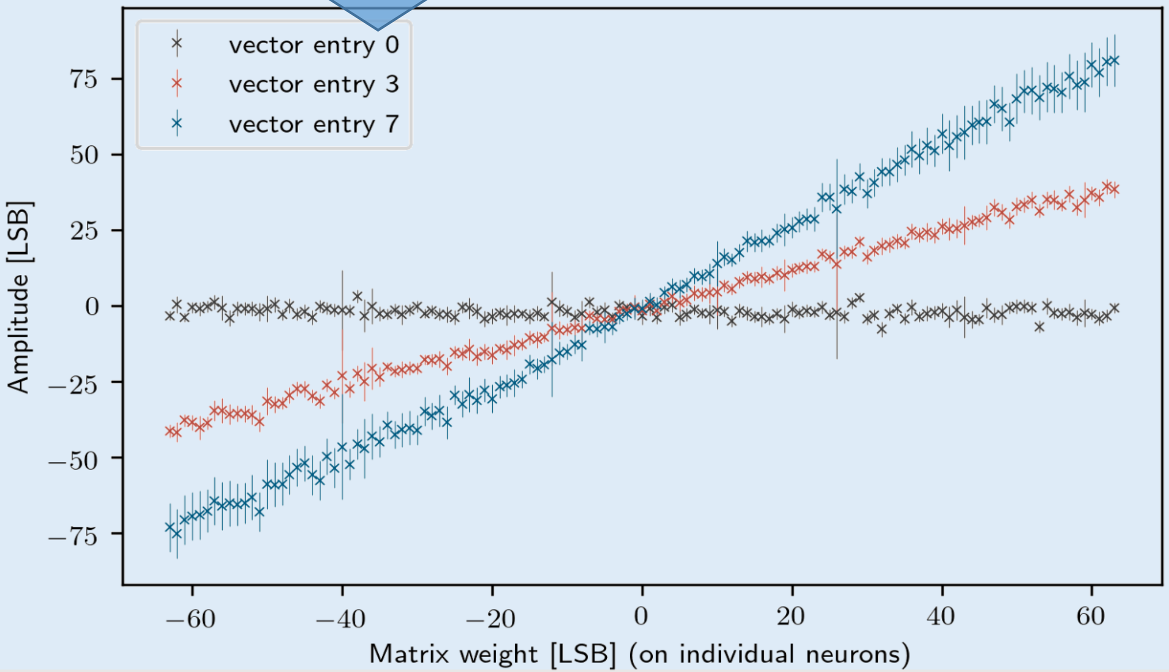
- Event-based direct IO
  - neuromorphic detectors
  - neuromorphic sensors
    - event-based cameras
    - bio-sensors
    - etc
- Classic IO (FPGA subsystem)

Analog vector-matrix multiplication

- allows rate-based modeling on same chip
- training with PyTorch and hardware-in-the-loop

work by Johannes Weis, Arne Emmel

Weis, J. et al. (2020): Inference with Artificial Neural Networks on Analog Neuromorphic Hardware. ITEM 2020, IoT Streams 2020.
Spilger, P. et al. (2020): hxtorch: PyTorch for BrainScaleS-2. ITEM 2020, IoT Streams 2020.



Benchmarks:

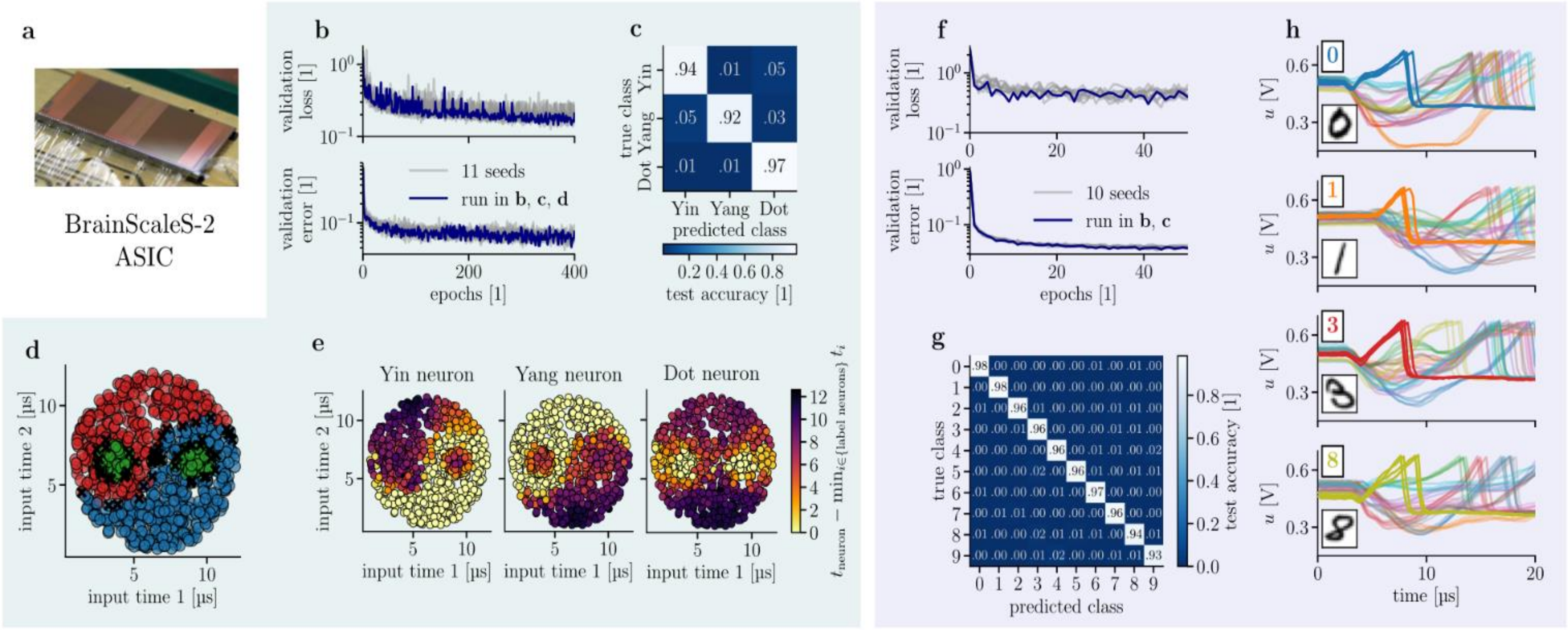
MNIST handwritten digits

- 3-layer CDNN
  - 98.5 % on CPU
  - 98.4 % on BSS-2

Human activity recognition

- 6 activities of daily living
- 3-layer CDNN
  - 89.7 % on CPU
  - 88.8 % on BSS-2

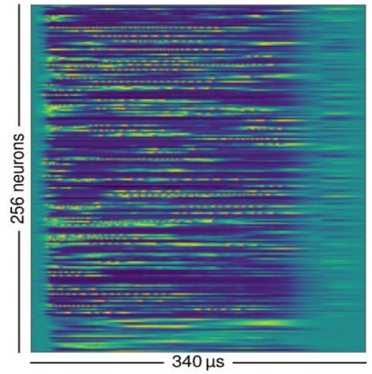
Fast and deep:
energy-efficient neuromorphic learning with first-spike times

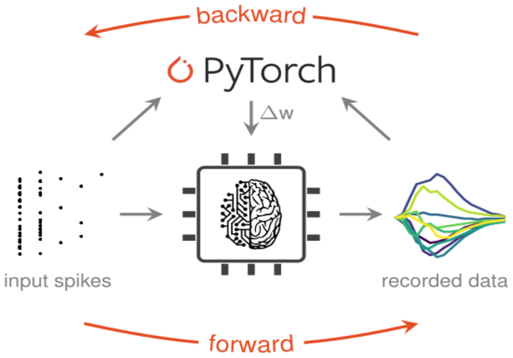


Lightning talk on Friday, Julian Göltz, Laura Kriener, et al. 2019, *https://arxiv.org/abs/1912.11443*

Training recurrent and multi-layer SNNs using surrogate gradients

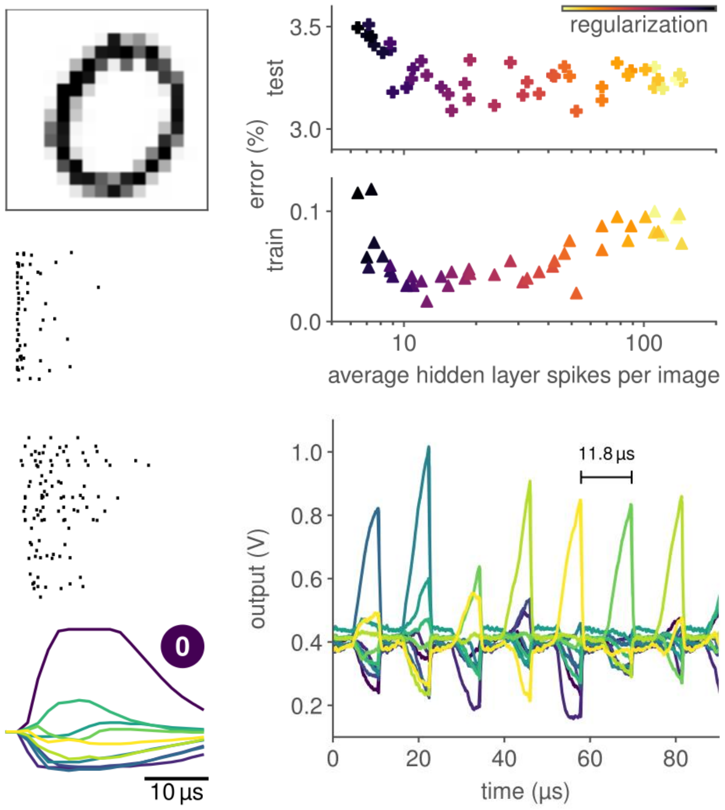
»zero«



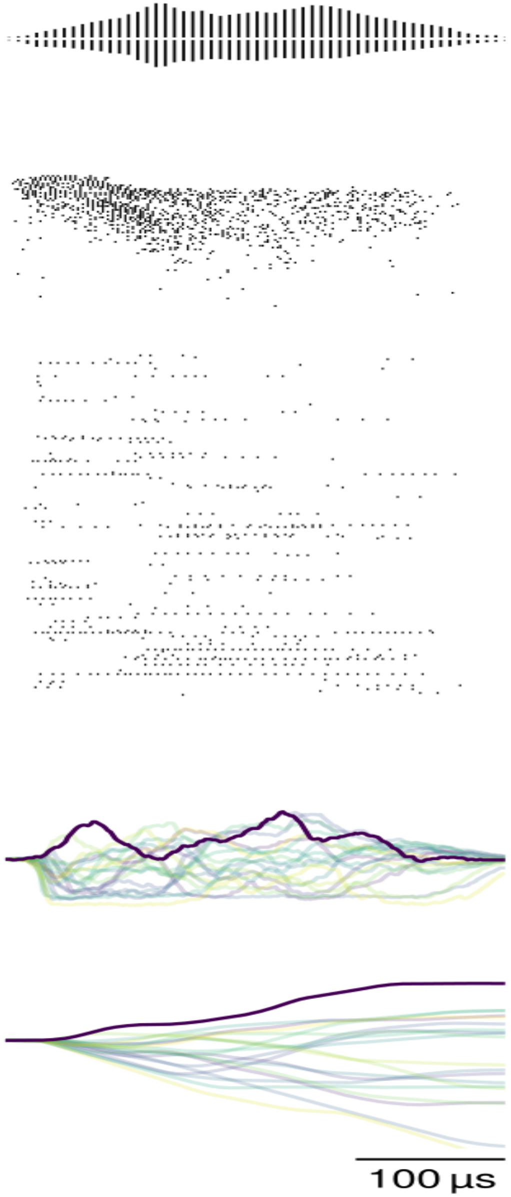


- Flexible in-the-loop training based on PyTorch
  - Feed-forward as well as recurrent SNNs
  - Arbitrary choice of loss function and regularization
  - Auto-differentiation for analog hardware
- Robust to fixed-pattern deviations of analog circuits

MNIST: 97.6 %, 85k images/s, 2.4 µJ/image



SHD: 80.6 %



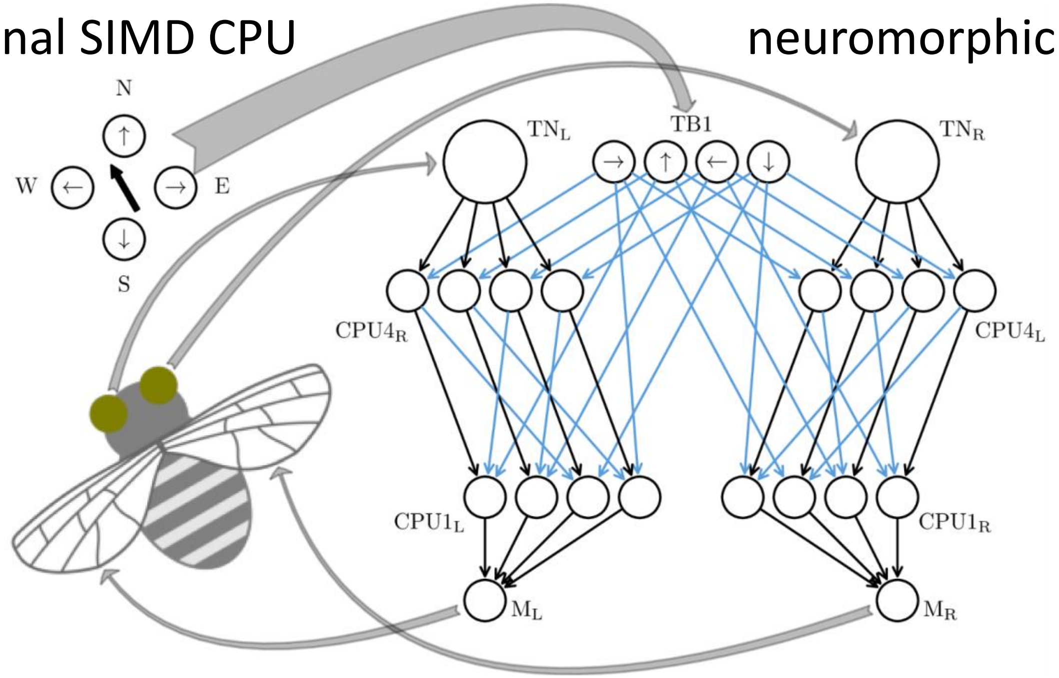
B. Cramer, S. Billaudelle, F. Zenke, et al. "Training spiking multi-layer networks with surrogate gradients on an analog neuromorphic substrate." *arXiv preprint arXiv:2006.072*

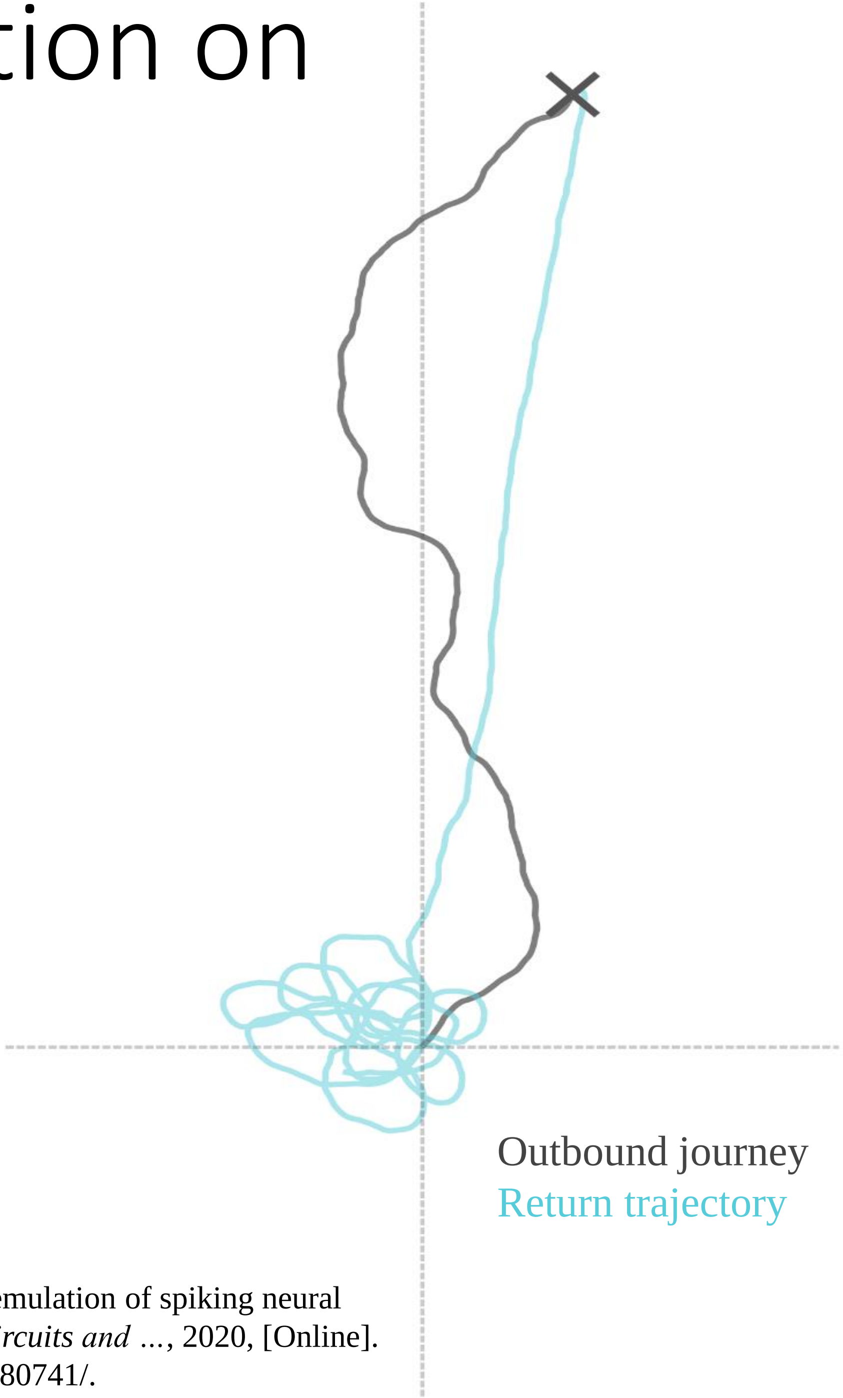
Analog model of insect navigation on HICANN-X

Neuromorphic insects navigate autonomously back to their nests after searching for food.

Body as digital model on internal SIMD CPU

analog neurons on neuromorphic core

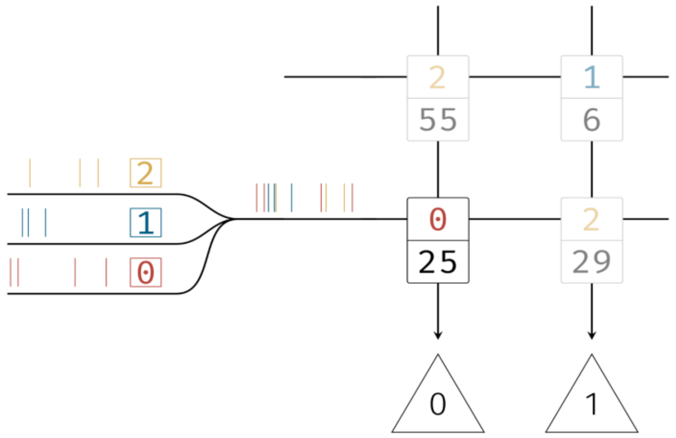


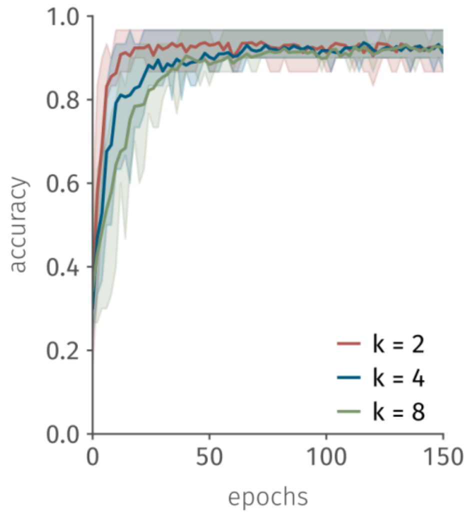


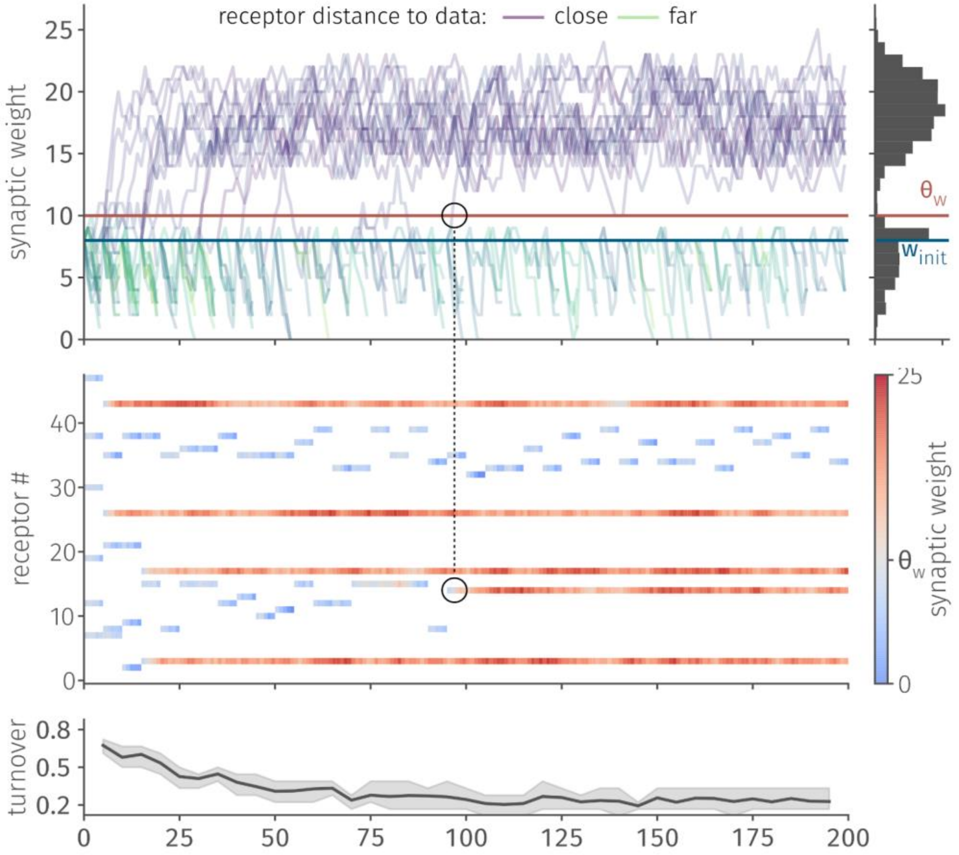
K. Schreiber, et al. “Insectoid path integration on accelerated neuromorphic hardware” *in preparation*
A. Leibfried: Migration from prototype to full-scale version

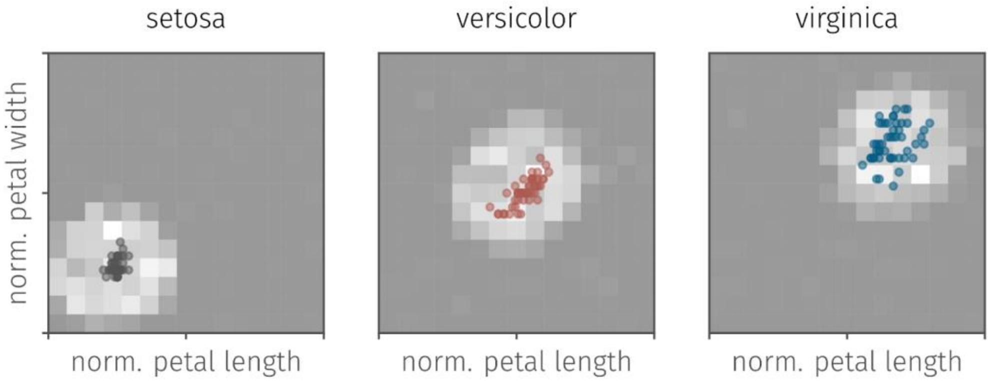
S. Billaudelle, Y. Stradmann, and K. Schreiber, “Versatile emulation of spiking neural networks on an accelerated neuromorphic substrate,” *on Circuits and …*, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9180741/.

Structural plasticity on BrainScaleS-2





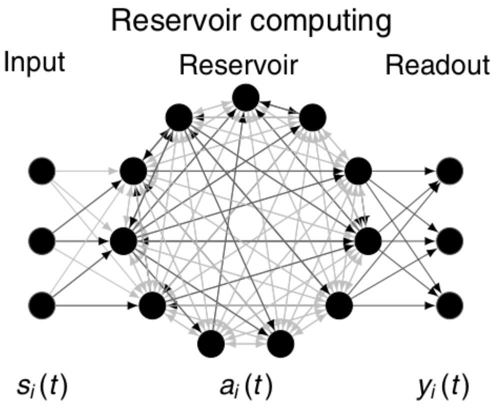


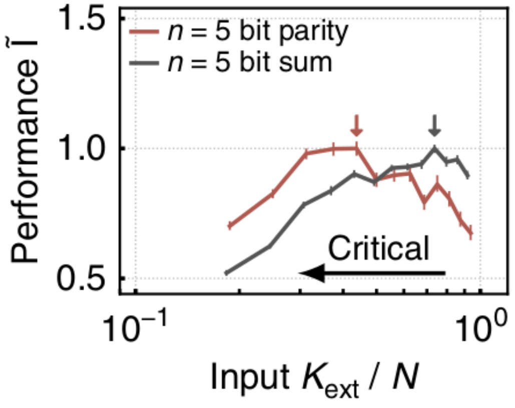


- On-chip structural plasticity
- Self-configuring receptive fields
- Efficient use of synaptic resources

S. Billaudelle, B. Cramer, et al. "Structural plasticity on an accelerated analog neuromorphic hardware system." Neural Networks 133 (2021): 11-20.

Control of criticality and computation in SNNs with plasticity

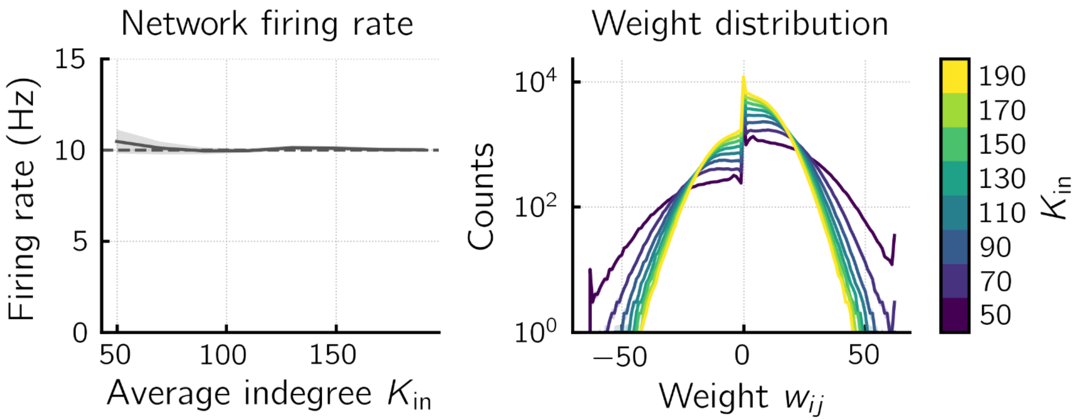


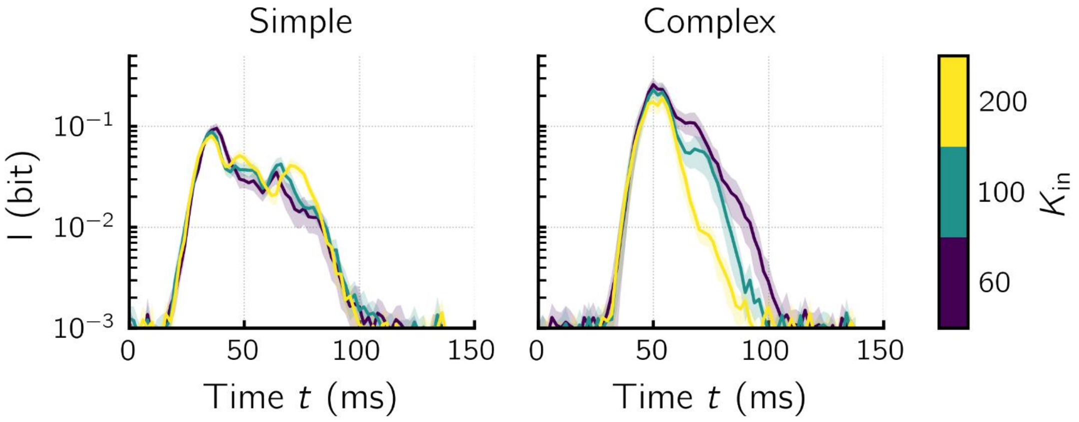


- Distance to a critical point of recurrent SNN was changed by adapting the input strength under homeostatic regulation
- Evaluating performance on a set of tasks of varying complexity at - and away from critical network dynamics shows:
  - Only complex task profit from critical dynamics
  - Simple tasks even suffer
- Collective network state has to be tuned to task requirements by changing the input strength
- Network then quickly self-organizes to desired state

B. Cramer, D. Stöckel, M. Kreft, M. Wibral, J. Schemmel, K. Meier, V. Priesemann "Control of criticality and computation in spiking neuromorphic networks with plasticity." *Nature communications* 11.1 (2020): 1-11.

- Hierarchy of time scales in homeostatically regulated neural networks in excitation-dominated regime
- Exploit full speedup by on-chip implementation with 512 LIF neurons
- Chip consumes only 100 mW during emulation
- Setup can be used to classify spatio-temporal pattern of varying complexity



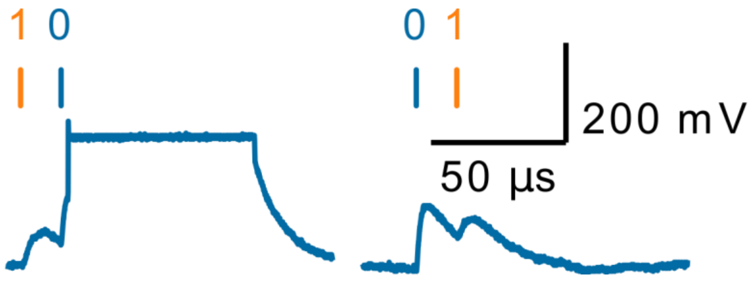


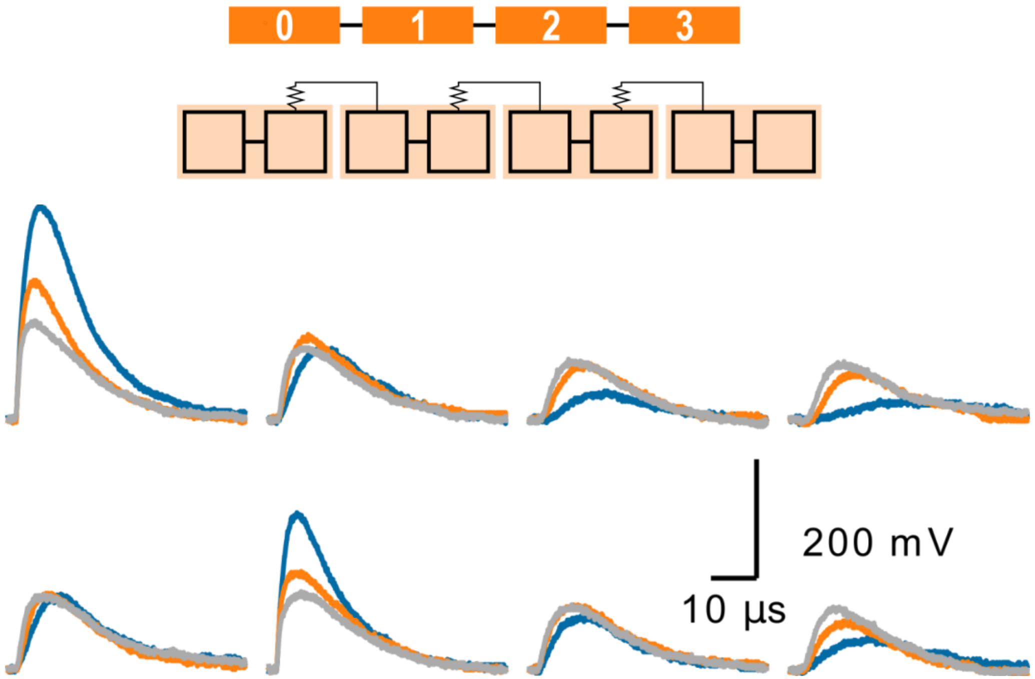
work by B. Cramer, M. Kreft, J. Zierenberg and V. Priesemann

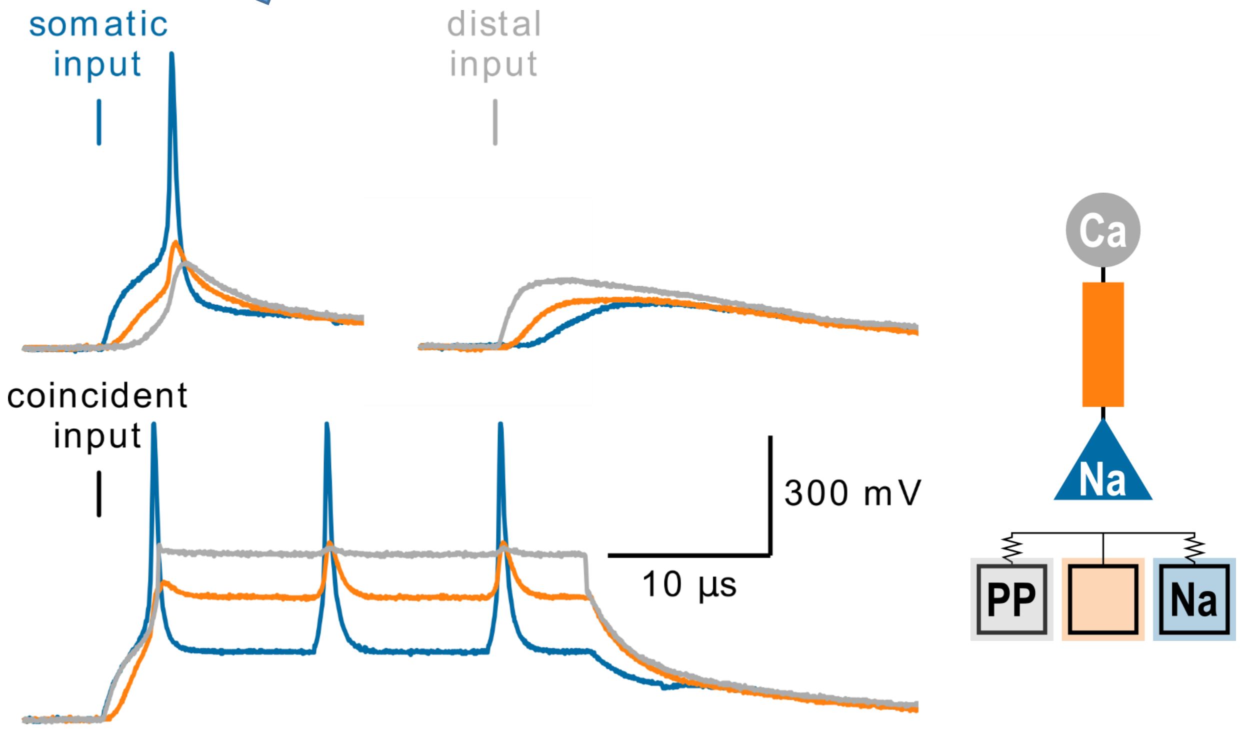
Multi-Compartment Neurons on BrainScaleS-2

- User-defined morphologies
- Parameter adjustable for each compartment
- Dendritic spikes
- No additional energy consumption
- Acceleration factor 1000

coincidence detection







work by Jakob Kaiser, Sebastian Schmitt - publication under review

Summary: analog neuromorphic computing

- successfully demonstrated gradient and plasticity-based training
- spike-based DNNs can be trained for sparseness and low latency
- rate-bases CDNN test results comparable to numerical solutions
- still a lot of software needed to maintain flexibility,
  user-friendliness has high priority → test it in the tutorials
- realization as analog co-processor provides a scalable solution
- BSS-2 demonstrates analog in-silico realization of in-memory computing for neuroscience and machine learning
- next step: combine BSS-1 wafer-scale technology with BSS-2 analog coprocessor architecture to achieve scaling in
  - size
  - speed
  - model-complexity
  - learning capabilities
  - energy