Neuromorphic Computing Architectures and Applications

PPAM2017 – Lublin (Poland) – September 10-13 2017

Karlheinz Meier
Ruprecht-Karls-Universität Heidelberg

meierk@kip.uni-heidelberg.de
@brainscales

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

The Brain

10^{15} connections (synapses)

Interacting with environment and other brains

10^{11} nodes (neurons)

Energy efficient

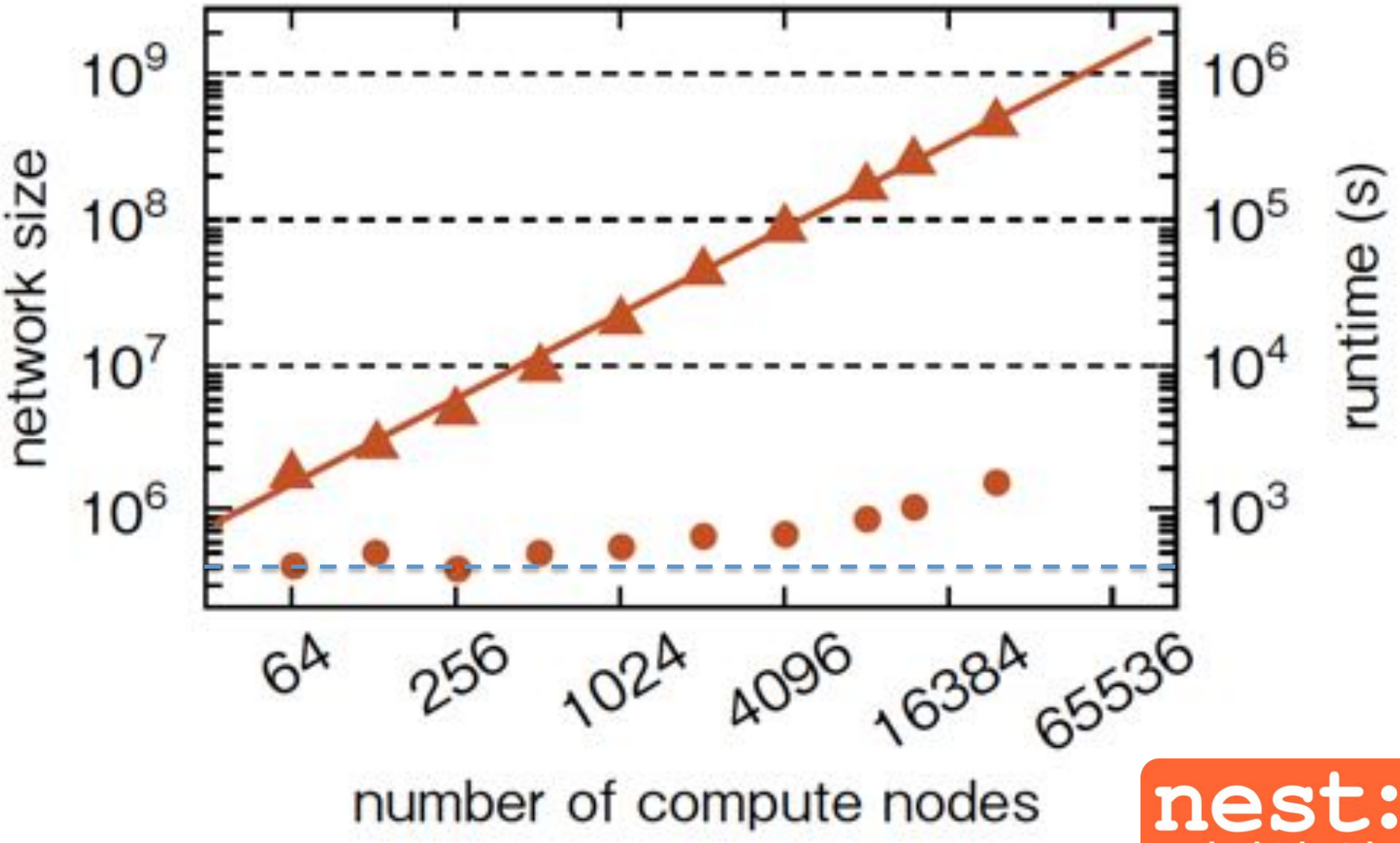
Self-learning

Fault tolerant

K-Computer, RIKEN Lab, 12.6 MW

Processor-to-Neural Cell Ratio 1 : 20.000

Simulation speed 1.520 : 1 compared to biological real-time



Diesmann, Proceedings of the 4th Biosupercomputing Symposium, Tokyo, 2012

nest:: initiative

Neuromorphic Computing

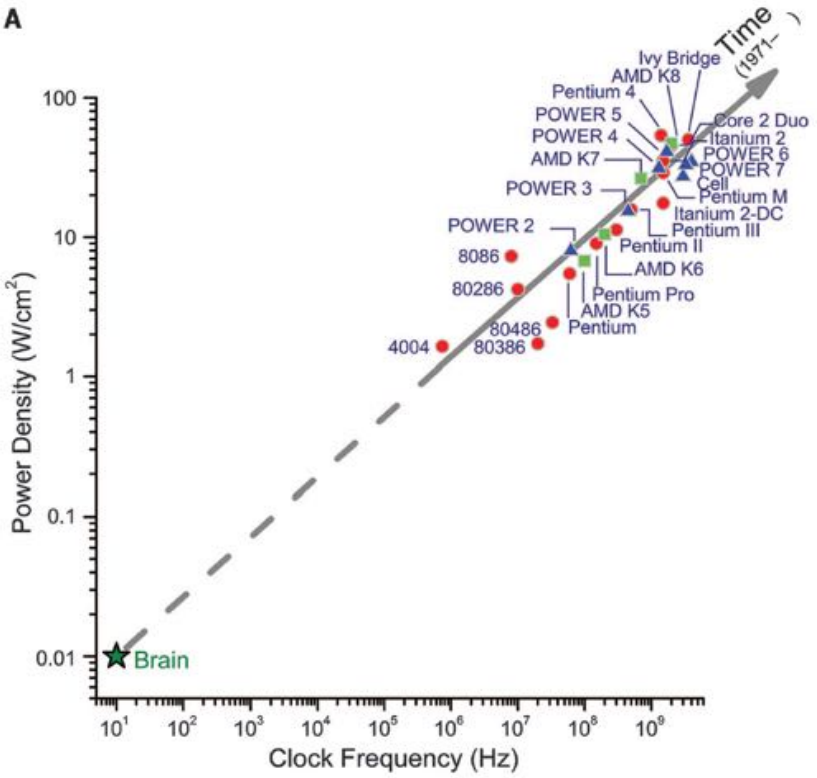
Implement some aspects of
structure and function
of biological circuits
as analogue or digital images
on electronics substrates

Structure
Cell Cores (Somas) – Networks (Axons and Dendrites) – Connections (Synapses)

Function
Local Processing – Communication – Learning

Assets of NM computing

- Energy efficiency
- Compactness
- Fault tolerance
- Speed
- Configuration and learning replace programming
- Scalability



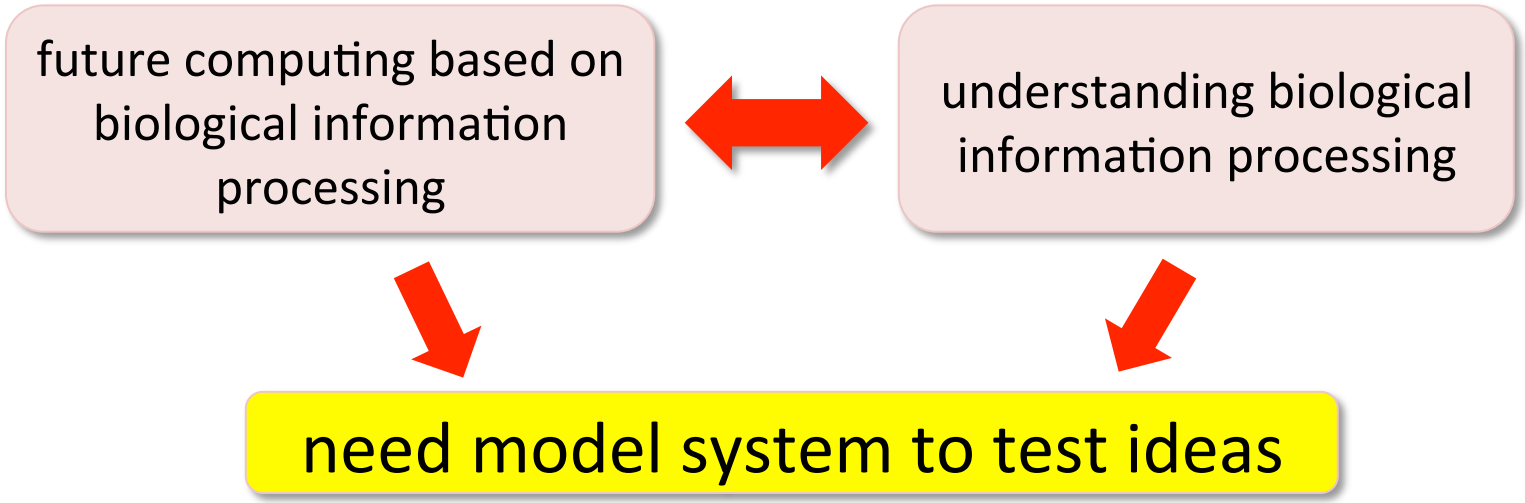
Larry Smarr, Calit

Challenges for NM Computing

- Connectivity
- Distributed memory concepts
- Configuration and learning

Conventional computing is moving away from the brain

What is the Goal ?



Two **fundamentally different** modeling approaches:

- **NUMERICAL MODEL (Turing)**

  represents model parameters as binary numbers

- **PHYSICAL MODEL (not Turing)**

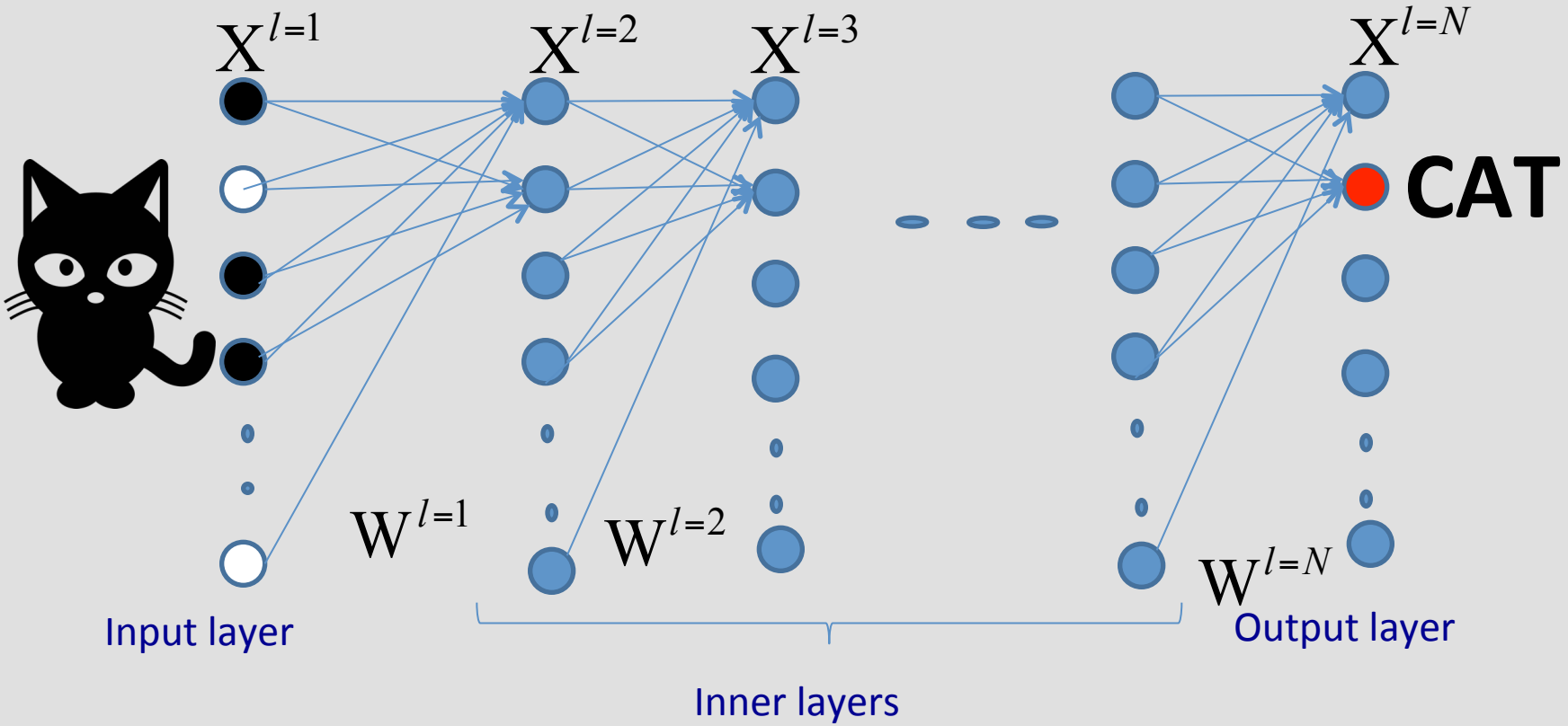
  represents model parameters as physical quantities

  → **voltage, current, charge** (like the biological brain)

can be combined to form a hybrid system

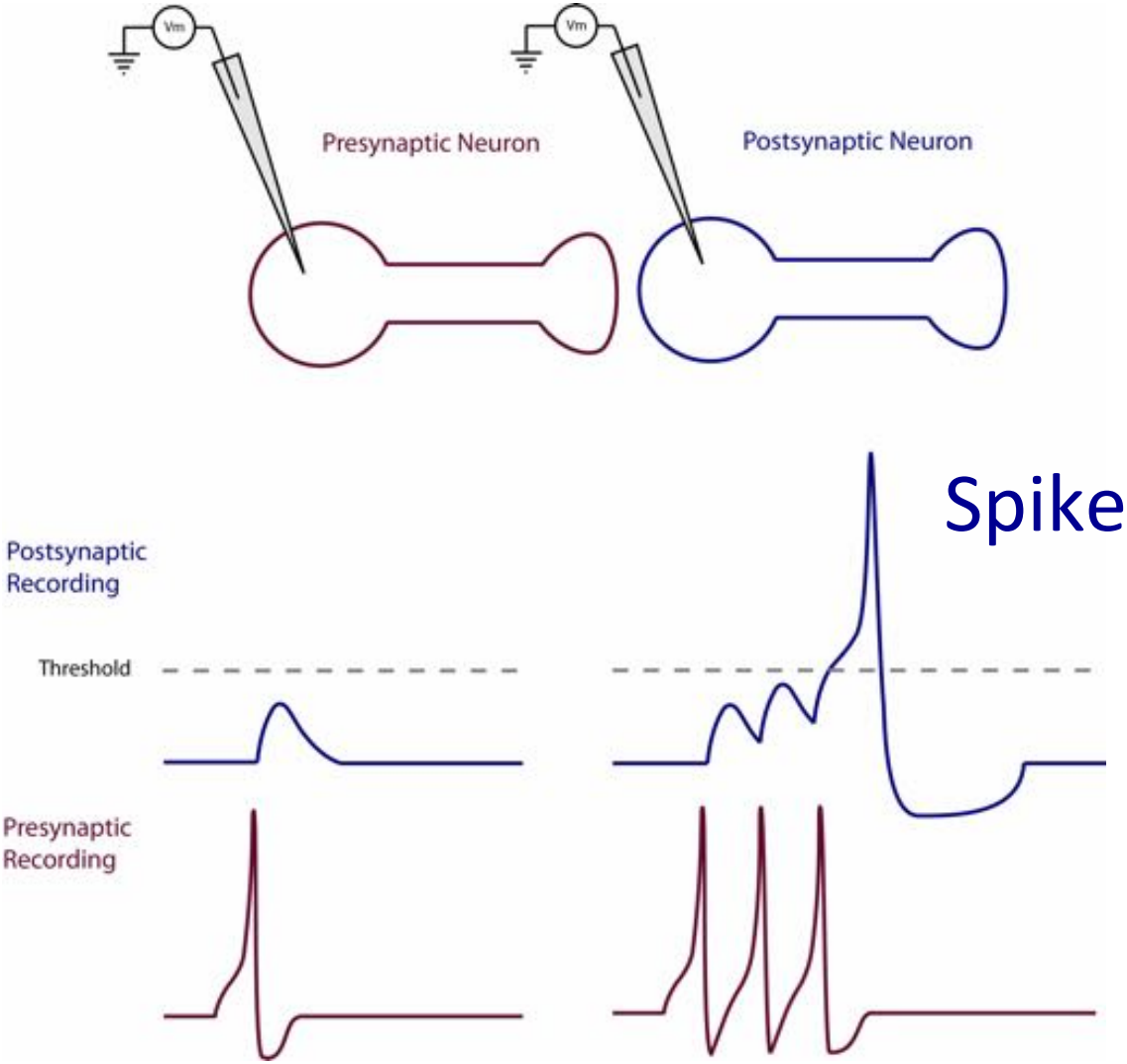
Artificial Neuronal Networks

ignore time evolution ….



Here : local, no recurrency *feed-forward*

Time and temporal integration

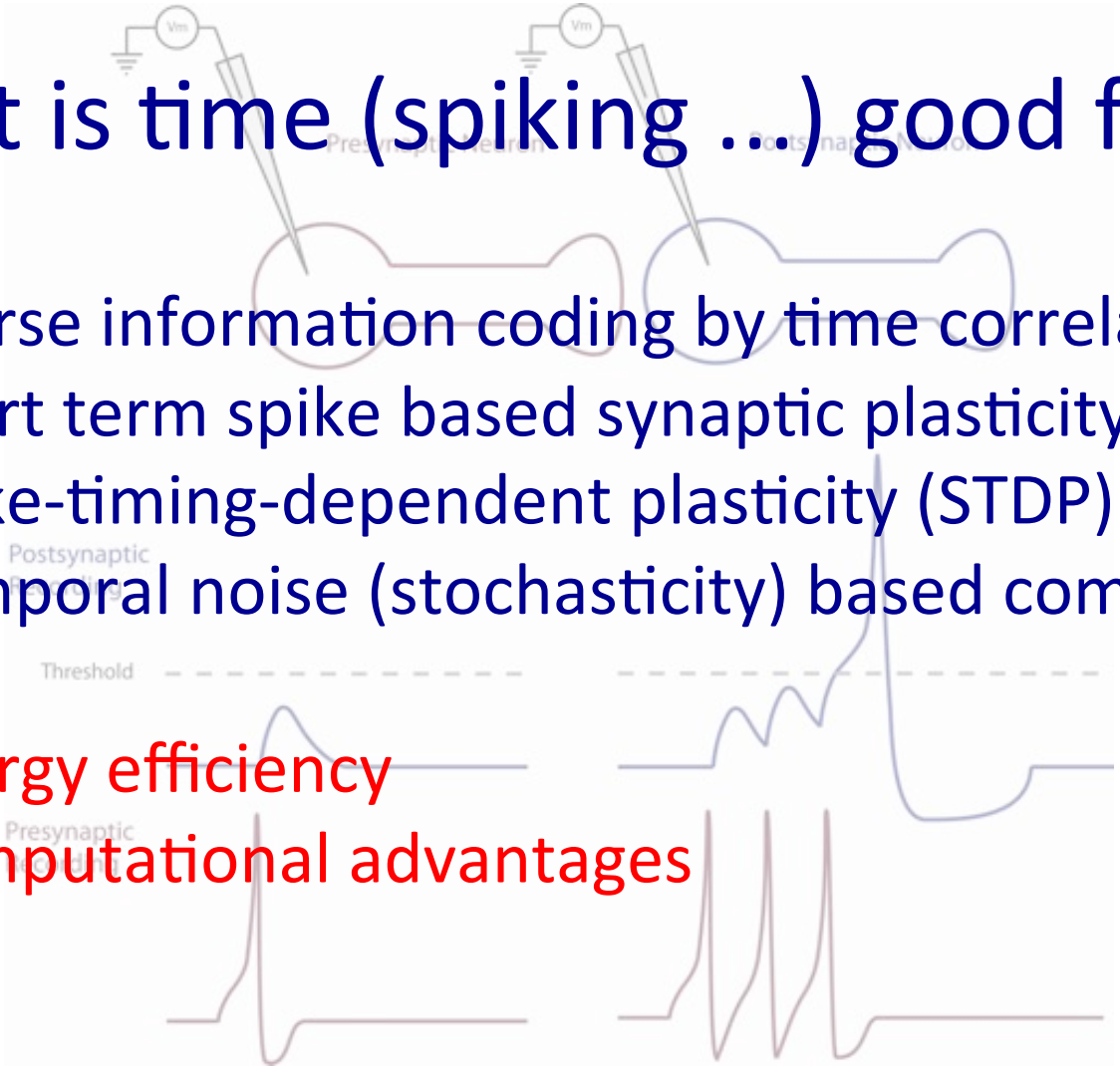


https://upload.wikimedia.org/wikipedia/commons/0/0a/Temporal_summation.JPG

What is time (spiking …) good for ?

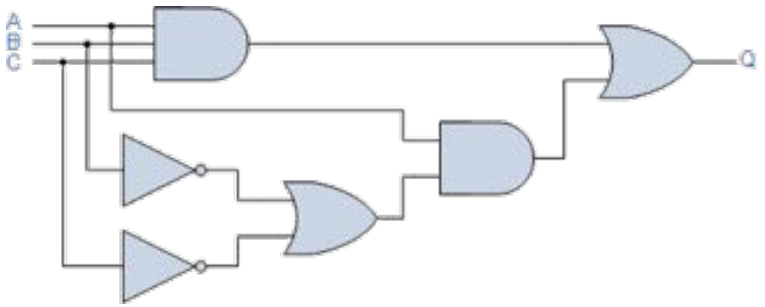
- Sparse information coding by time correlations
- Short term spike based synaptic plasticity (STP)
- Spike-timing-dependent plasticity (STDP)
- Temporal noise (stochasticity) based computing

- Energy efficiency
- Computational advantages



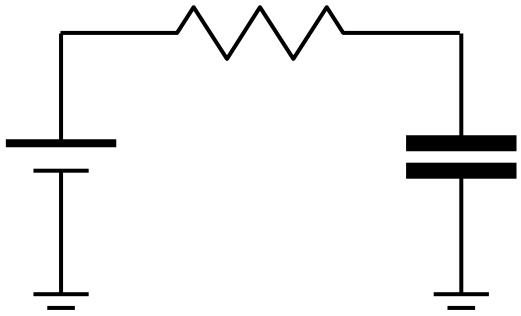
Digital

- Discrete values of physical variables
- Computation by Boolean algebra
- One wire one bit of information
- Signal restored after gate



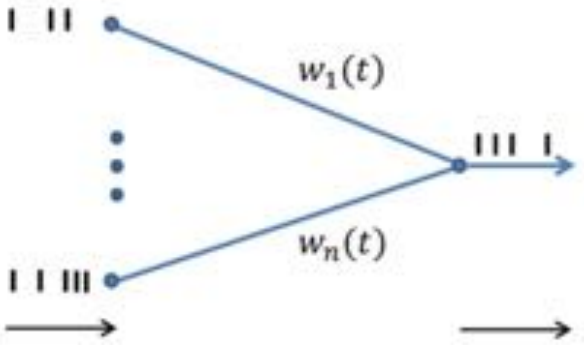
Analog

- Continuous values of physical variables
- Computation by component physics
- One wire many bits of information
- Signal not restored after stage



Nature / mixed-signal

- Local analogue computation
- Binary communication by spikes
- Signal restoration



Large-scale Neuromorphic Computing – compare

➢ Commodity microprocessors	SpiNNaker, HBP	Soft-binary-code
➢ Custom fully digital	TrueNorth, IBM	Hard-binary-code
➢ Custom Mixed-Signal	BrainScaleS, HBP	Physical model

Anything in common ?

+ Massively parallel (close to perfect weak scaling)
+ Asynchronous communication
+ Configurability
- Limited flexibility and complexity in neural models

COMPLEMENTARITY OF APPROACHES ESSENTIAL !

SpiNNaker
Biologically Inspired Massively Parallel Architectures

IBM®

BrainScaleS
ScaleS

HBP Neuromorphic Computing Concepts

MANY-CORE NUMERICAL MODEL SYSTEM

0.5 – 1 Million ARM processors – address-based, small packet, asynchronous communication – real-time simulation

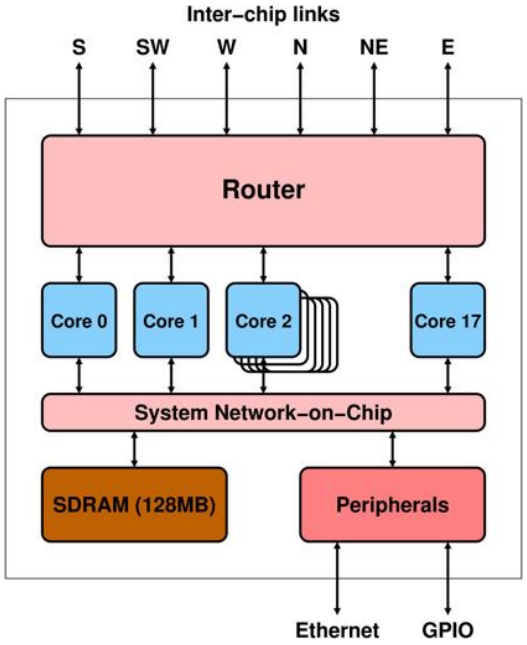
Location : Manchester (UK)

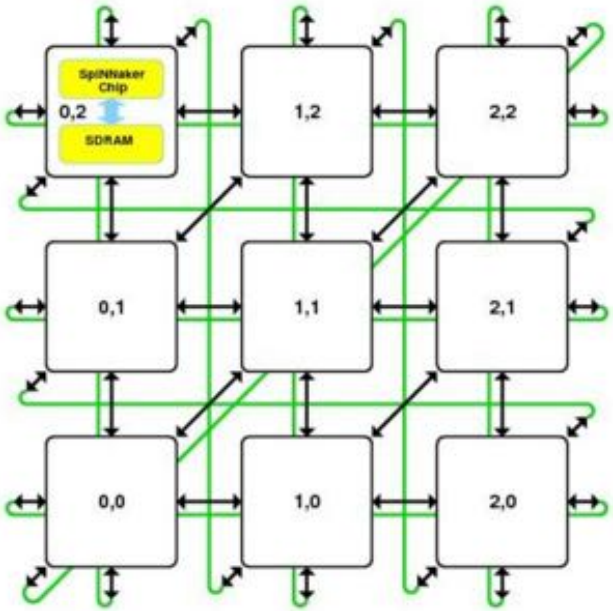
PHYSICAL MODEL SYSTEM

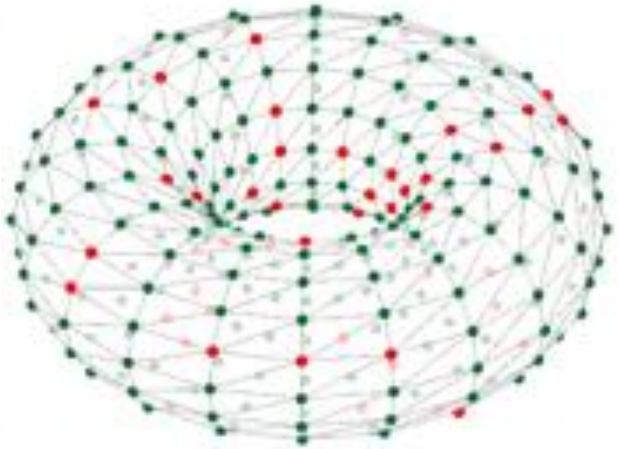
Local analog computing with 4 Million neurons and 1 Billion synapses – binary, asynchronous communication – x 10 000 accelerated emulation

Location : Heidelberg (Germany)

SpiNNaker : Many Core System







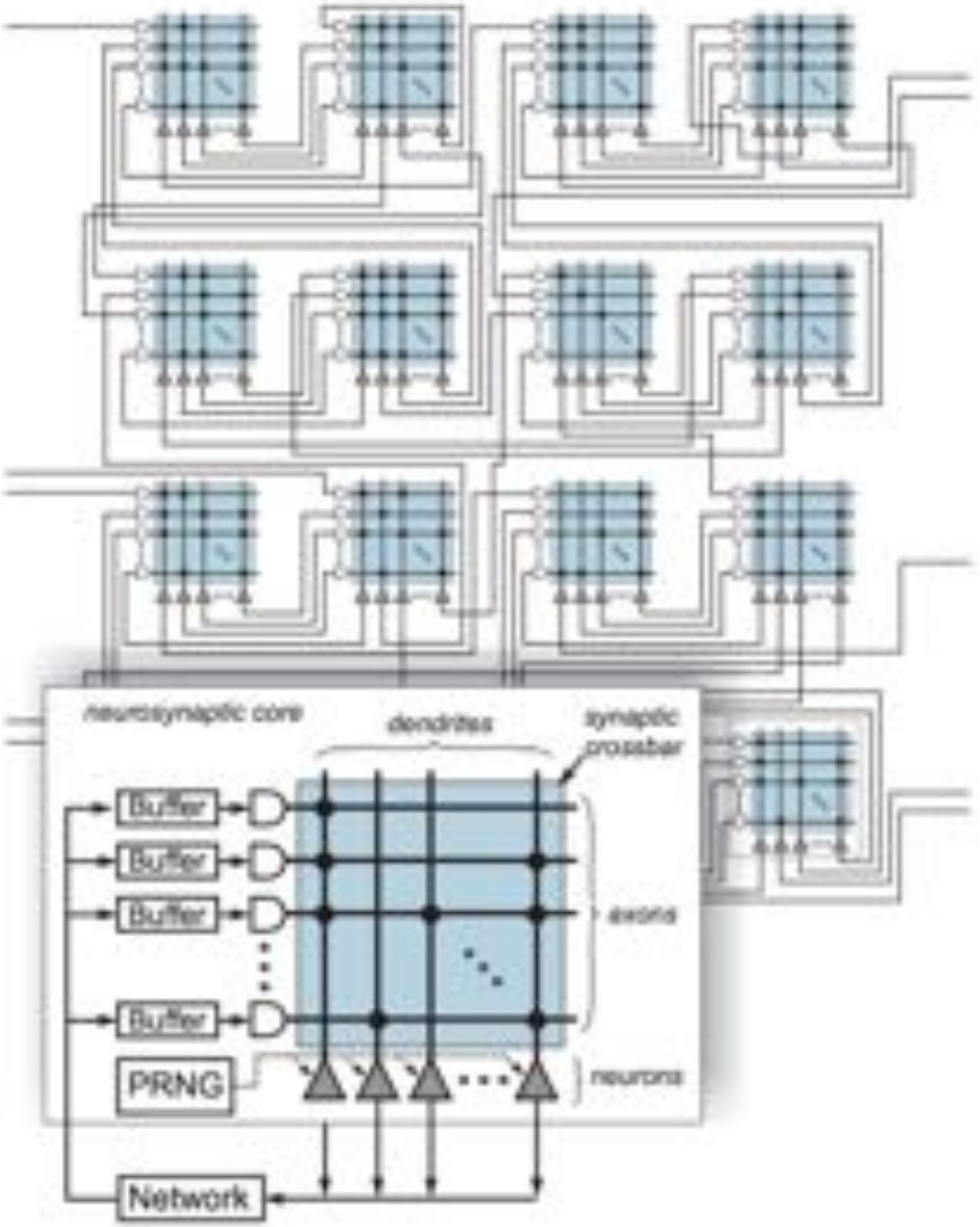
Connecting many cheap processors by spike-optimized network

- 18 *ARM 968* Cores per chip
- Integer Arithmetic
- 200 MHz Processor Clock
- Shared system RAM on die
- 128 Mbyte SDRAM stacked on die
- Each Chip 6 bi-directional links
- 6 million spikes / s / link
- Real Time Simulator

A drastic approach to weak scaling

IBM Almaden Group

- Fully custom, fully digital design
- Exploit econnomy of scale.
  28nm, Samsung
- 1 Million hardwired LIF neurons
- 256 Million 1-bit static cross bars (synapses)
- 4096 neurosynaptc cores on a 64x64 grid
- Direct axon-neuron connectivity only internal to neurosynaptic core
- 5.4 Billion transistors / chip
- No local plasticity



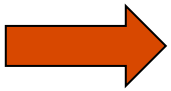
Imam, Nabil, et al. "Implementation of olfactory bulb glome
in a digital neurosynaptic core." *Frontiers in neuroscience* 6 (

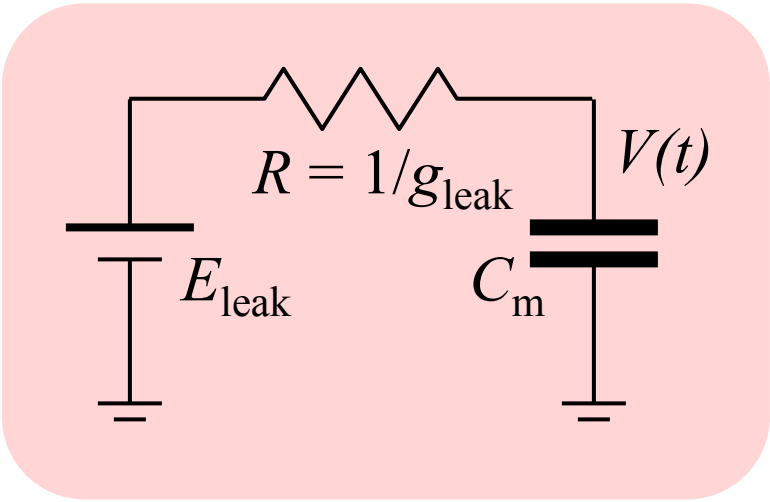
Physical Model System

Continuous Time Integrating Neural Cell Membrane

(+ non-linearity)

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





	g_{leak} [S]	C_m [F]
Biology(*)	10^{-8}	10^{-10}
VLSI	10^{-6}	10^{-13}

(*) Brette/Gerstner, J. Neurophysiology, 2005

$$c_m \frac{dV}{dt} = -g_{leak}\left(V - E_l\right) + \sum_k p_k g_k \left(V - E_x\right) + \sum_l p_l g_l \left(V - E_i\right)$$

$p_{k,l}(t)$ exponential onset and decay (post-synaptic potential shape)

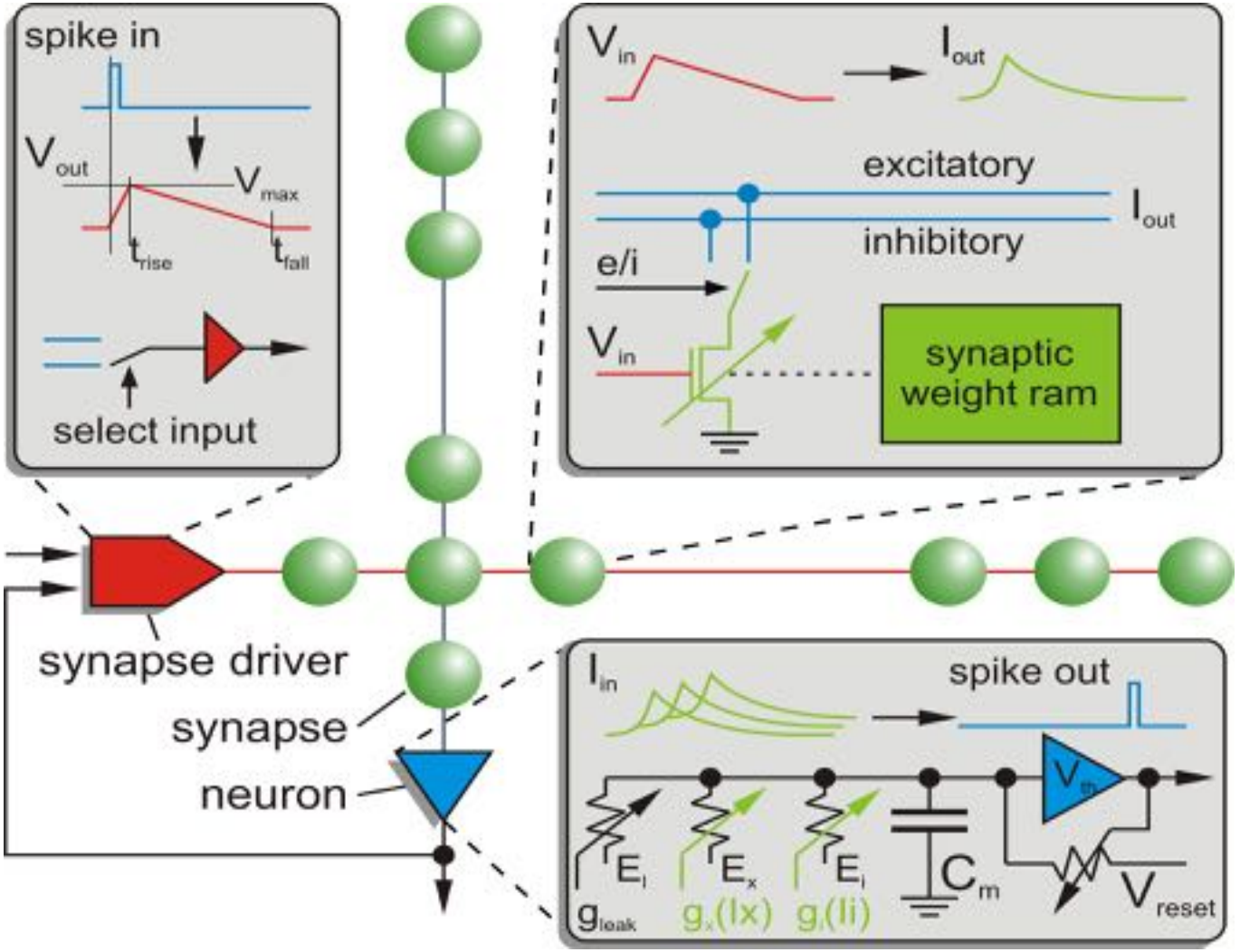
$g_{k,l}$ 0 to g_{max} (“weights”)

effective membrane time-constant c_m / g_{total} is time-dependent

„Time“ is imposed by internal physics, not by external control

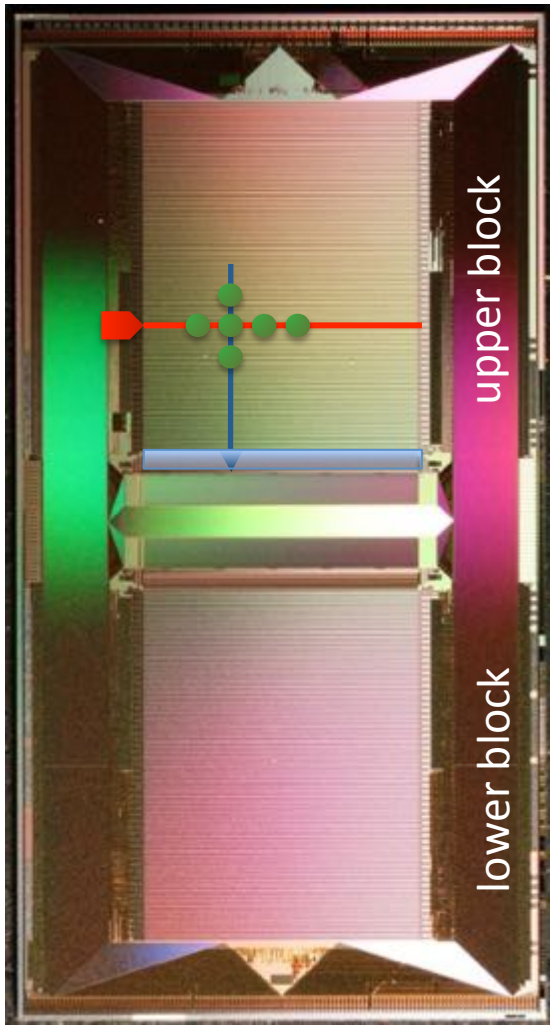
Implementation example with synaptic inputs and neuron non-linearity

mixed-signal : analog cores, binary communication



A New VLSI Model of Neural Microcircuits Including Spike Time Dependent Plasticity, Johannes Schemmel, Karlheinz Meier, Eilif Muller, Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04), IEEE Press, pp. 1711-1716, 2004

1st generation BrainScaleS Chip

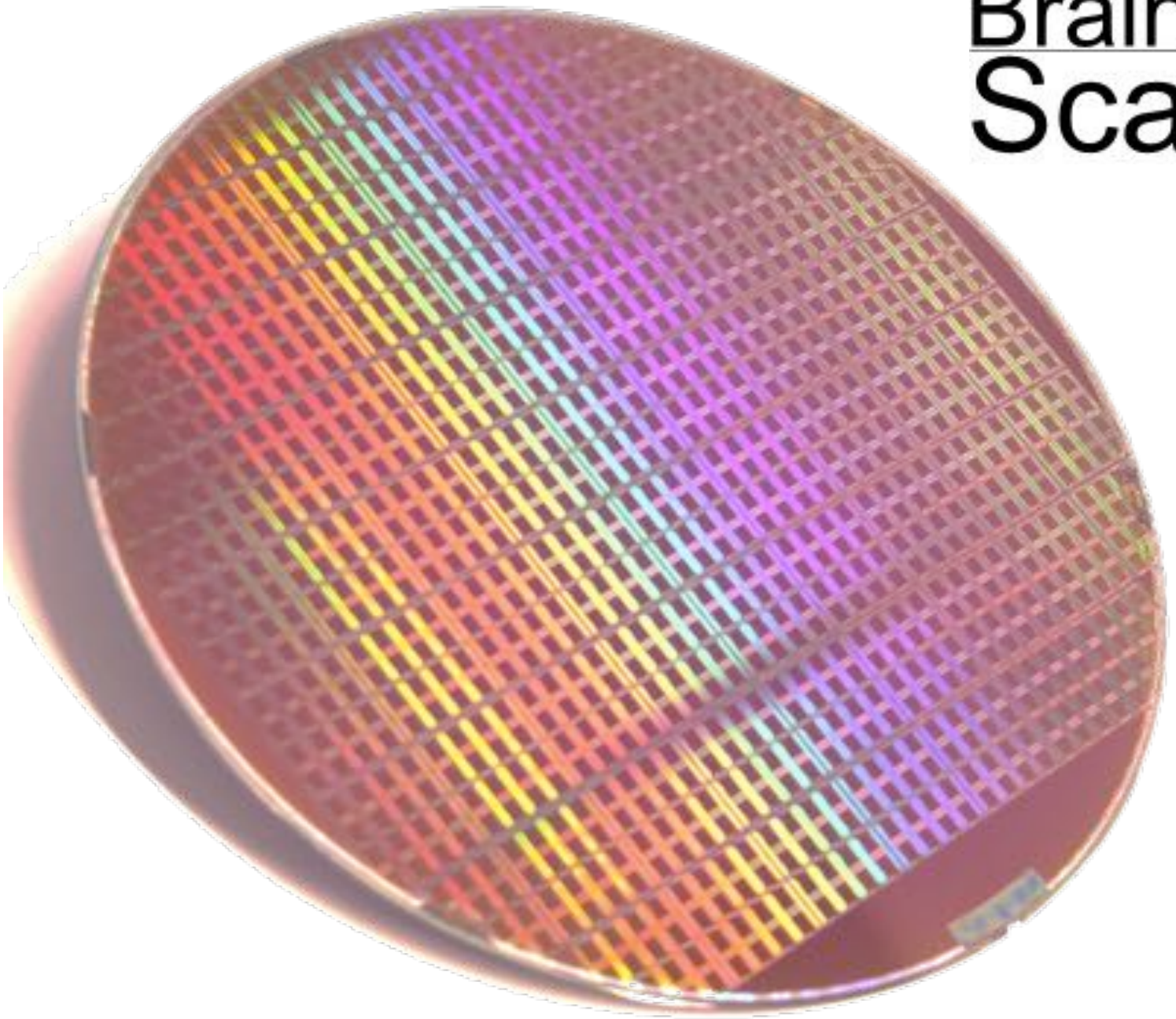


$$C\frac{dV}{dt} = -g_L(V - E_L) + g_L\Delta_T \exp\left(\frac{V - V_T}{\Delta_T}\right) + I - w, \quad (1)$$

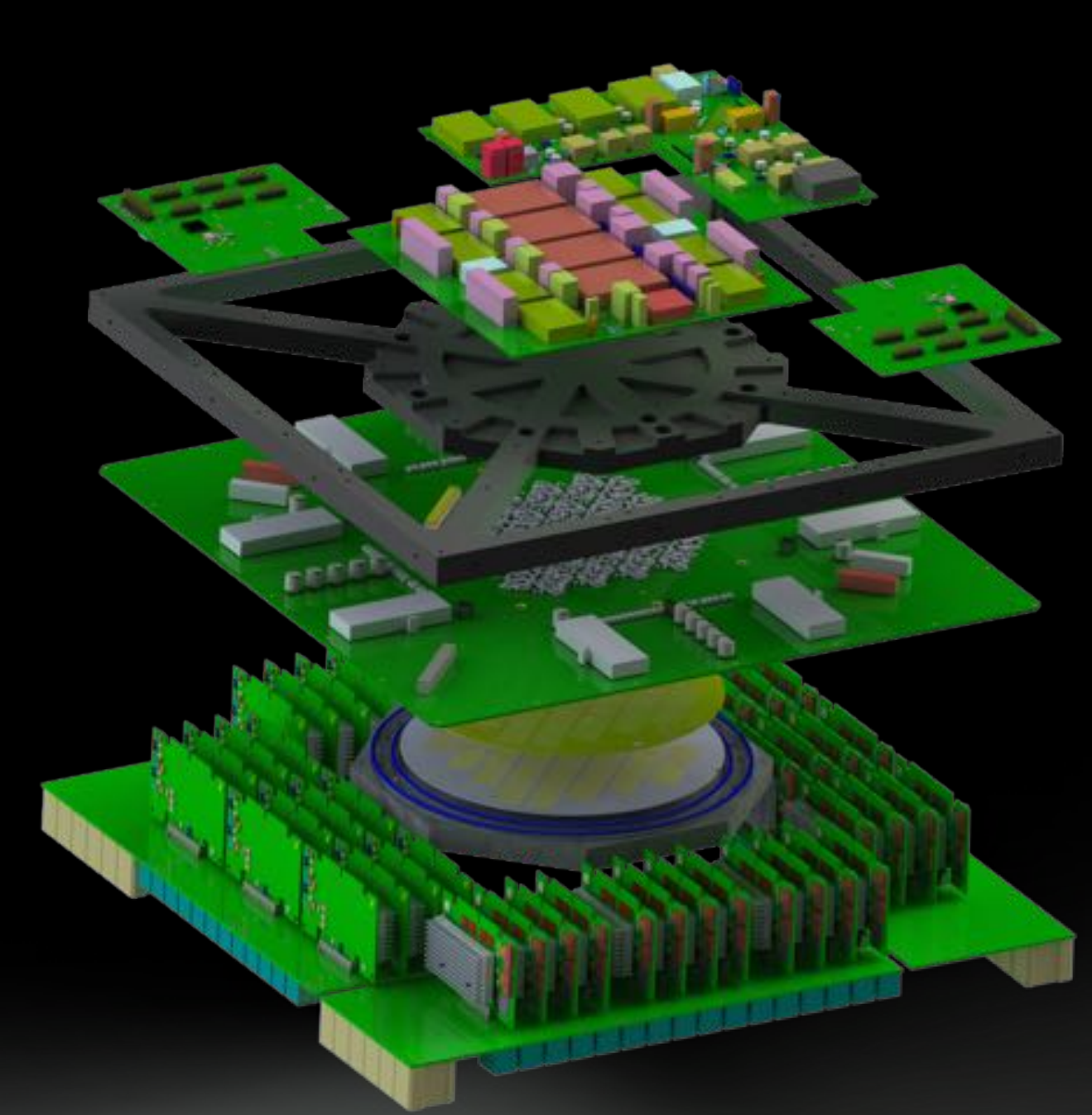
$$\tau_w\frac{dw}{dt} = a(V - E_L) - w. \quad (2)$$

- 180nm CMOS technology
- 10x5 mm die size
- 512 AdeX neurons
- 100.000 plastic synapses
- Up to 16.000 synaptic inputs per neuron
- Short term depression and potentiation
- Spike-timing-dependent plasticity

- X10.000 acceleration w.r.t. real-time
- Wafer scale integration



BrainScaleS



Physical Model, local analogue computing, binary continuous time communication

Wafer-Scale Integration of 200.000 neurons and 50.000.000 synapses on a single 20 cm wafer

Short term and long term plasticity, 10.000 faster than real-time

Wafer-scale integration of analog neural networks, J. Schemmel, J, Fieres and K. Meier
In : Proceedings of IJCNN (2008), IEEE Press, 431

BrainScaleS

Configuration Space 40 MB for a full Wafer

Scope	Name	Type	Description
Neuron circuits (A)			membrane voltage
Synapse line drivers (B)			amp
Synapses (B)			
STP related (C)			on
STDP related (D)	n/a	i_t	Bias current controlling delay for presynaptic correlation pulse (for calibration purposes)
	$A_{+/-}$	s_t	Two voltages dimensioning charge accumulation per (anti-)causal correlation measurement
	n/a	s_t	Two threshold voltages for detection of relevant (anti-)causal correlation
	τ_{STDP}	g	Voltage controlling STDP time constants

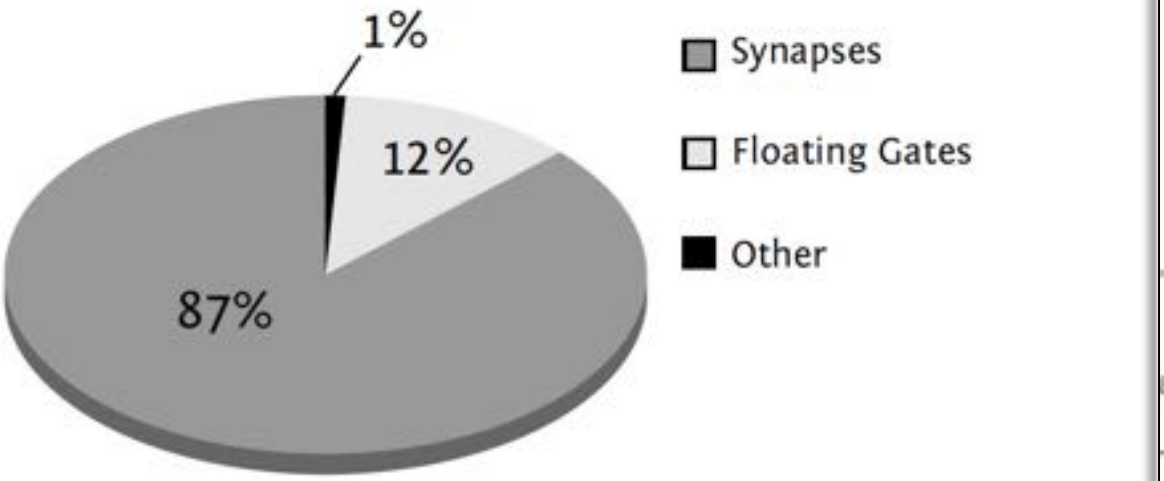
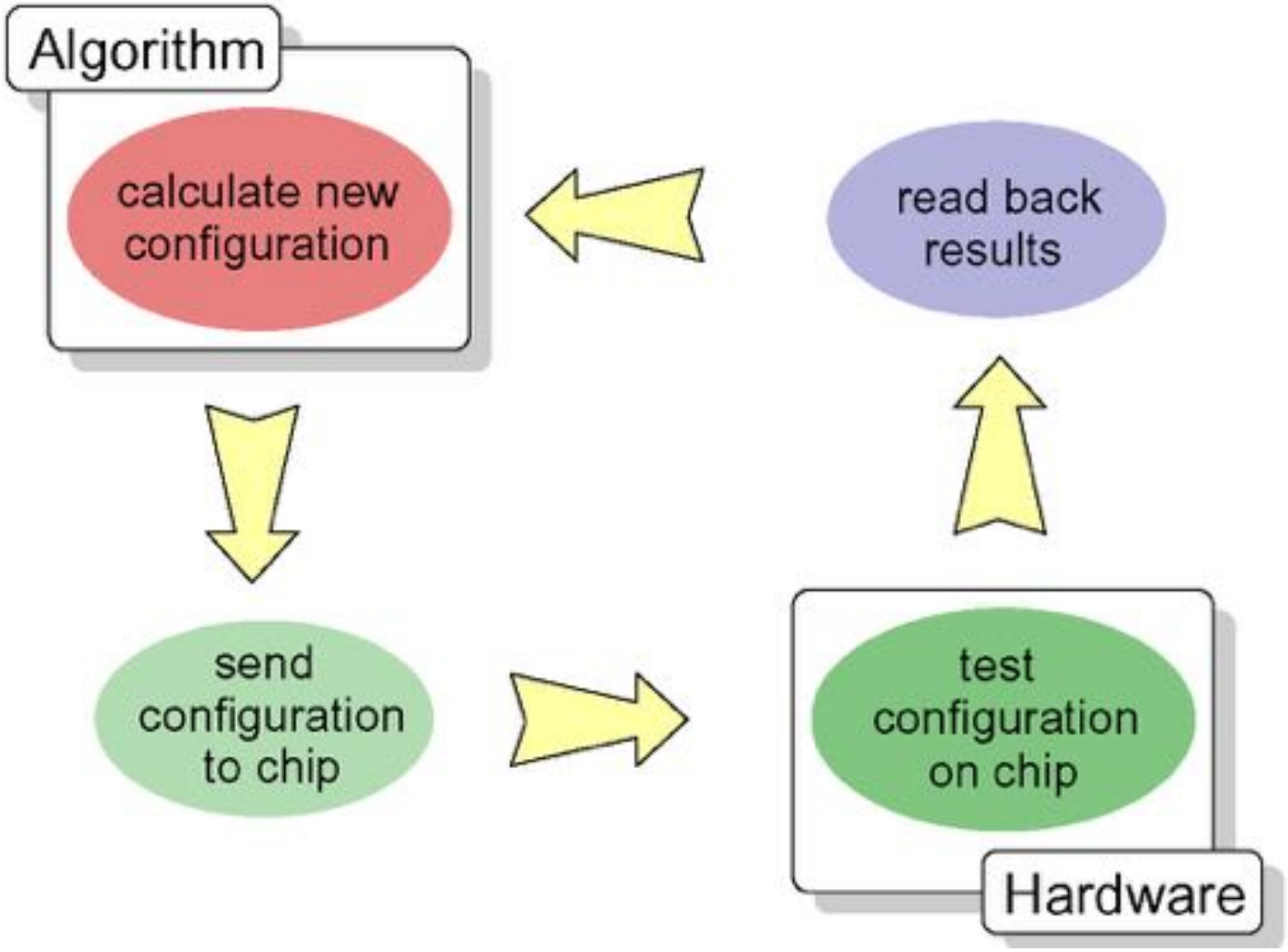
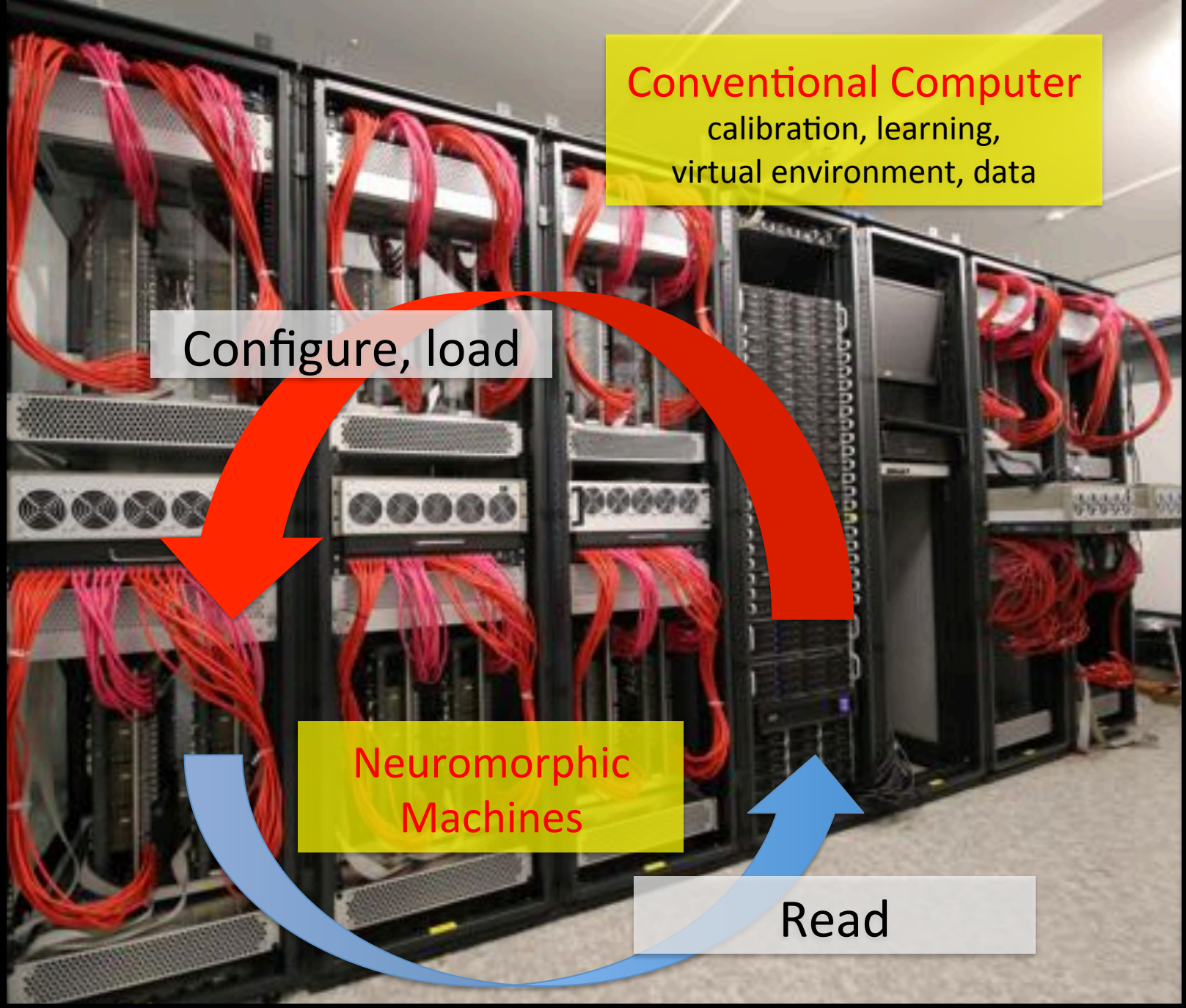


Fig. 4: Sector diagram of the parameter space to configure one HICANN chip. For a full wafer, the configuration data volume is 44 MB large.

Hardware-In-the-Loop

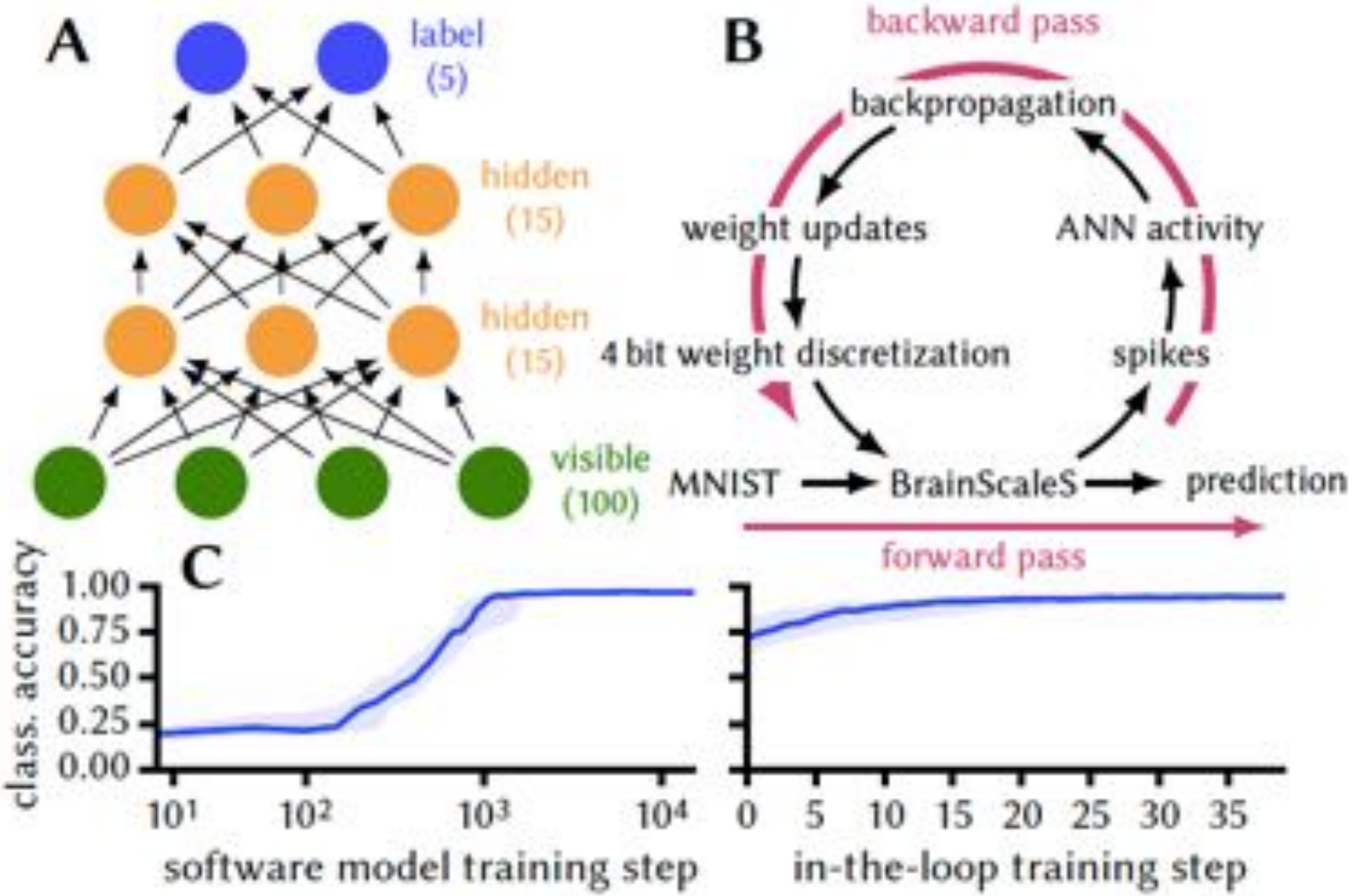



Conventional Computer
calibration, learning,
virtual environment, data
Configure, load
Neuromorphic
Machines
Read

4 Application Examples

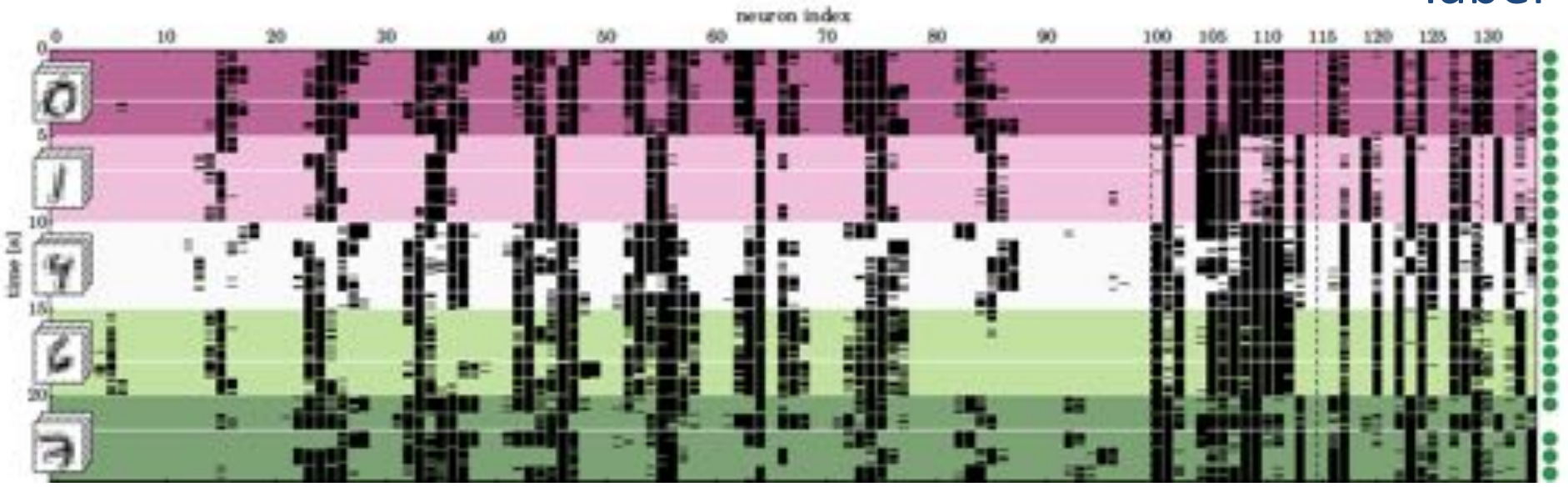
- Feed-forward, deep, spiking network (deterministic)
- Boltzmann machine, stochastic computing (stochastic)
- Reverse engineered biological circuit with learning (supervised)
- Reverse engineered biological circuit with learning (unsupervised)

Feed-forward, rate-based. 4-layer spiking network
MNIST classification on a physical model machine
performance before and after hardware in-the-loop learning



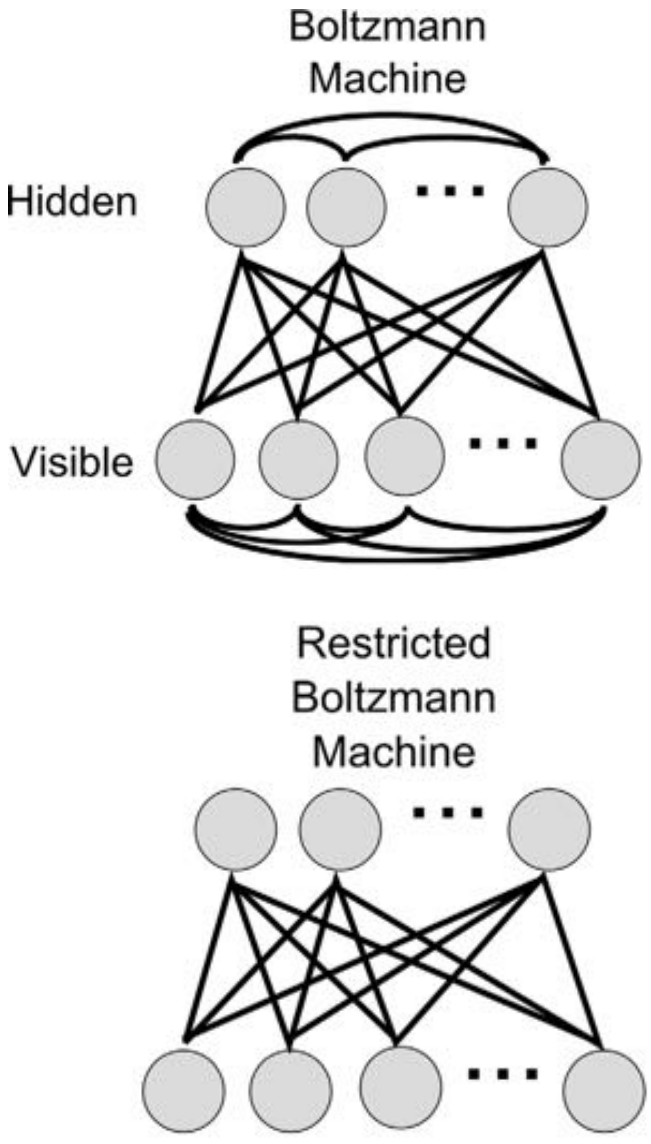
Sebastian Schmitt et al., accepted ISCAS 2017

MNIST classification on a physical model machine
Neuronal firing activity after hardware in-the-loop learning



input

2 x hidden



Boltzmann Machines

Networks of symmetrically connected stochastic nodes k

State of nodes described by vector of binary random variables z_k (0,1)

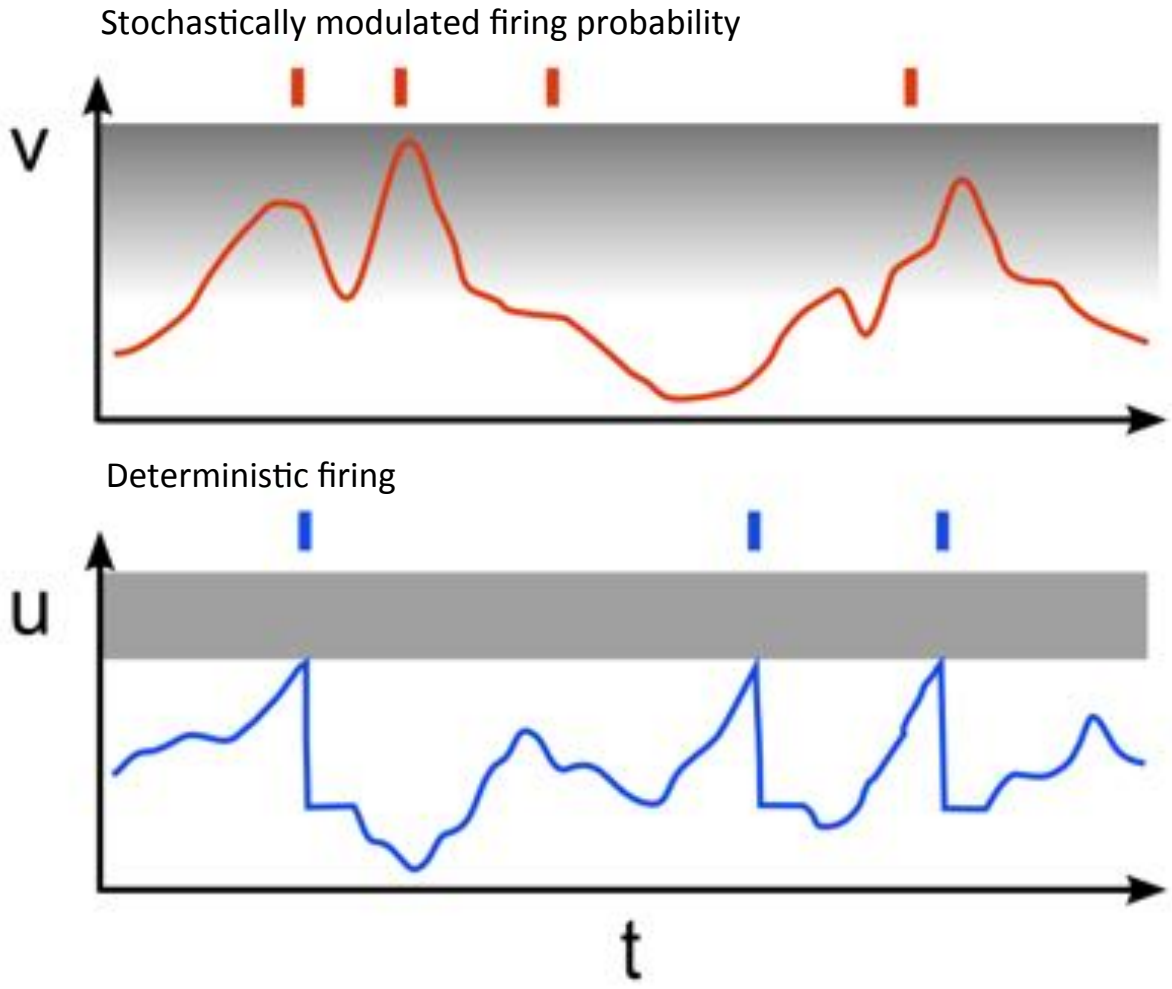
Probability for state-vector converges to a target Boltzmann-distribution

$$p(\vec{z}) = \frac{1}{Z} \exp\left[-E\left(\vec{z}\right)\right]$$

With an energy function

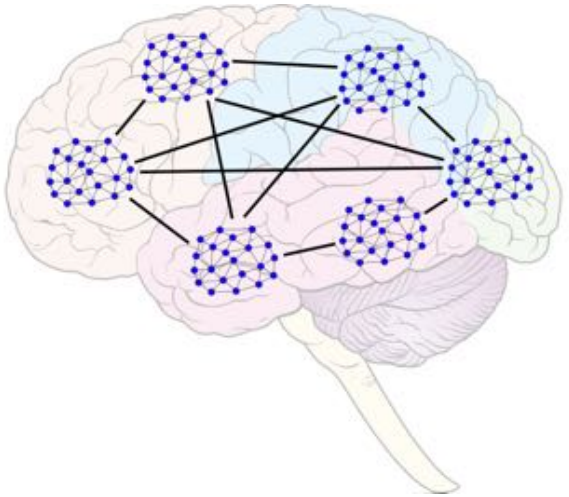
$$E(\vec{z}) = -\frac{1}{2}\sum_{i \neq j} w_{ij} z_i z_j - \sum_i b_i z_i$$

WHAT FOR ? Learn internal stochastic model of input space – Generate or discriminate



Stochasticity from

- External noise source
- Internal noise source
- Ongoing network activity

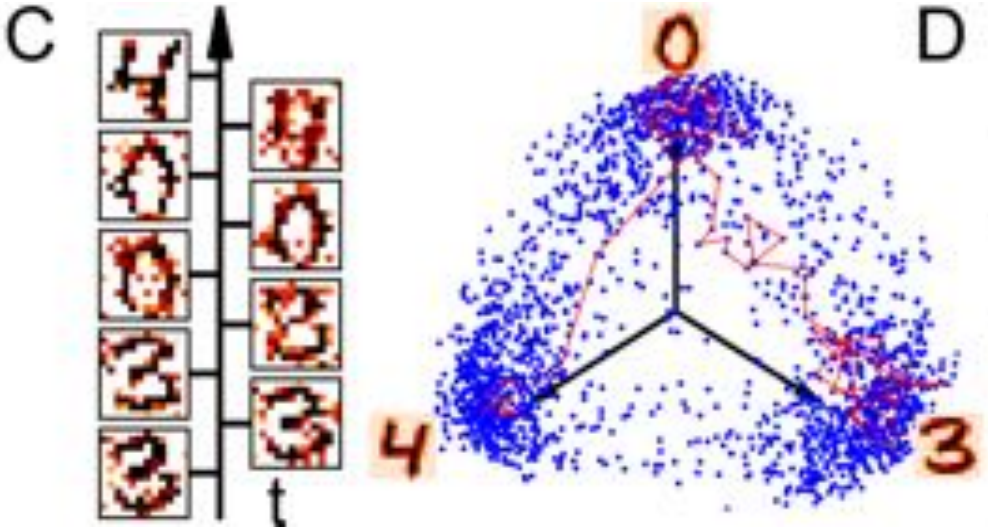


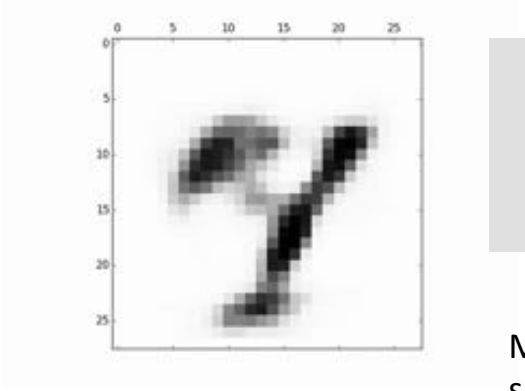
M. A. Petrovici, J. Bill, I. Bytschok, J. Schemmel, and K. M.:
Stochastic inference with spiking neurons in the high-conductance state. Physical Review E 94, 2016

LEARNING by adjusting LOCAL connections in BOLTZMANN MACHINES

- Clamp visible units to particular pattern – reach thermal equilibrium
- Incremement interaction between any 2 nodes that are both on

- Generate from stored probability distribution
- Infer from clamped input



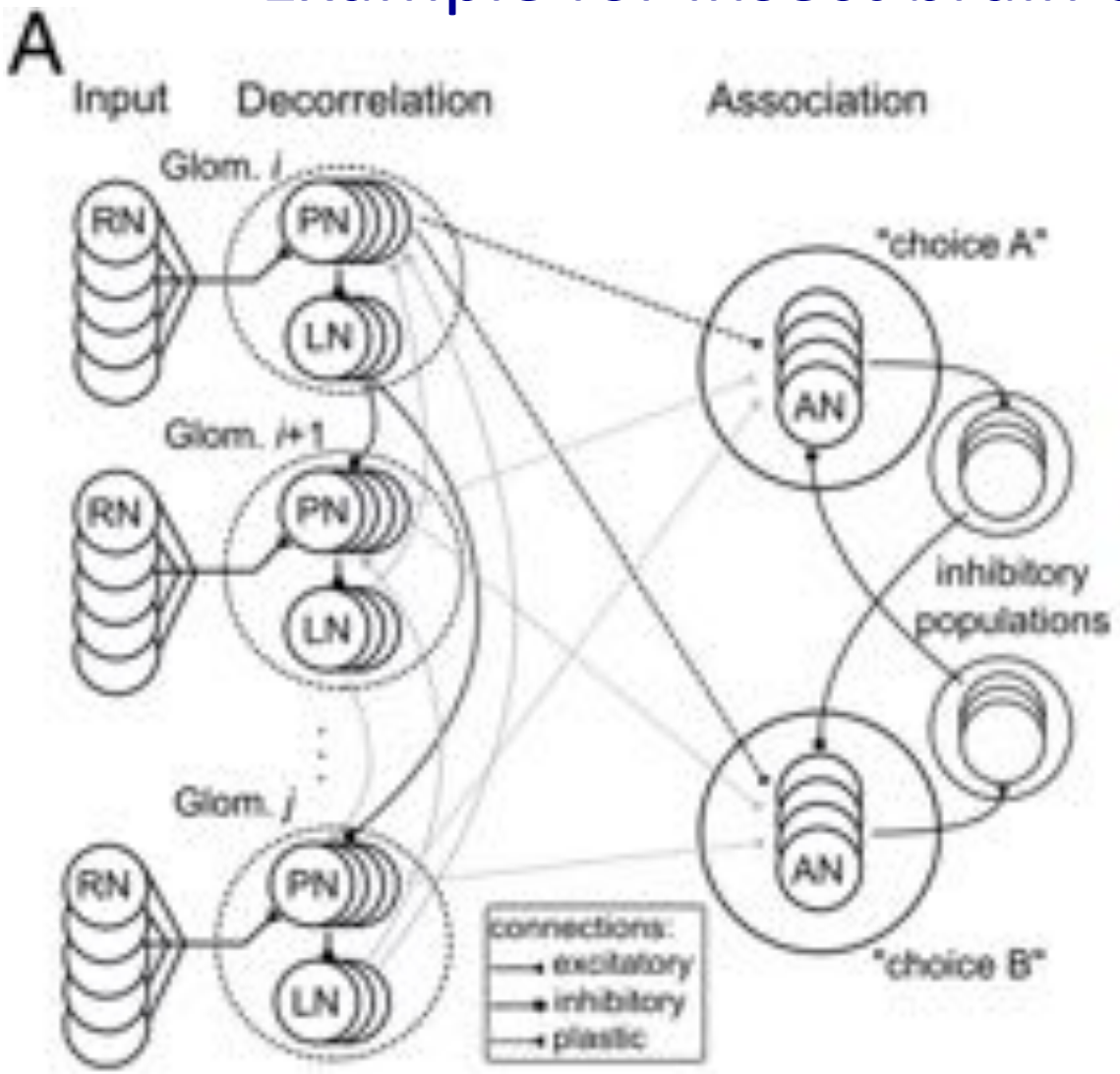


Free running
„Dreaming“
Generative

Inferring
Input incompatible with 0
Discriminative

M. A. Petrovici, J. Bill, I. Bytschok, J. Schemmel, and K. M.: Stochastic inference with spiking neurons in the high-conductance state. Physical Review E 94, 2016

Example for insect brain derived circuit



3 Layer Spiking Neuron Network derived from Insect Olfactory System

L I : Receptor Neurons

L II : Decorrelation through lateral inhibition (Glomeruli)

L III : Association (Soft WTA through strong inhibitory populatuions)

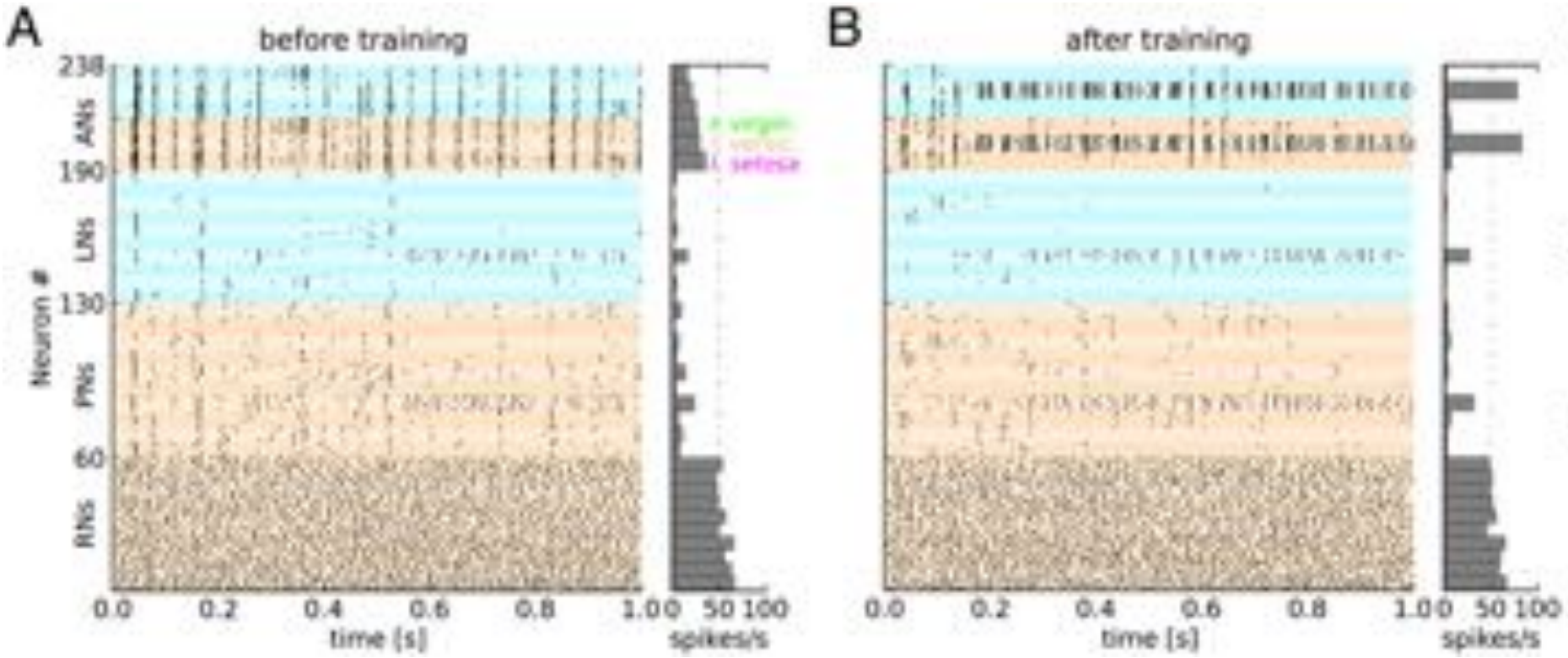
Supervised Learning

Synaptic Projections from Layer 2 to Layer 3

Schmuker, M. et al., "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

Neuronal firing activity before and after learning

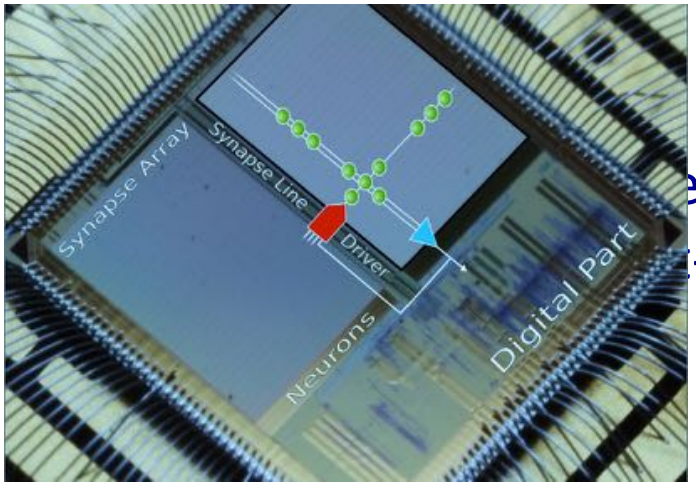
Application in generic multivariate data classification

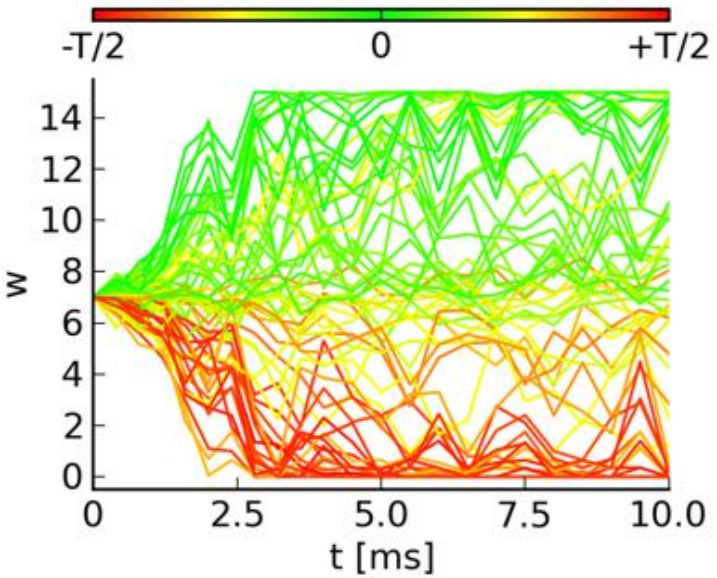


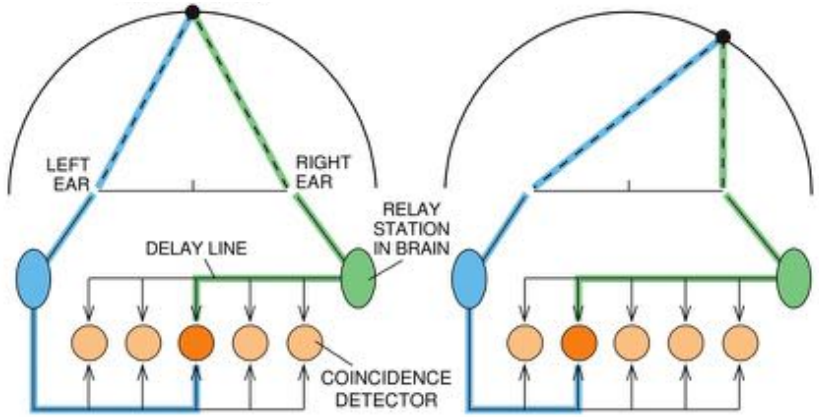
Schmuker, M.et al., "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

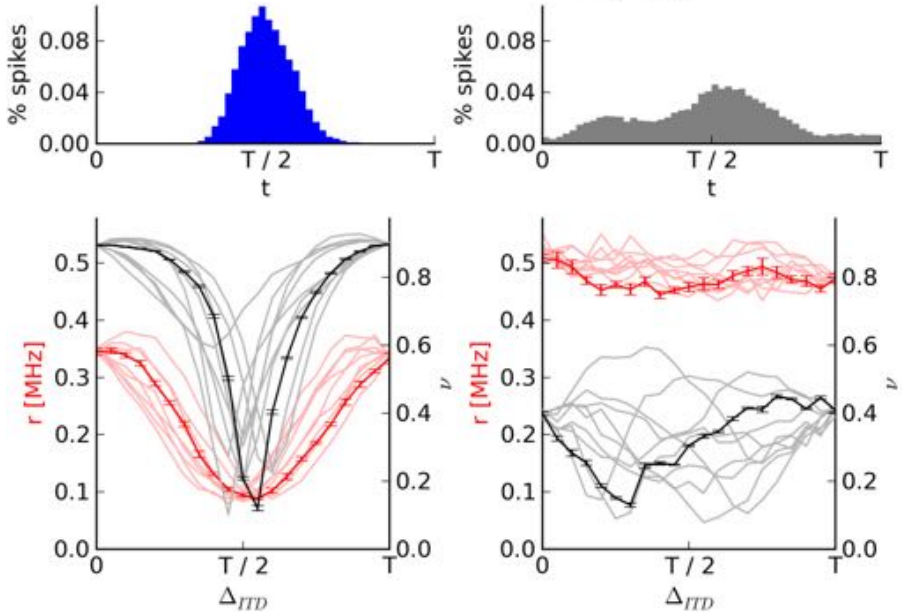
Barn owl derived echolocation by precision phase detection

Application of on-chip learning via spike-timing dependent plasticity (STDP)

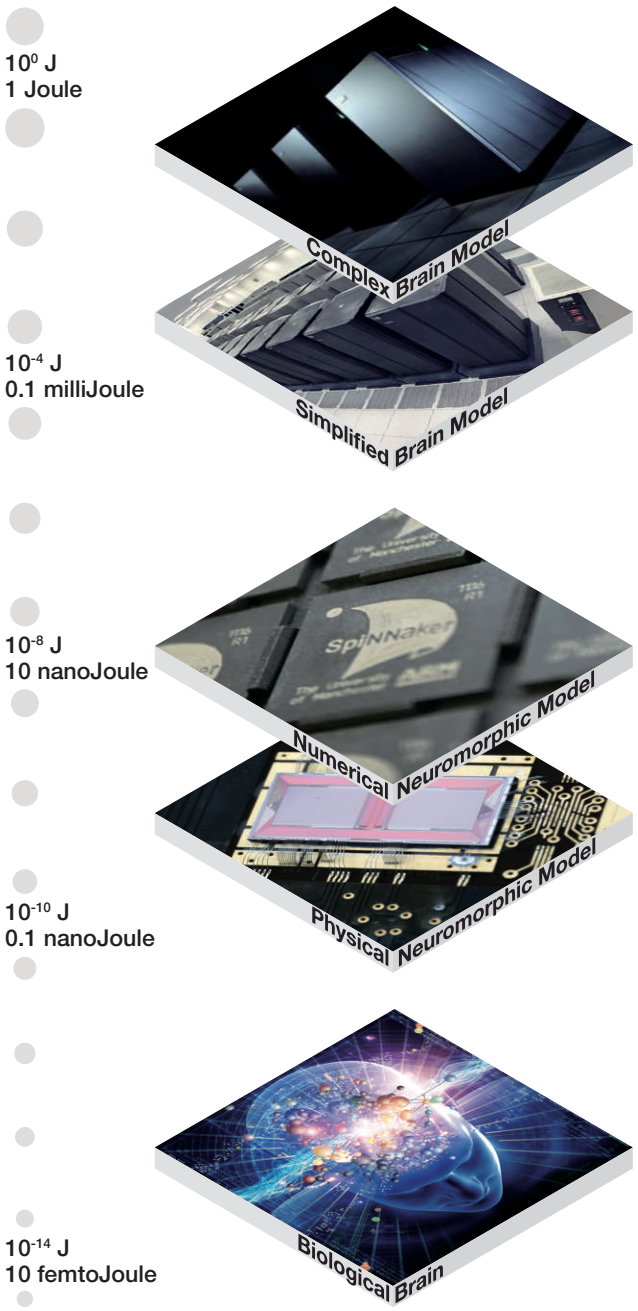








T. Pfeil, A.-C. Scherzer, J. Schemmel and K. Meier,
Neuromorphic Learning towards Nano Second Precision,
Proceedings 2013 International Joint Conference on Neural Networks
Dallas, TX, USA: IEEE Press, 2013, pp. 869-873.



Energy*Scales*

Energy used for a synaptic transmission

Filling the Gap

- Typically 10.000.000 times more energy efficient than state-of-the art HPC (comparable model)
- 10.000 less efficient than biology

From : HBP project report

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			

2nd generation

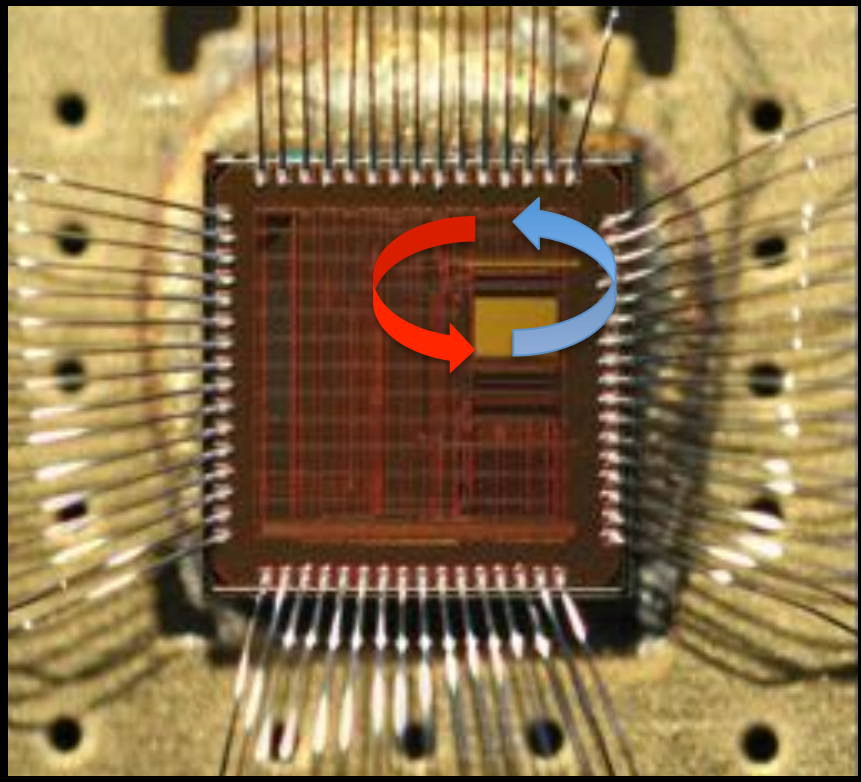
- Hybrid plasticity with on-chip processor (PPU): on-chip loops, time-scales from ms to years
  - Input : timing correlations, rates, membrane potentials, external signals
  - Change : synaptic weights, neuromodulation, network structure
- Structured neurons
  - Multicompartment neurons
  - Active, non-linear dendrites, backpropagating APs
  - NMDA, Ca plateau potentials

Initial applications

- *Neural backpropagation learning*
- *HTM temporal prediction*

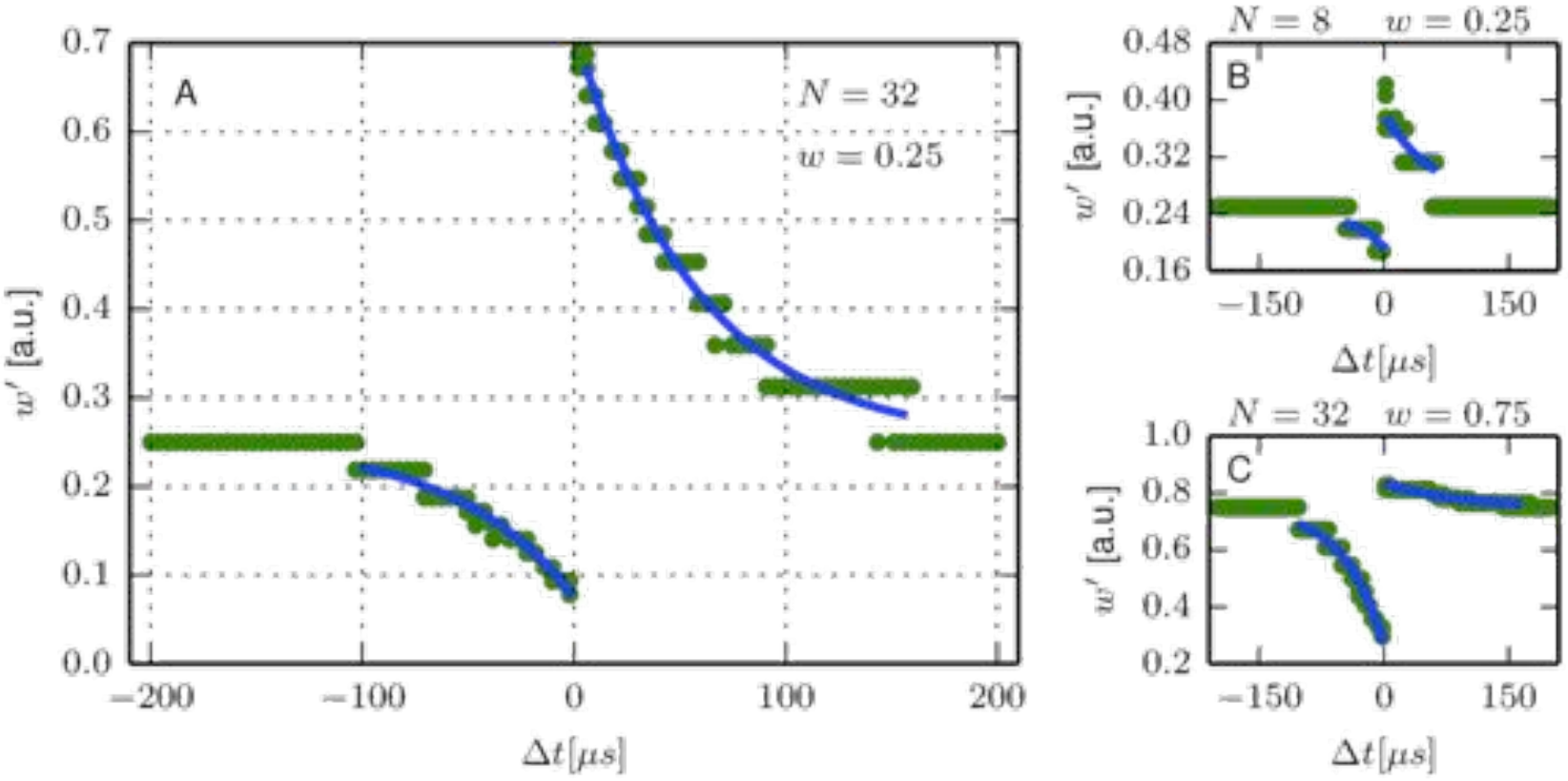
BrainScaleS-2

65 nm prototype chip in the lab

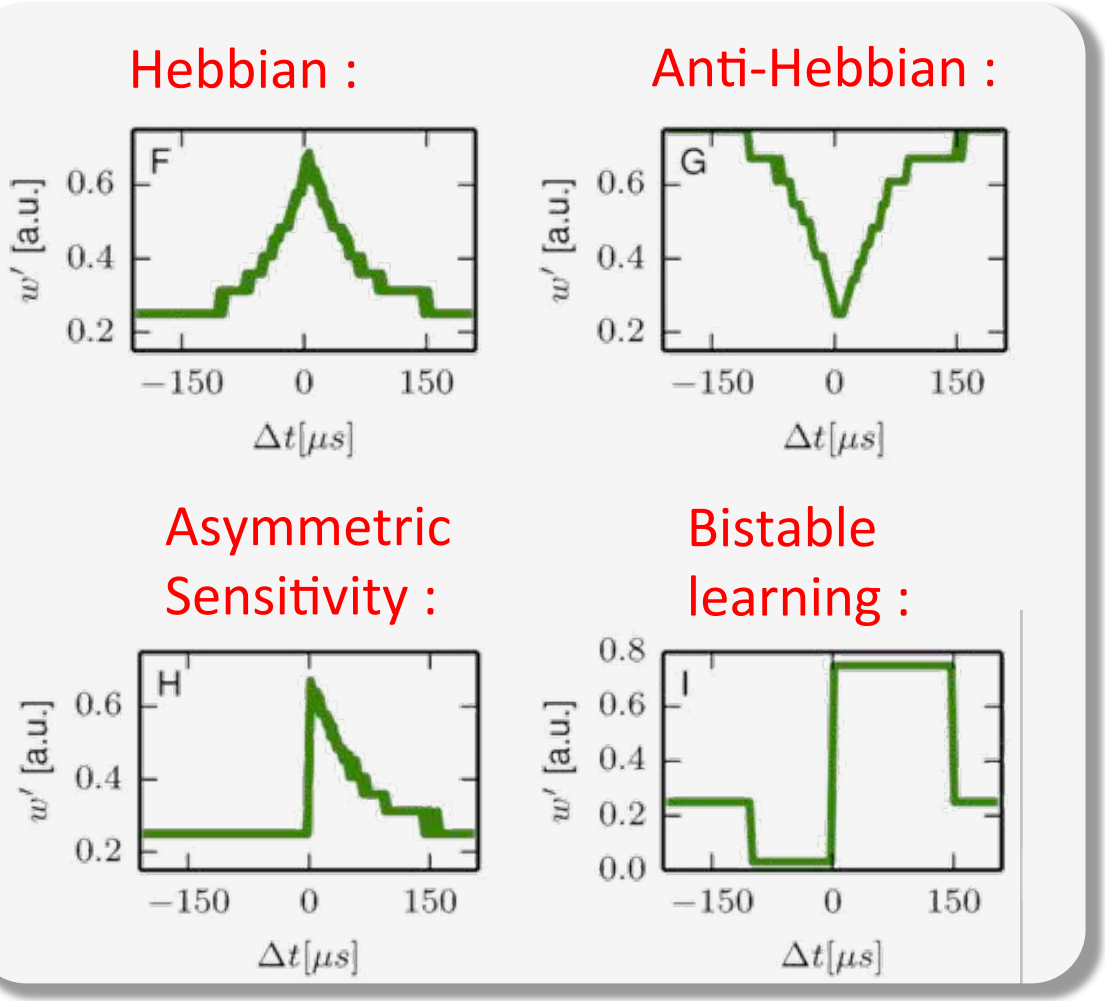


- Public evaluation system by mid-2018
- Full-size prototypes by mid-2020

PPU : Measurement Results for Multiplicative STDP Rule



PPU : Measurements demonstrating variable STDP rules

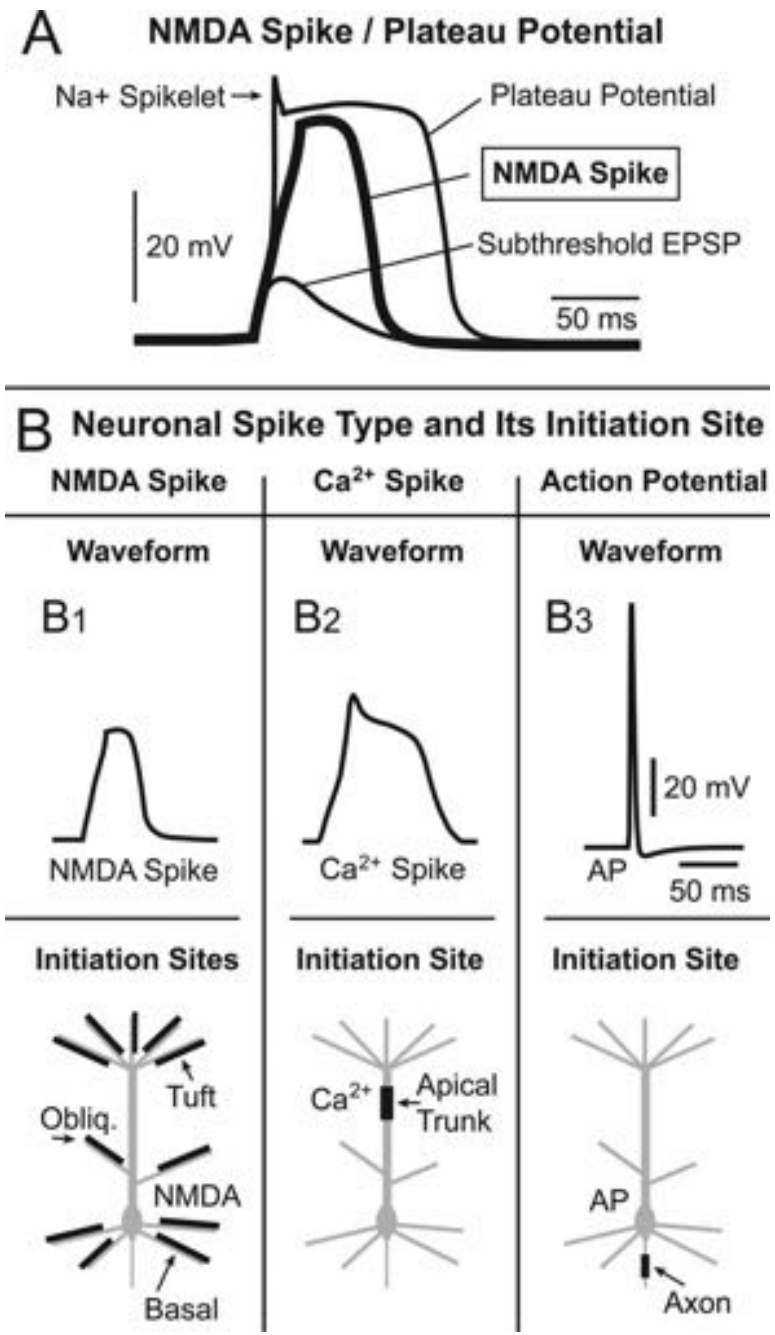


- So far only variations of the STDP PPU code
- PPU also supports :
  - Gated plasticity
  - Reinforcement learning
  - Neuromodulation
  - Re-wiring on the fly
  - structural plasticity
  - anything you can code
  - any time scale

Biological finding

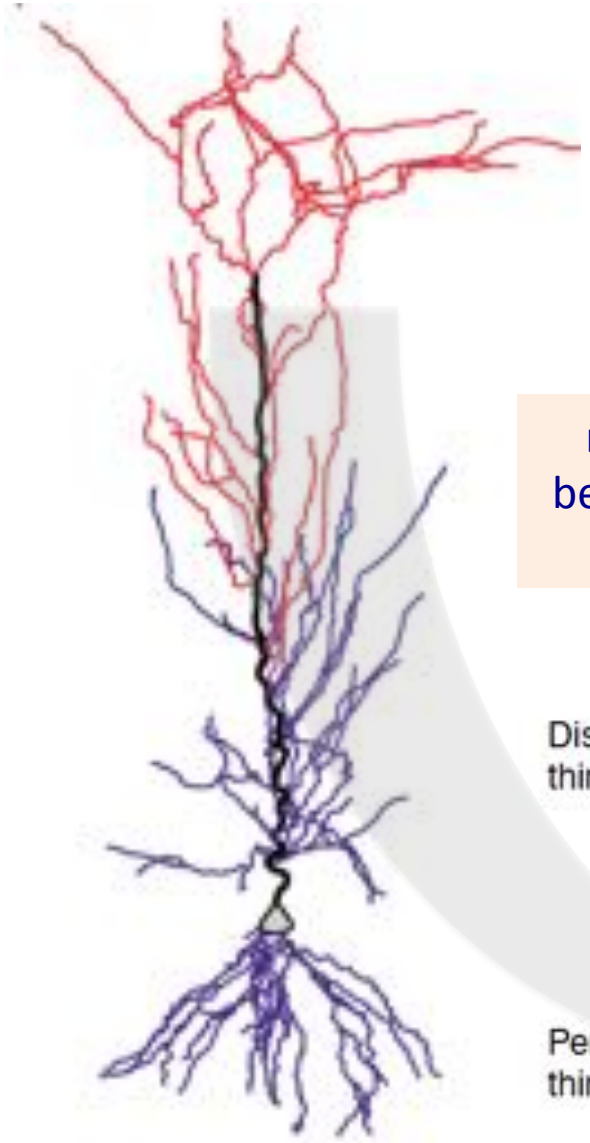
Three spike types in structured neurons

- Na spikes perform network communication
- NMDA plateau potentials create non-linear dendrites
- Ca spikes add coincidence detection between basal and distal inputs



The decade of the dendritic NMDA spike.
Antic SD, Zhou WL, Moore AR, Short SM, Ikonomu KD.
J Neurosci Res. 2010 Nov 1;88(14):2991-3001. doi: 10.1002/jnr.22444. Review., PMID: 20544831

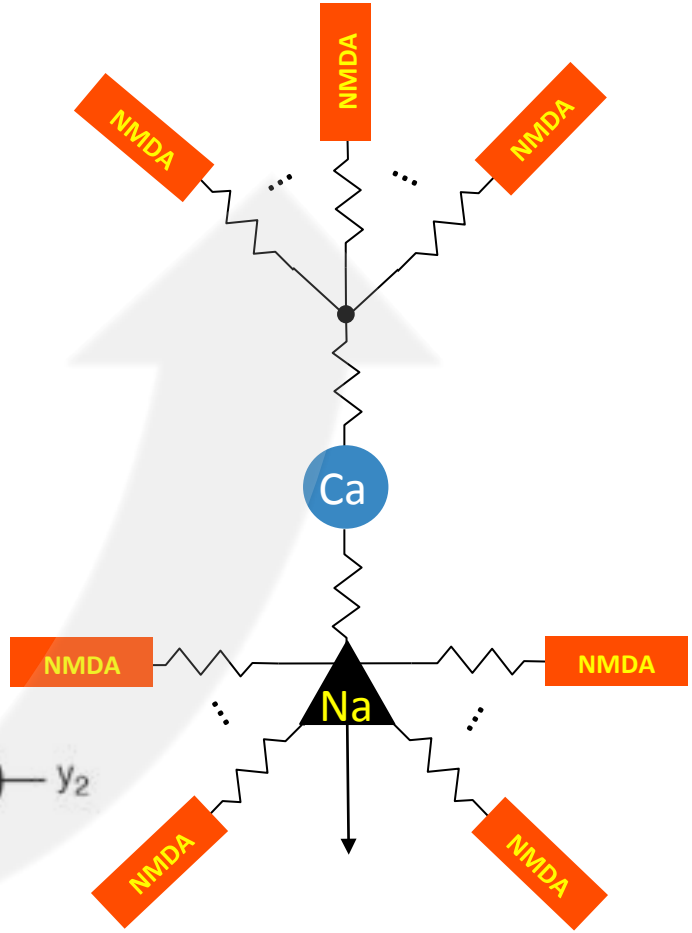
CA1 pyramidal cell



Active dendrites

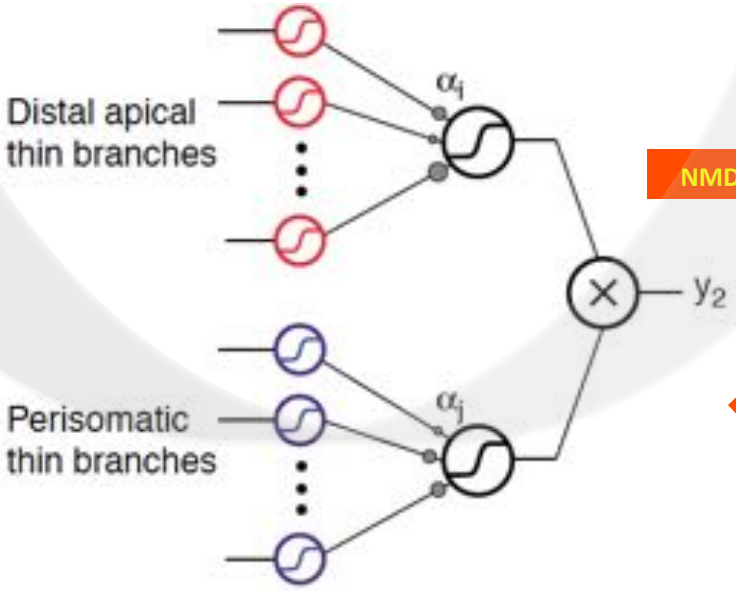
From biology
to abstraktion
to CMOS

Non-linear dendritic NMDA integration
and Ca coincidence
in CMOS hardware



Schemmel et al., arxiv
1703.07286, 2017

multiplicative interaction
between proximal and distal
integration regions



Dendrites: bug or feature?, Michael Häusser and Bartlett Mel, 2013

Final Thoughts

- After 10 years of development available neuromorphic hardware systems have reached a high degree of maturity, in-use for non-expert use cases e.g. in established machine learning tasks

- 2nd generation physical model systems start to offer very advanced accelerated local learning capabilities and exploitation of dendritic computation

Goal : Build a continuously learning cognive machine

neuromorphic.eu

Groundbreaking Ceremony
European Institute for Neuromorphic Computing (EINC)
Heidelberg (Germany), May 5th 2017

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

KTS
KLAUS TSCHIRA STIFTUNG GEMEINNÜTZIGE GMBH

DIETMAR HOPP STIFTUNG • gefördert durch DH • DIETMAR HOPP STIFTUNG

Supported by

The Klaus Tschira Foundation (KTS)
The Dietmar Hopp Foundation
The Wild Foundation