

# Mixed-signal accelerated Systems Progress, Results, Plans

Karlheinz Meier  
Ruprecht-Karls-Universität Heidelberg

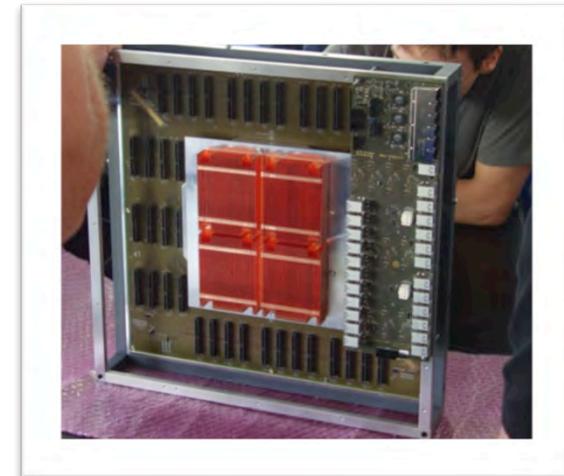
Neuro-Inspired Computational Elements Workshop  
Santa Ana Pueblo, NM, February 24, 2015



**UNIVERSITÄT  
HEIDELBERG**  
ZUKUNFT  
SEIT 1386

# 10 Rationales for the Physical Model System

- **Mixed-Signal** (Local analog computation, binary spike communication)
- Driven by **architecture**, not devices (180nm CMOS)
- High Neuron **Input Count** (>10.000)
- **Configurability** (cell parameters, connections) -> Universality
- **Scalability** : ChipScale ( $10^5$ ) -> WaferScale ( $10^8$ ) -> Systems ( $>10^9$ )
- **Acceleration** x10.000, consistent time constants (1 day compressed to 10 seconds)
- Short-term und long-term **Plasticity**
- **Upgradability** with unchanged system architecture
- **Hybrid Operation**, closed loop experiments
- Non-Expert User **Access**

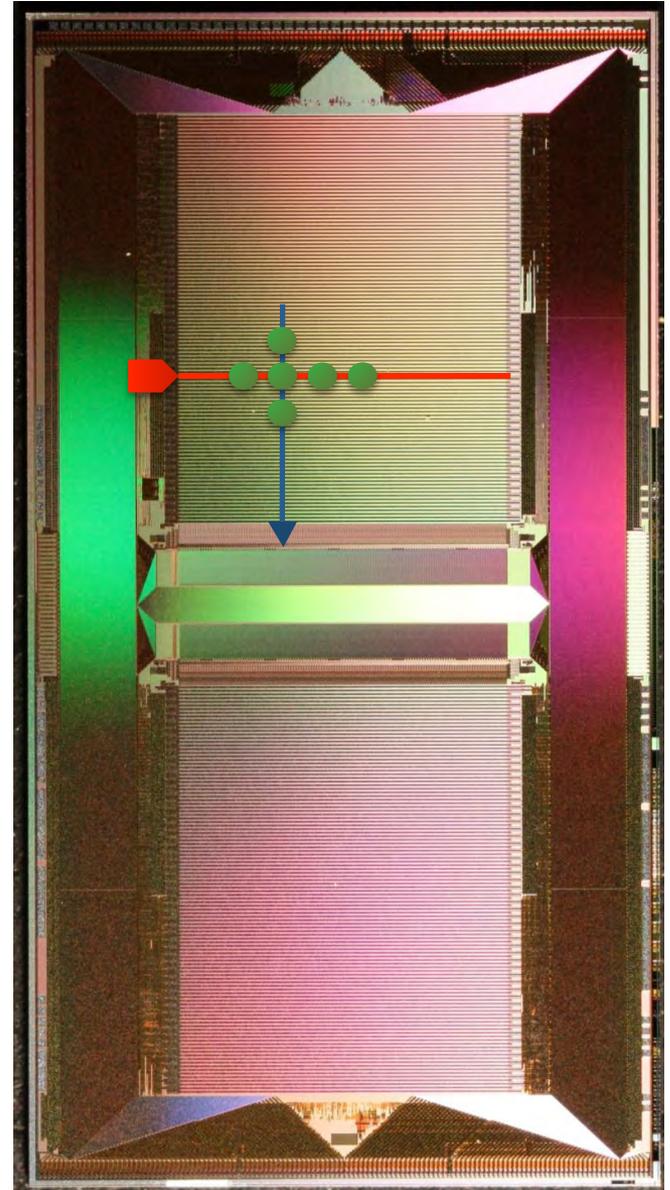


**Objective : Exploit configurability and acceleration**

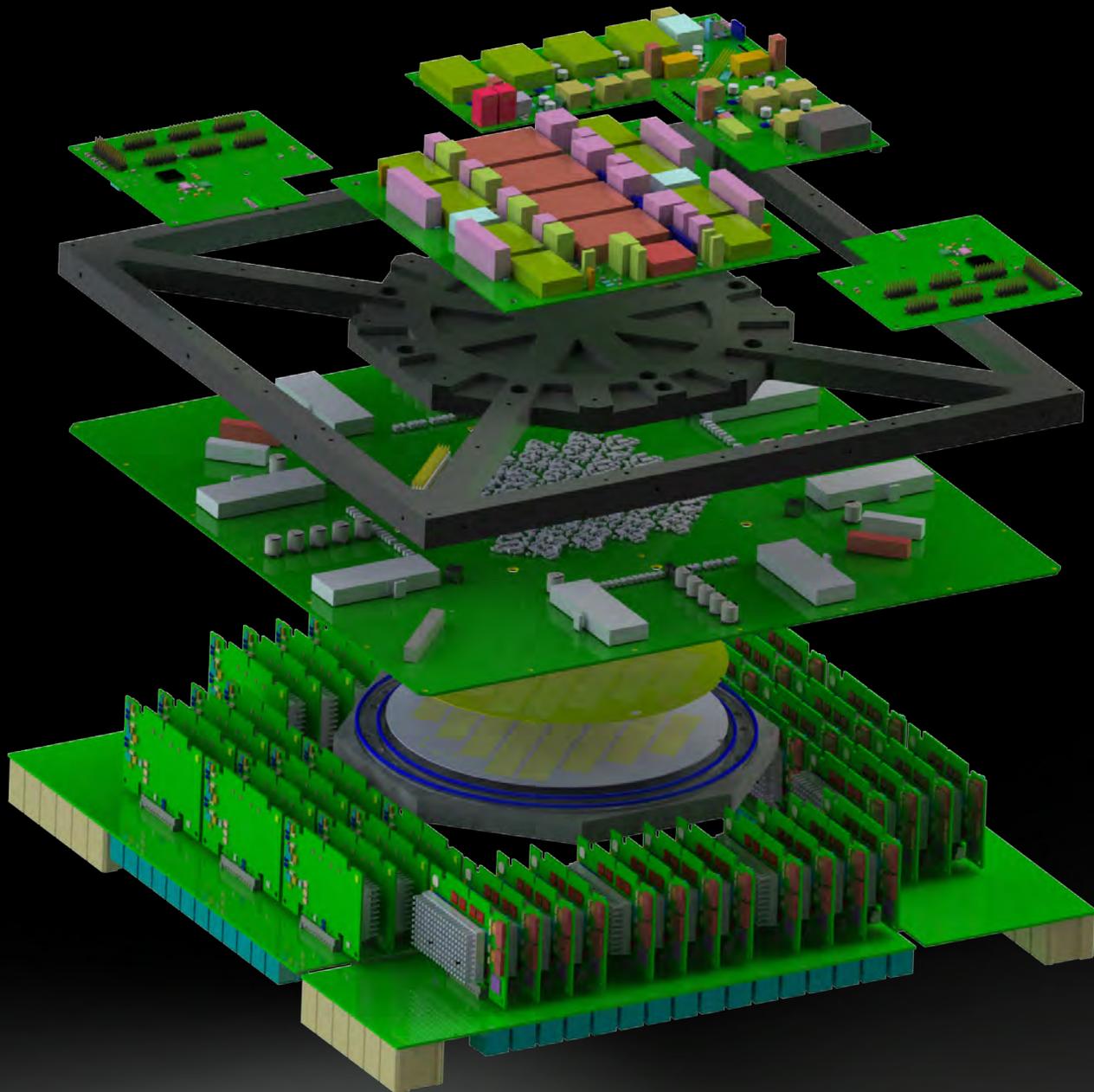
- rapid exploration of large parameter spaces
- cover short and long timescale circuit dynamics
- perform computing in the presence of spatial and temporal noise

# Neuromorphic ASIC HICANN

- continuous time analog neuron circuits based on the adaptive-exponential I&F neuron model (AdEx)
- configurable neuron size :  
up to 14000 pre-synaptic inputs
- accelerated emulation :  
acceleration factor  $10^4$
- 512 neurons and 114k synapses per HICANN ASIC



Photograph of the HICANN ASIC



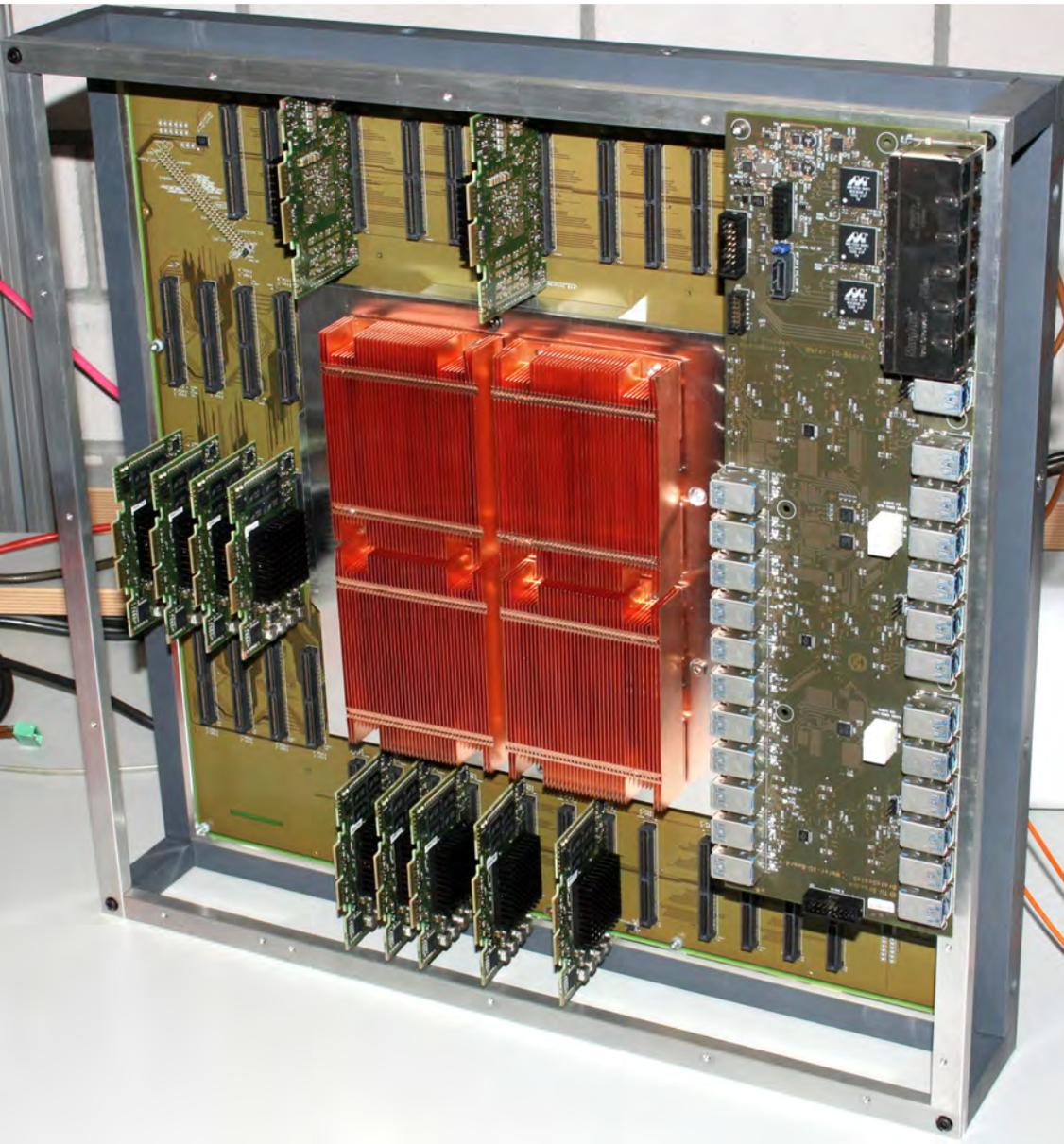
Physical Model, local  
analogue computing,  
binary continuous time  
communication

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



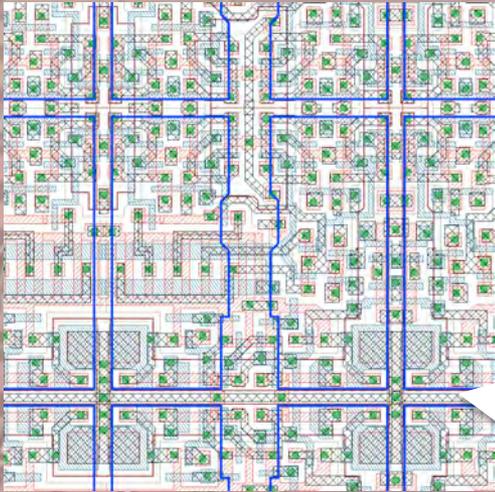
# Status February 2015



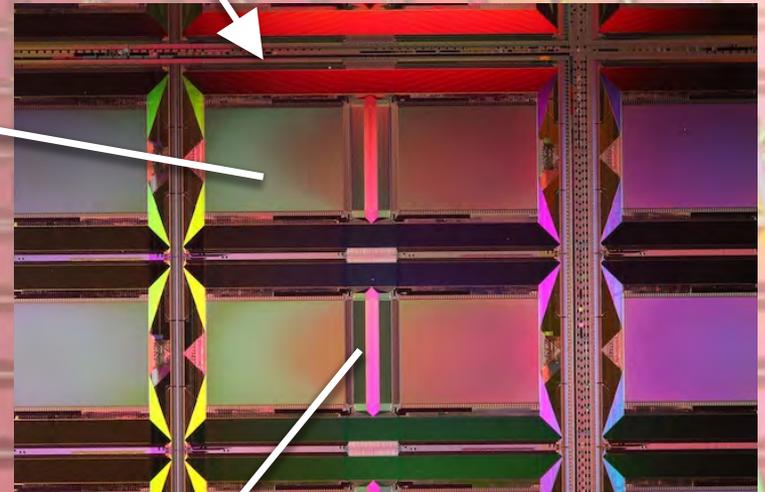
- 2 wafer pre-prototype systems in 24/7 operation
- Routine remote job entry via Slurm
- Final prototype wafer module finished
- 30 wafer modules under construction
- Delivery to project in July

Wafer Scale  
Integration on 8 inch  
CMOS Wafer (180nm)

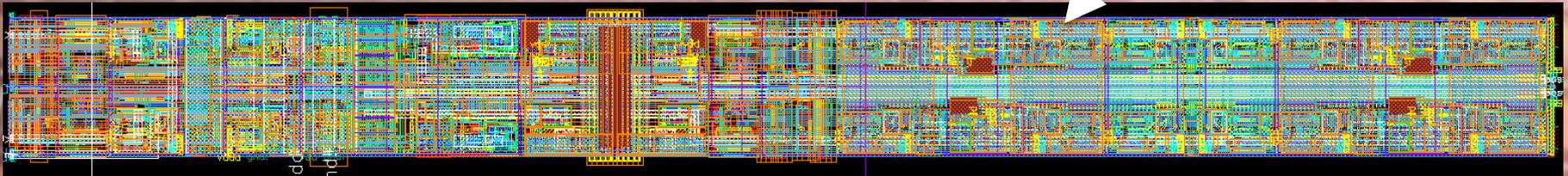
High Input Count  
**Network Chips**, 400  
Instances on Wafer,  
Length Scale 1 cm  
network routing



**Plastic Synapses**,  
50.000.000 Million  
Instances on Wafer,  
Length Scale 10  $\mu\text{m}$   
4-bit SRAM Weights  
STD, STF , STDP



**AdEx** Neurons, 200.000 Instances on Wafer, Length Scale 300  $\mu\text{m}$ ,  
16.000 synaptic Inputs per Neuron, Analog Floating Gate Parameter Storage  
Poisson Noise Generators



# Configuration Space 40 MB for a full Wafer

Scope	Name	Type	Description
Neuron circuits (A)	n/a	i	Two digital configuration bits activating the neuron and readout of its membrane voltage
Synapse line drivers (B)	$t_r$		
Synapses (B)			
STP related (C)	$\tau_{tr}$		
STDP related (D)	n/a	i <sub>l</sub>	Bias current controlling delay for presynaptic correlation pulse (for calibration purposes)
	$A_{+/-}$	s <sub>l</sub>	Two voltages dimensioning charge accumulation per (anti-)causal correlation measurement
	n/a	s <sub>l</sub>	Two threshold voltages for detection of relevant (anti-)causal correlation
	$\tau_{STDP}$	g	Voltage controlling STDP time constants

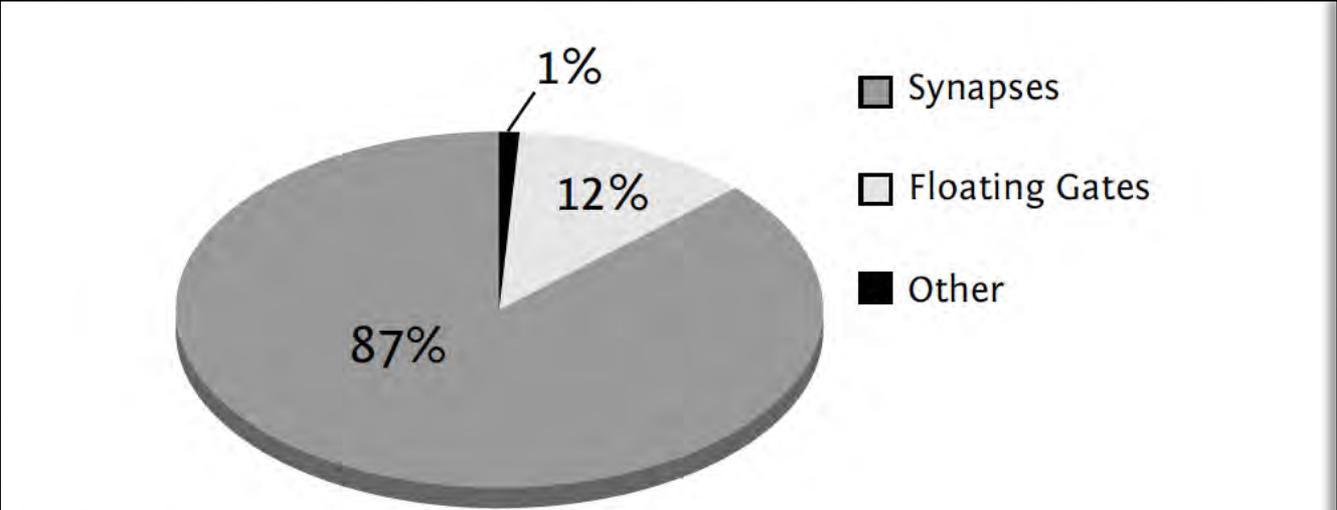
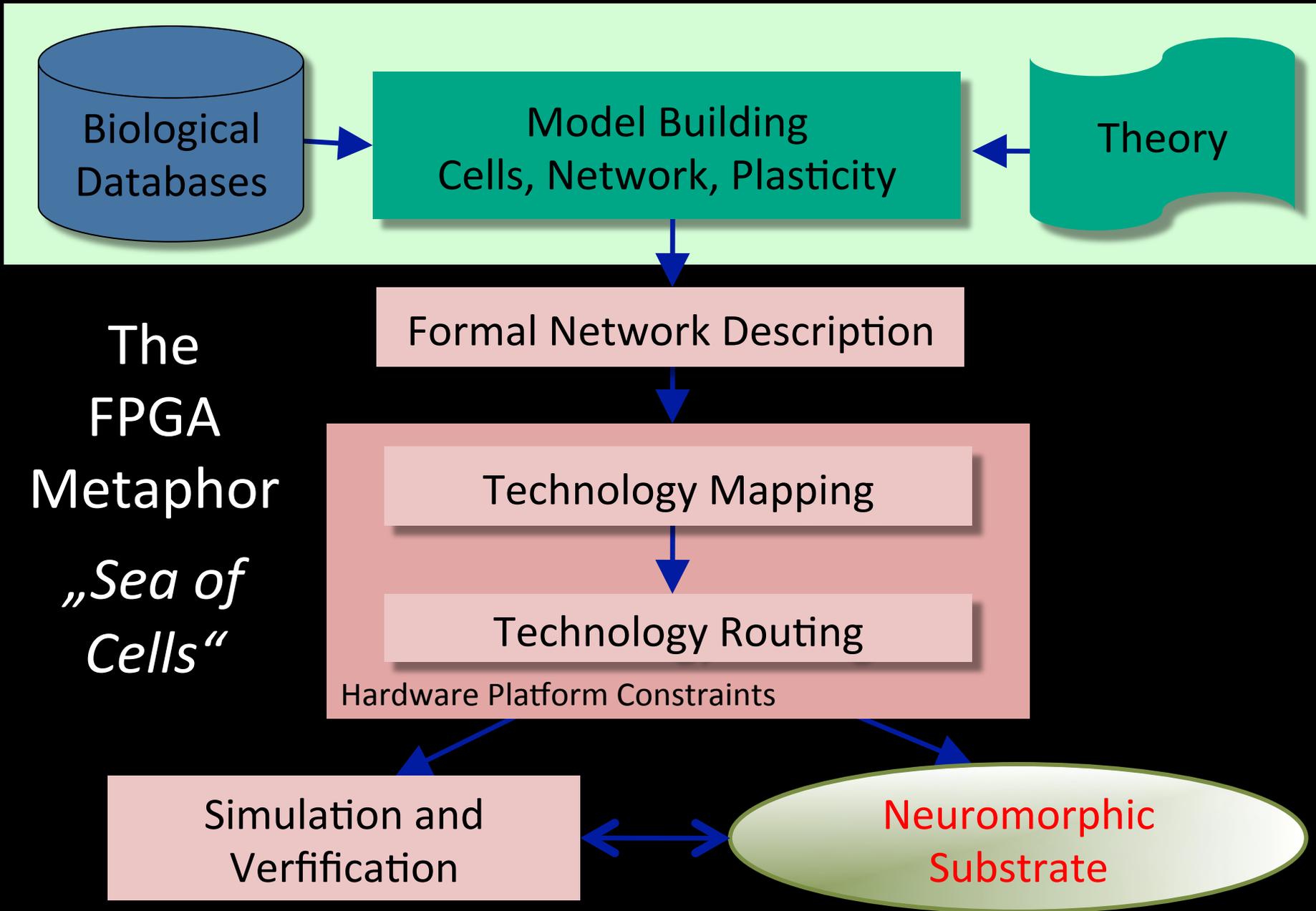


Fig. 4: Sector diagram of the parameter space to configure one HICANN chip. For a full wafer, the configuration data volume is 44 MB large.



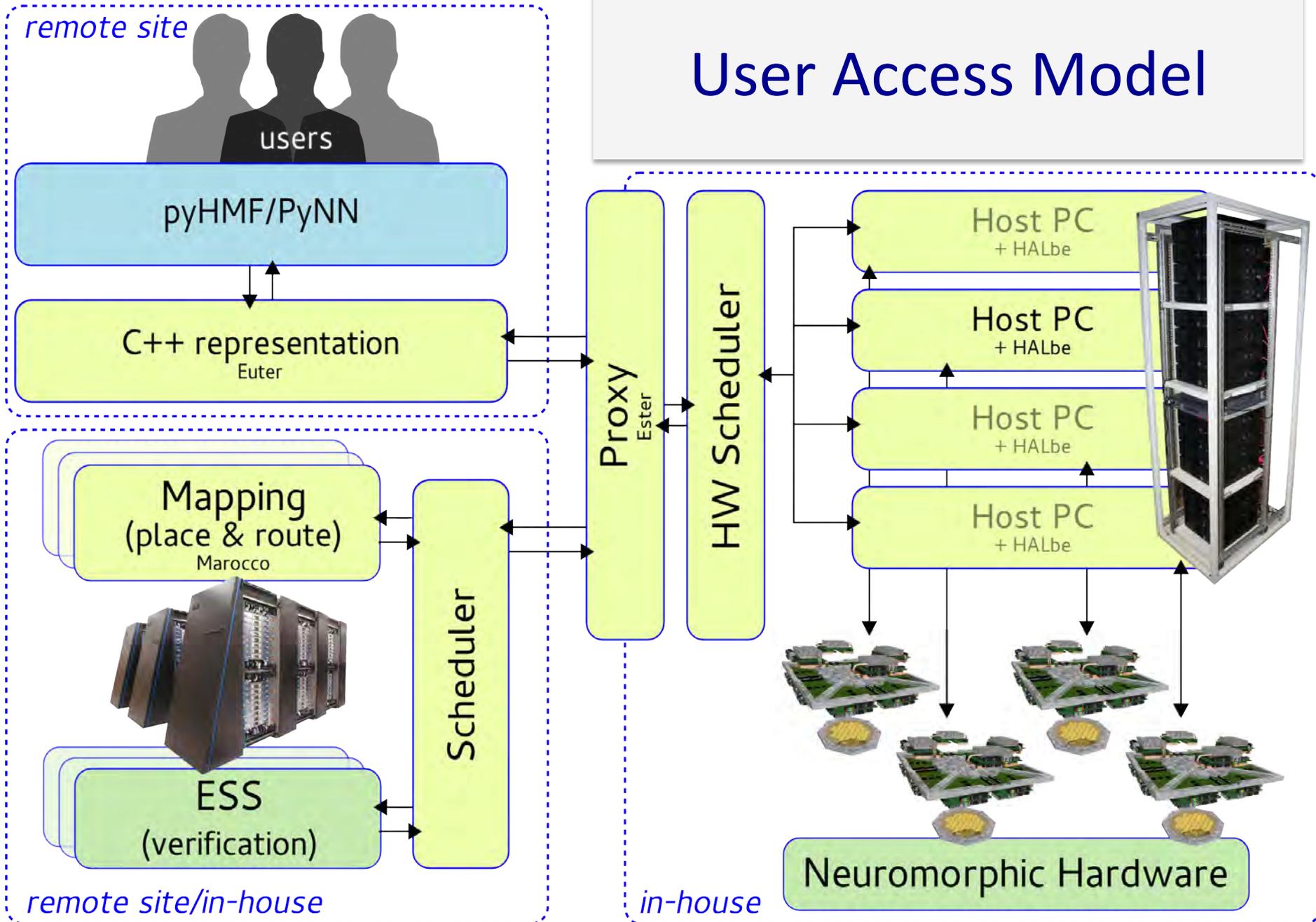
# The HBP Physical Model System

20 Wafer Modules, remote access through HBP portal

Low latency, high bandwidth, local cluster, closed loop hybrid operation



# User Access Model



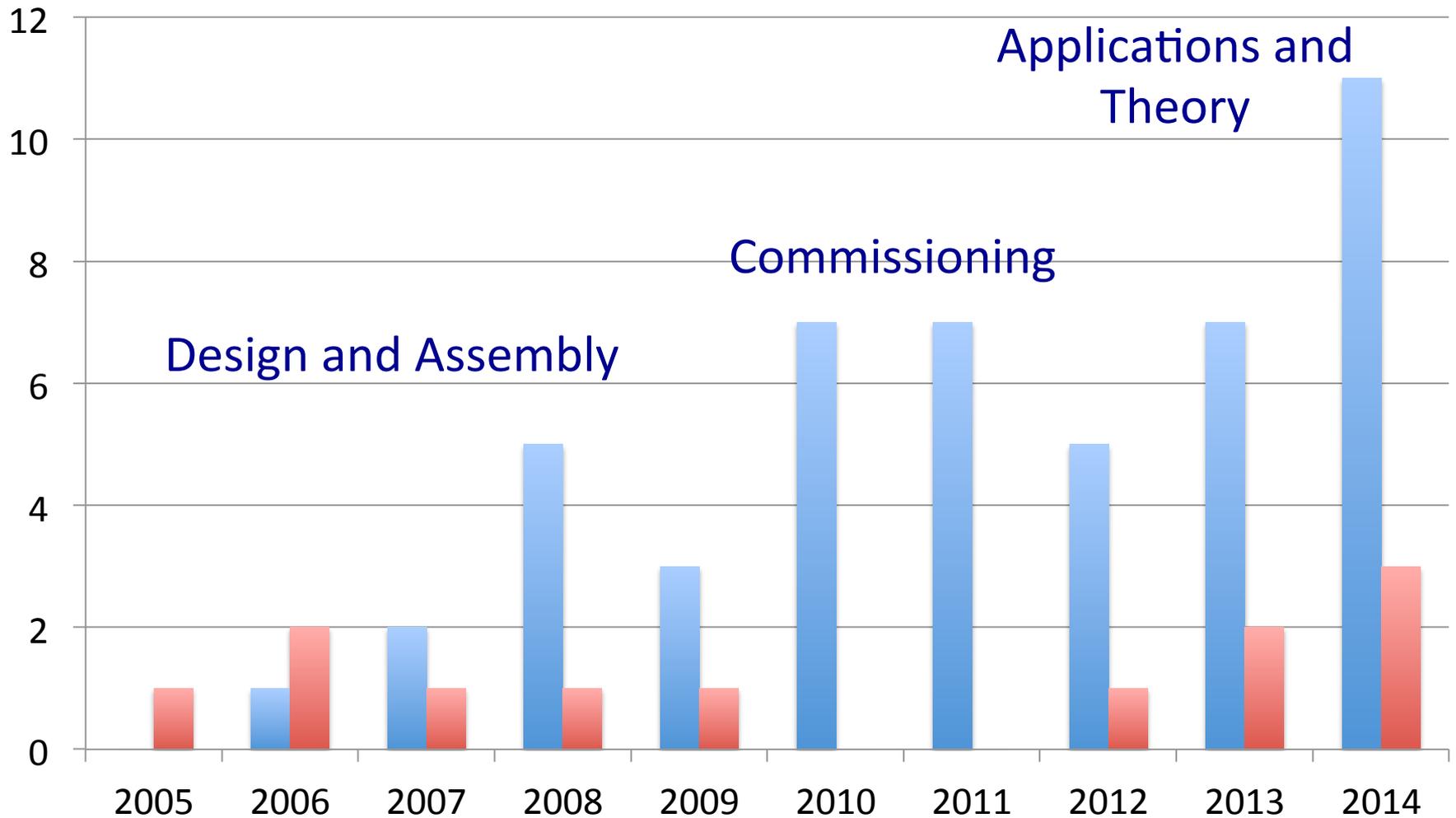
Increasing number of use cases and applications covering a wide spectrum of network types

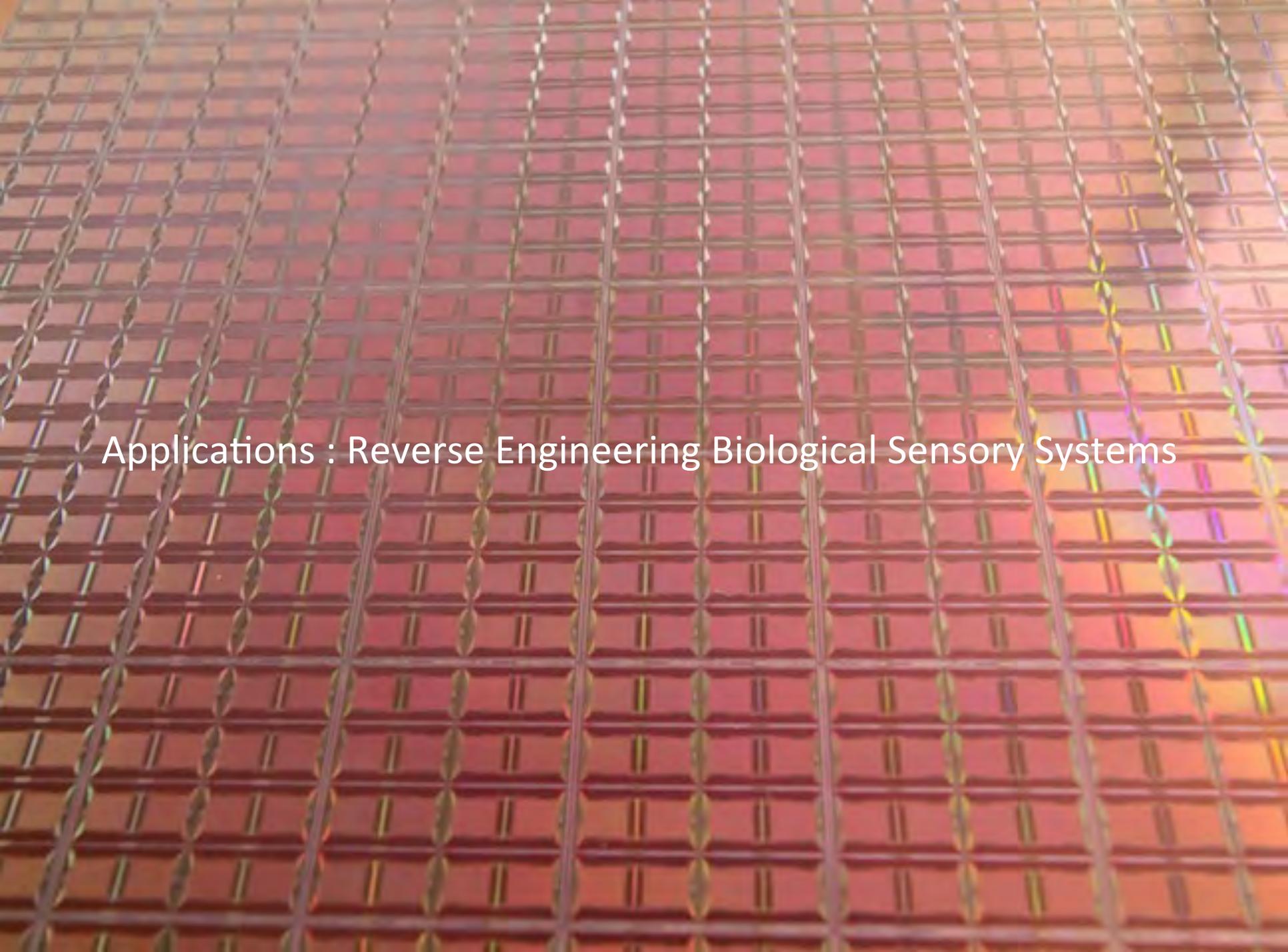
## Exploiting Substrate UNIVERSALITY

- *Canonical circuits (synfire chains, WTA, attractor circuits)*
- *Balanced random networks*
- *Liquid computing, temporal pattern identification*
- *Minicolumn Layer 2/3 circuits*
- *Closed-loop hybrid control systems \**
- *Multivariate classification*
- *Echolocation, applying STDP*
- *Decorrelation through inhibitory feedback \**
- *Stochastic inference through neural sampling \**
- *Bayesian networks as Boltzmann machines of LIF neurons \**

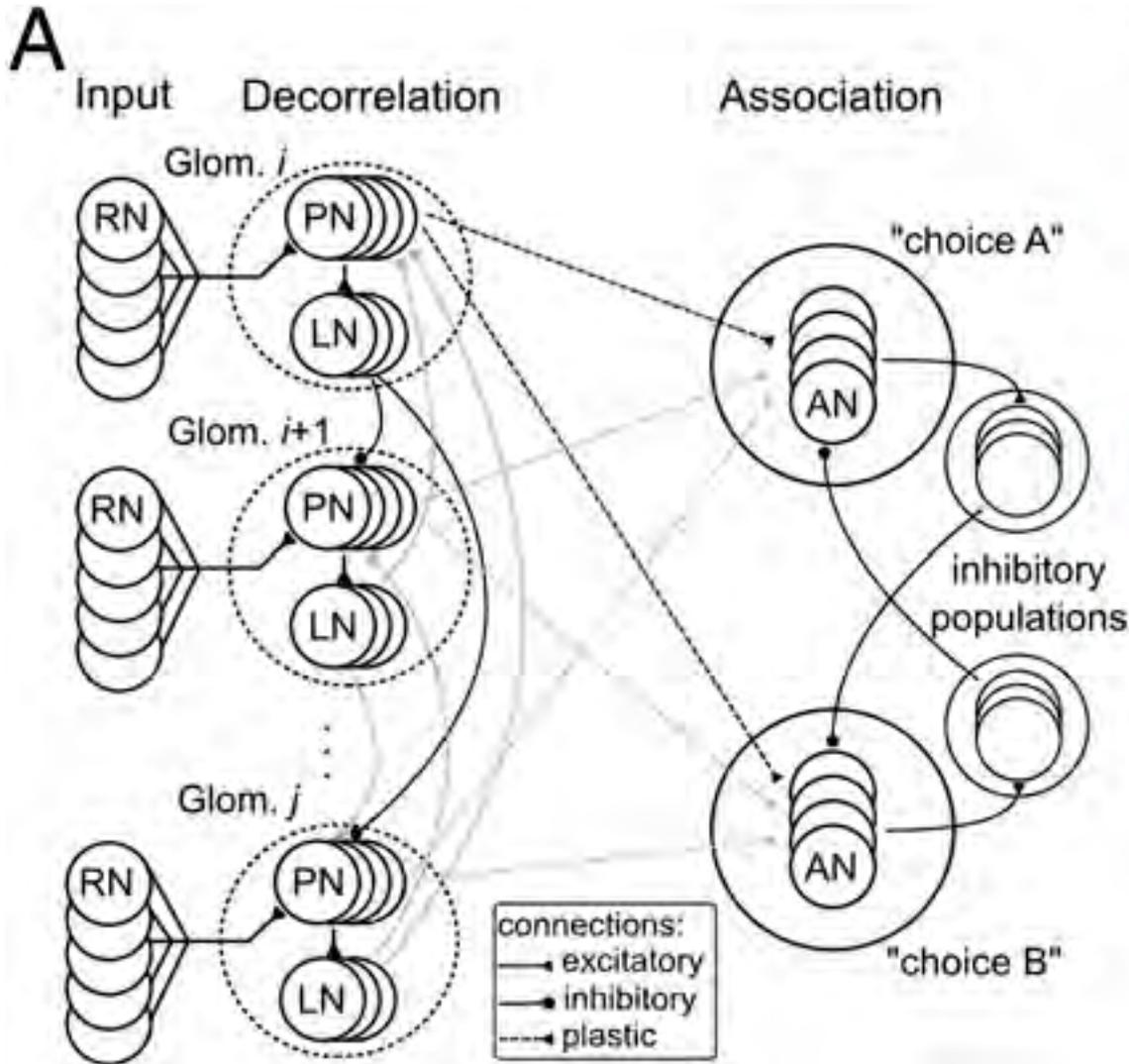
*\* New since NICE II (2014)*

# Ba/Ma and PhD Theses submitted





Applications : Reverse Engineering Biological Sensory Systems



## 3 Layer Spiking Neuron Network derived from Insect Olfactory System

**L I** : Receptor Neurons

**L II** : Decorrelation through lateral inhibition (Glomeruli)

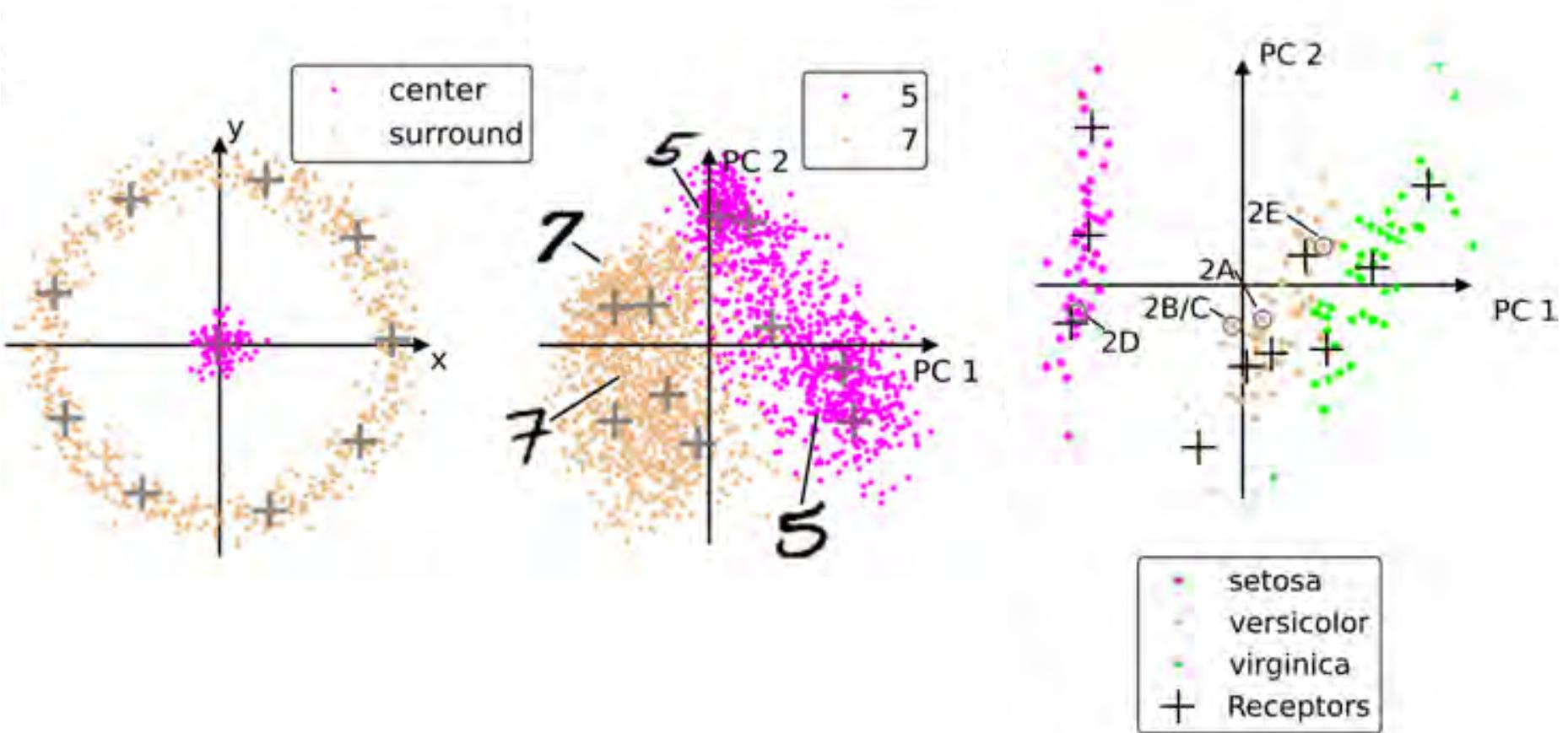
**L III** : Association (Soft WTA through strong inhibitory populations)

## Supervised Learning

Synaptic Projections from Layer 2 to Layer 3

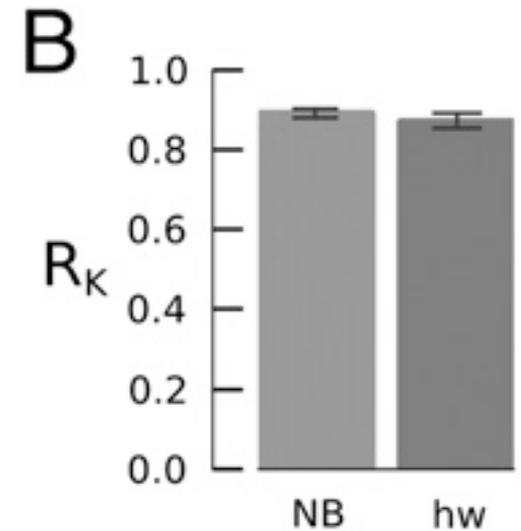
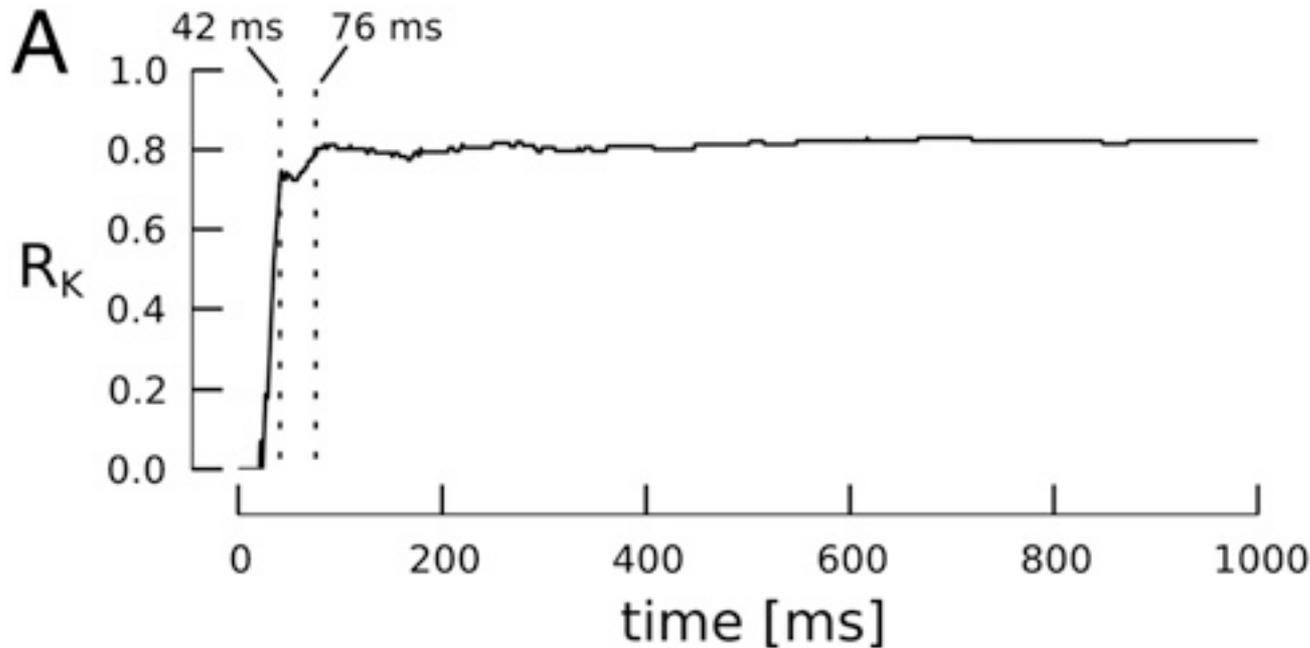
Schmuker, Michael, Thomas Pfeil, and Martin Paul Nawrot. "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

# Typical Datasets



Schmuker, Michael, Thomas Pfeil, and Martin Paul Nawrot. "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

# Classification Performance compared to Software Bayesian Classifier with 5-fold cross-validation



Performance equivalent to classical neural networks, but :

- Energy efficiency x 1.000.000 (0.1 nJ vs. 0.1 mJ per syn. transm.)
- Response time  $\div$  10.000 (10  $\mu$ s vs. 100 ms)

# “Noise, Anomalies and Faults”

- **Temporal** noise (at frequencies small compared to relevant system time constants)  
(Largely) **incoherent**
  - Not user controllable
    - thermal** noise (membrane fluctuations)
    - shot noise** (discrete charge carrier statistics) (membrane fluctuations)
  - User controllable
    - on-purpose **integrated pseudo or quantum** noise sources (spikes)
    - Irregular network **network activity** („spikes, sea of noise“)

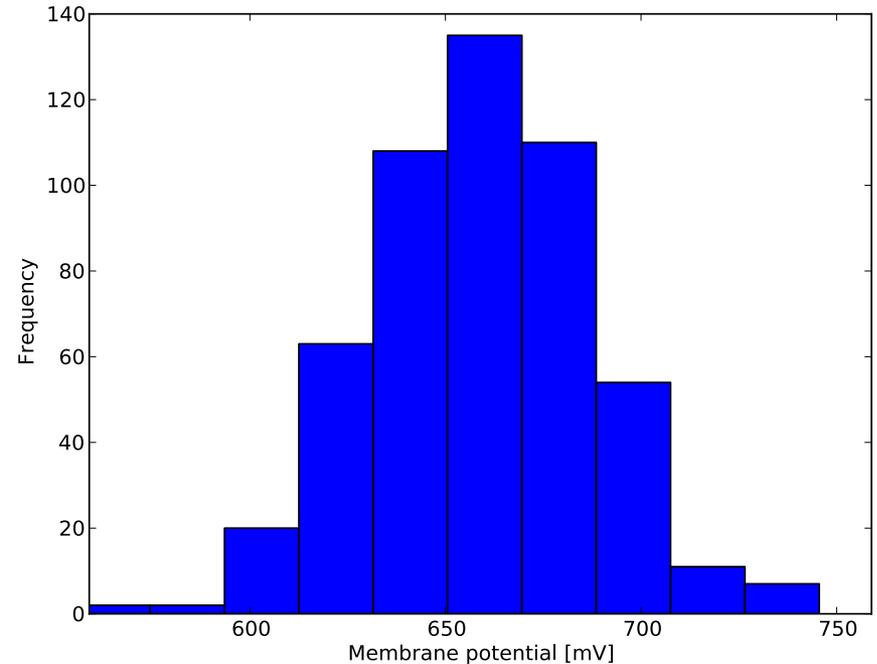
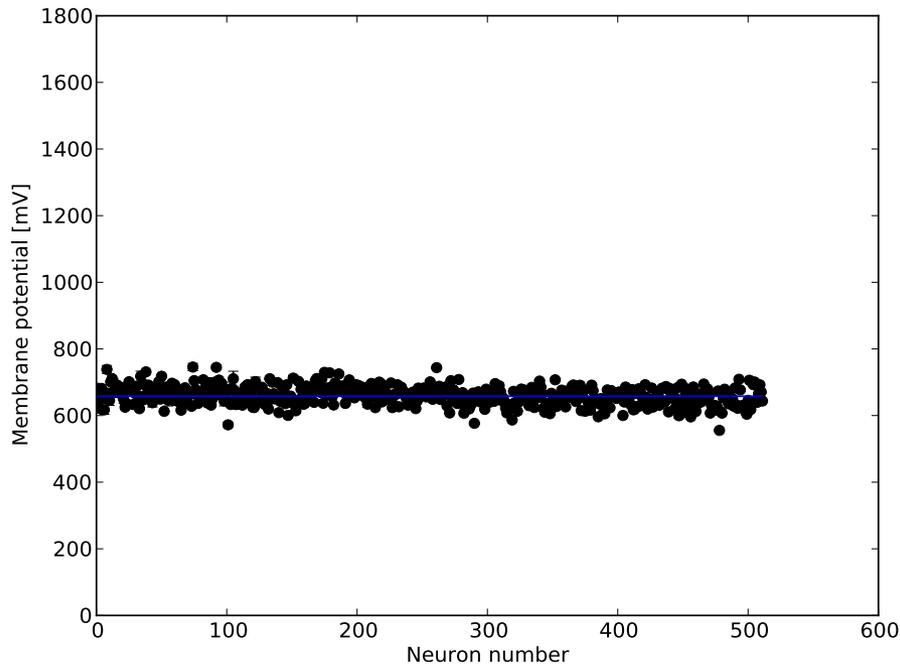
(Largely) **coherent**

  - Somewhat user controllable
    - Cross-talk** from digital control signals to analogue membranes
- **Spatial** (fixed-pattern) noise
  - Static** device mismatch caused by the (small feature size) VLSI production process
  - To some extend user controllable (hard)
    - Neuron parameters** (voltages and timing), “Calibration”
    - Synapse parameters** (voltages and timing), “Calibration”
  - Not practically user controllable
    - Trial-to-trial** variations from analogue parameter setting
    - Dead or faulty** components

# Full chip $E_L$ calibration

BrainScaleS Project  
M.O. Schwarz PhD Thesis

Target :  $E_L = 655$  mV

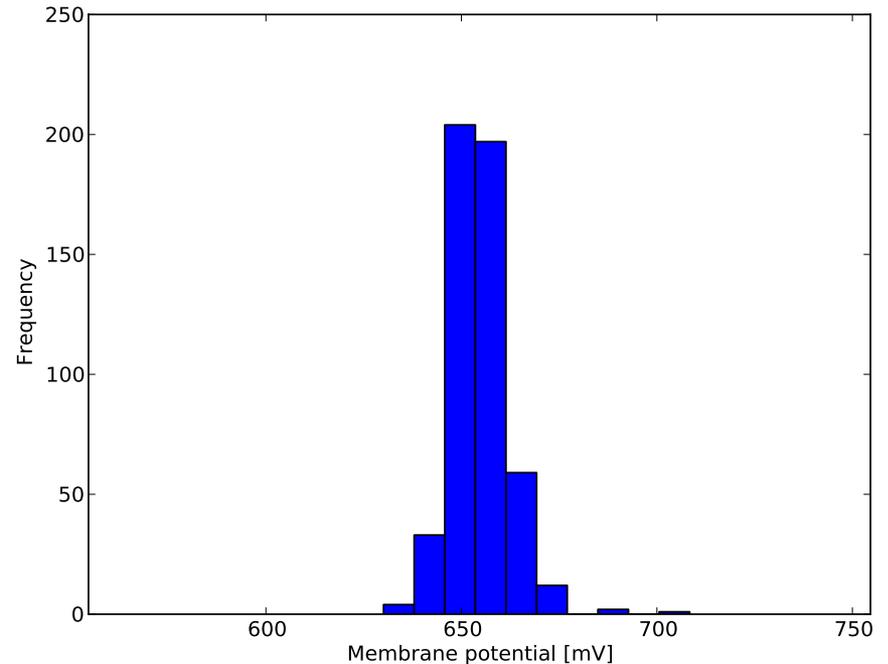
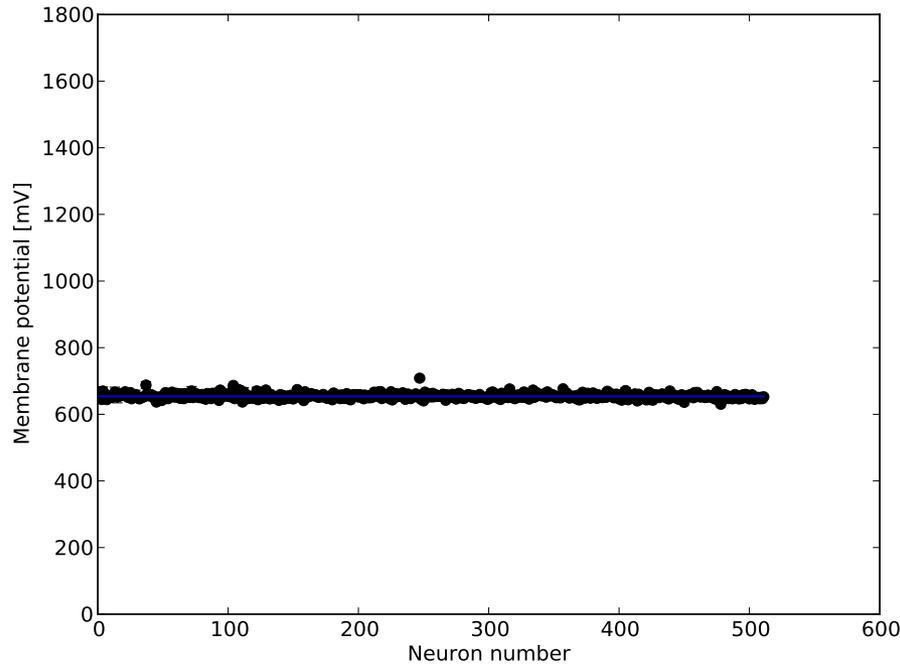


Mean = 658.9 mV | Std = 28.5 mV

# Full chip $E_L$ calibration

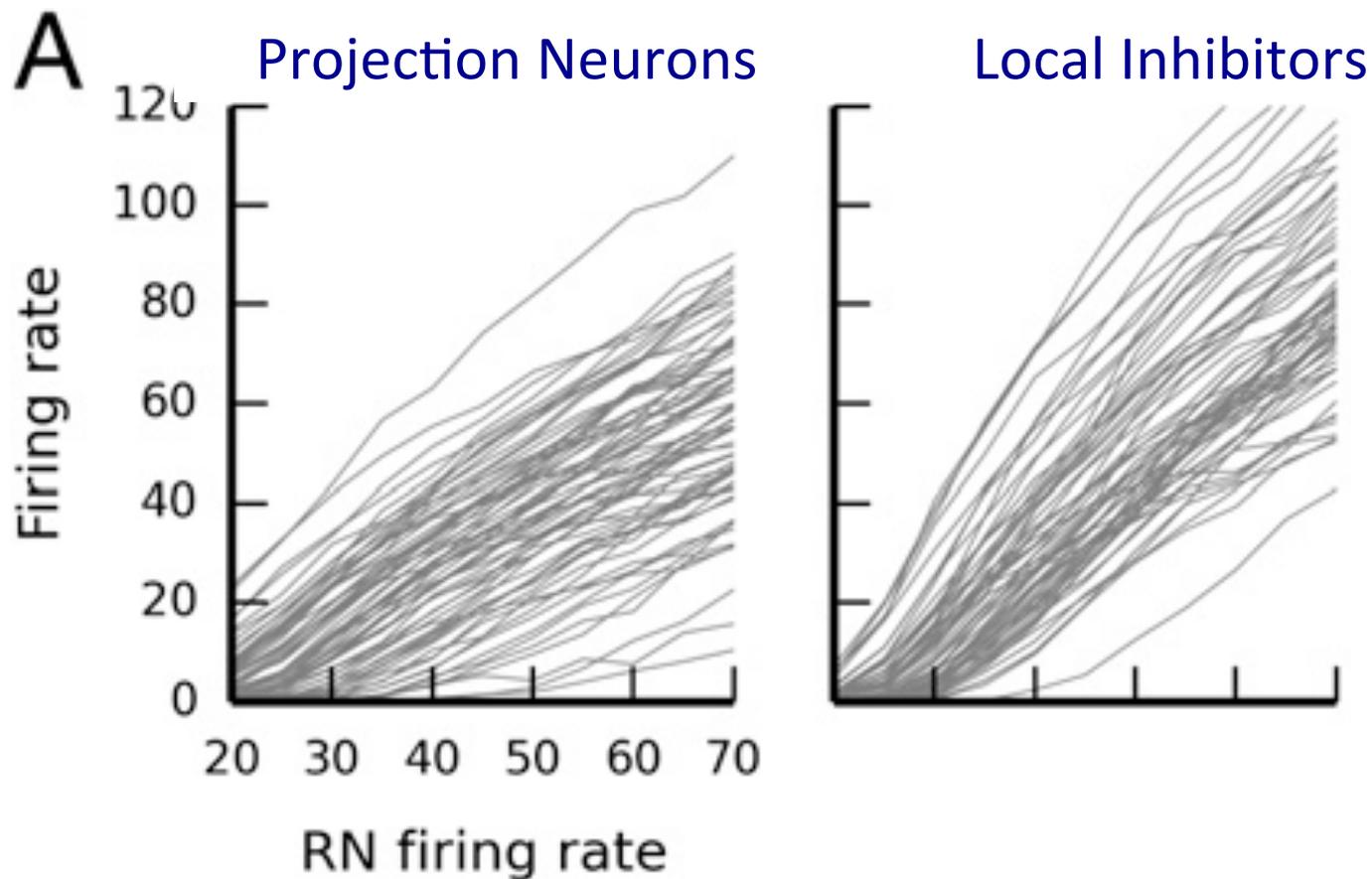
BrainScaleS Project  
M.O. Schwarz PhD Thesis

Target :  $E_L = 655$  mV

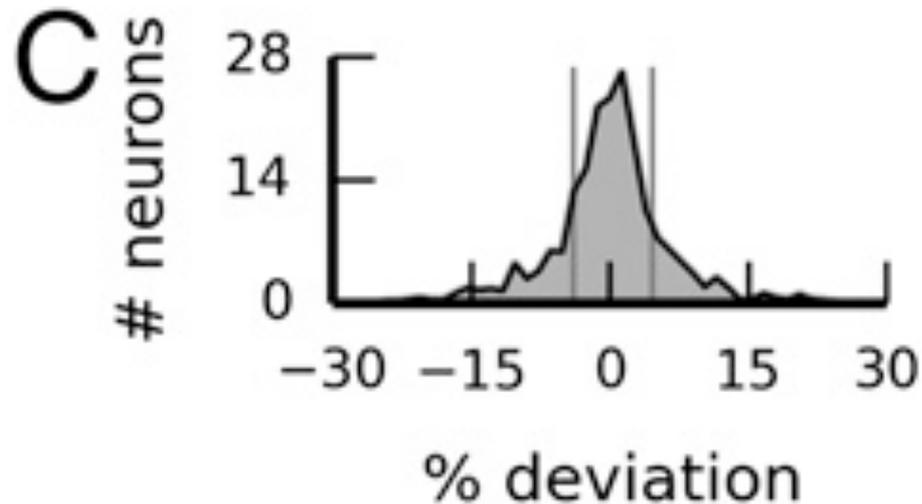
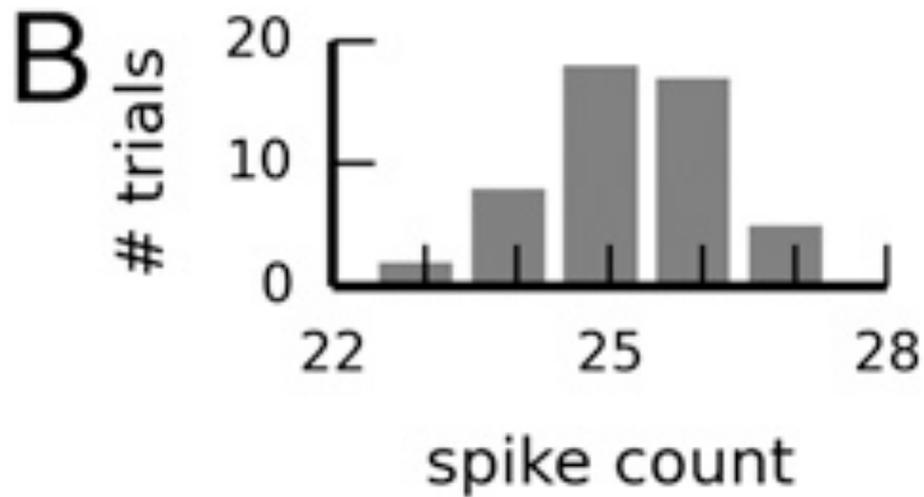


Mean = 654.3 mV | Std = 5.1 mV

# Static Electronic Device Variations impact neuron response



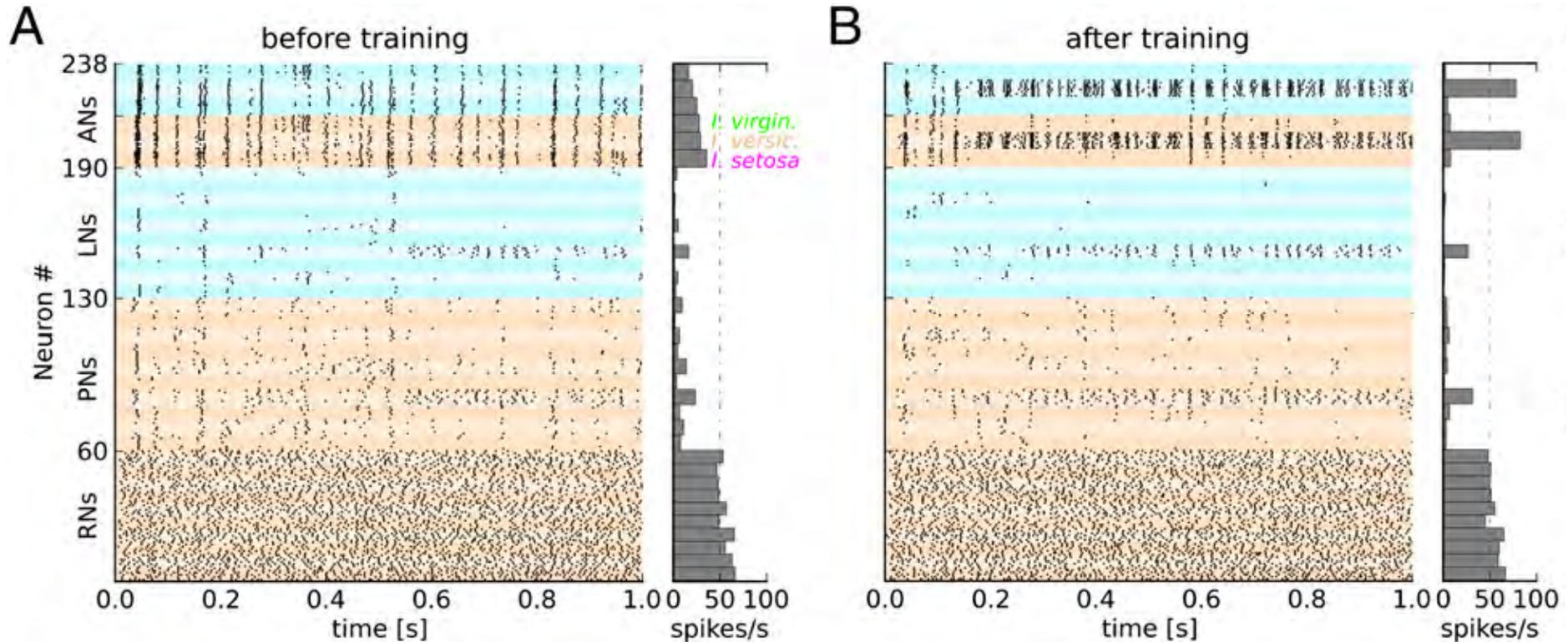
Schmuker, M. et al., "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.



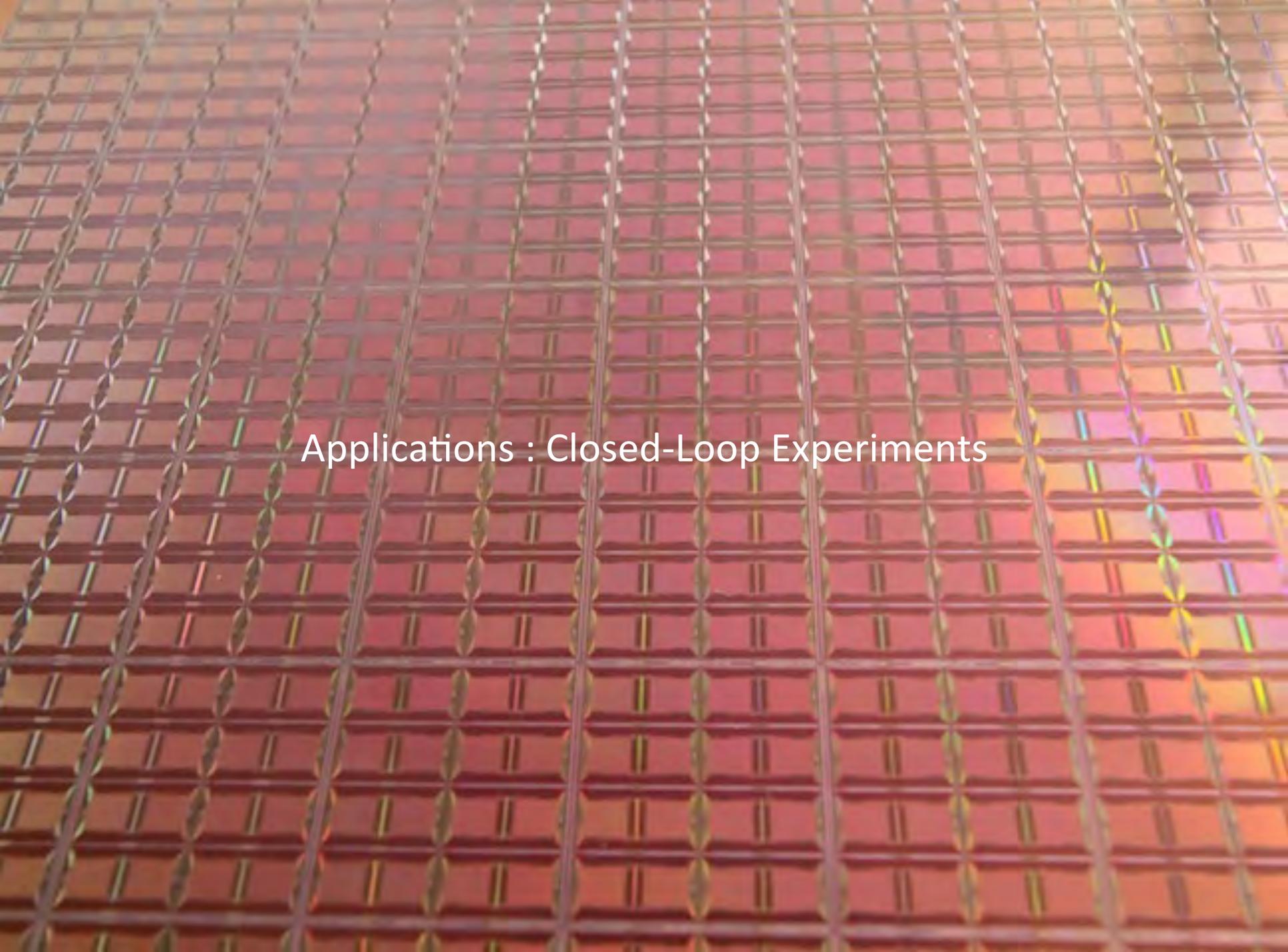
Temporal  
Electronic Device  
Variations  
impact neuron  
response

Schmuker, Michael, Thomas Pfeil, and Martin Paul Nawrot. "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.

# Neuromorphic Network Activity before and after supervised Learning

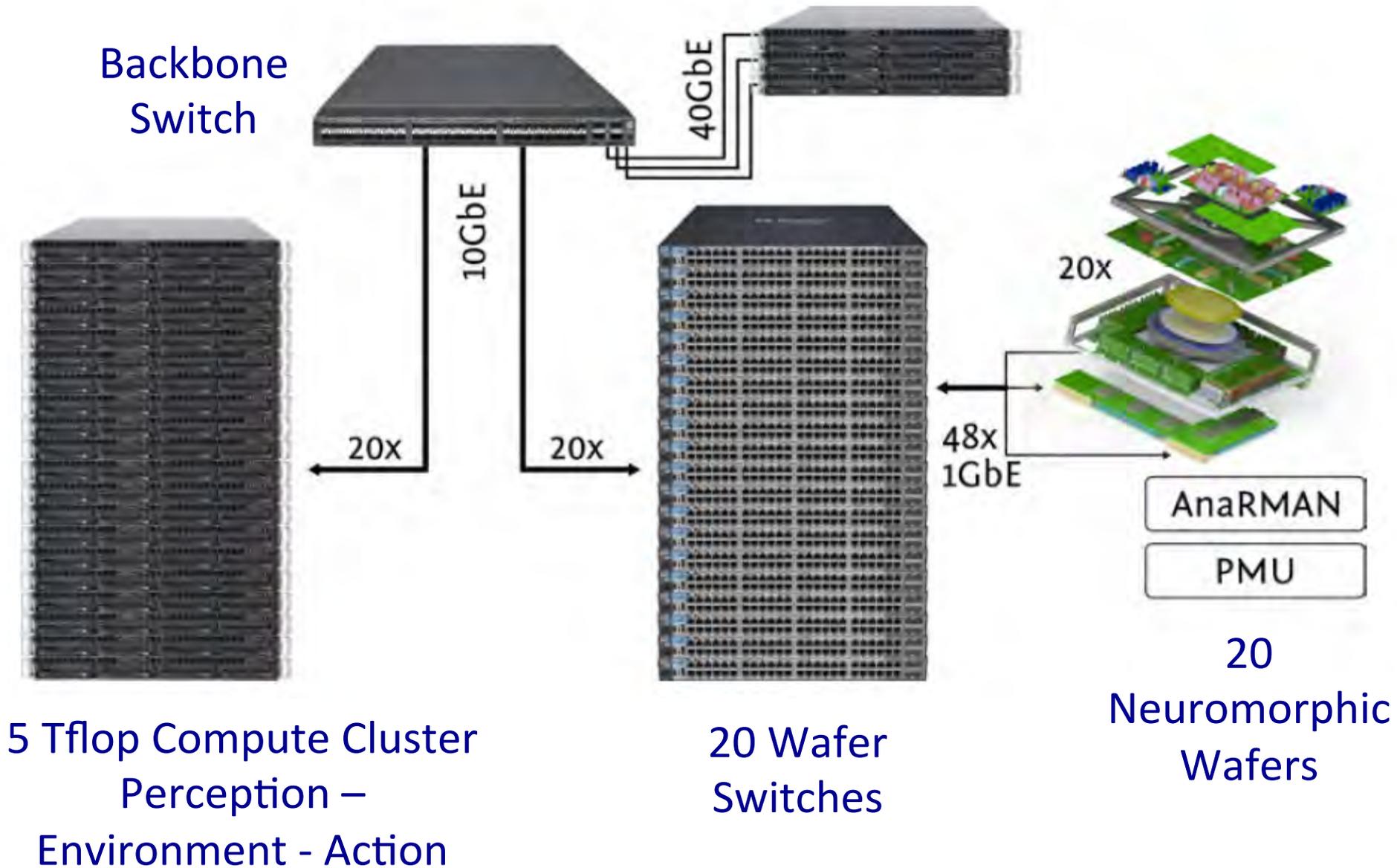


Schmuker, Michael, Thomas Pfeil, and Martin Paul Nawrot. "A neuromorphic network for generic multivariate data classification." *Proceedings of the National Academy of Sciences* (2014): 201303053.



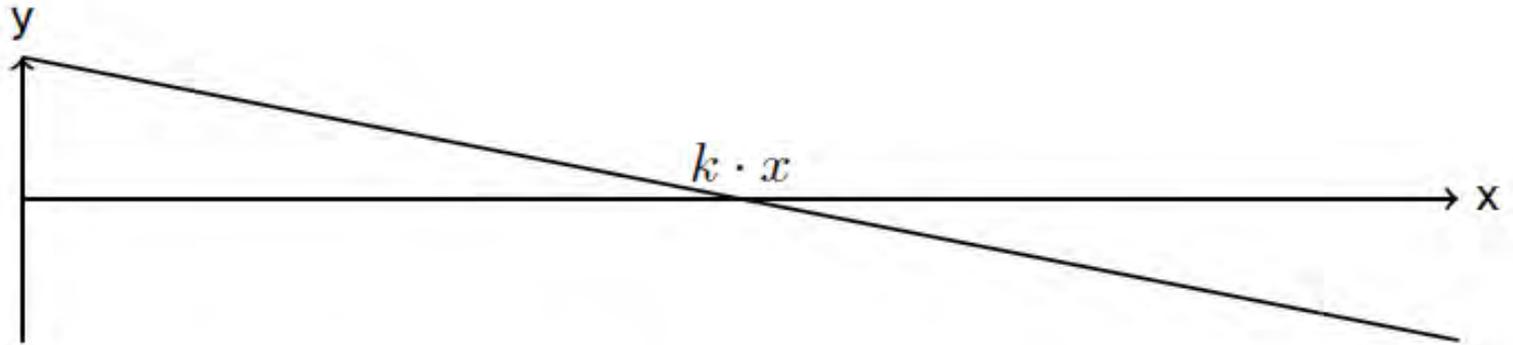
Applications : Closed-Loop Experiments

# Closed Loop Infrastructure



## Experiment Setup

- ▶ 1-dimensional space containing a movable object
- ▶ exhibits “force”  $k$  pulling towards  $x_{\text{center}} = 0.5$



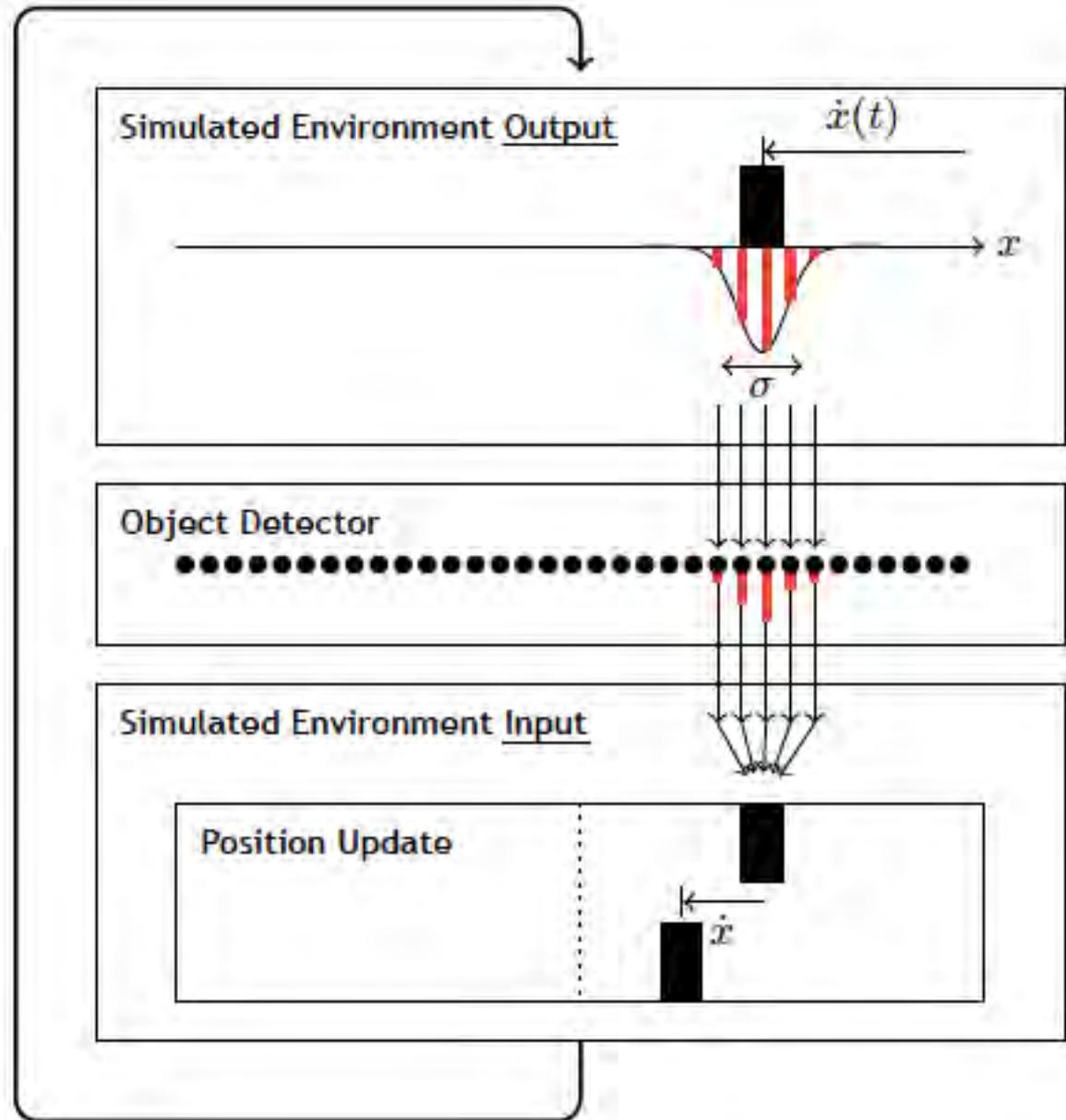
- ▶ object's position is updated after time delay  $\tau$

$$\rightarrow \dot{x}(t) = -k \cdot x(t - \tau)$$

- ▶  $x(t) = e^{W(-k \cdot \tau) / \tau \cdot t}$

Closed Loop  
Sensory  
Motor  
Control

First  
Experiments



# Tracking an Object in a simulated Force Field $k$

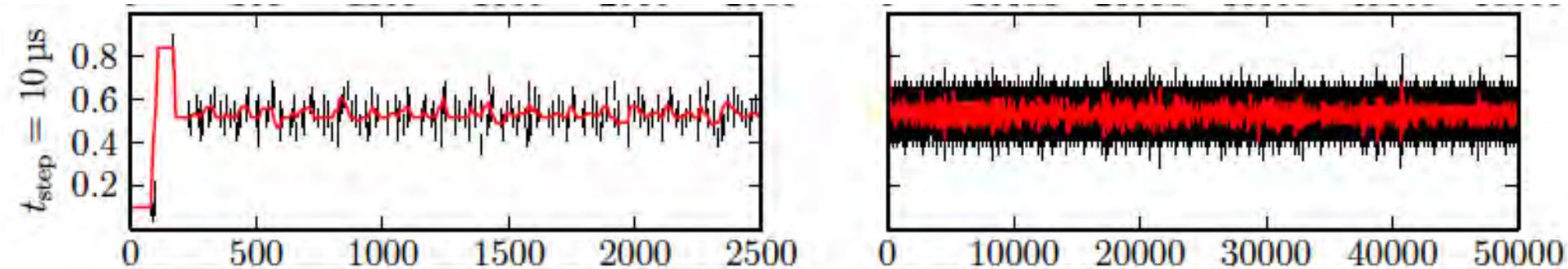
$$\dot{x}(t) = -k \cdot x(t - \tau)$$

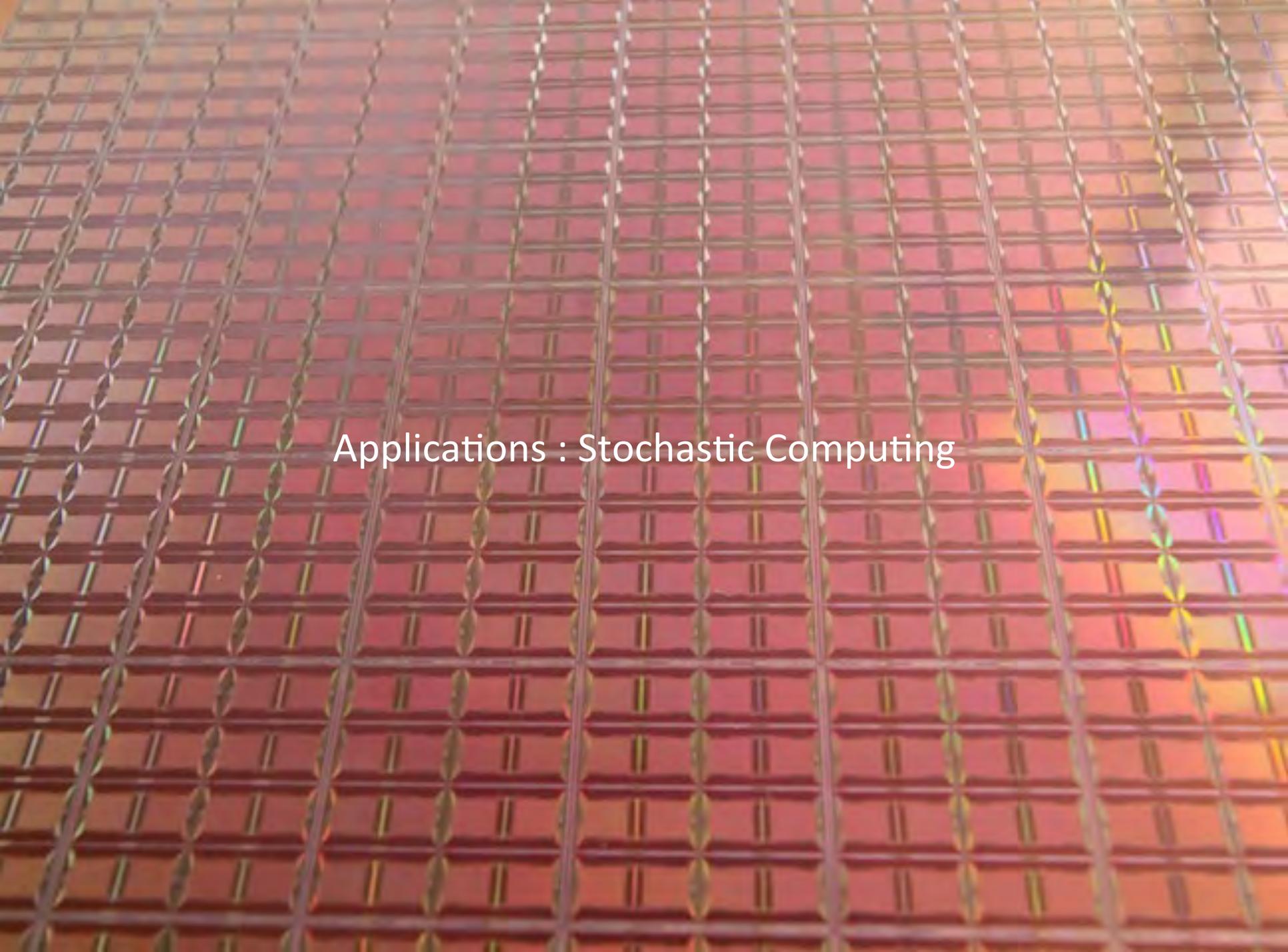
Low Latency, High Bandwidth, Closed Loop, Hybrid Operation of a single Neuromorphic Wafer with a Compute Cluster

$(8.5 \pm 0.4) \mu\text{s}$  one-way latency

Wire-speed performance between host and 8 wafer FPGAs  $((846.7 \pm 1.2) \text{ MB/s})$

Neuromorphic Detector vs. actual **simulated Position**





Applications : Stochastic Computing

# Work based on 4 prerequisites

Abstract model neurons can be understood as **Markov-Chain MC sampling from target distributions** (*Büsing et al. 2011*)

Abstract model neurons can be replaced by **stochastic LIF neurons** on neuromorphic hardware platforms (*Petrovici et al. 2013*)

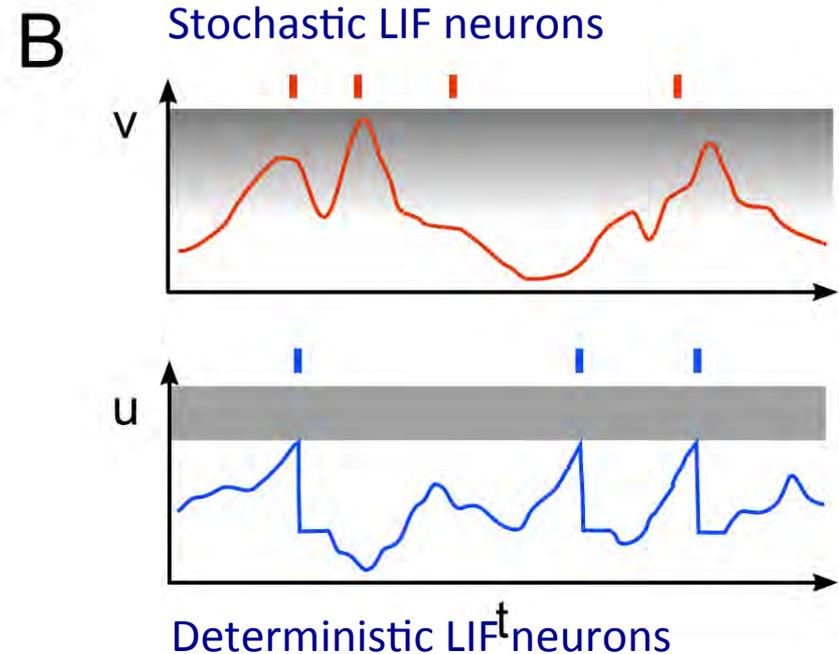
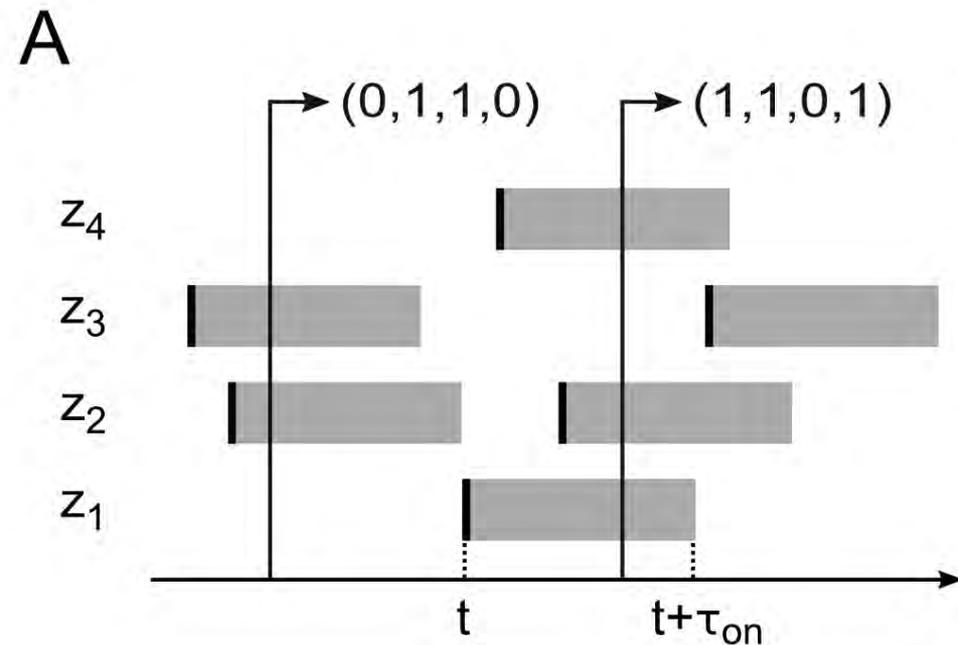
Networks of LIF neurons can represent probability distributions in **any space of binary random variables** and perform stochastic inference in this space (*Probst et al. 2015*)

Underlying graphical models (e.g. Bayesian) can be **transferred to Boltzmann machines** (*Probst et al. 2015*)

*Ideally suited for energy/time efficient and resilient implementation on neuromorphic systems*

# Stochastic Inference with deterministic spiking neurons

## State of binary random variables defined through refractory period



Interpretation of spike patterns as samples of a binary random vector  $z$   
Variable  $z_k$  is active for duration  $\tau_{on}$  after a spike of neuron  $k$ .

Stochastic Inference with deterministic spiking neurons

Petrovici et al., arXiv:1311.3211 [q-bio.NC], submitted to Phys. Rev.

# Formulation of an inference problem as a Bayesian network and translation to a Boltzmann machine with spiking LIF neurons

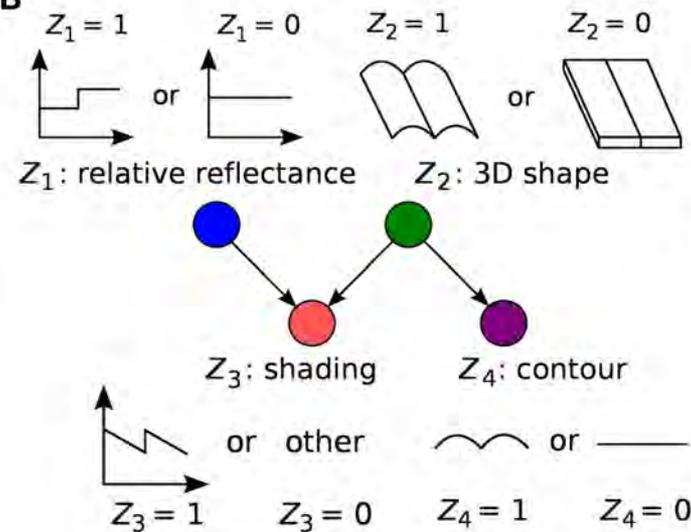
## Concrete Example : Knill-Kirsten Illusion

A



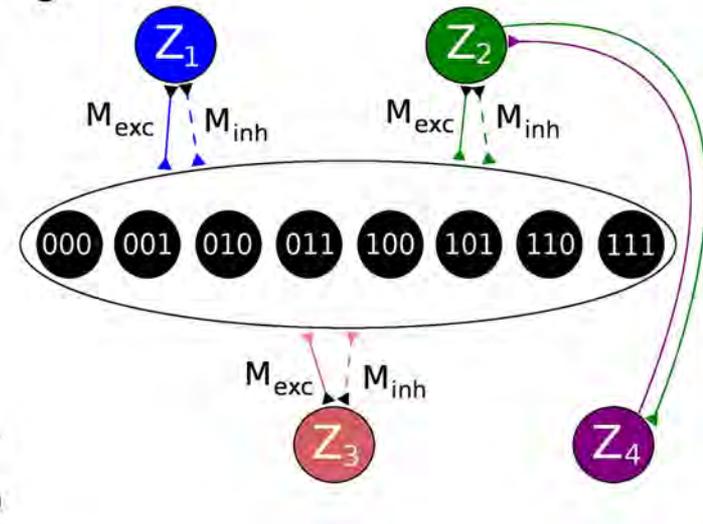
Knill-Kersten Illusion  
(1991)

B



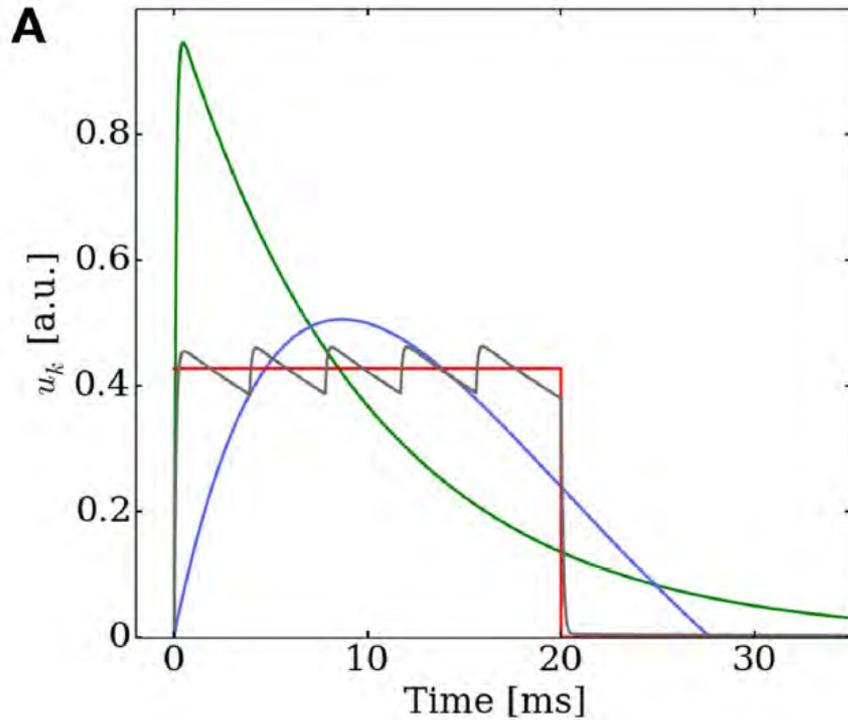
Translation to a Bayesian network with four binary random variables

C

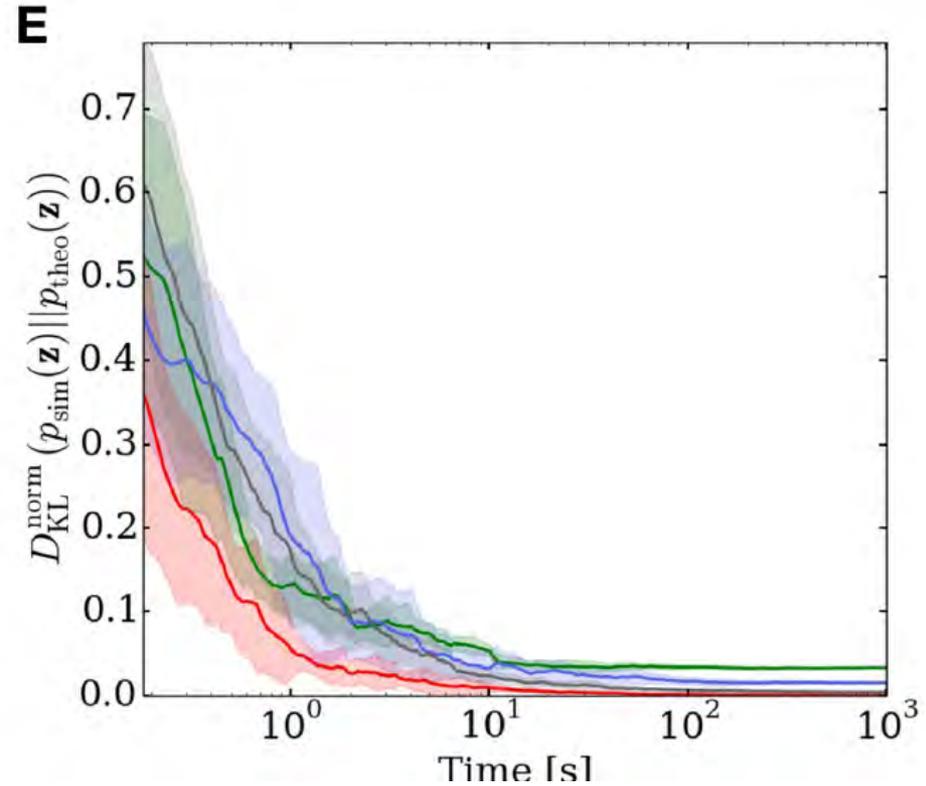


Representation of the Bayesian network as a Boltzmann machine

# Convergence to target depends on effective inter-neuron coupling



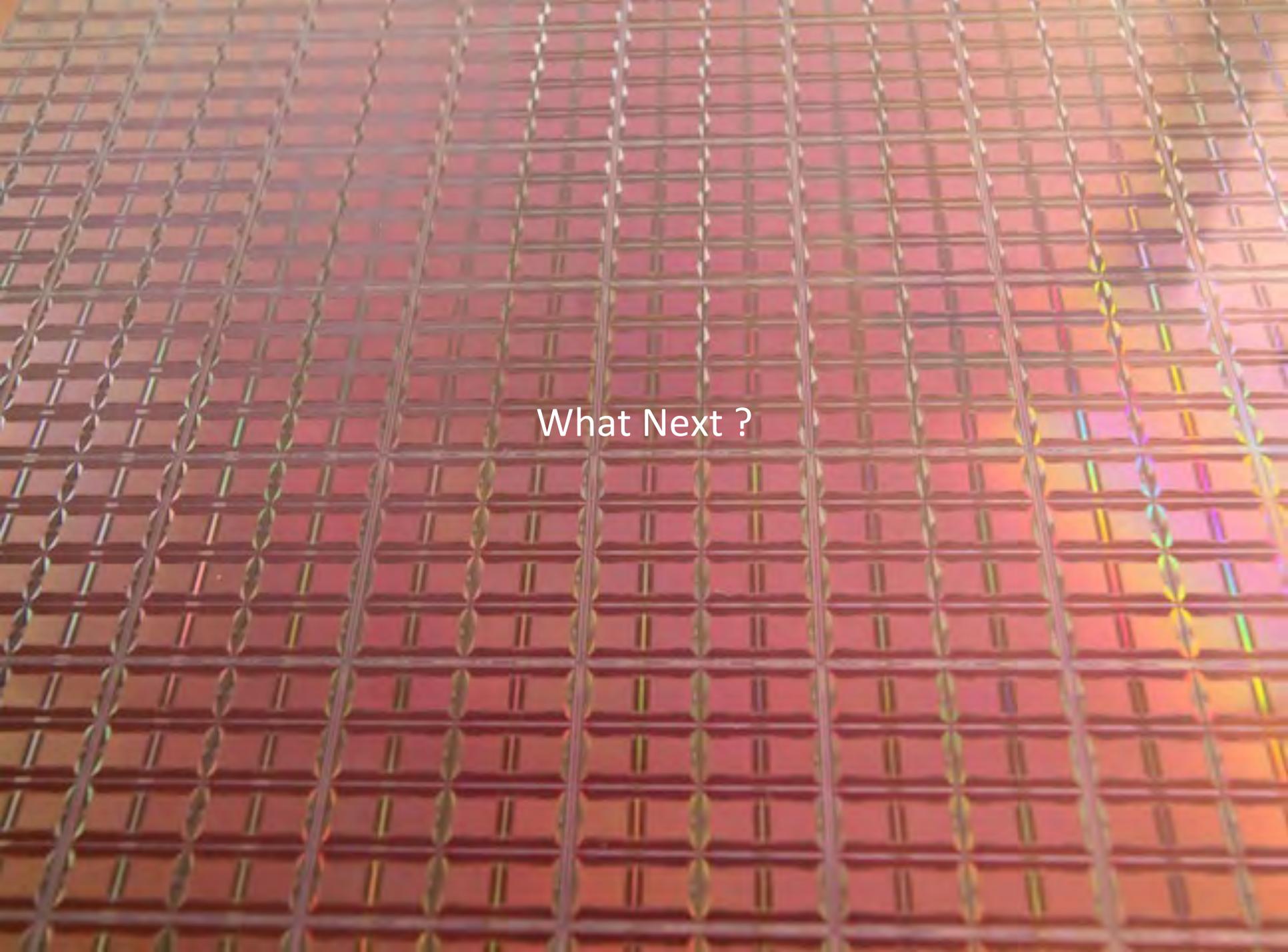
Shape implementation of Post-Synaptic Potentials (PSPs)



Convergence towards the unconstrained equilibrium distributions compared to the target distribution

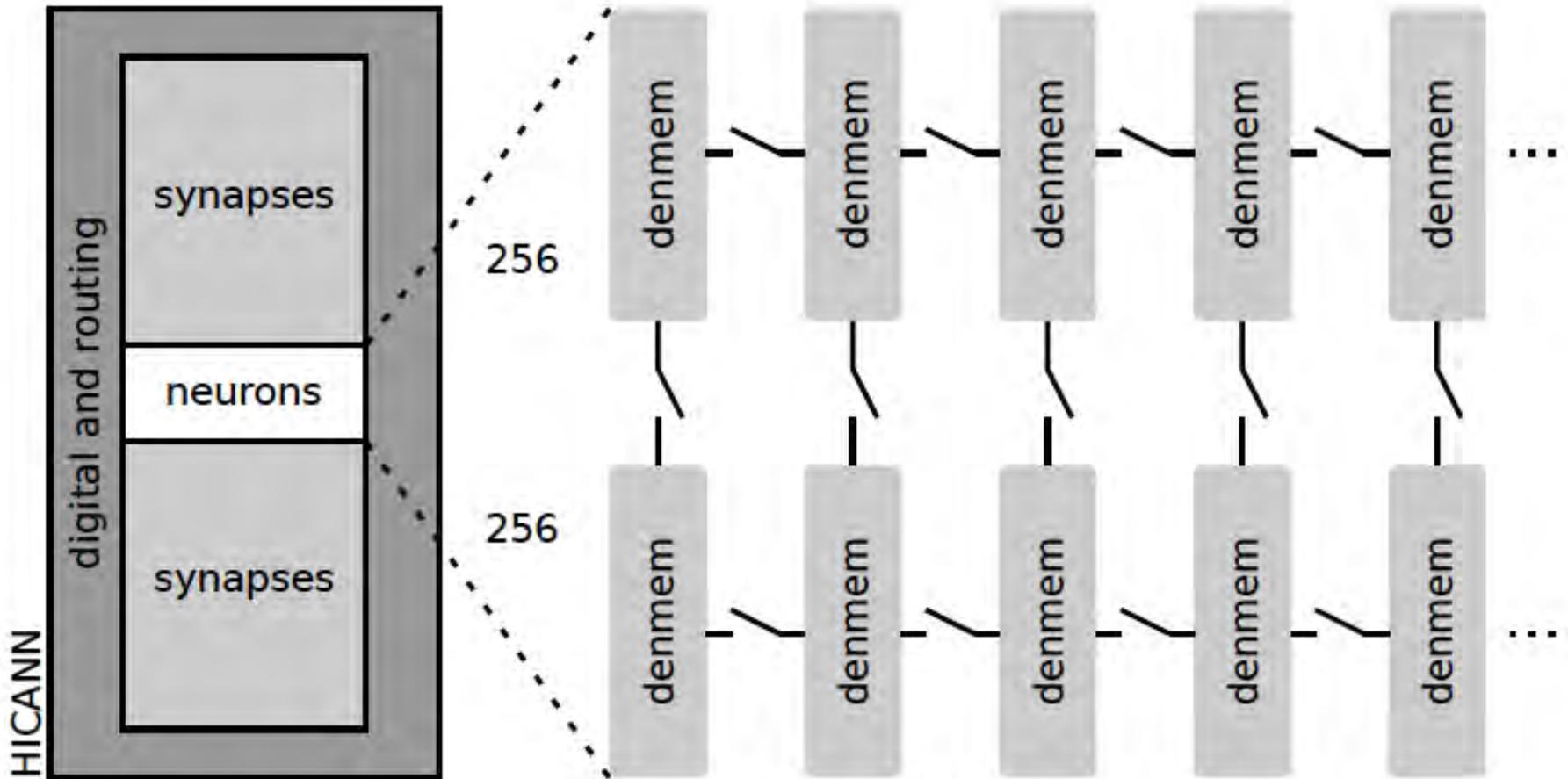
*All times in bio-time scales - slower in simulation – faster in emulation*

Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons  
Probst et al., Frontiers in Computational Neuroscience, February 2015 (9)

The background of the image is a close-up, slightly angled view of a woven metal mesh. The mesh consists of a grid of small, diamond-shaped openings. The surface of the metal has a fine, granular texture. A rainbow-like iridescent sheen is visible, particularly on the right side, where the colors transition from purple and blue to yellow and orange. The overall lighting is soft and even, highlighting the metallic texture and the subtle color variations.

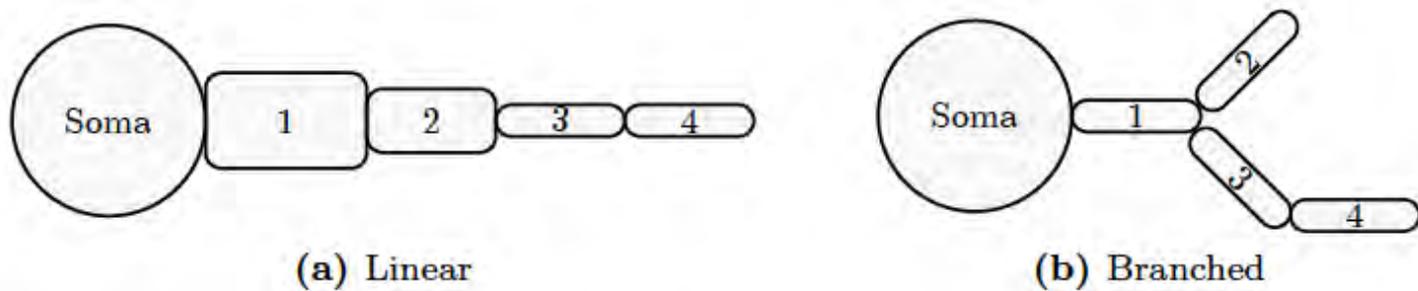
What Next ?

# Building Blocks for Synaptic Input – Prepared for Multi-Compartment Upgrades



Millner, Sebastian, et al. "Towards biologically realistic multi-compartment neuron model emulation in analog VLSI." *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 2012.

# Passive Compartment Concept, Size and Branching – PhD Sebastian Millner



**Figure 6.2:** Two artificial sample neurons. The dendrites are passive except for conductance based synapses.

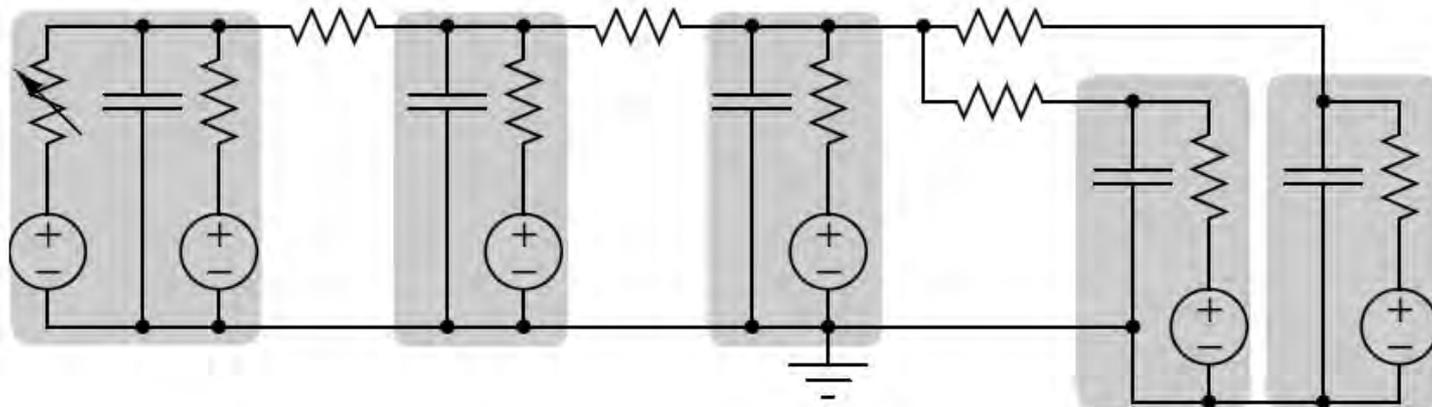
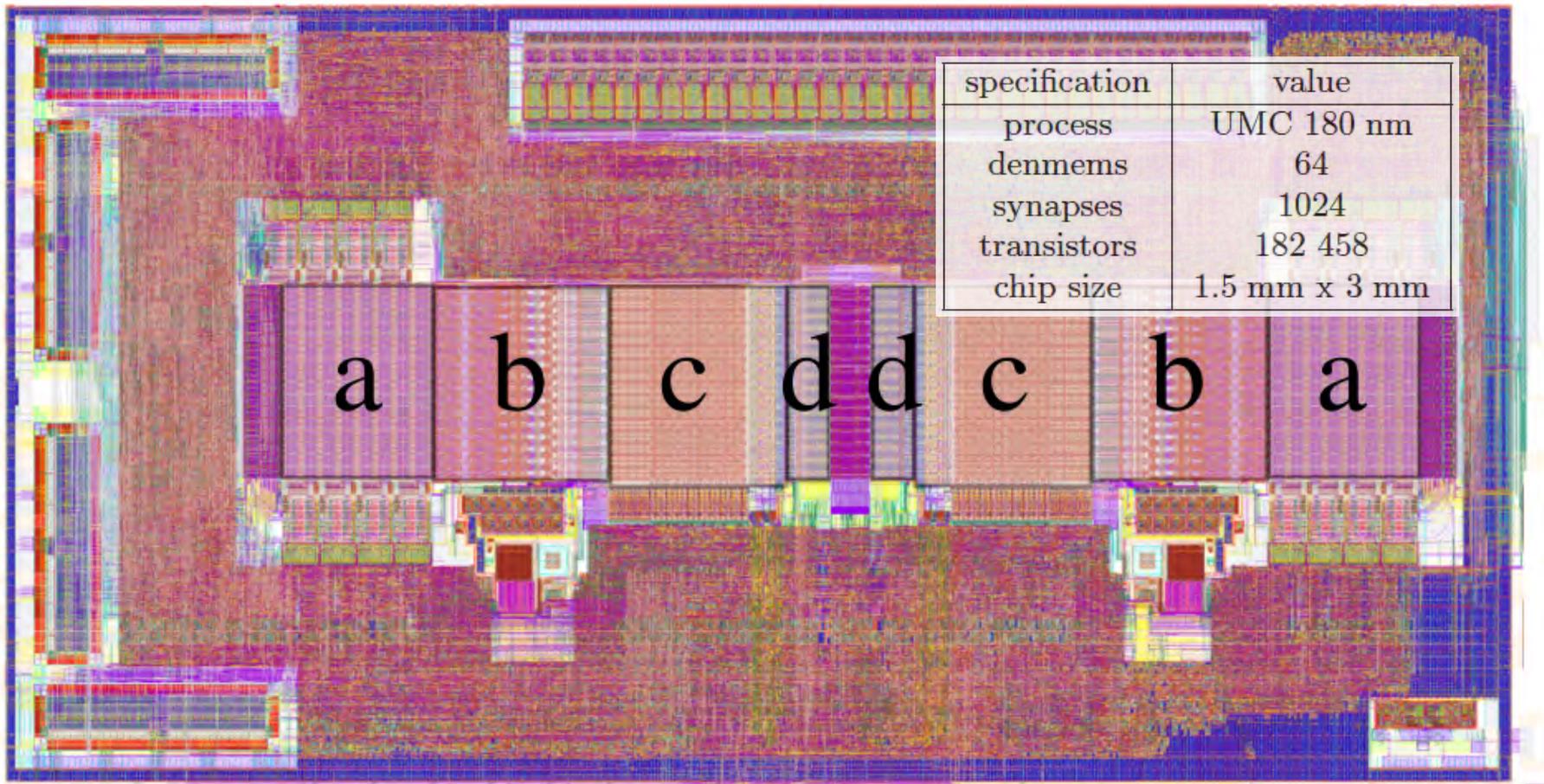


Fig. 1: Simplified schematic of a compartmental model with passive dendrites and one branching point.



## Structured Neurons

Passive features :Dendritic branching, signal dispersion

Active features : Dendritic Spikes, back propagating action potentials

Millner, Sebastian, et al. "Towards biologically realistic multi-compartment neuron model emulation in analog VLSI." *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. 2012.

# Passive Signal Dispersion for Signal Injection at different Distances from Soma Measurement vs. Simulation

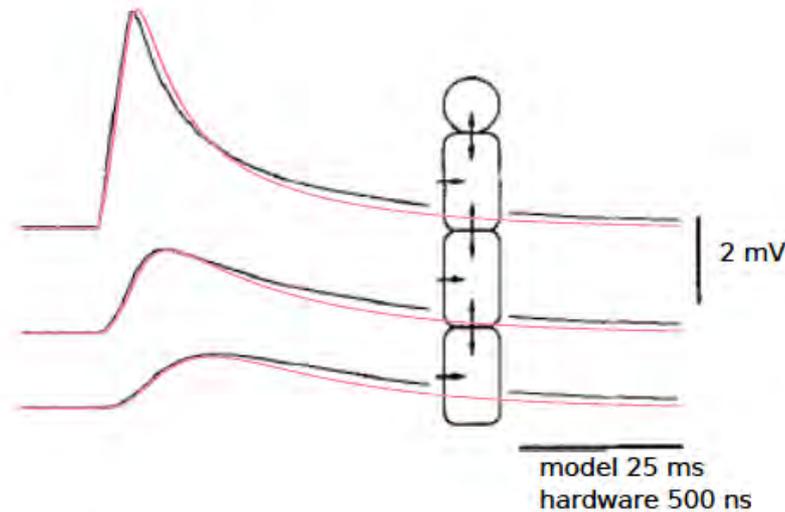
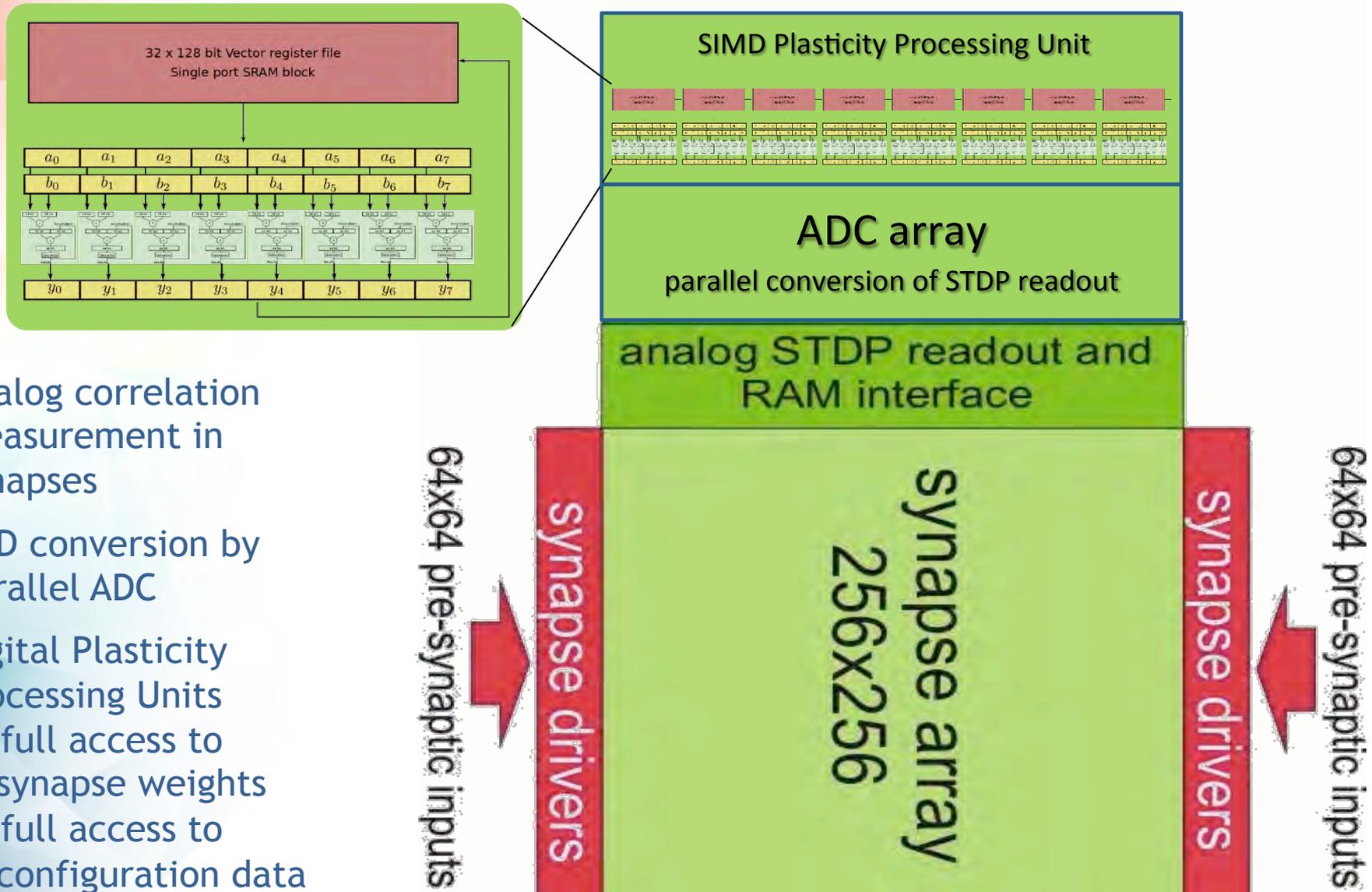


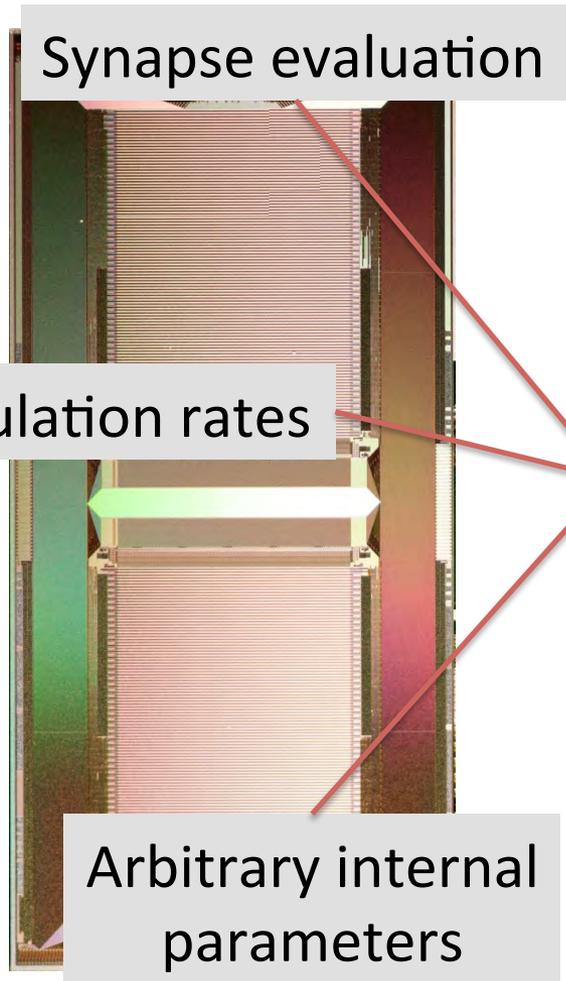
Fig. 4: Comparison between somatic membrane response to synaptic input at different parts of the dendrite in computer simulation (black) from [17] and a simulation of our hardware emulation (red)

# Plasticity : Hybrid Scheme Provides Flexibility

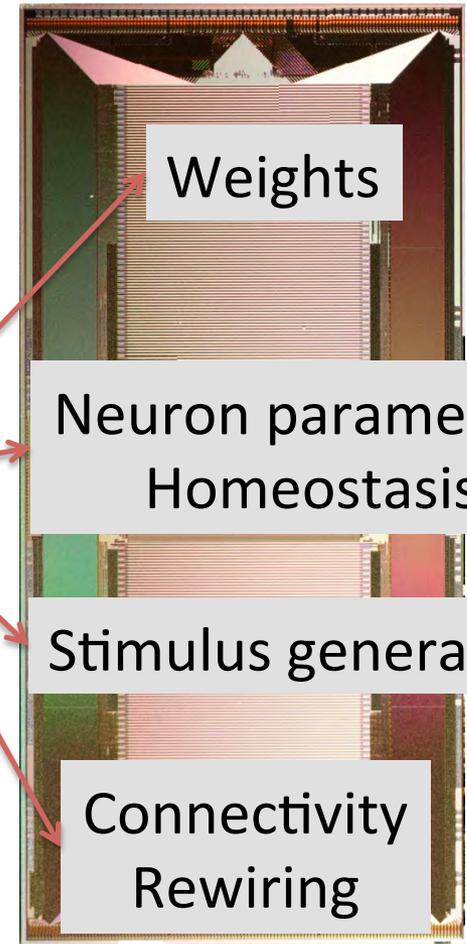
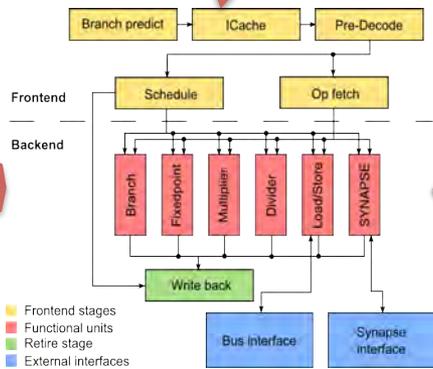


# Observables

# Controls



External rewards and controls

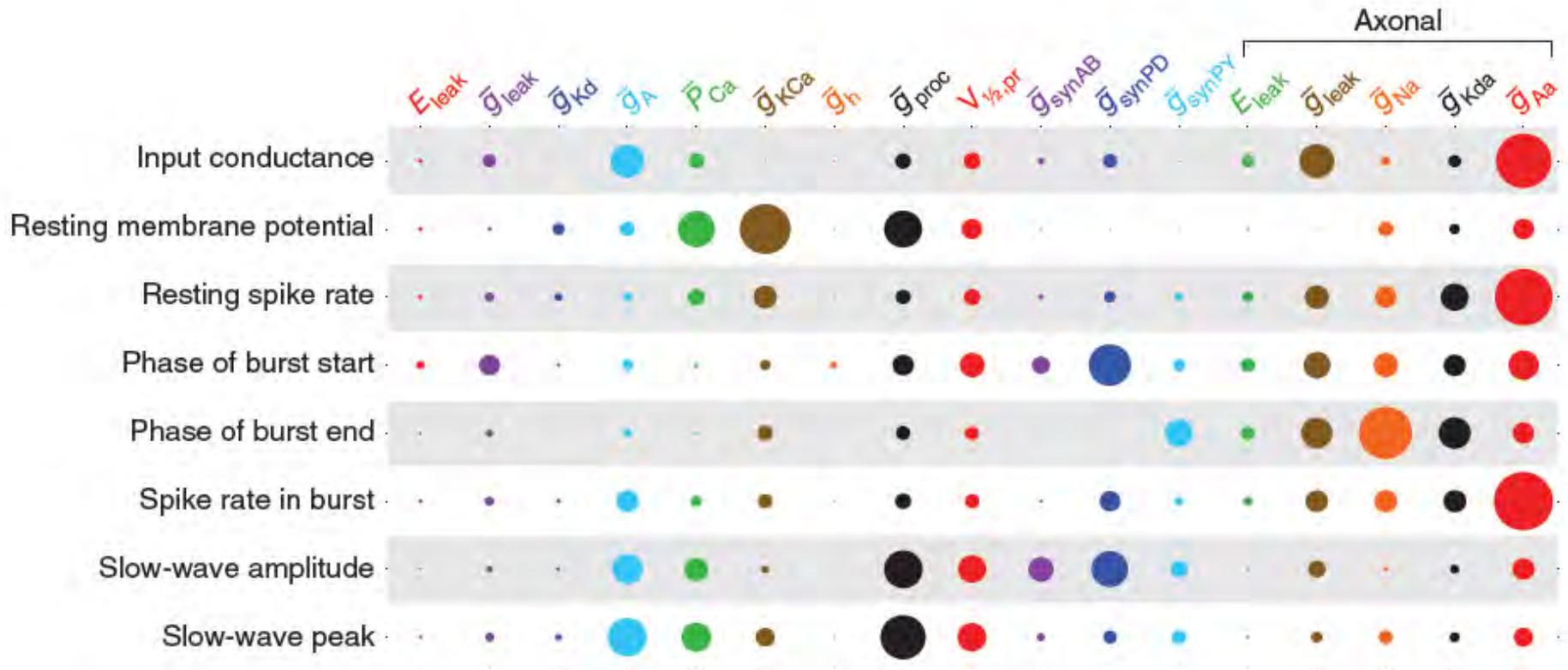
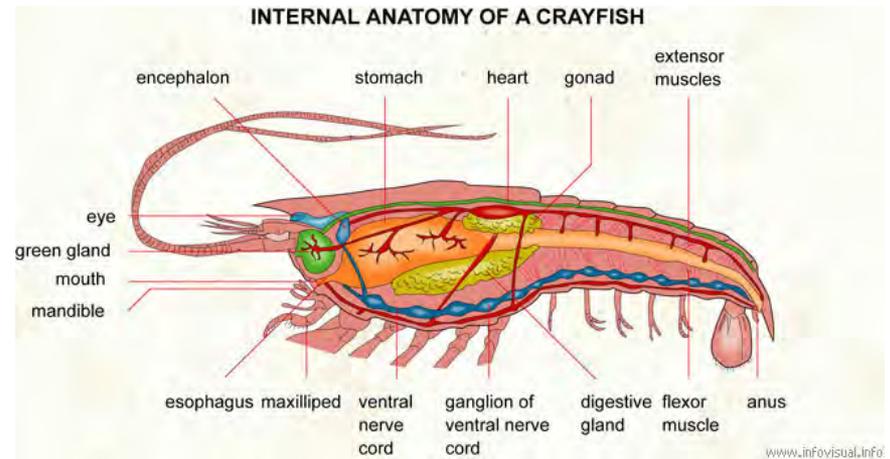


Essential : Any timescale > 100  $\mu$ s (bio) is accessible

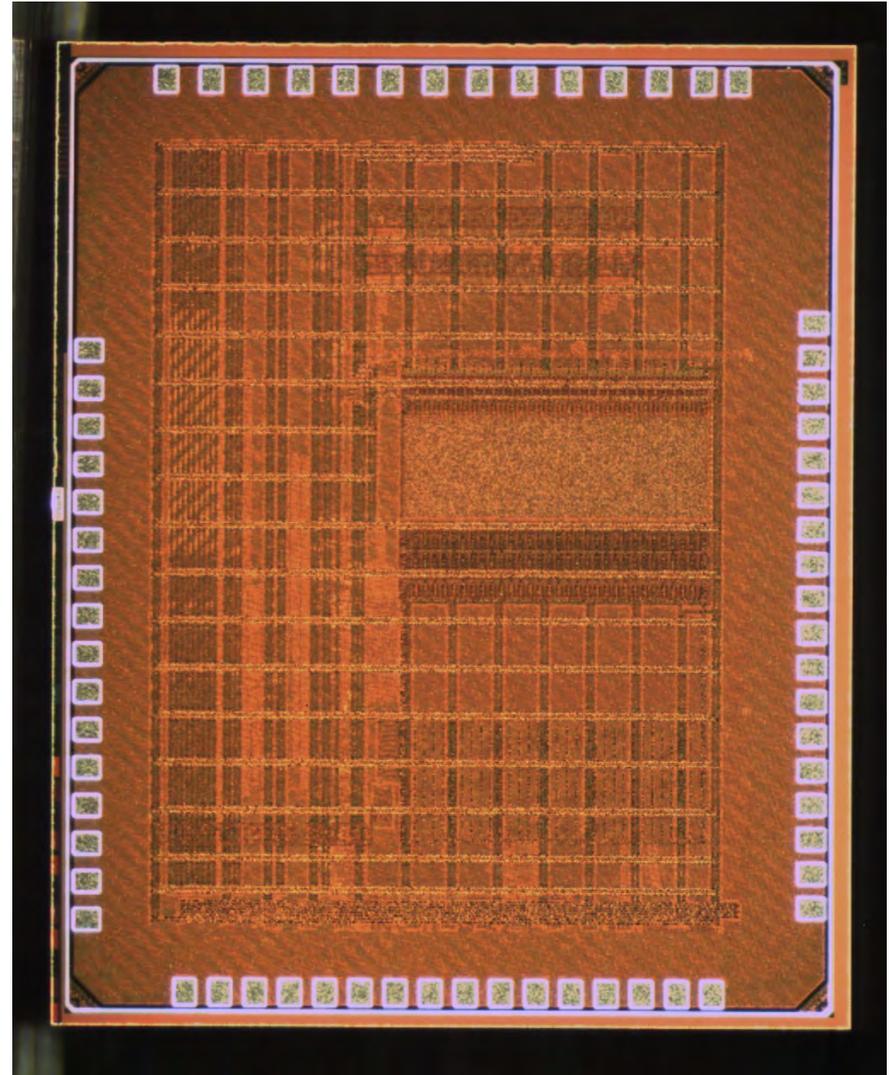
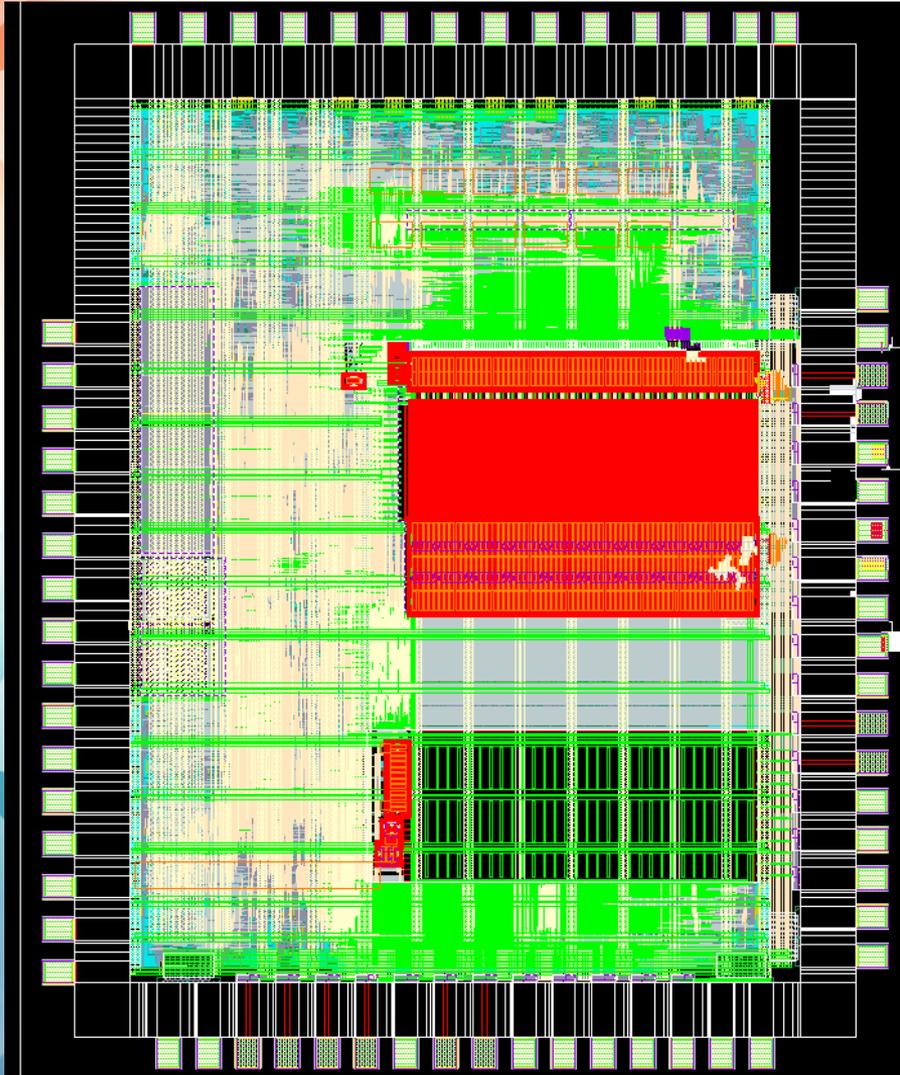
# Homeostasis

Crustacean stomatogastric ganglion

400.000 de-generate solutions  
for a model control network  
with 17 Cell Parameters



# February 2015 : Prototype ready (65nm)



# TimeScales

	Nature	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	Jahr	1000 Years	3000 s
<i>12 Orders of Magnitude</i>			
Evolution	> Millenia	> 1000 Millenia	> Months
<i>&gt; 15 Orders of Magnitude</i>			

