The BrainScaleS physical model machine From commissioning to real world problem solving

5th Neuro Inspired Computational Elements Workshop

NICE 2017

Karlheinz Meier
Ruprecht-Karls-Universität Heidelberg

meierk@kip.uni-heidelberg.de
@brainscales

Why brain inspired computing ?

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

need model system to test ideas

Two **fundamentally different** modeling approaches:

- **NUMERICAL MODEL (Turing)**

  represents model parameters as binary numbers

- **PHYSICAL MODEL (not Turing)**

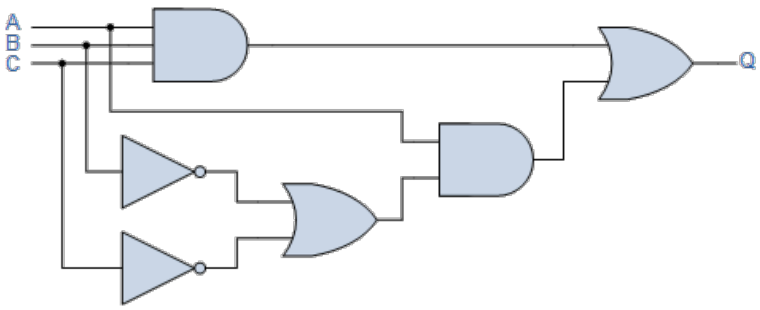
  represents model parameters as physical quantities

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

can be combined to form a hybrid system

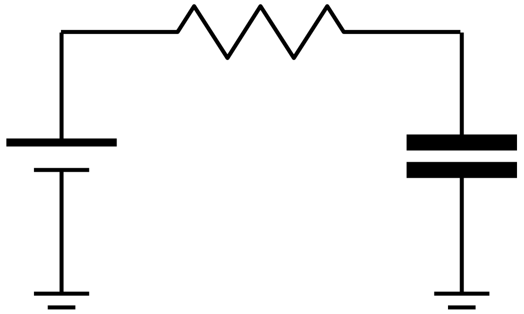
Digital

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



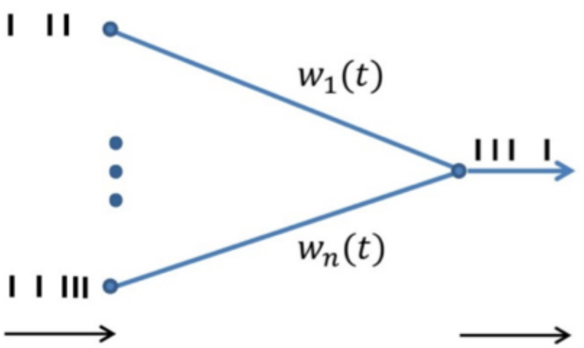
Analog

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

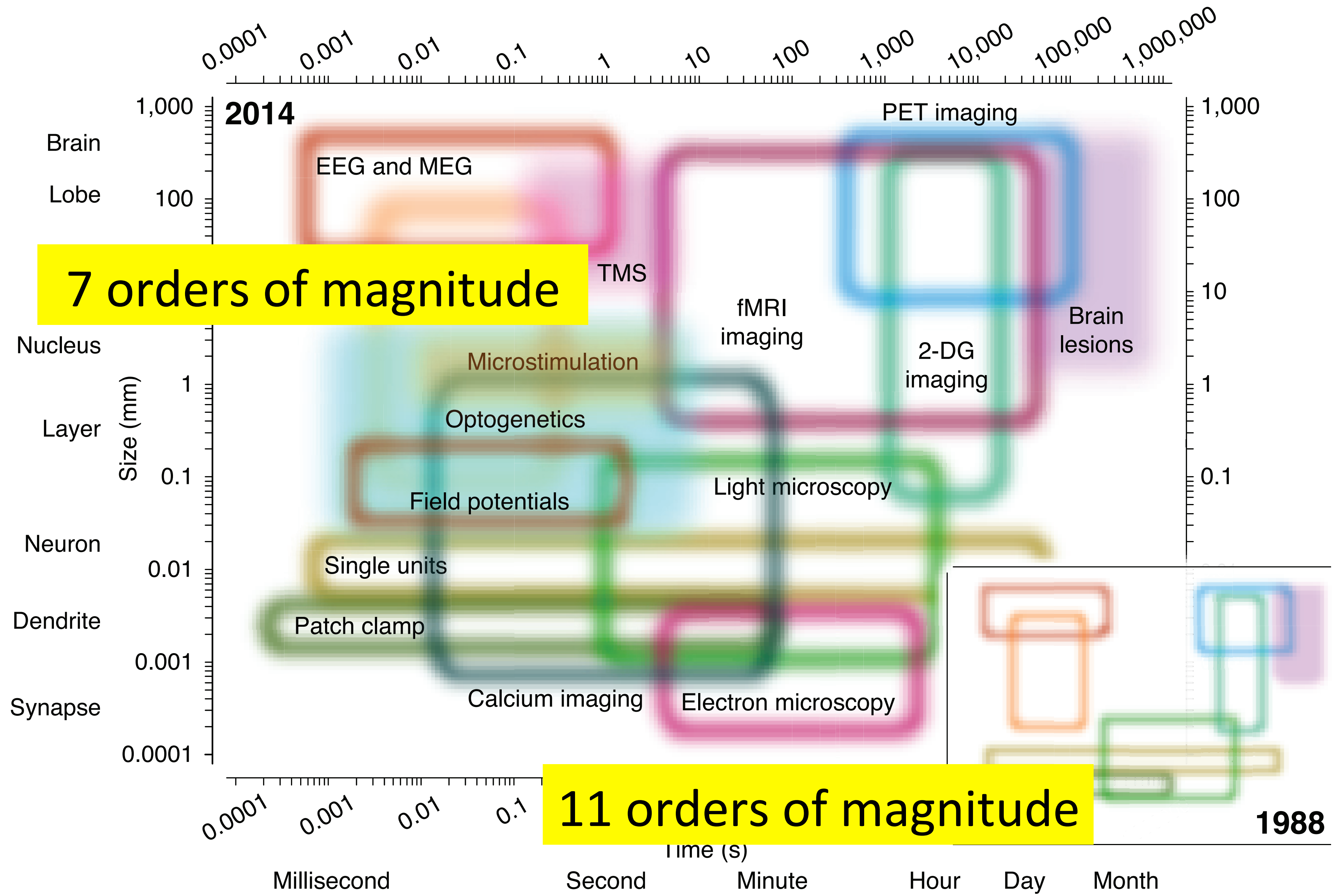


Nature / mixed-signal

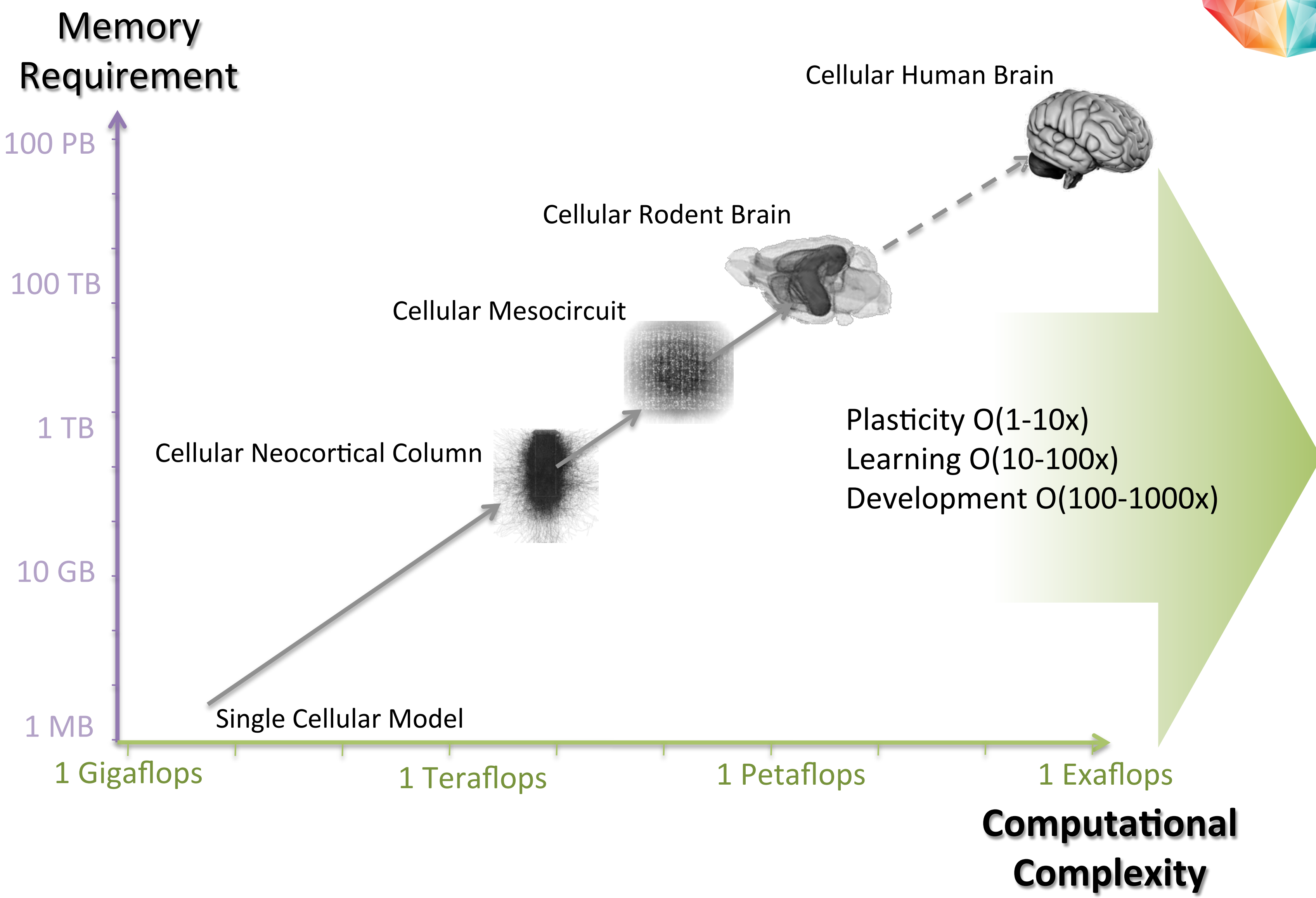
- Local analogue computation
- Binary communication by spikes
- Signal restoration



Modern Neuroscience : Access to multiple Scales in Space and Time



Sejnowski et al, Nature Neuroscience, 2014



Subcellular detail and plasticity require advances in strong scaling !

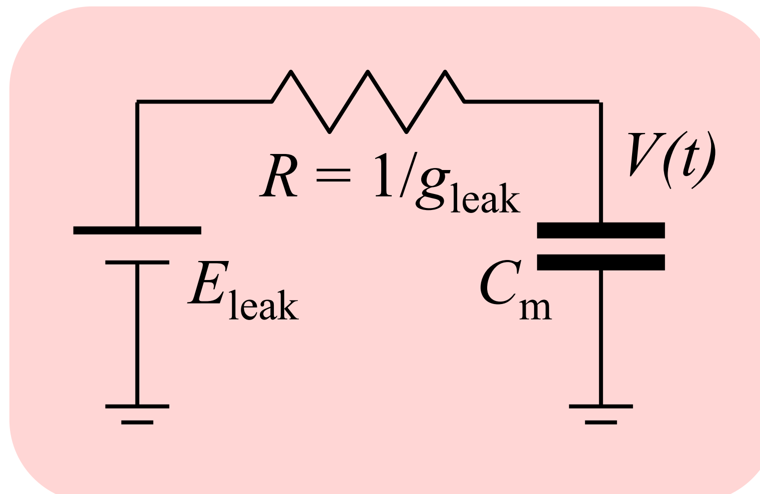
TimeScales	Nature	Simulation
Causality Detection	10^{-4} s	0.1 s
Synaptic Plasticity	1 s	1000 s
Learning	Day	1000 Days
Development	Year	1000 Years
12 Orders of Magnitude		
Evolution	> Millenia	> 1000 Millenia
> 15 Orders of Magnitude		

Physical Model System

Continuous Time Integrating Neural Cell Membrane

(+ non-linearity)

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



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

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

$$c_{\mathrm{m}} \frac{dV}{dt} = -g_{\mathrm{leak}} \left(V - E_{\mathrm{l}} \right) + \sum_k p_k g_k \left(V - E_{\mathrm{x}} \right) + \sum_l p_l g_l \left(V - E_{\mathrm{i}} \right)$$

$p_{k,l}(t)$ exponential onset and decay (PSP shape)

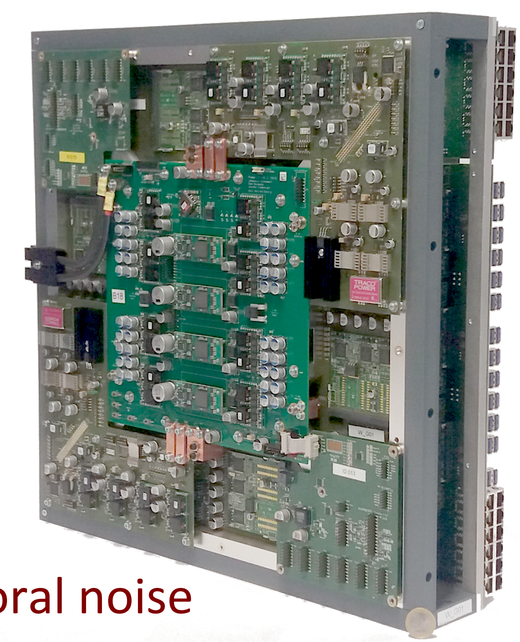
$g_{k,l}$ 0 to $g_{\max}$ ("weights")

effective membrane time-constant $c_{\mathrm{m}} / g_{\mathrm{total}}$ is time-dependent

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

10 Rationales for the Physical Model System

- Mixed-Signal (Local analog computation, binary spike communication)
- Driven by architecture, not devices (180nm & 65nm 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

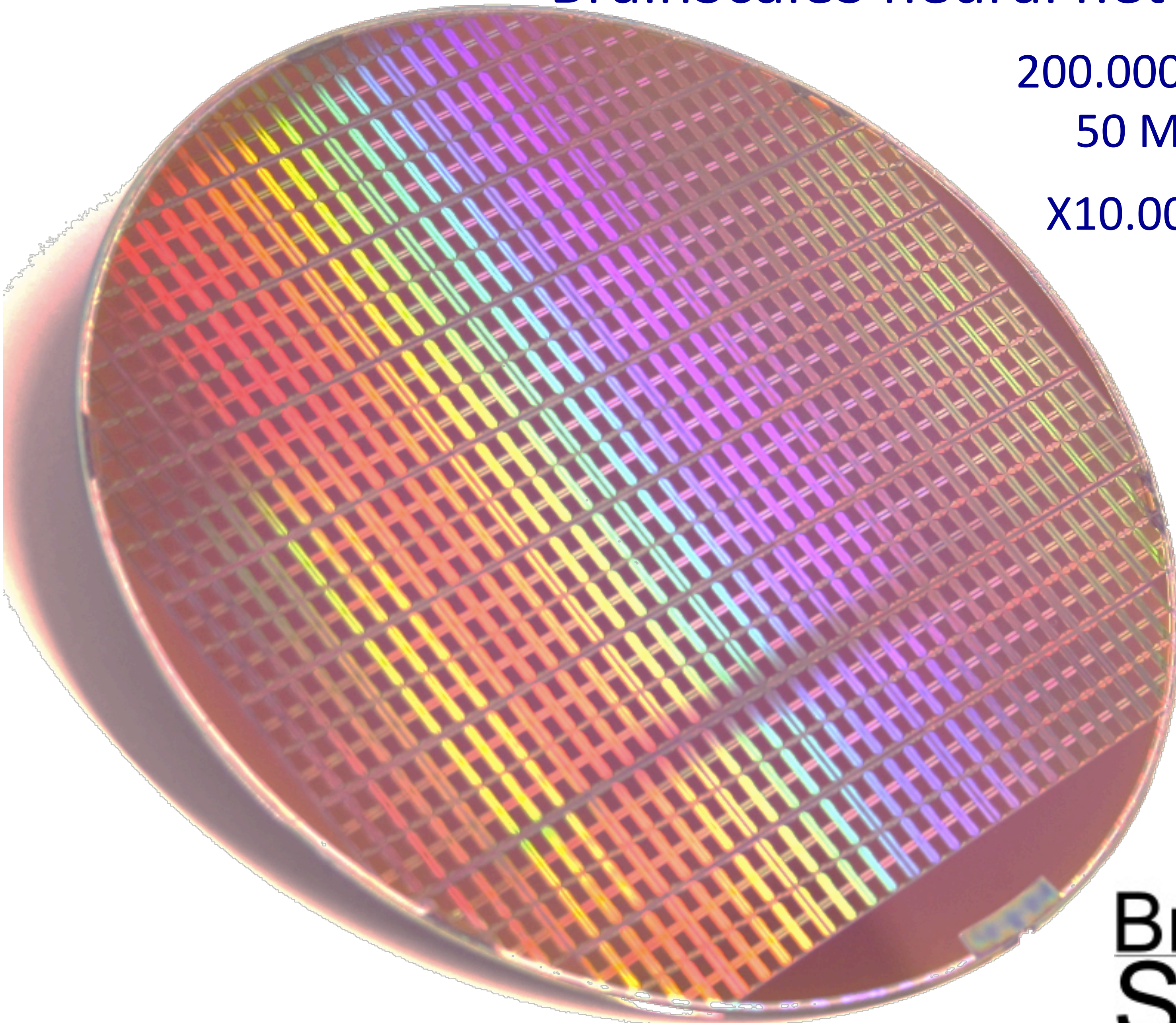
- rapid exploration of large parameter spaces
- cover short and long timescale circuit dynamics
- perform computing in the presence of spatial and temporal noise

BrainScaleS neural network wafer

200.000 AdEx neurons

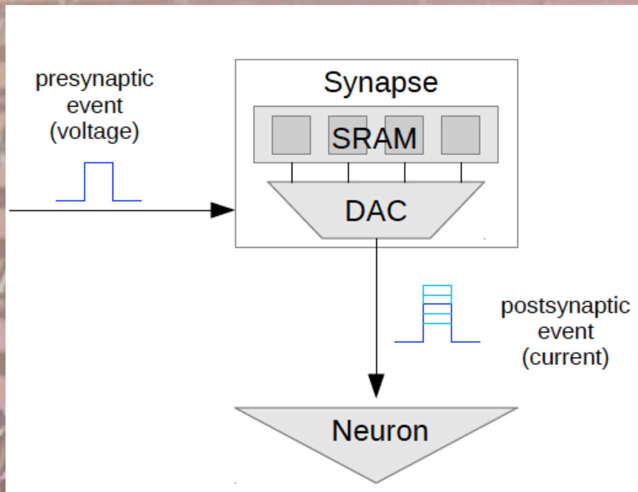
50 Million synapses

X10.000 acceleration

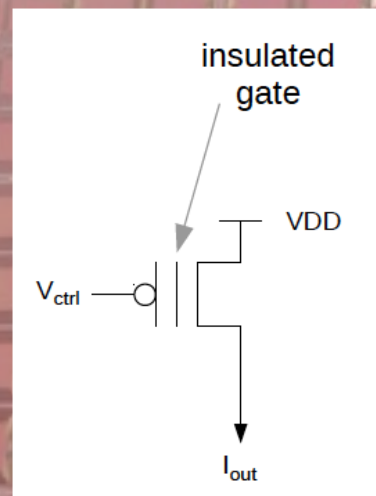


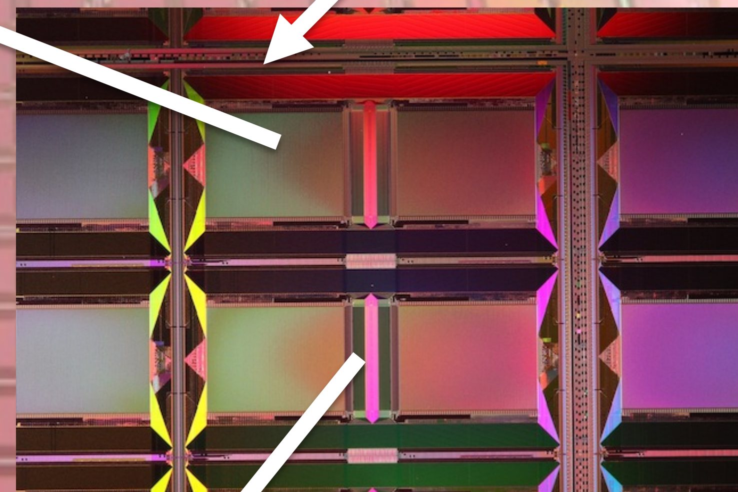
Multi-Scale Circuit Structure on an 8 inch CMOS Wafer (180nm)

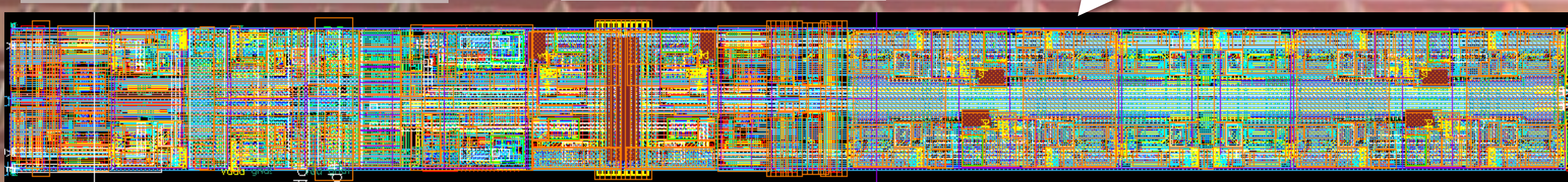
High Input Count **Network Chips**, 400 Instances on Wafer, Length Scale 1 cm network routing



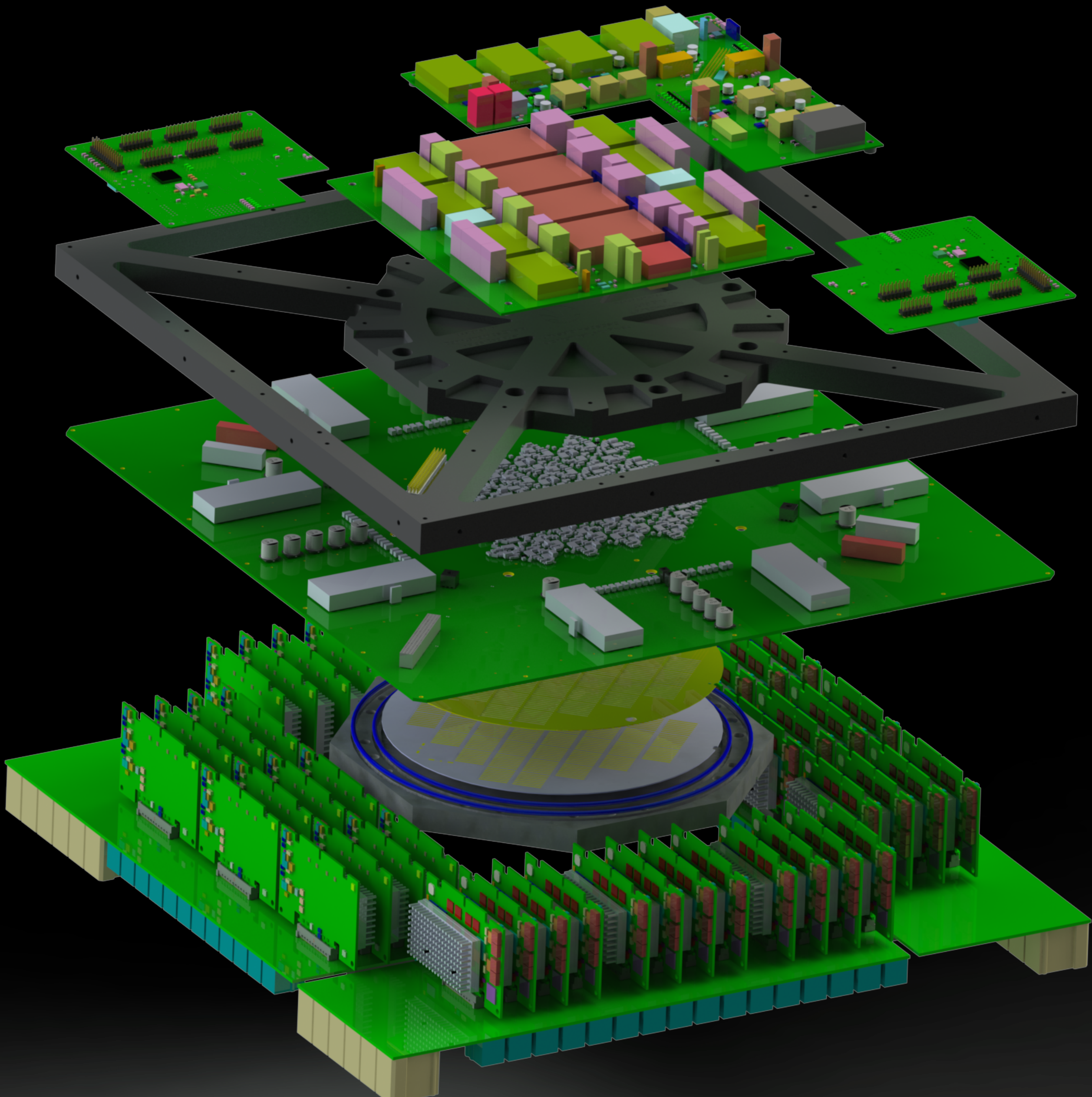
Plastic Synapses, 50.000.000 Million Instances on Wafer, Length Scale 10 µm, **volatile, fast,** 4-bit SRAM Weights







AdEx Neurons, 200.000 Instances on Wafer, Length Scale 300 µm,
NON-volatile, slow, Analog Floating Gate Parameter Storage
Poisson Noise Generators

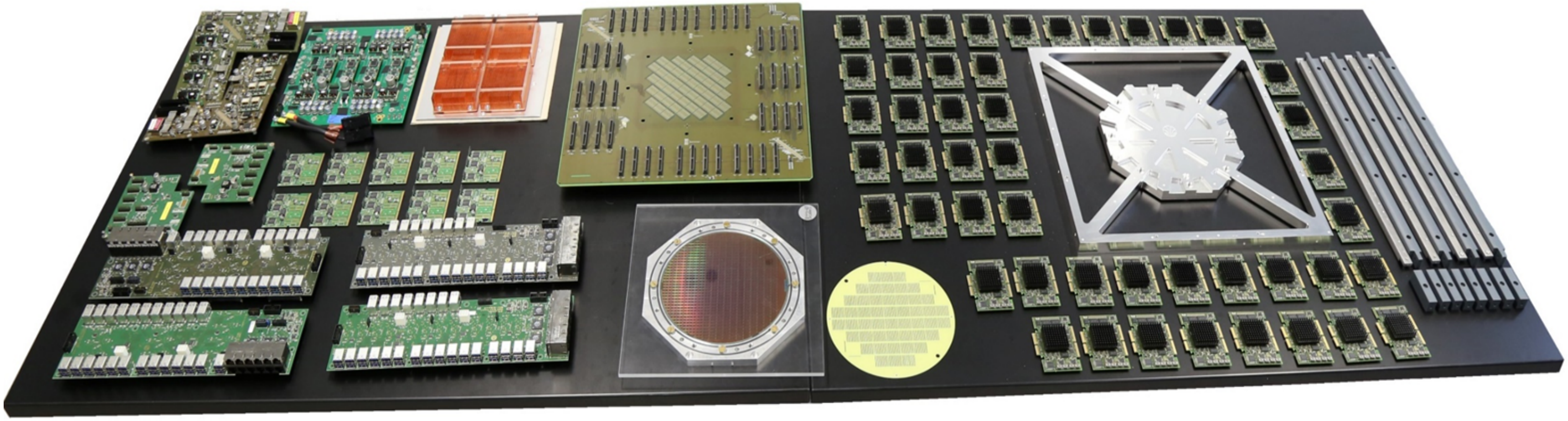


Physical Model, local analogue computing, binary continuous time communication

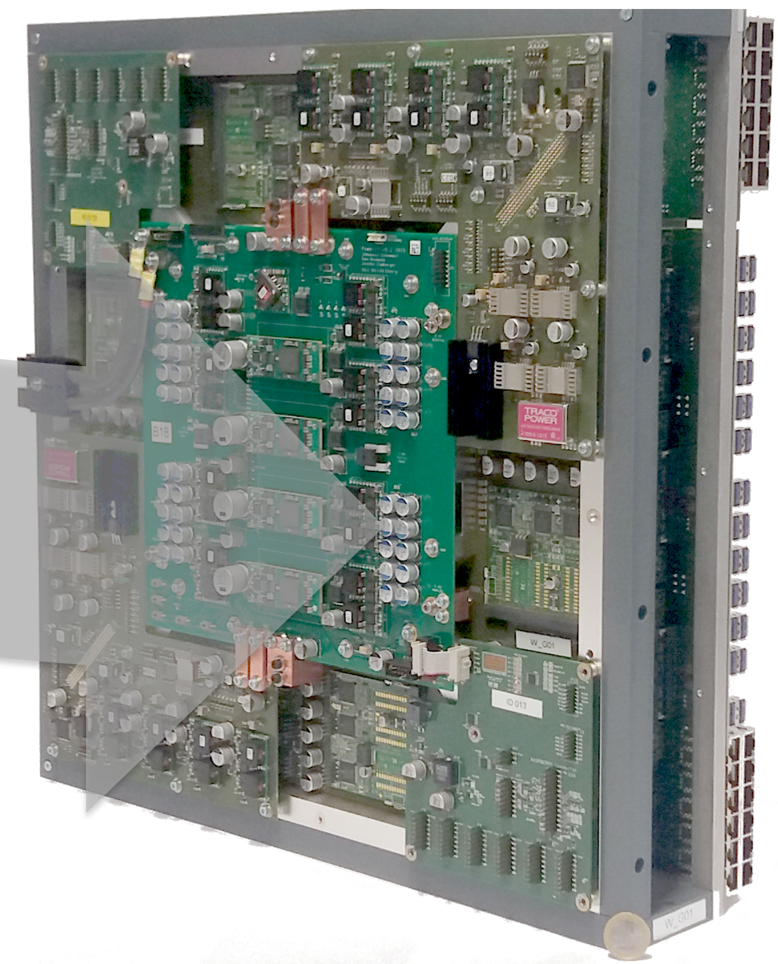
Wafer-Scale Integration of 200.000 neurons and 50.000.000 synapses on a single 20 cm wafer

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

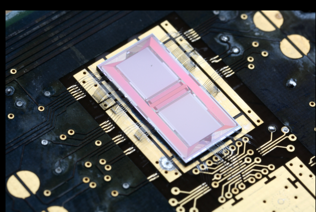
Wafer-scale integration of analog neural networks, J. Schemmel, J, Fieres and K. Meier
In : Proceedings of IJCNN (2008), IEEE Press, 431

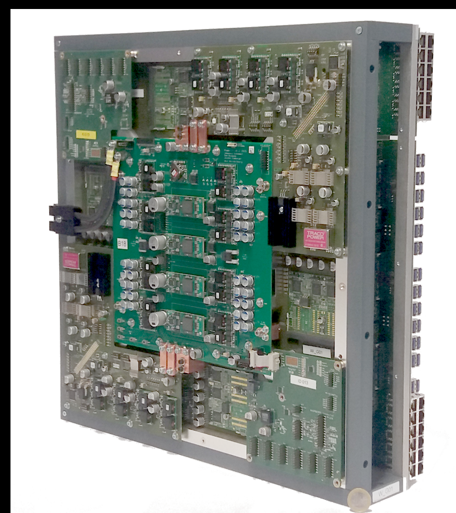


x 20 : 2500 PCBs



Scaling up





500 n / 100k s

200k n / 50m s

4m n / 1b s

Big machine in commissioning phase since March 30th 2016
Part the Human Brain Project (HBP) platform system

Configuration Space 40 MB for a full Wafer

Scope	Name	Type	Description
Neuron circuits (A)	n/a	$\mathsf{i_n}$	Two digital configuration bits activating the neuron and readout of its membrane voltage
	g_l	$\mathsf{i_n}$	Bias current for neuron leakage circuit
	τ_refrac	$\mathsf{i_n}$	Bias current controlling neuron refractory time
	E_l	$\mathsf{s_n}$	Leakage reversal potential
	E_inh	$\mathsf{s_n}$	Inhibitory reversal potential
	E_exc	$\mathsf{s_n}$	Excitatory reversal potential
	V_th	$\mathsf{s_n}$	Firing threshold voltage
	V_reset	$\mathsf{s_n}$	Reset potential
Synapse line drivers (B)	n/a	$\mathsf{i_l}$	Two digital configuration bits selecting input of line driver
	n/a	$\mathsf{i_l}$	Two digital configuration bits setting line excitatory or inhibitory
	$t_\mathrm{rise}, t_\mathrm{fall}$	$\mathsf{i_l}$	Two bias currents for rising and falling slew rate of presynaptic voltage ramp
	g_i^max	$\mathsf{i_l}$	Bias current controlling maximum voltage of presynaptic voltage ramp
Synapses (B)	w	$\mathsf{i_s}$	4-bit weight of each individual synapse
STP related (C)	n/a	$\mathsf{i_l}$	Two digital configuration bits selecting short-term depression or facilitation
	U_SE	$\mathsf{i_l}$	Two digital configuration bits tuning synaptic efficacy for STP
	n/a	$\mathsf{s_l}$	Bias voltage controlling spike driver pulse length
	$\tau_\mathrm{rec}, \tau_\mathrm{facil}$	$\mathsf{s_l}$	Voltage controlling STP time constant
	I	$\mathsf{s_l}$	Short-term facilitation reference voltage
	R	$\mathsf{s_l}$	Short-term capacitor high potential
STDP related (D)	n/a	$\mathsf{i_l}$	Bias current controlling delay for presynaptic correlation pulse (for calibration purposes)
	$A_{+/-}$	$\mathsf{s_l}$	Two voltages dimensioning charge accumulation per (anti-)causal correlation measurement
	n/a	$\mathsf{s_l}$	Two threshold voltages for detection of relevant (anti-)causal correlation
	τ_STDP	g	Voltage controlling STDP time constants

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		... amp
Synapses (B)			
STP related (C)	τ_{r}		... on
		s_l	Short-term capacitor high potential
STDP related (D)	n/a	i_l	Bias current controlling delay for presynaptic correlation pulse (for calibration purposes)
	$A_{+/-}$	s_l	Two voltages dimensioning charge accumulation per (anti-)causal correlation measurement
	n/a	s_l	Two threshold voltages for detection of relevant (anti-)causal correlation
	τ_{STDP}	g	Voltage controlling STDP time constants

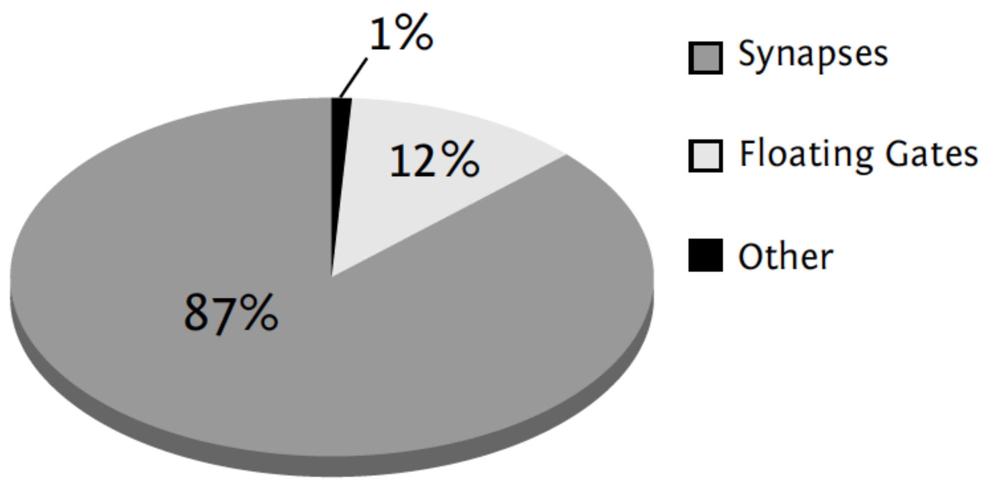
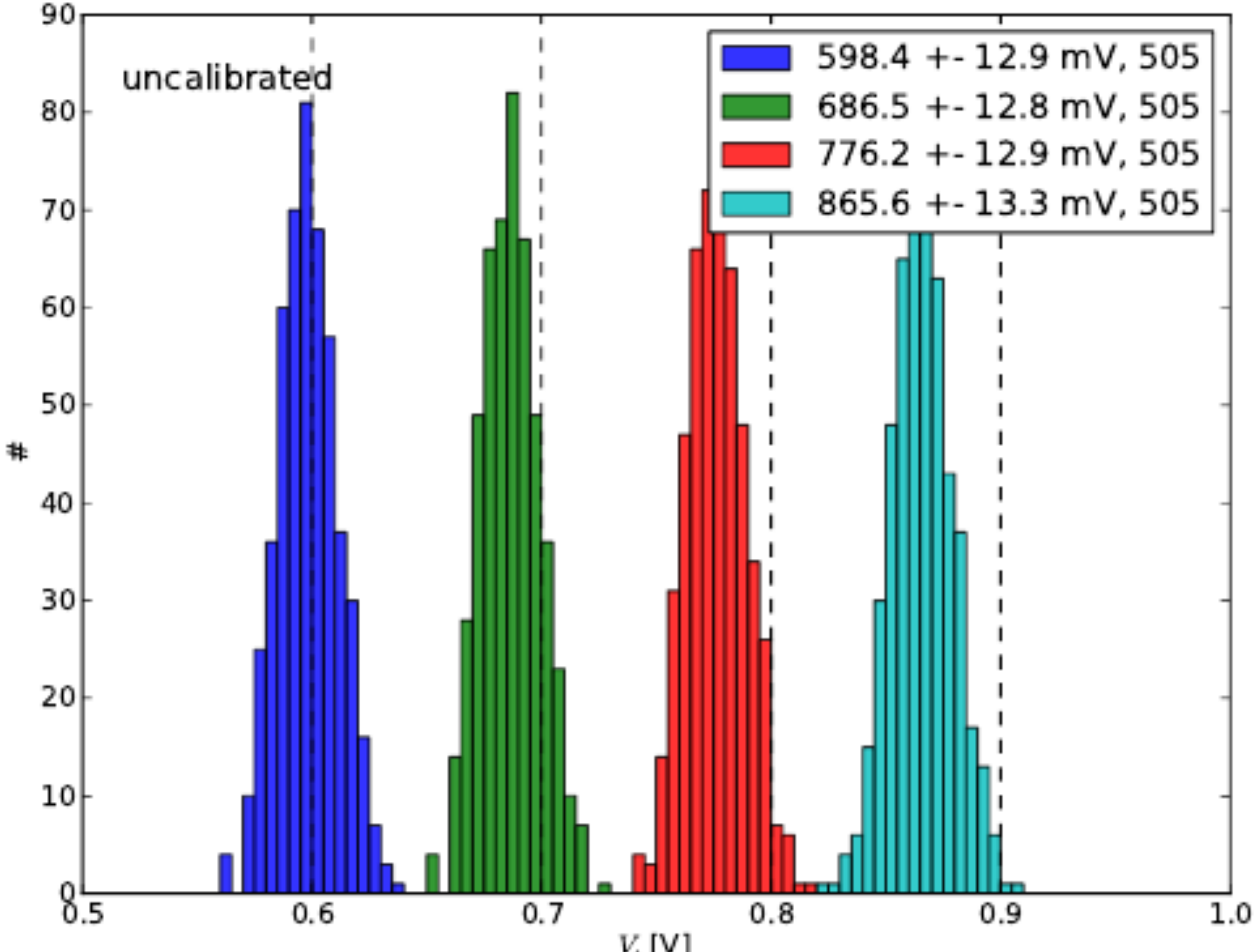


Fig. 4: Sector diagram of the parameter space to configure one HICANN chip. For a full wafer, the configuration data volume is 44 MB large.

Challenge and Opportunity : Variability

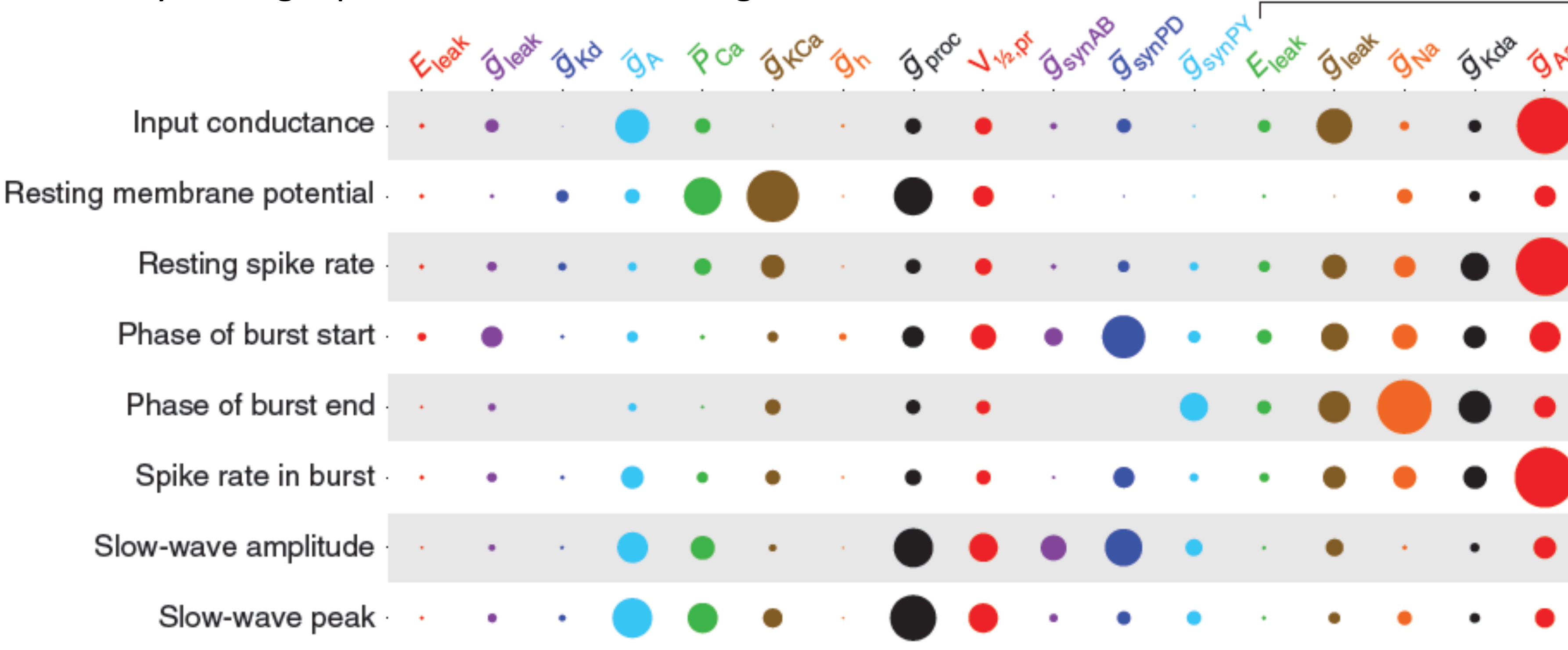


Pyloric rhythm of the crustacean stomatogastric ganglion

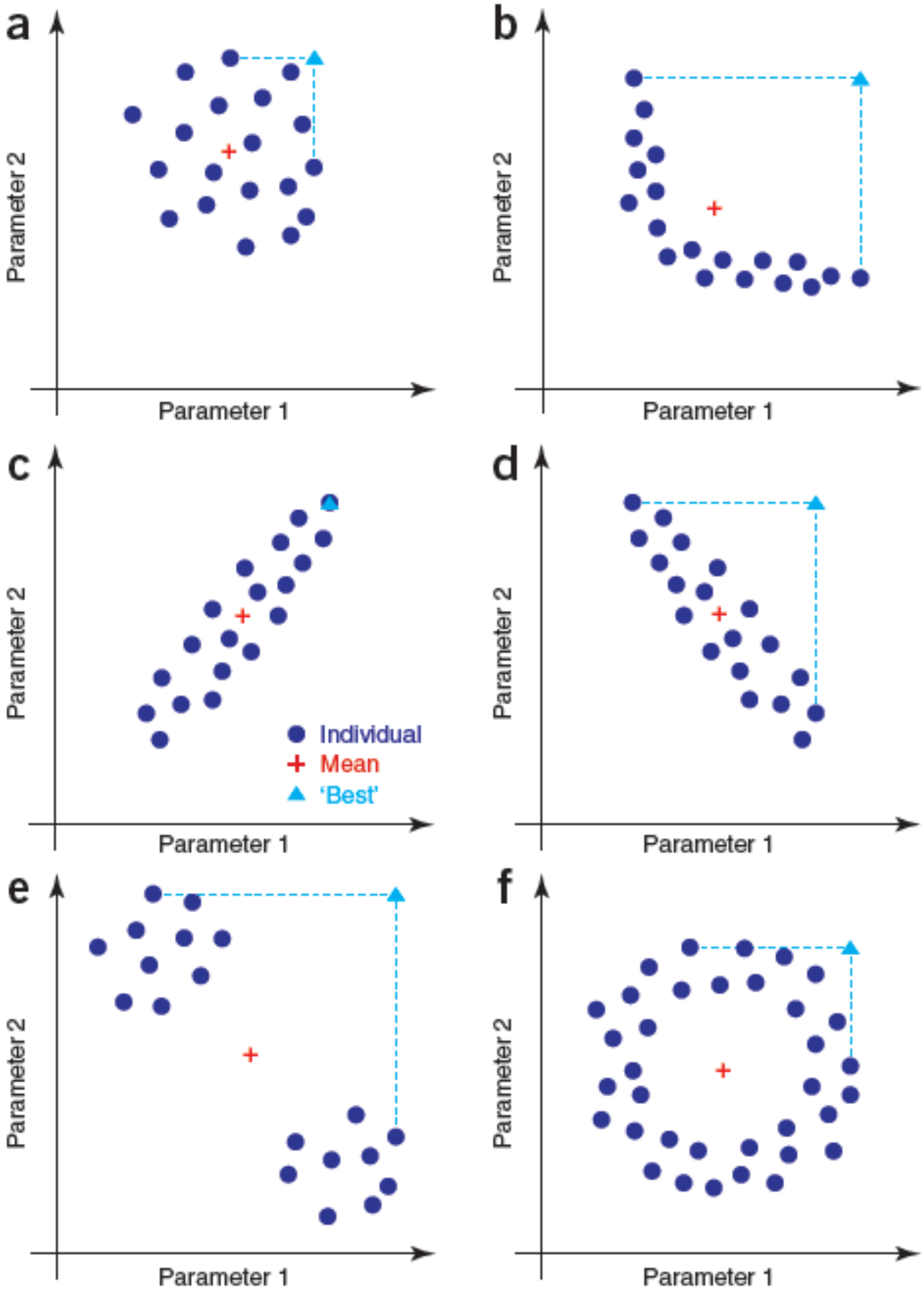
20.000.000 model networks created with 17 random cell parameters, fixed connectivity (Neuron)

400.000 networks found with „identical (de-generate)“ timing behaviour in measured biological range

Sensitivity of single parameters within „de-generate“ solutions



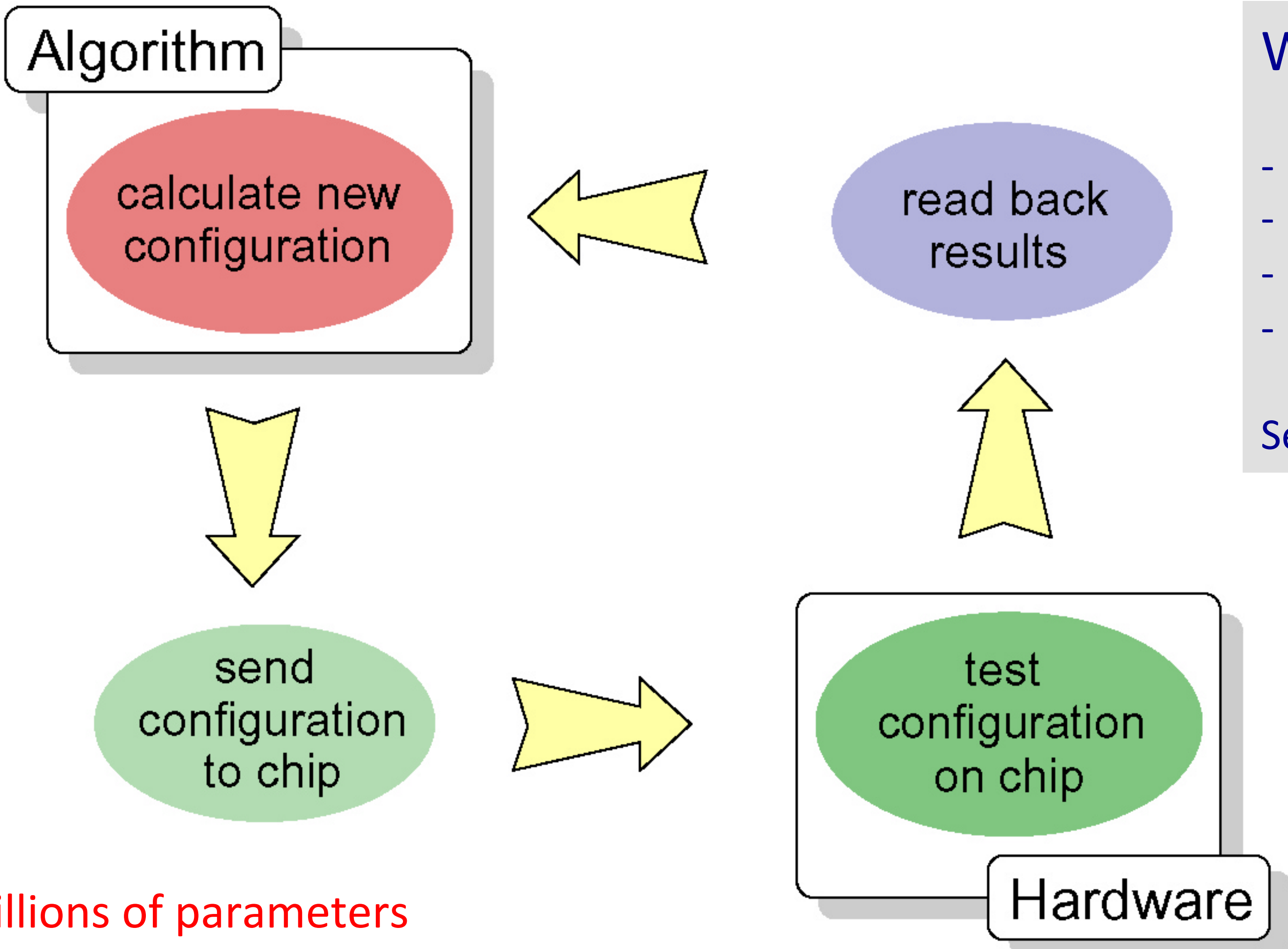
Marder, Taylor Nature Neuroscience 14, Nr 2, 2011



Variability has to be at the right place ...

Marder, Taylor Nature Neuroscience 14, Nr 2, 2011

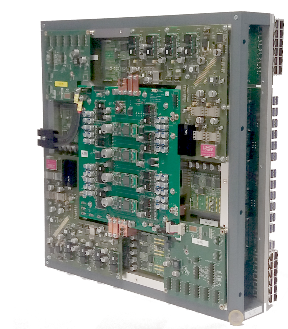
Hardware-In-the-Loop



What for ?

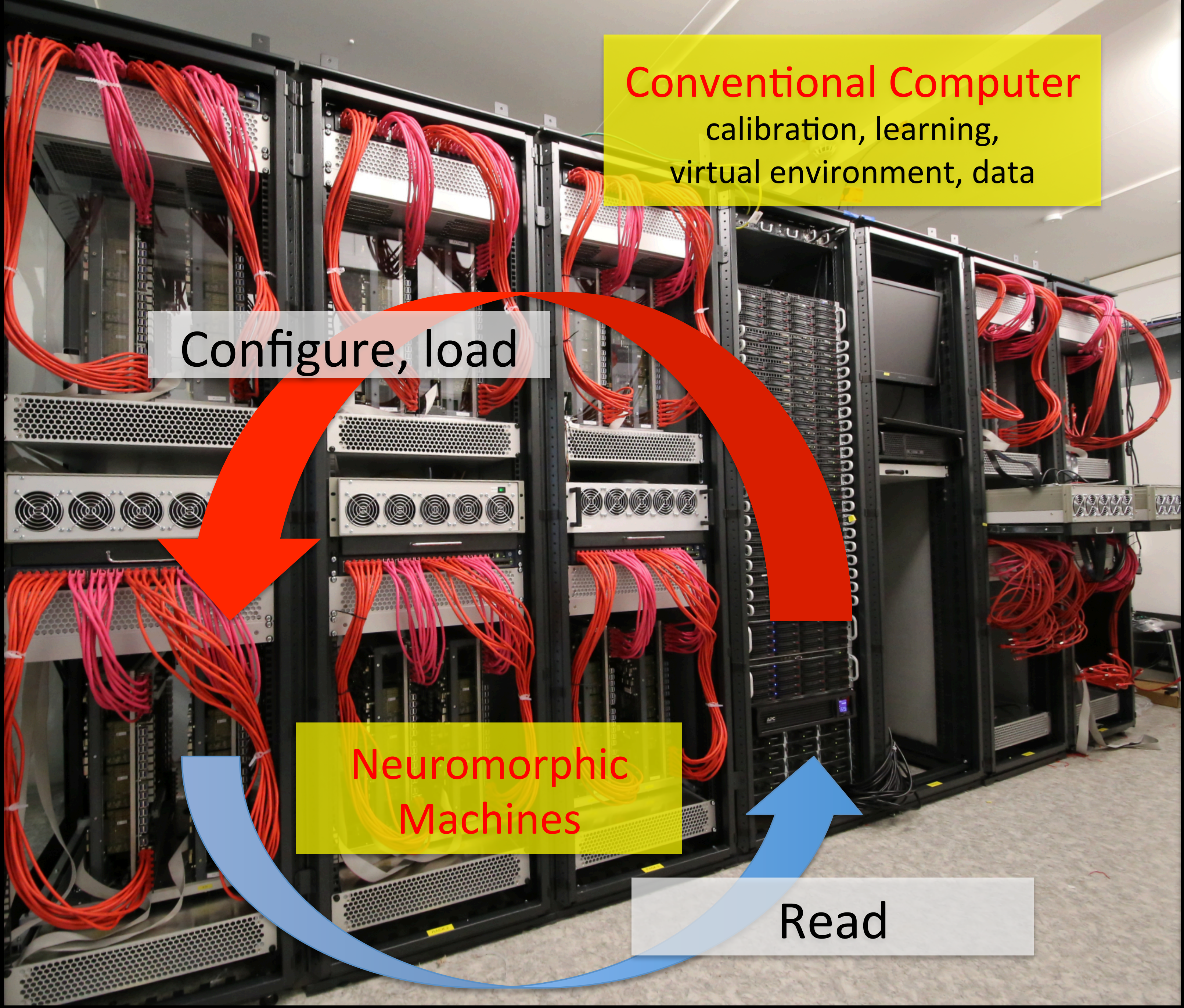
- Calibration
- Learning
- Environment
- Data

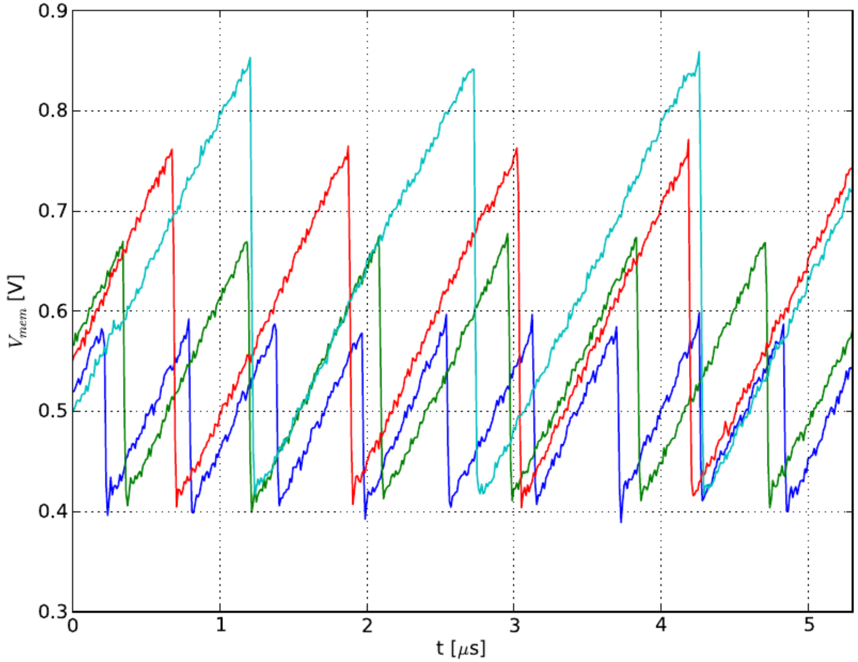
Separated ?

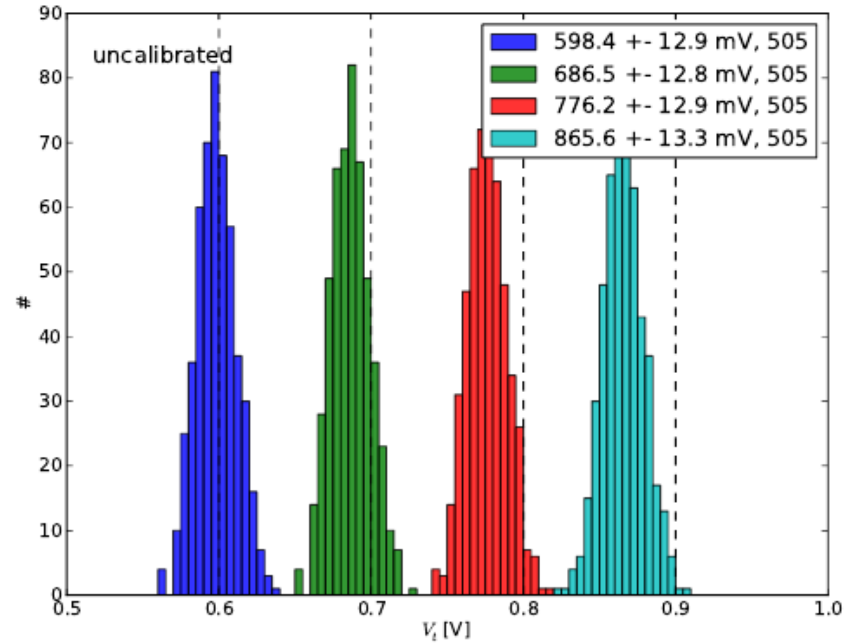


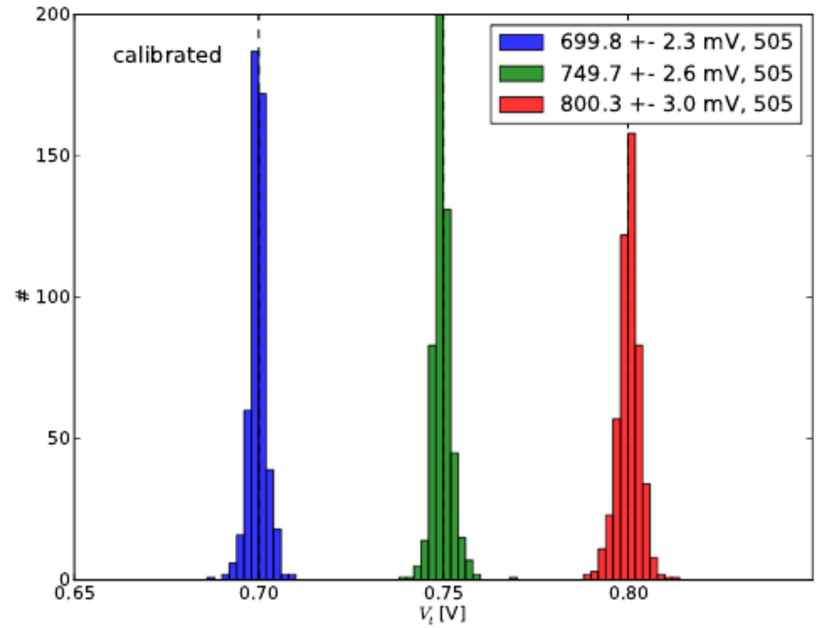
Millions of parameters

- network topology
- neuron sizes and parameters
- synaptic strengths


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







Calibration

Make BrainScaleS like a digital simulator ?

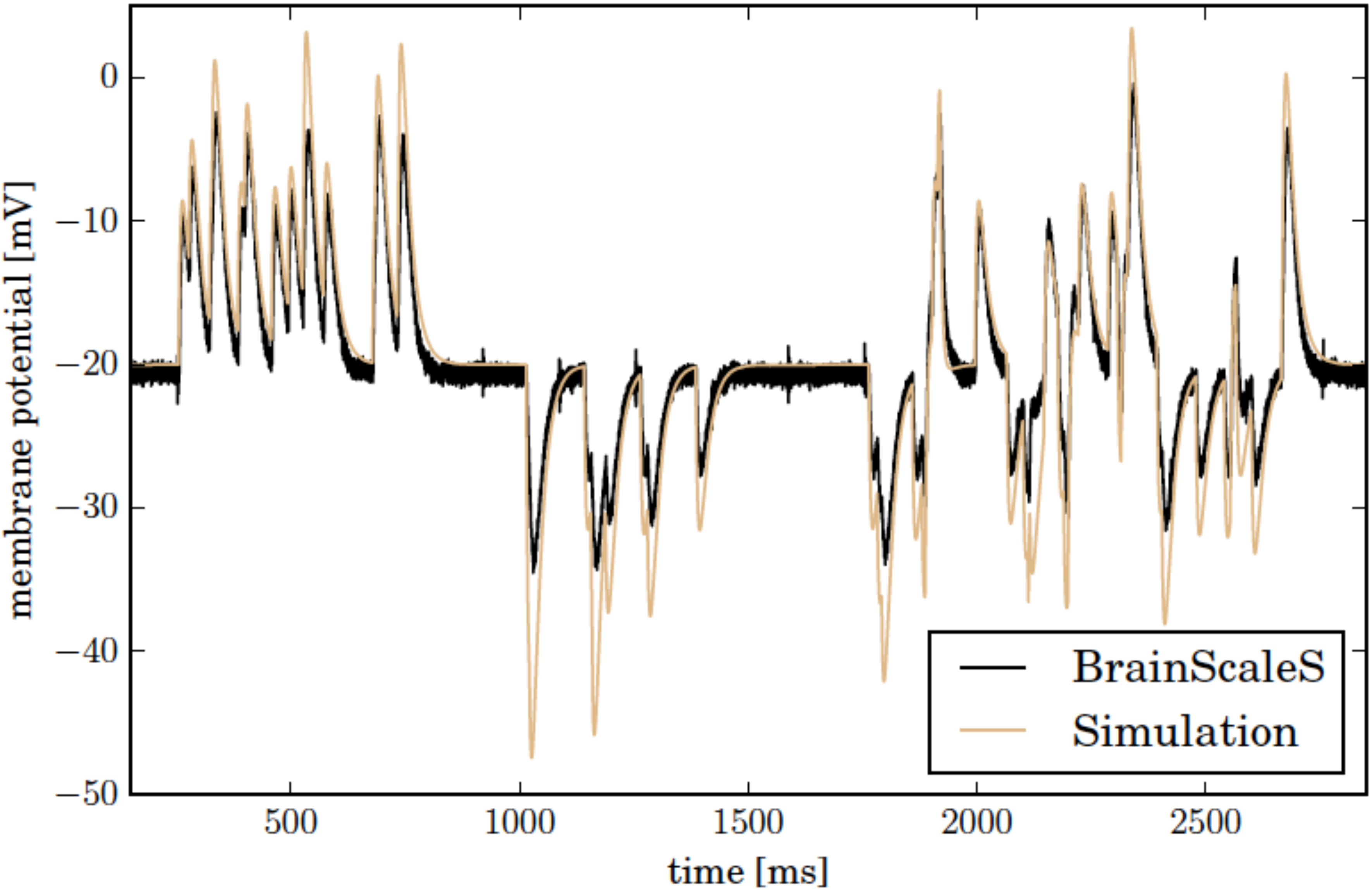
OR

Put variabiity at the right place !

By hand ? – By self learning !

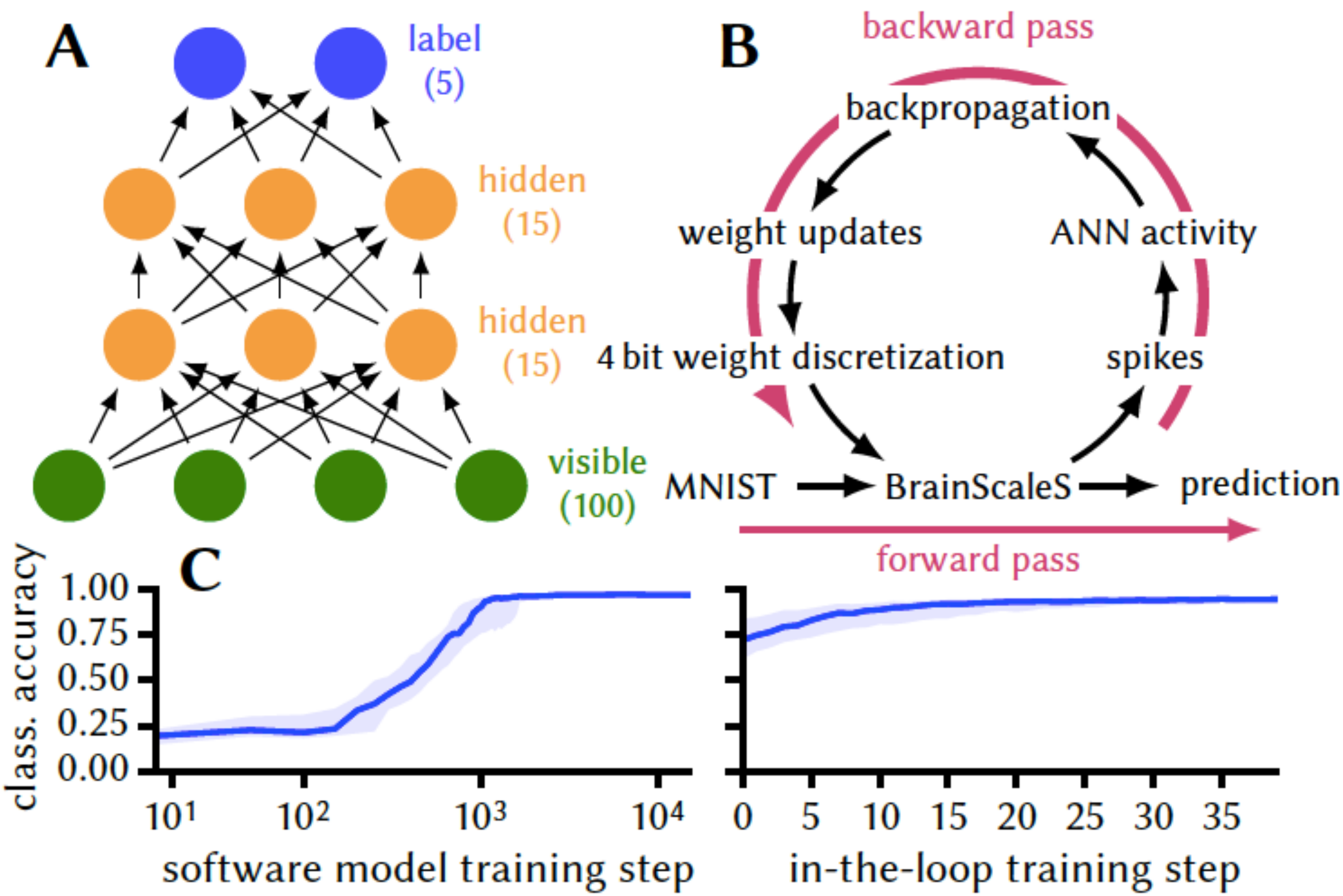
Sebastian Schmitt, Paul Müller

After hardware in-the-loop calibration



Sebastian Schmitt et al., accepted IJCNN 2017

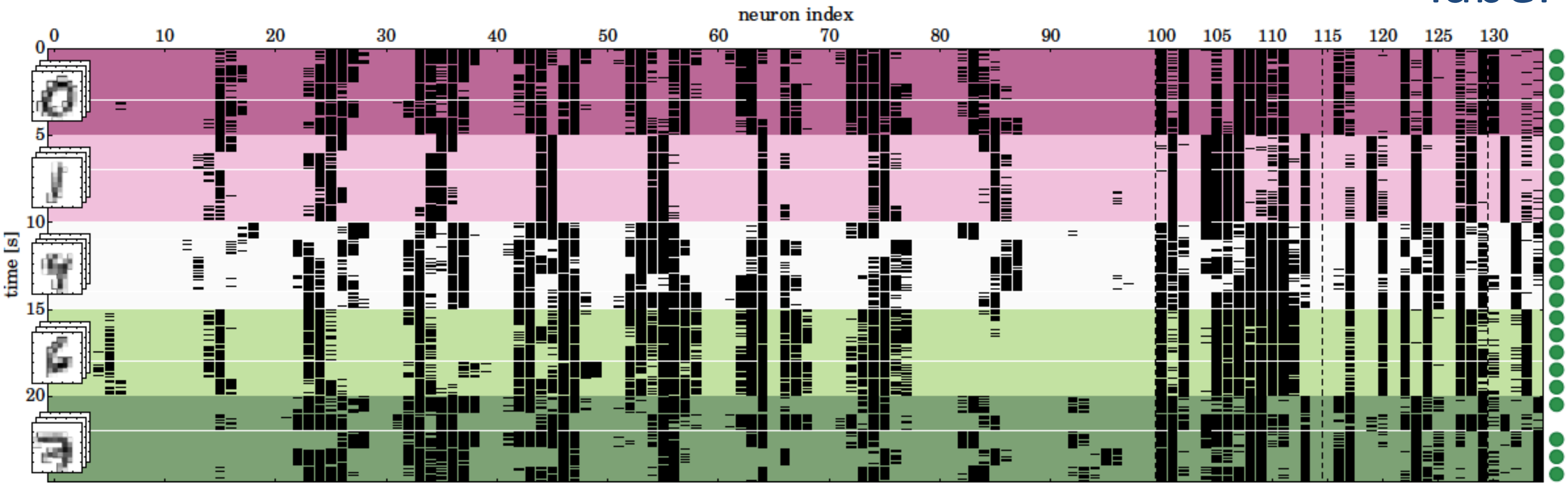
Feed-forward, rate-based. 4-layer spiking network MNIST classification on a physical model machine performance before and after hardware in-the-loop learning



Sebastian Schmitt et al., accepted IJCNN 2017, ISCAS 2017

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

label



input

2 x hidden

Sebastian Schmitt et al., accepted JCNN 2017, ISCAS 2017

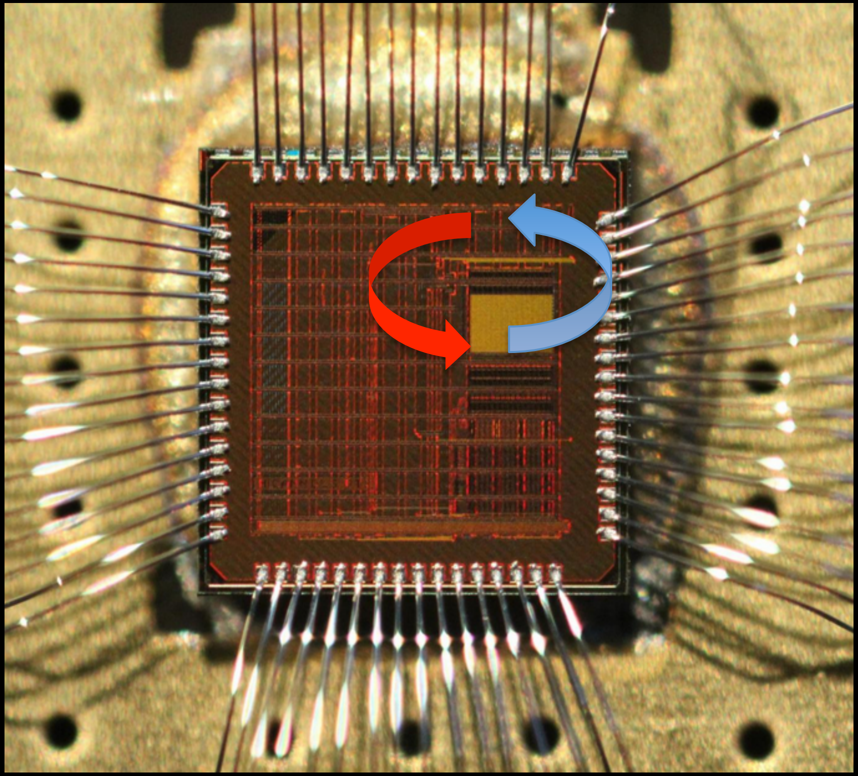
TimeScales	Nature + Real-time	Simulation	Accelerated Model
Causality Detection	10^{-4} s	0.1 s	10^{-8} s
Synaptic Plasticity	1 s	1000 s	10^{-4} s
Learning	Day	1000 Days	10 s
Development	Year	1000 Years	3000 s
12 Orders of Magnitude			
Evolution	> Millenia	> 1000 Millenia	> Months
> 15 Orders of Magnitude			

BrainScaleS-2

62 nm prototype chip in the lab

New key features

- Improved parameter storage
- Hybrid plasticity by on-chip processor : on-chip loops
  - Input : timing correlations, rates, membrane potentials, external signals
  - Change : synaptic weights, network topology, neuron parameters
- Structured neurons
  - NMDA plateau potentials create non-linear dendrites
  - Calcium spikes for coincidence detection between basal and distal inputs
  - Na spikes (action potentials) communicate with other neurons



- Evaluation system by mid-2018
- Full-size prototypes and wafer masks by mid-2020

Final Thoughts

- After 10 years of development the BrainScaleS large scale physical hardware system is being commissioned and delivers first results

- Fully non-Turing, physical model computing can solve 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 cognitive machine

Eric Müller

DEMO : Neuromorphic Hardware In-The-Loop: Training a Deep Spiking Network on the BrainScaleS Wafer-Scale System

Johannes Schemmel

Training and Plasticity Concepts of the BrainScaleS Neuromorphic Systems