VLSI Implementations of Very Large Scale Neuromorphic Circuits

Johannes Schemmel

Electronic Vision(s) Group
Kirchhoff Institute for Physics
Ruprecht-Karls-Universität, Germany

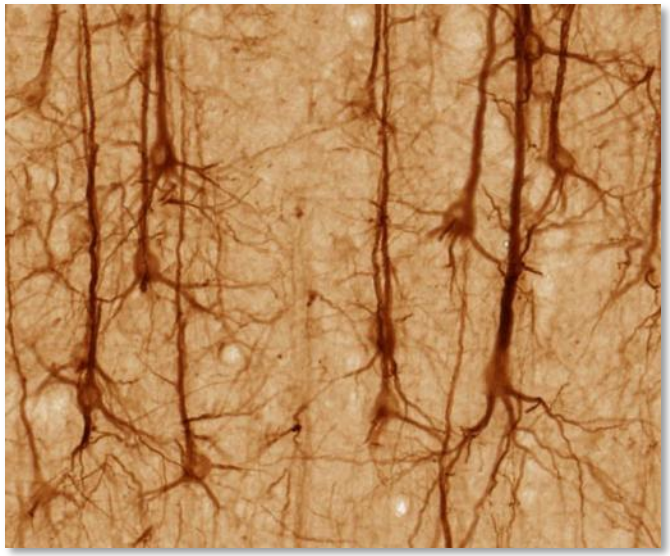
The Heidelberg FACETS team (in no particular order):

Karlheinz Meier, Sebastian Millner, Dan Husmann, Andreas Grübl, Johannes Schemmel, Maurice Güttler, Holger Zoglauer, Matthias Hock, Simon Friedmann, Stefan Philipp, Daniel Brüderle, Mihai Petrovici, Moritz Schilling, Johannes Bill, Bernhard Kaplan, Erik Müller, Andreas Hartel

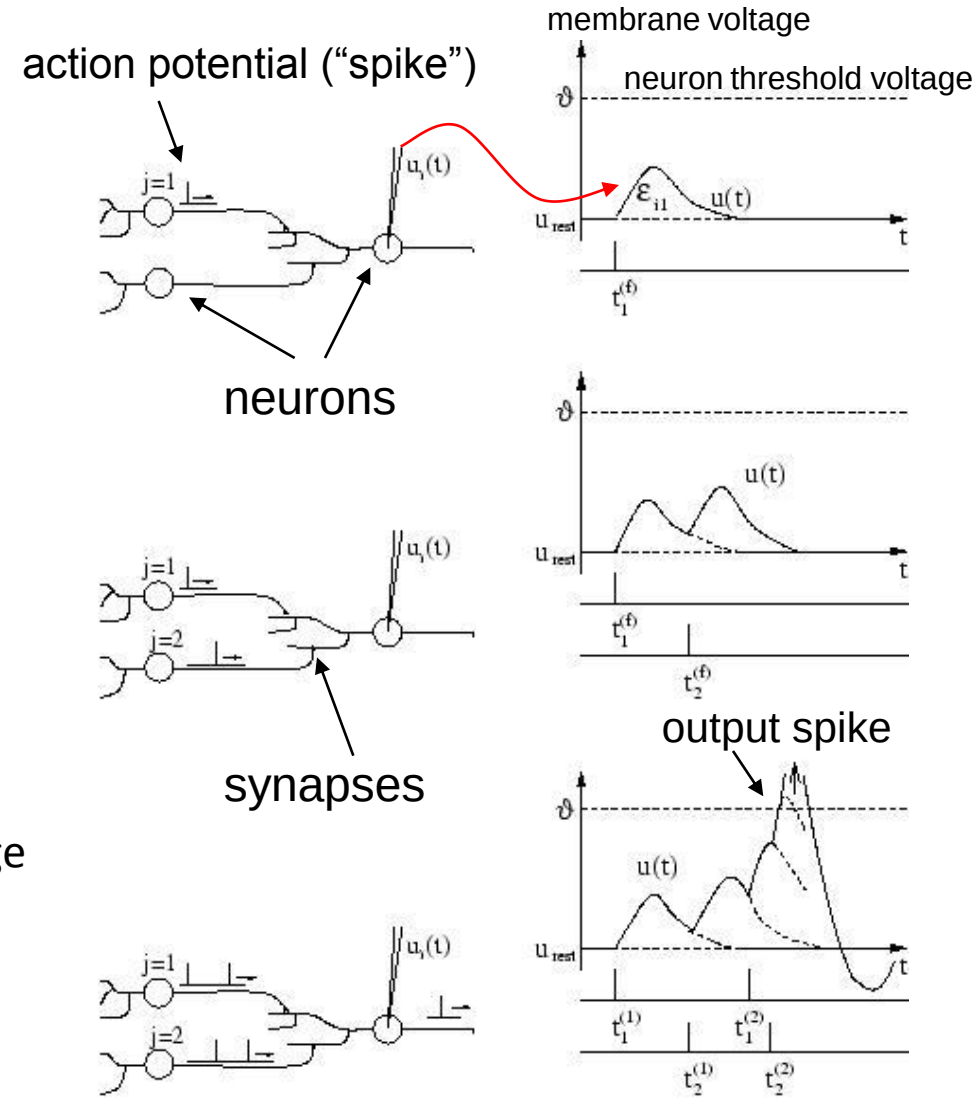
Why Build Artificial Neural Systems?

- we want to understand the biological computing paradigm
  - limited experimental access
  - artificial neural systems act as a model for biology
- new applications → robotics, ambient intelligence, user interfaces, etc

Basics of Neural Communication



© brainmaps.org



- 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

Modeling approaches for Artificial Neural Systems

starting point: mathematical description

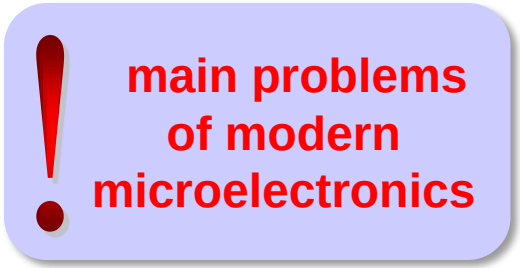
methods:

- **analytical treatment**
  proof of general properties and limits
- **numerical solution (high performance computing)**
  flexibility, parallel objects not obvious
- **physical model (neuromorphic hardware)**
  artificial nervous system
  artificial parallel object = biological objects
- ***biological model***
  “custom-made biological nervous system”

Limits of Numerical Models

computers use too much resources

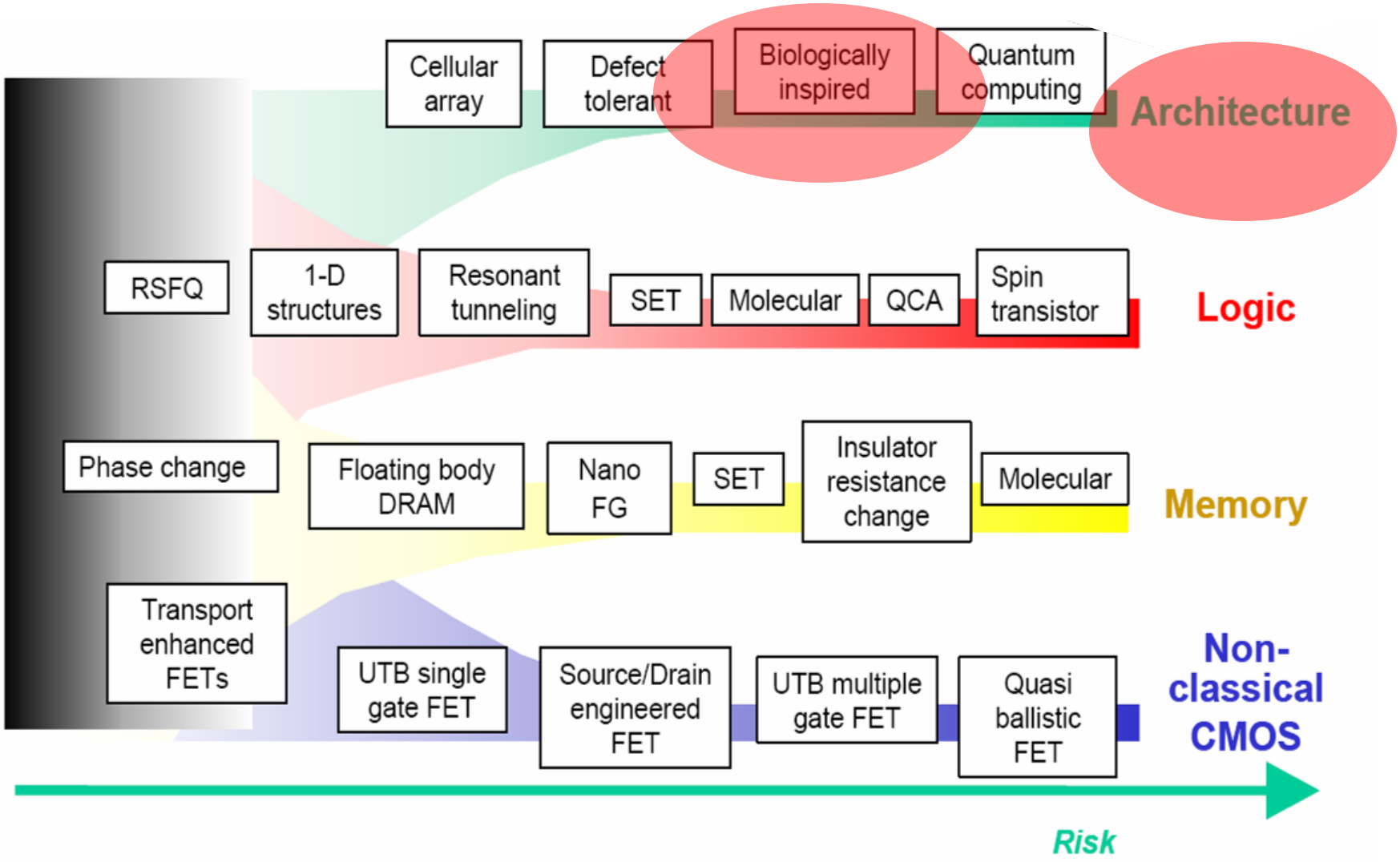
- loss of fault tolerance inherent to neural systems
- power consumption of the simulation layer



biologically inspired architectures preserve the fault tolerance and low power consumption of neural systems at the device level

→ physical model

International Technology Roadmap for Semiconductors (ITRS)



The FACETS Project

FACETS
Fast Analog Computing with Emergent Transient States

FACETS - From Neurobiology to new Computing Architectures

U Bordeaux, CNRS (Gif-sur-Yvette and Marseille), U Debrecen, TU Dresden, U Freiburg, SCCH Hagenberg, TU Graz, U Heidelberg, EPFL Lausanne, U London, U Plymouth, INRIA Sophia-Antipolis, KTH Stockholm

An Integrated Project in the 6th Framework Programme
Information Society Technology - Future Emergent Technologies
FP6-2004-IST-FETPI

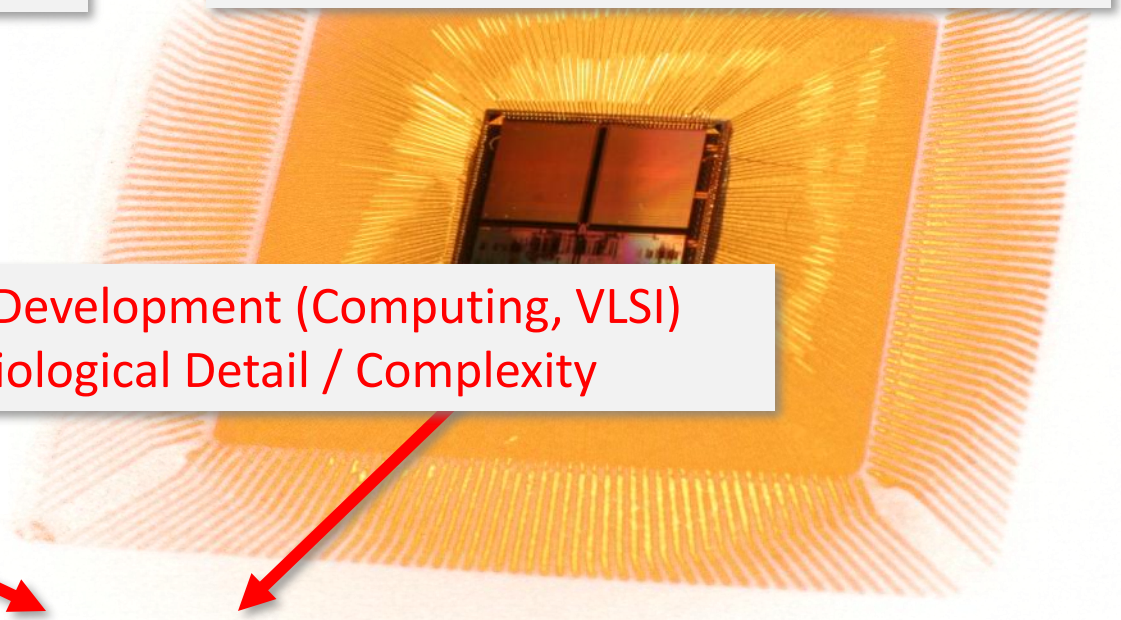
FACETS : Basic Idea, methodological approach and goals

Neurobiology : Structural and Functional Investigation of the Neocortical Microcircuit and the Circuit Elements in-vivo and in-vitro

Modelling : Virtual Microcircuits on State-of-the-Art Computers

Hardware : Physical models in analog and mixed-signal VLSI systems

```
area      = 20000.  # (µm²)
tau_m     = 20.     # (ms)
cm        = 1.      # (µF/cm²)
g_leak    = 5e-5    # (S/cm²)
if benchmark == "COBA":
    E_leak    = -60.   # (mV)
elif benchmark == "CUBA":
    E_
v_thre
v_rese
t_refr
v_mean
tau_exc   = 5.      # (ms)
tau_inh   = 10.     # (ms)
```



Methodology : Tool Development (Computing, VLSI)
Reduction of Biological Detail / Complexity

Common Goal : Study non-classical universal computing solutions

Benchmarking (biology vs. Modelling vs. Hardware with visual tasks in VI)

Essentials for Neuromorphic Hardware Systems

Biological Neural Computation

- Connectivity — 10^{11} Neurons, 10^{15} Synapses in Neocortex
  10.000 Synapses per Neuron on average
- Diversity — Categories and Parameters of Neurons
- Plasticity — Long Term, Short Term, Local, Global
- Timing — Time constants, delays, correlations

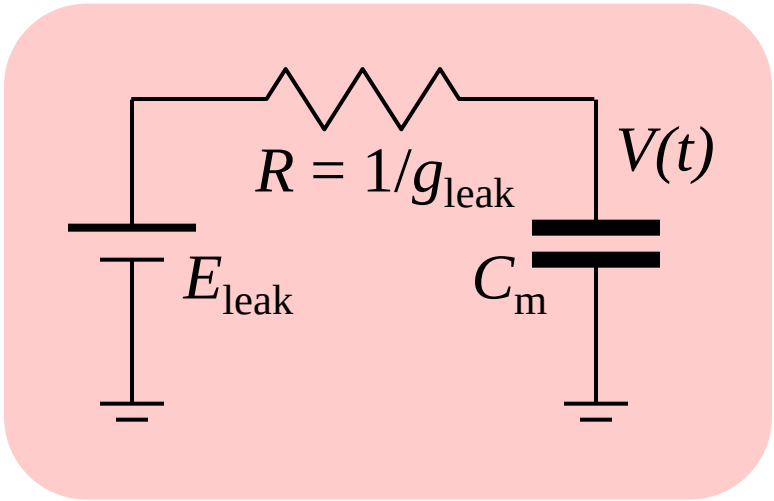
Neuromorphic Hardware Systems

- Connectivity — Efficient data protocols, 2D-3D connection technology
- Diversity — Configurability (distributed memory)
- Plasticity — Local and global dynamic and static memory
- Timing — Control time constants, delays and time correlations
- SCALABILITY — Learn from small systems – Approach large scales
  Bandwidth, delays, power, cost, fault tolerance

Continuous Time Integrating Membrane Model

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

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



	ΔV [V]	g_{leak} [S]	C_{m} [F]	(gV)/C [V/s]
Biology(*)	10^{-2}	10^{-8}	10^{-10}	10^{0}
VLSI	10^{-1}	10^{-6}	10^{-13}	10^{6}

Inherent speed gap:

10^6 Volt/second

→ accelerated neuron model

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

Properties of the Facets Neuron Model

Accelerated continuous time:

- acceleration factor 10^3 to 10^5
- high speed operation for statistics and parameter searches
- minimum area leads to high synapse numbers
- No deep sub-threshold operation - in principle all analog parameters could be well controlled
  → automatic transfer of experiments (PyNN)

FACETS Stage 1 : Conductance-based Network Model

current source, no voltage dependence

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

membrane current

leakage current

sum over excitatory synapse currents k

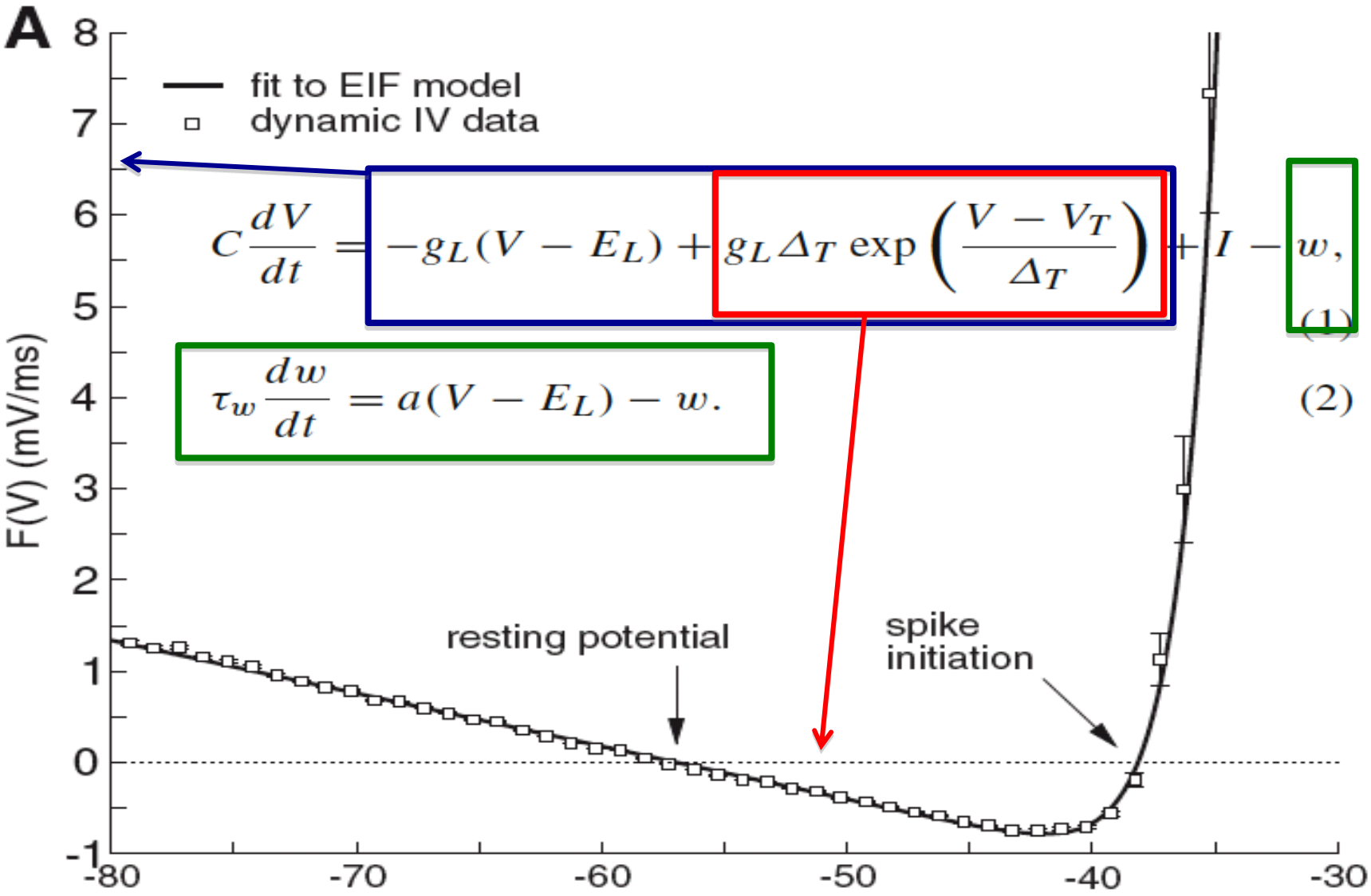
sum over inhibitory synapse currents l

Voltage dependent part, changes membrane conductance

Synapses:

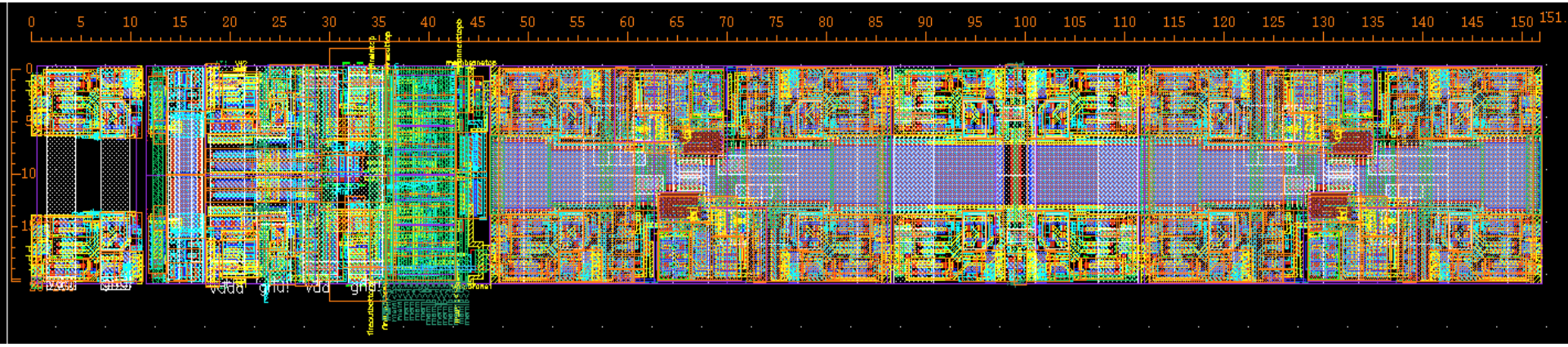
- $p_{k,l}(t)$ exponential onset and decay (spike shape)
  $g_{k,l}$ 0 to g_{max} with 4 bit (8 bit) resolution
- effective membrane time-constant $c_{\mathrm{m}}/g_{\mathrm{total}}$ is time-dependent

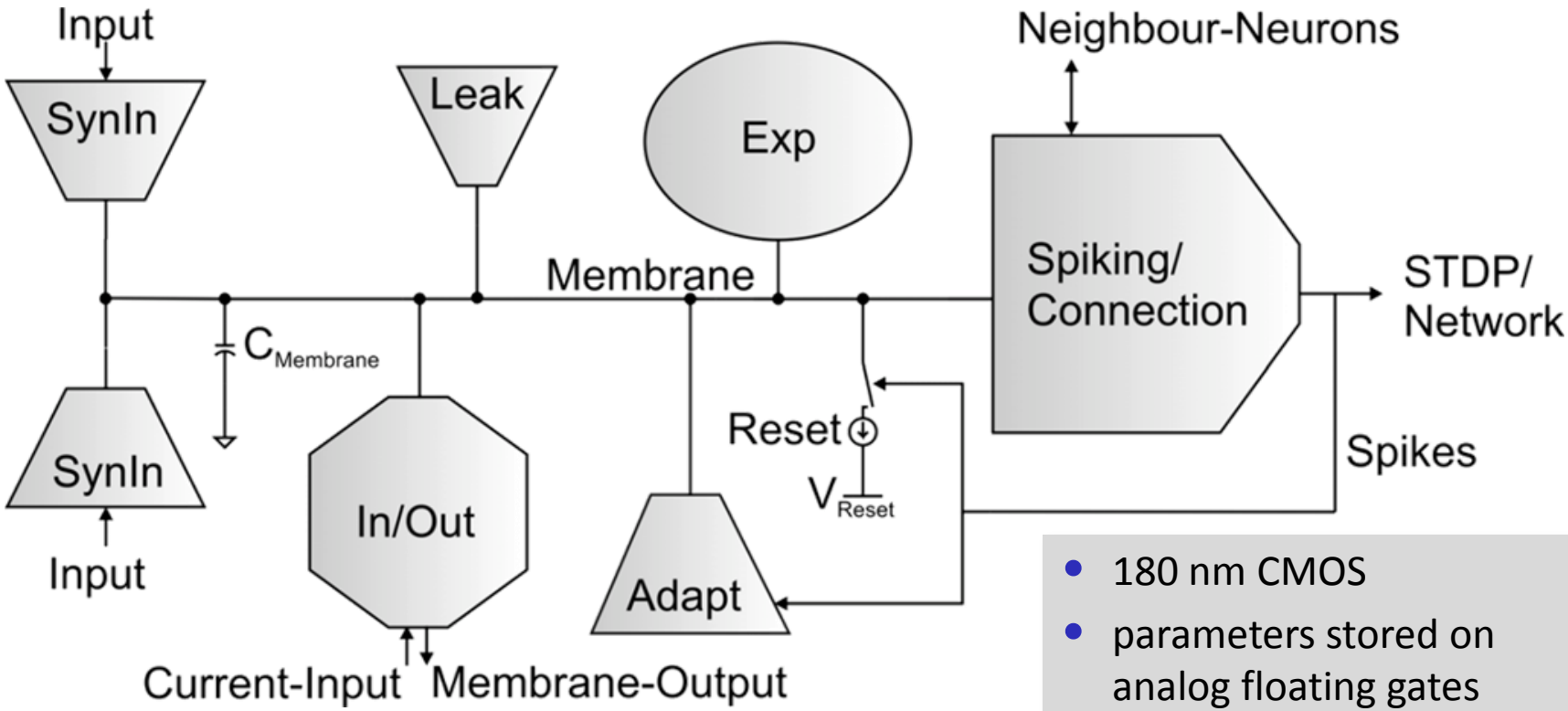
Stage 2: FACETS Adaptive-Exponential Integrate-and-Fire Neuron Model



R. Naud et al.:*Versatility and Relevance of the adaptive-exponential integrate-and fire-model*, Biol Cybern (2008) 99:335–347

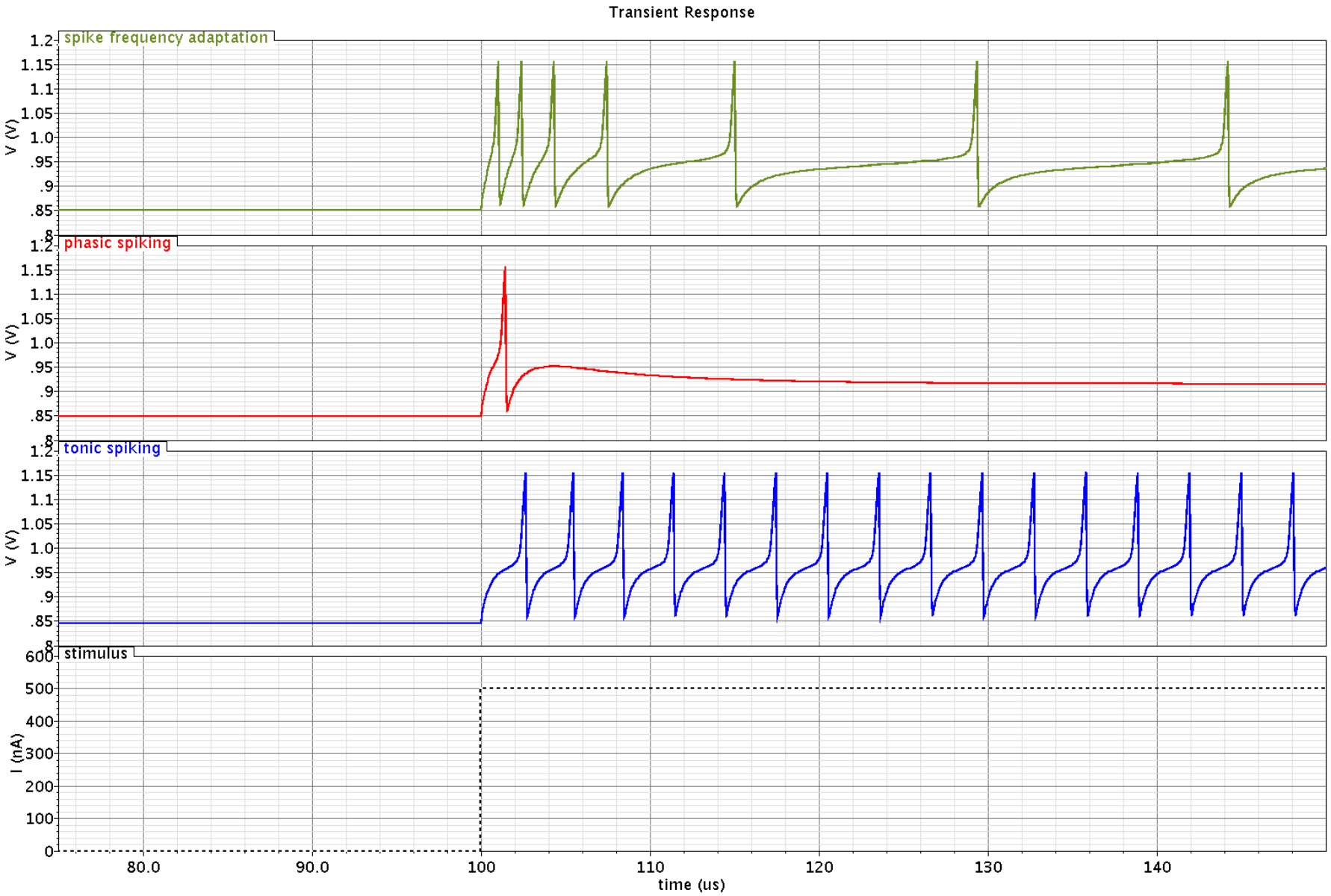
CMOS VLSI Implementation of the AdExp Integrate-and-Fire Neuron



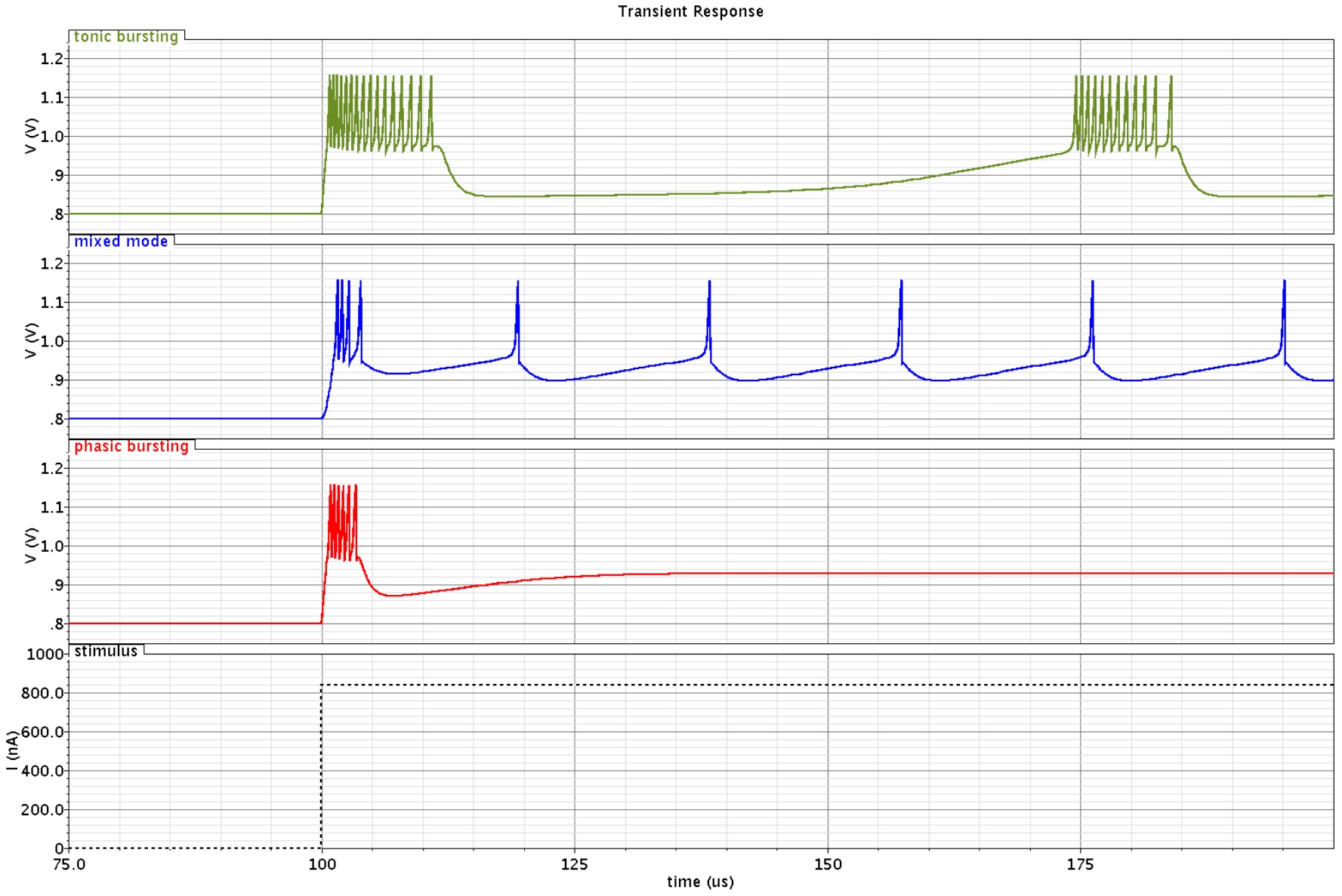


- 180 nm CMOS
- parameters stored on analog floating gates

Spike Firing Modes of the AdExp VLSI Neuron

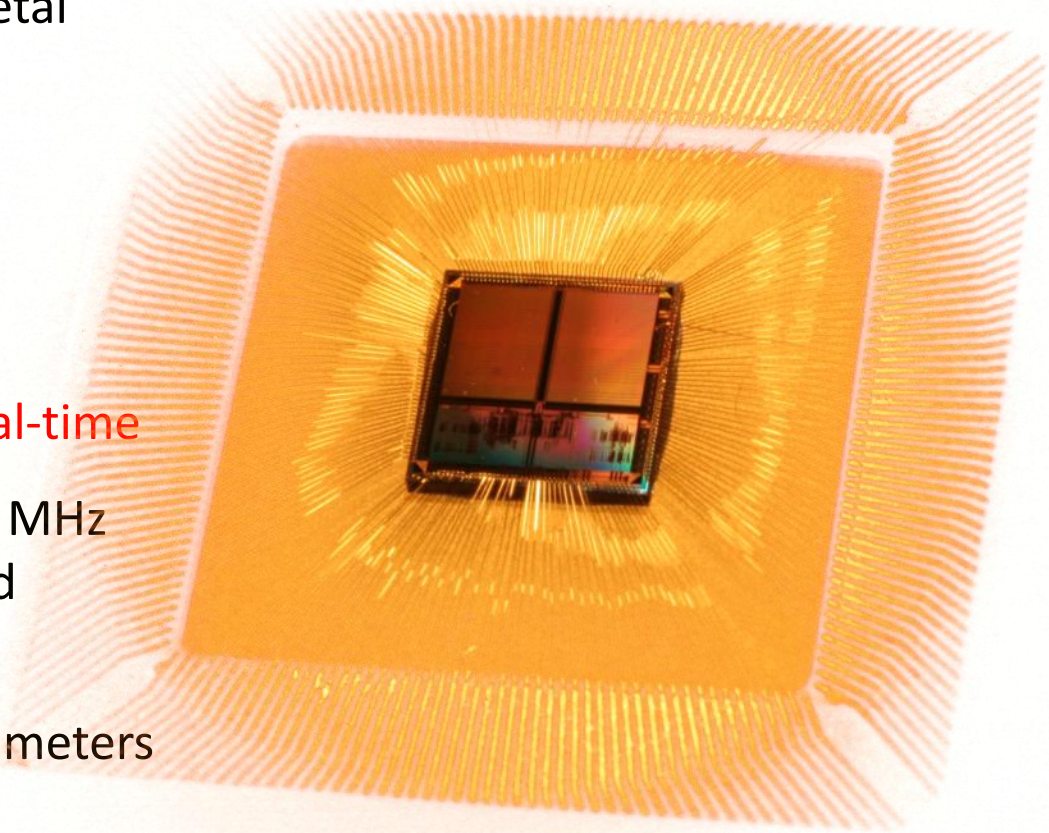


Burst Firing Modes of the AdExp VLSI Neuron



Stage 1 : Chip Specifications

- technology: UMC 180 nm, 6 metal layers, 1 polysilicon layer
- chip size: 5 x 5 mm^2
- 384 neurons, 100k synapses
- scale factor up to 100k :
  10 ns chip-time equals 1 ms real-time
- fast analog outputs (about 400 MHz bandwidth) to monitor selected membrane potentials
- internal storage for model parameters (about 4k values)



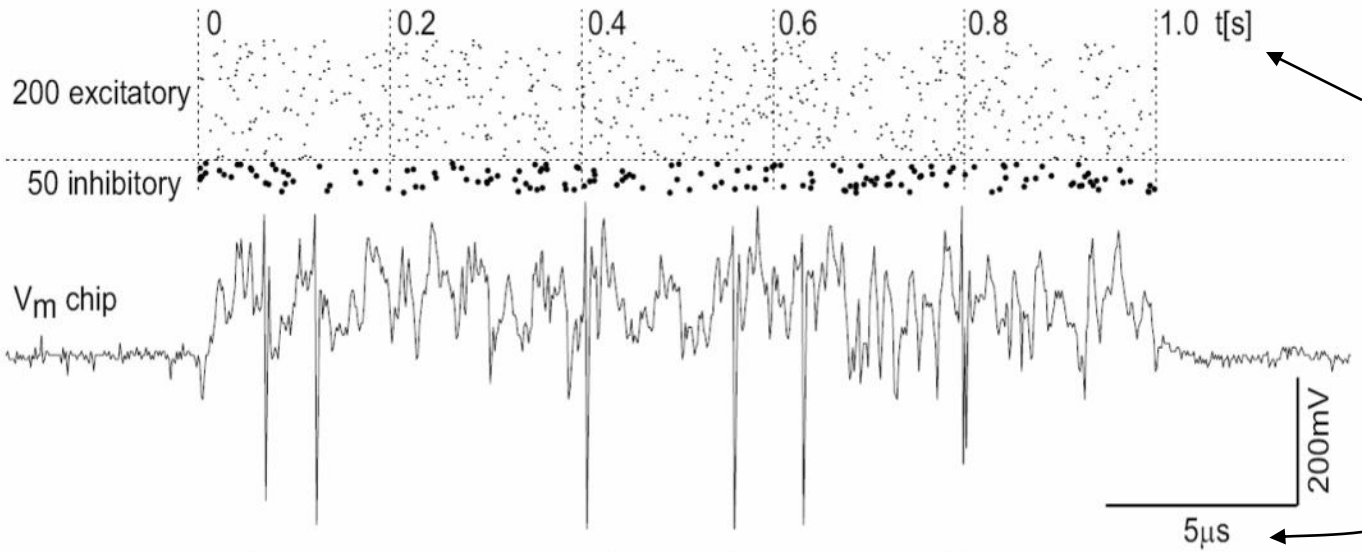
A New VLSI Model of Neural Microcircuits Including Spike Time Dependent Plasticity, Johannes Schemmel, Karlheinz Meier, Eilif Muller, Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04), IEEE Press, pp. 1711-1716, 2004

Stage 1 : Circuit Features

- fully analog network core
- continuous time network operation
- short-term synaptic depression and facilitation: analog on-chip
- Spike Time Dependent Plasticity measurement in each synapse, weight update performed digitally
- programmable model parameters (individually or group-wise):
  - reversal potentials: excitatory, inhibitory and leakage (E_x, E_i, E_l)
  - threshold voltage level V_{th} and comparator speed
  - reset potential (V_{reset}) and leakage conductance (g_{leak})
  - synapse parameters: rise time, fall time, maximum conductance (t_{rise}, t_{fall}, $g_{k,l\,max}$)

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

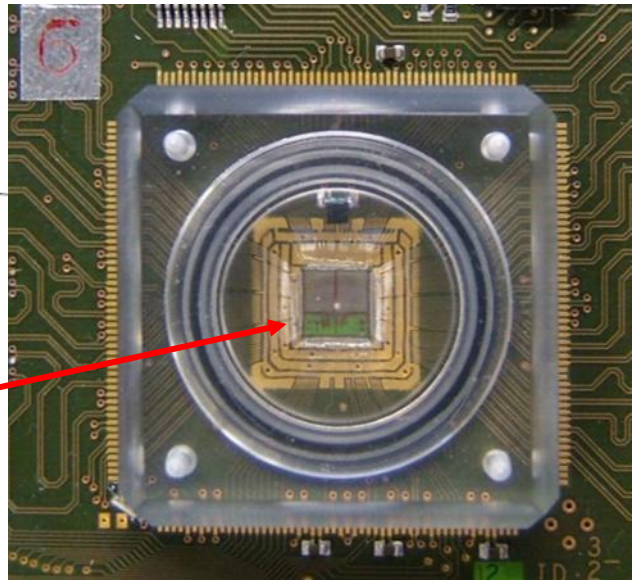
Response to Poisson-Distributed Input Spike Trains: Hardware vs. Simulation



biological time

chip time

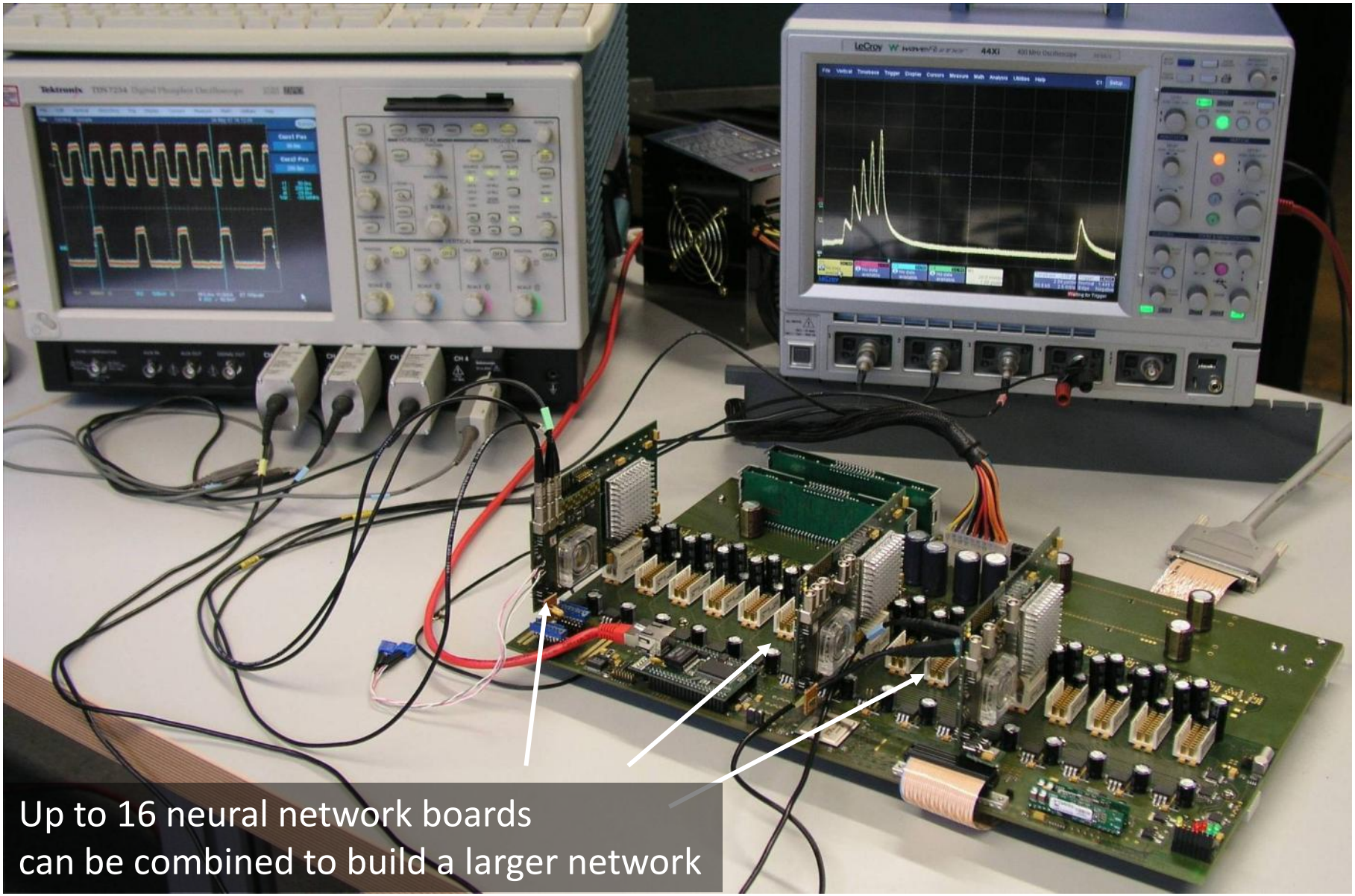
1cm



"Spikey" Chip

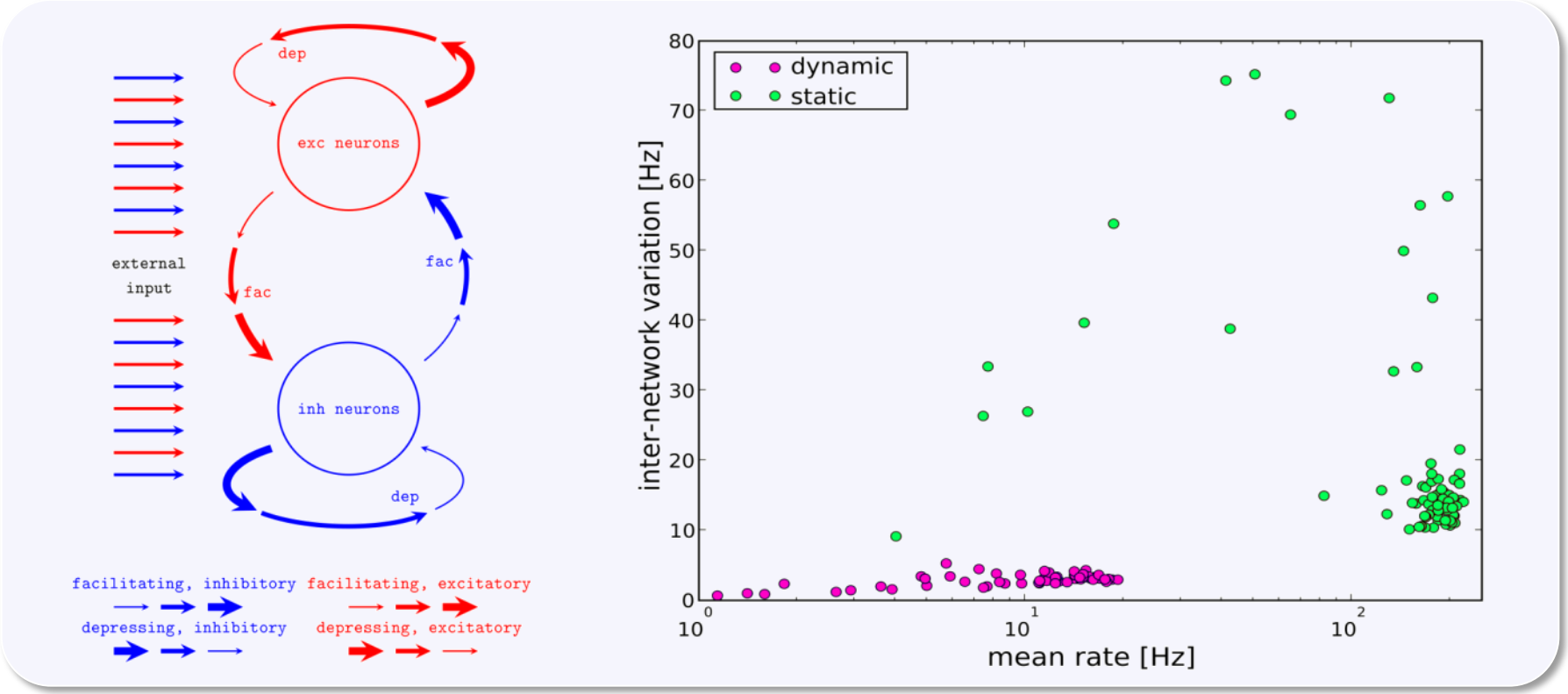
see Schemmel et. al., IJCNN 2006, ISCAS 2007

Stage 1 Setup

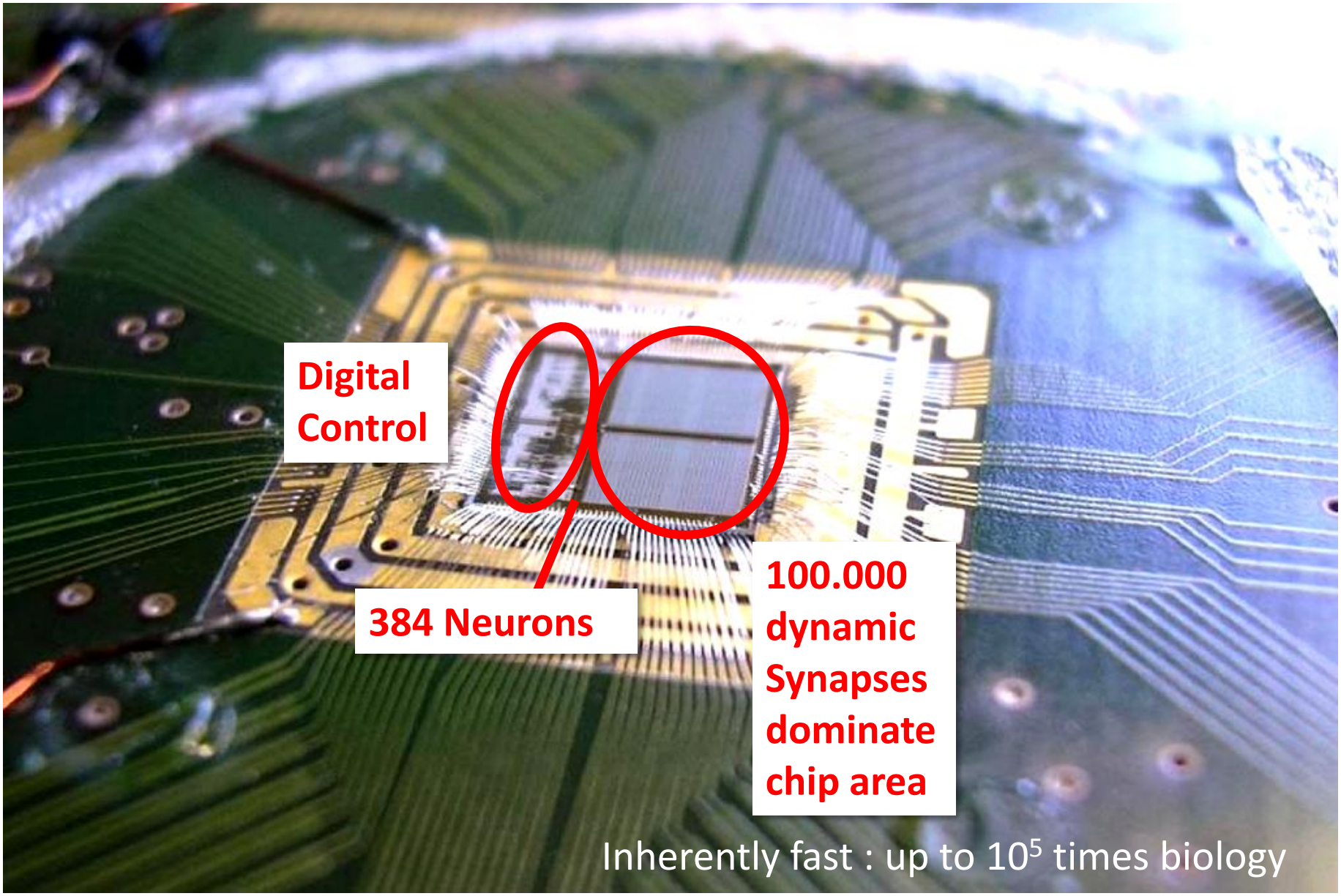


Example: Study of self-stabilizing network architectures

- Based on short-term synaptic plasticity, to counterbalance transistor level hardware variations
- Various parameter sets, multiple repetitions with new random connections each
- Once with static, once with dynamic synapses
- Result: Stable average firing rates, increased run-to-run reliability

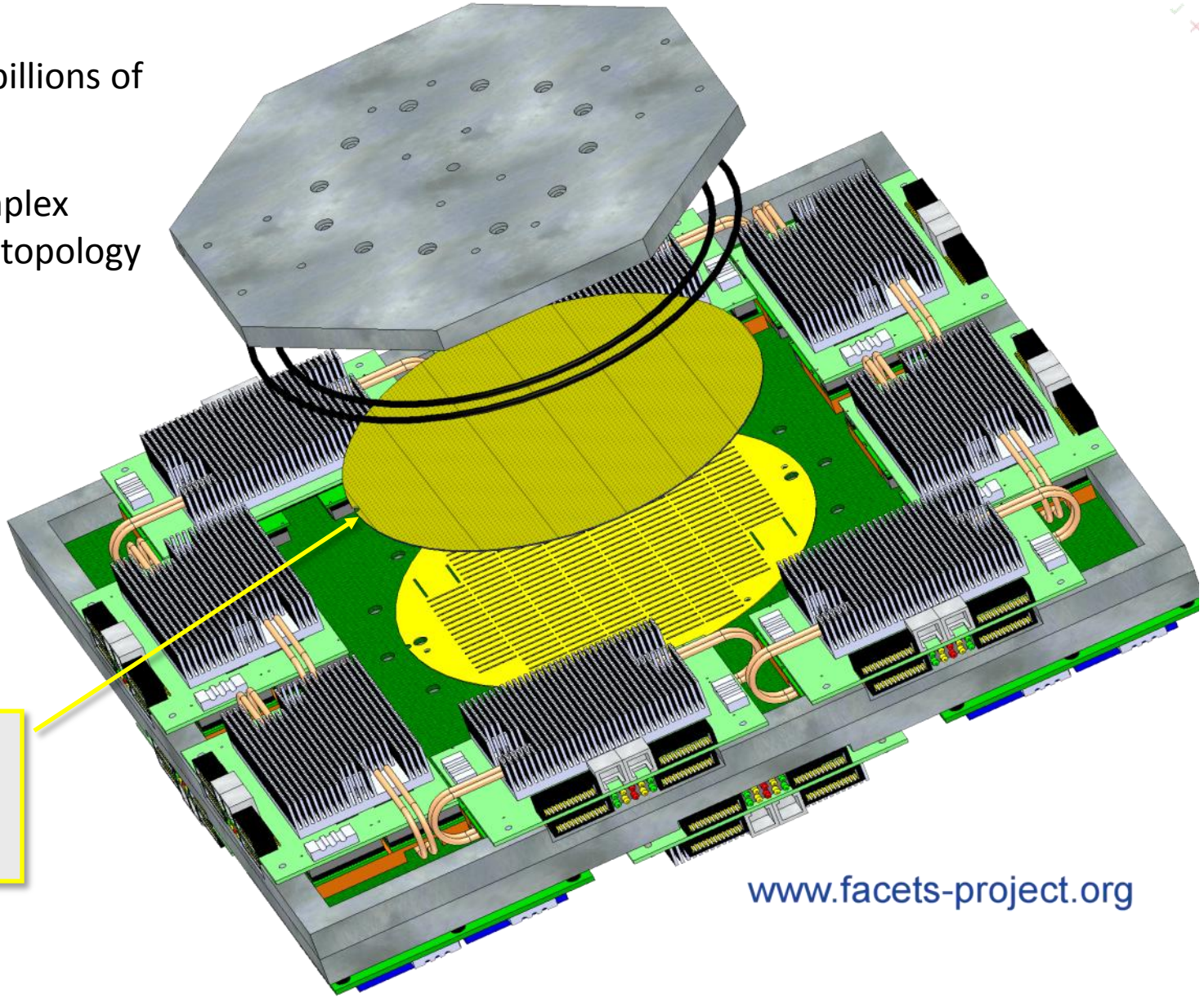


FACETS mixed-signal VLSI System Stage 1 (Chip based)



Solution: The FACETS Wafer-Scale System

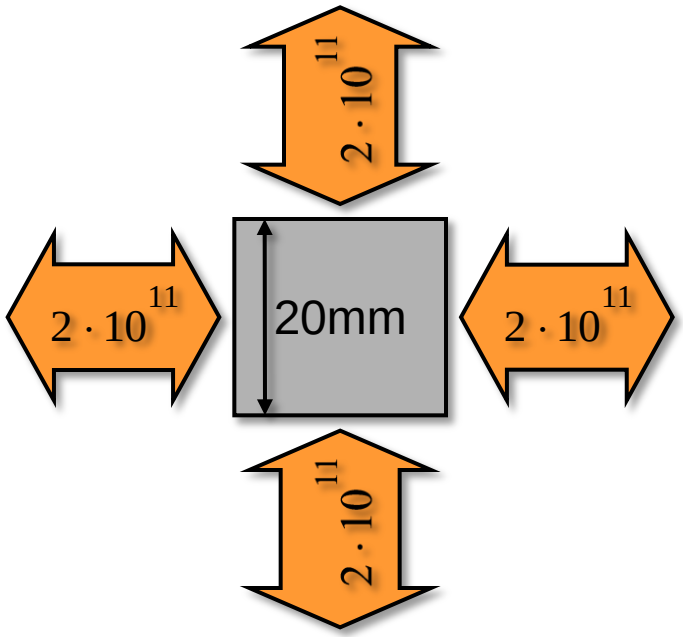
- Scalable up to billions of synapses
- but: needs complex programmable topology



20cm wafer with 40 million synapses

Connection Requirements

network die with maximum reticle size of 20*20 mm^2
→ 1 Msynapses



- biological connection density of 1000-10000 inputs per neuron
- only two dimensions for wiring (the brain uses three)
- acceleration factor of 10^4, biological event rate 20 Hz

→ resulting signal density: approx. 10^{10} events/(mm*s)

Routing Properties

spatial communication bandwidth 10^{10} events/(mm*s)

this seems possible, but

events must be routed from synapse A to synapse B

routing properties:

- constant propagation delay
- routing topology changes only slowly with time (long term plasticity)
- programmable topology for different experiments and neural circuits

solution: one connection per synapse
→ **connection-based routing**

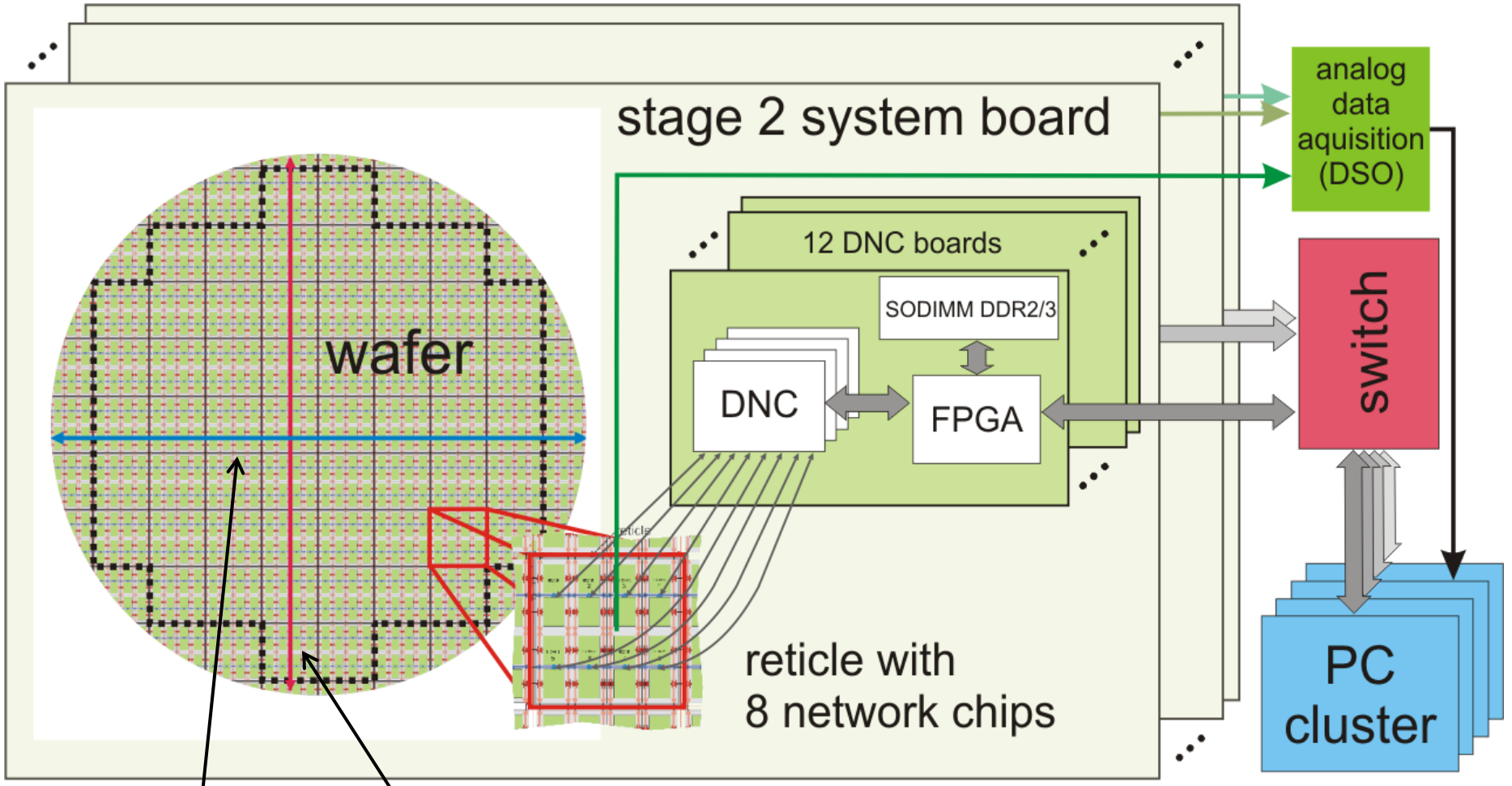
resulting connection density: 10^4 connections/mm

not possible with traditional bonding techniques

→ wafer scale integration allows a connection pitch close to 5µm

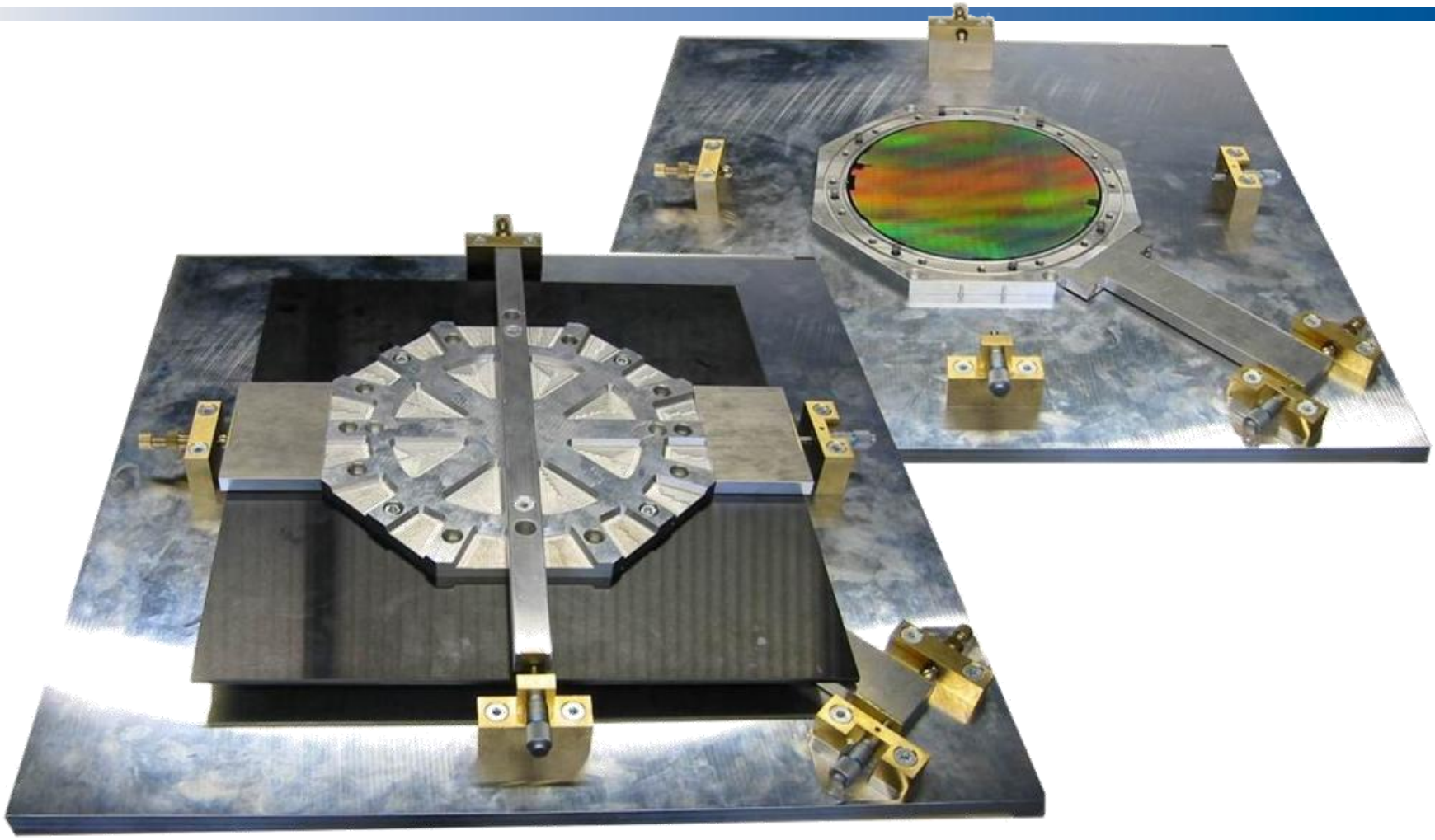
remaining gap: use time multiplexing

Overview of the Wafer-Scale System



horizontal and vertical connections cover the whole wafer across reticle boundaries

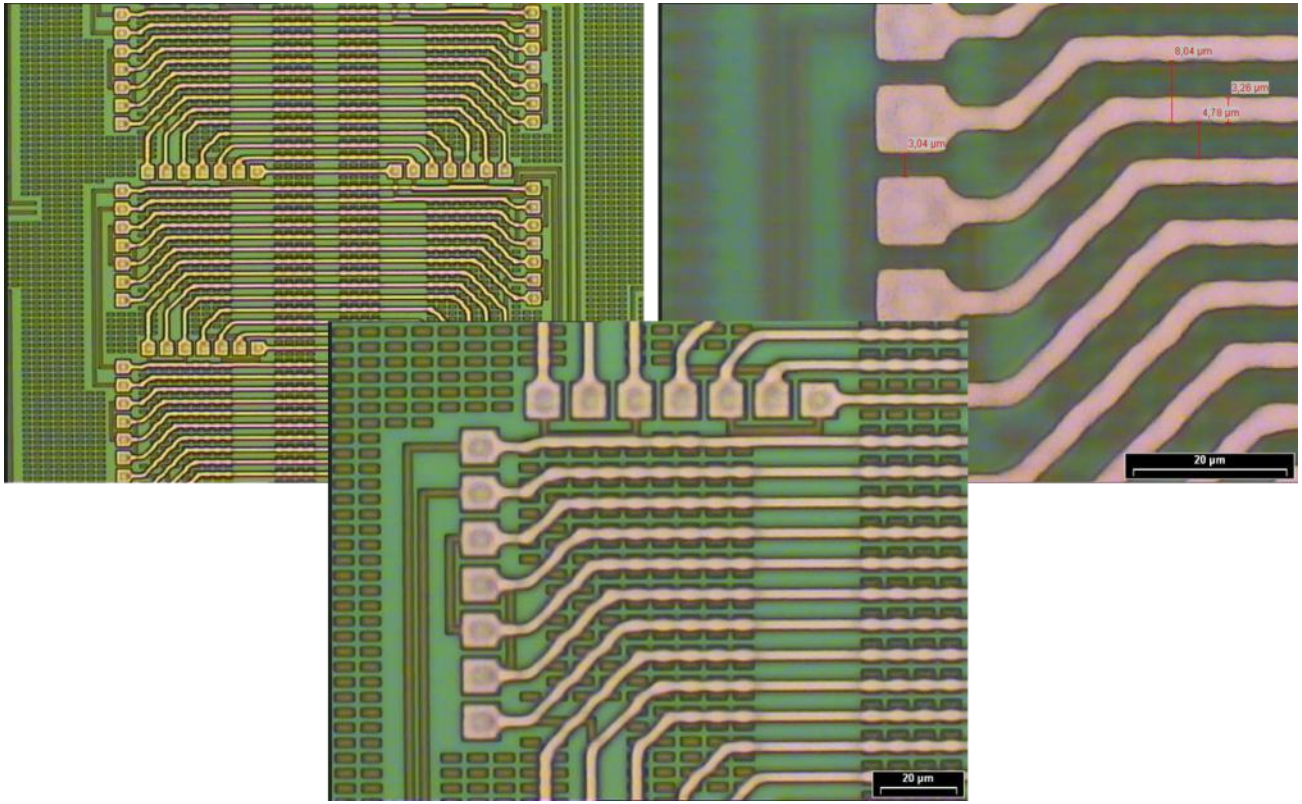
Wafer-PCB Alignment and Assembly Facility



- Large (approx.10.000) number of vertical elastomeric contacts demand precise alignment between wafer and PCB with the peripheral electronics / power delivery. Pitch 200 µm.
- Adjustment accuracy better than 50 µm over the 8 inch wafer

Post Processing Results

- The measured yield for 8 µm pitch test structures was 100%
- The pitch needed for the final system is 9.5 µm
- With larger pad separation even smaller structures could be realized



Microscopic views of an 8 µm pitch post processing structure

General Concept of Wafer-Scale Communication

problem: communication bottleneck between wafers

consider 10Hz biological firing rate at a an acceleration factor of 10^4

assume we implement about 10^5 synapses on a 5x10 mm^2 die

→ 10^{10} action-potentials/s

imagine an AER protocol transmitting the target synapse number (14 bit)

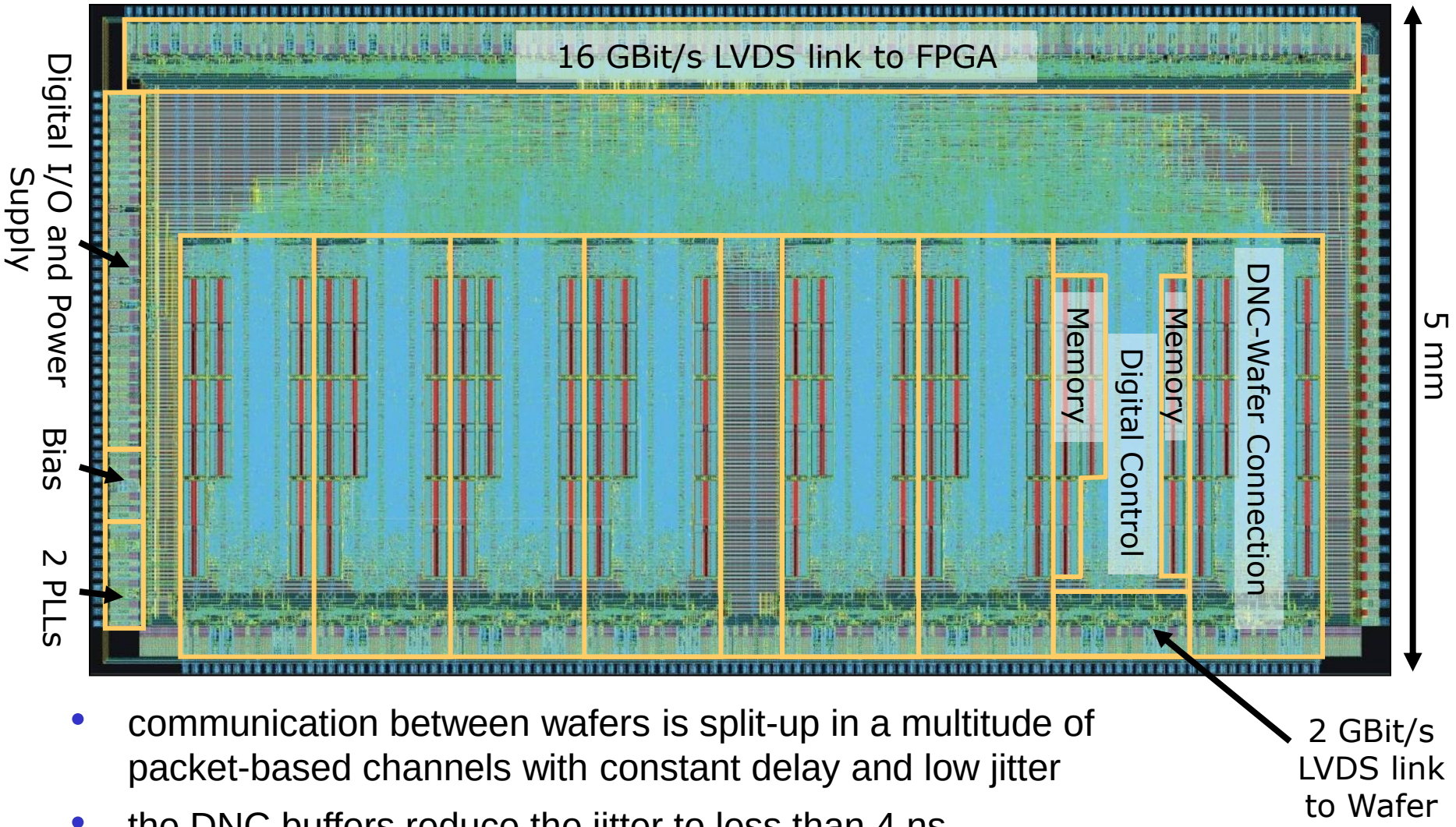
→ about 100 Gbit/s (very optimistic considering the collision probability)

idea: a neighboring chip needs about the same pre-synaptic signals

→ imagine a fast connection between two chips:
each chip would need only half the bandwidth to the rest of the world

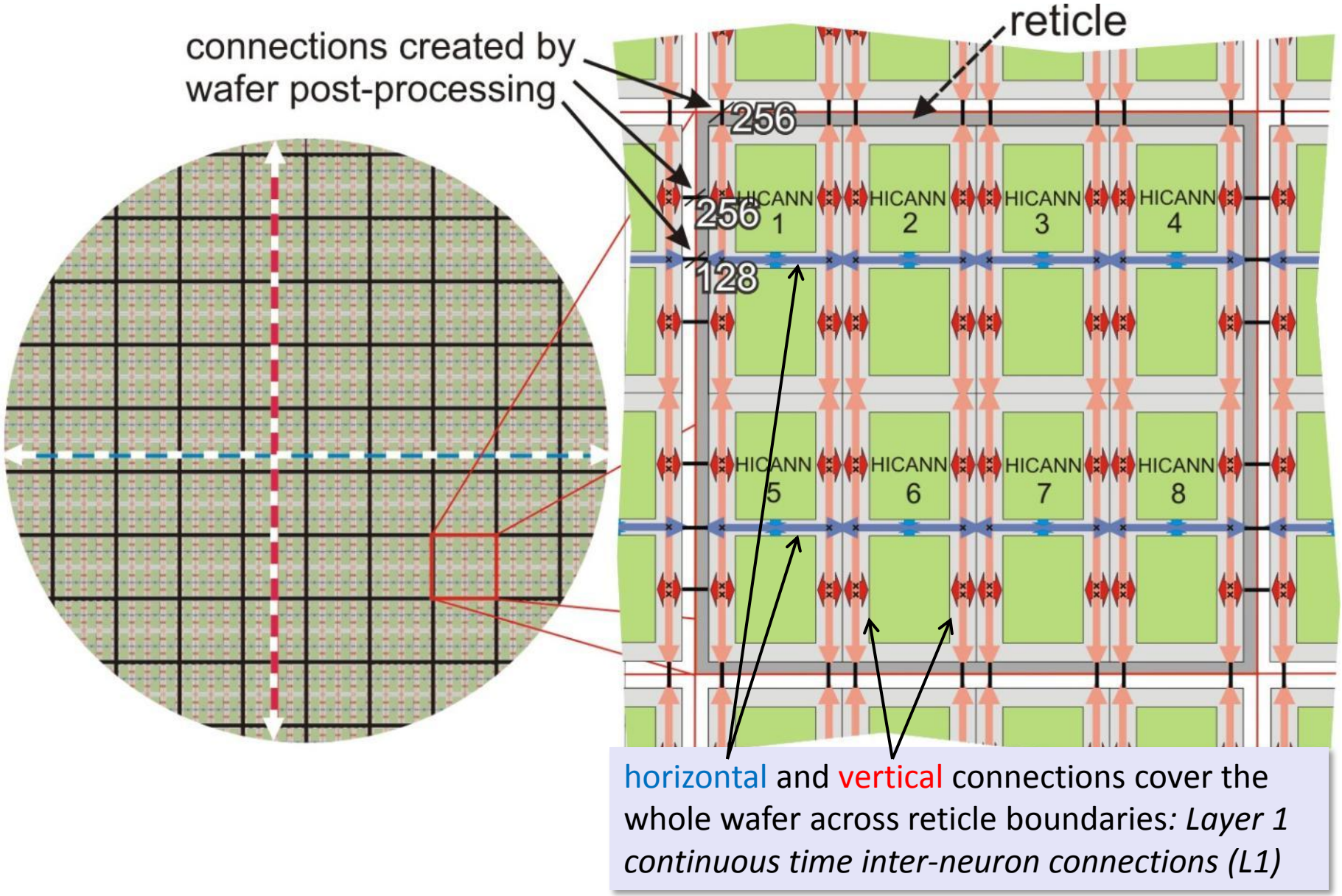
→ a wafer: 350 chips, external bandwidth reduced to GBits/s

Digital Network Chip (DNC, TU Dresden)



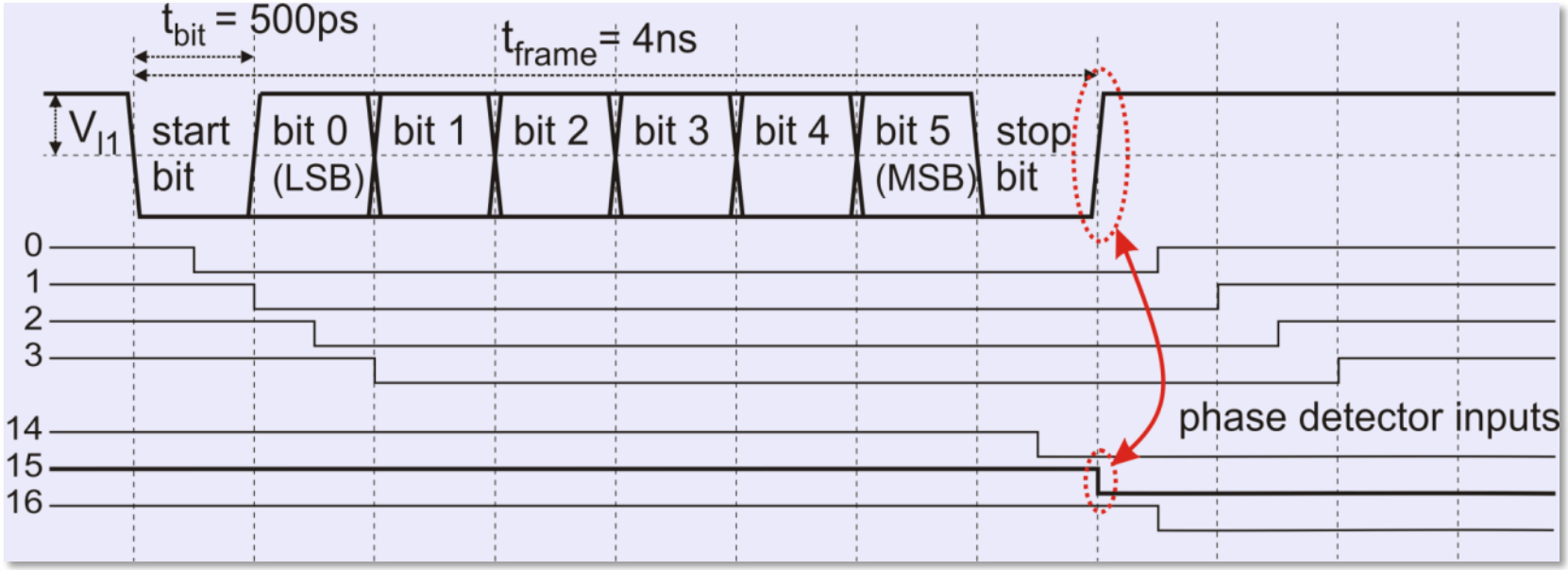
- communication between wafers is split-up in a multitude of packet-based channels with constant delay and low jitter
- the DNC buffers reduce the jitter to less than 4 ns
- the packets are routed in-between the wafers by the FPGAs

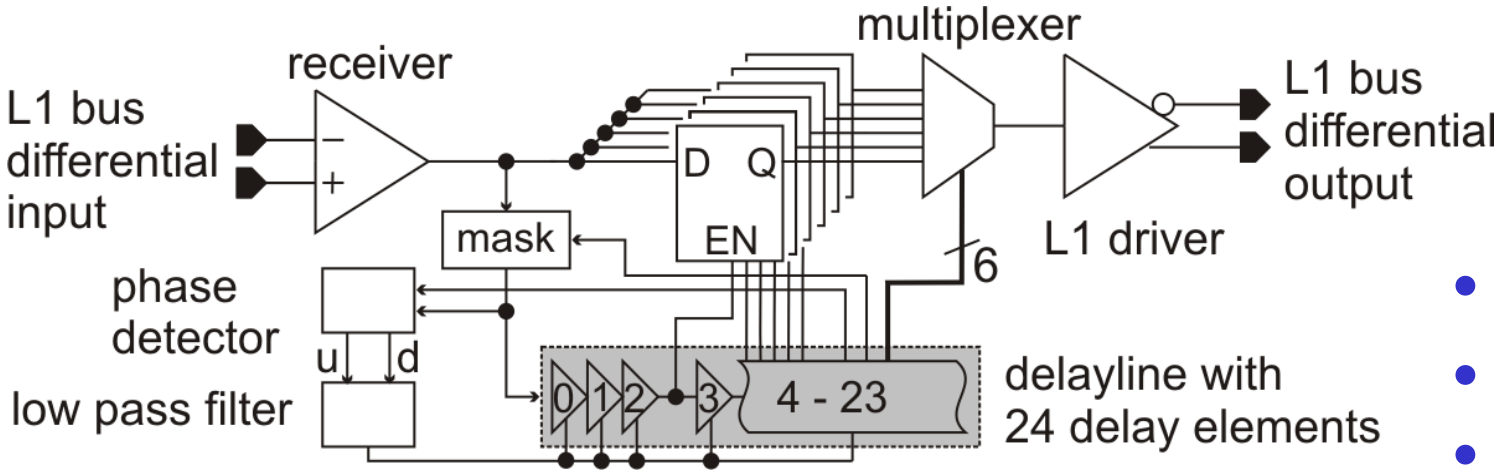
Structure of a FACETS Wafer



horizontal and vertical connections cover the whole wafer across reticle boundaries: *Layer 1 continuous time inter-neuron connections (L1)*

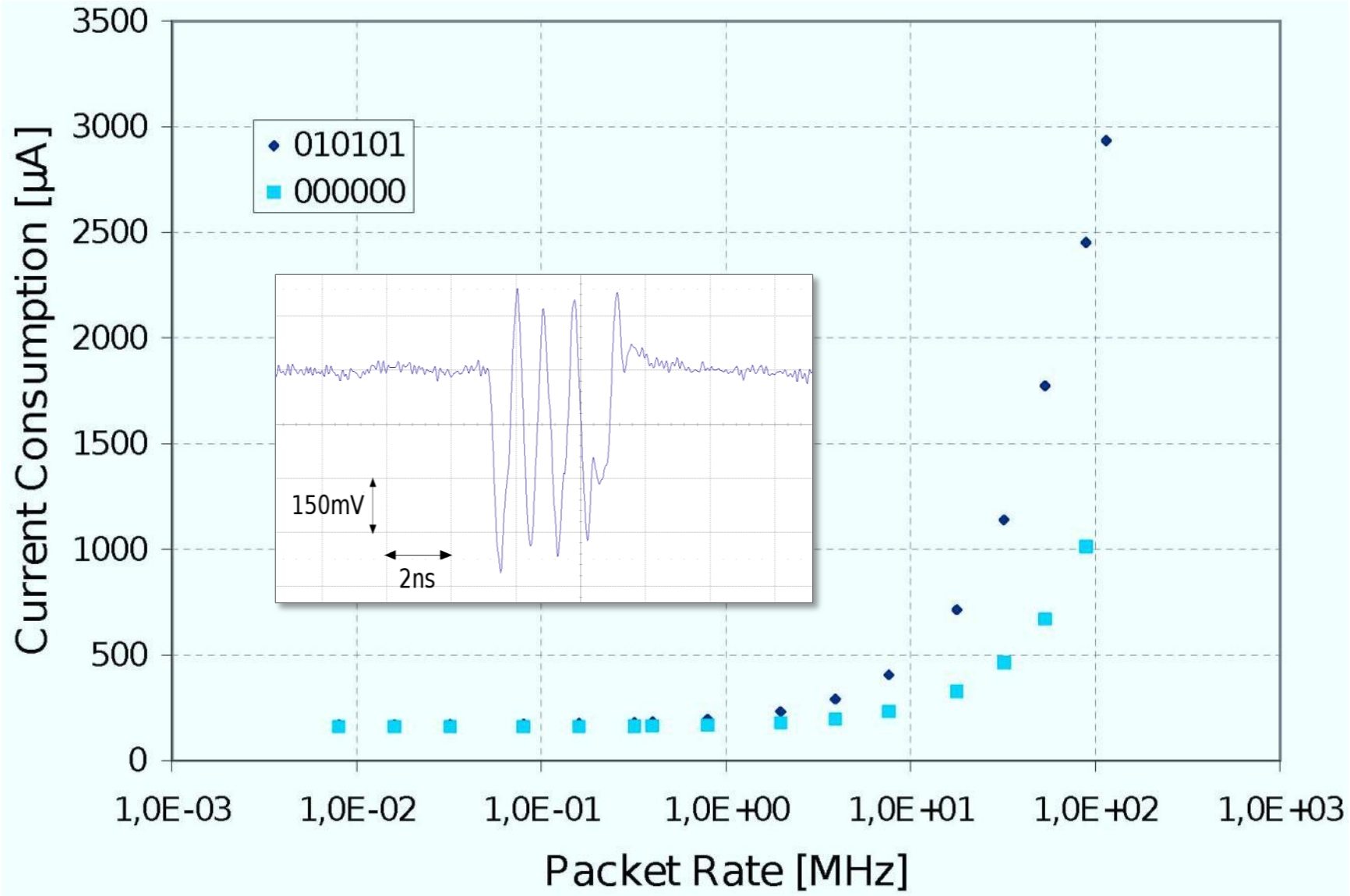
Low-Power Neuron to Neuron Communication



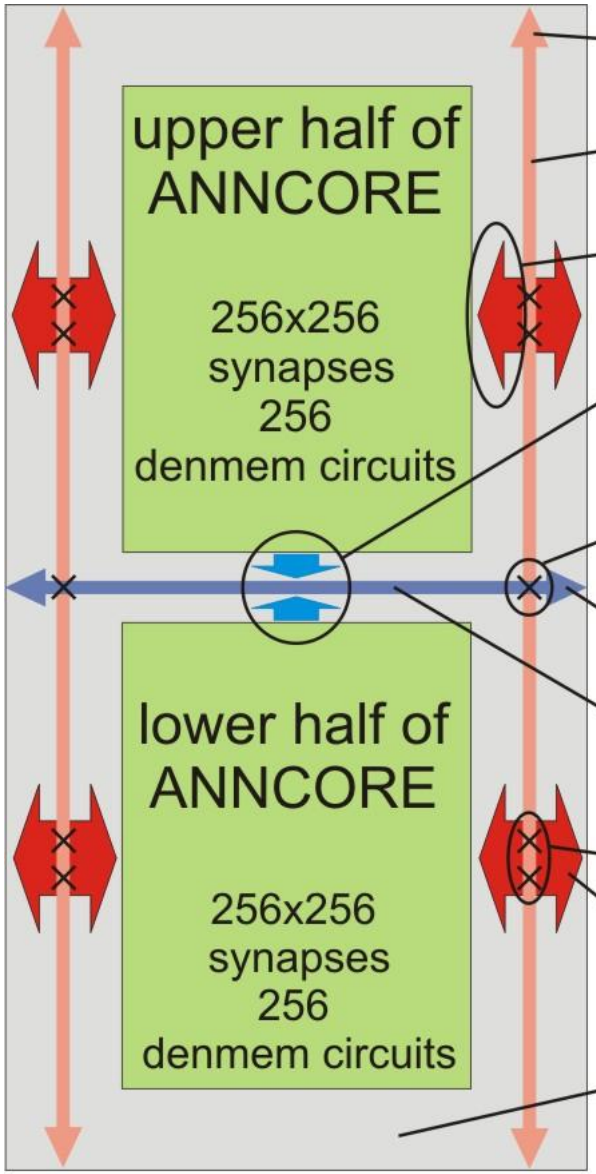


- 180 nm CMOS
- 166x38 µm^2
- up to 2 Gbit/s

Power Consumption of the Serial Repeater Circuits



High Input Count Analog Neural Network (HICANN)



64 repeaters with timing recovery

128 vertical L1 bus lanes

64 inputs to synapse drivers

outputs from 8 neuron-L1 interfaces

sparse crossbar between horizontal and vertical L1 busses

32 repeaters with timing recovery

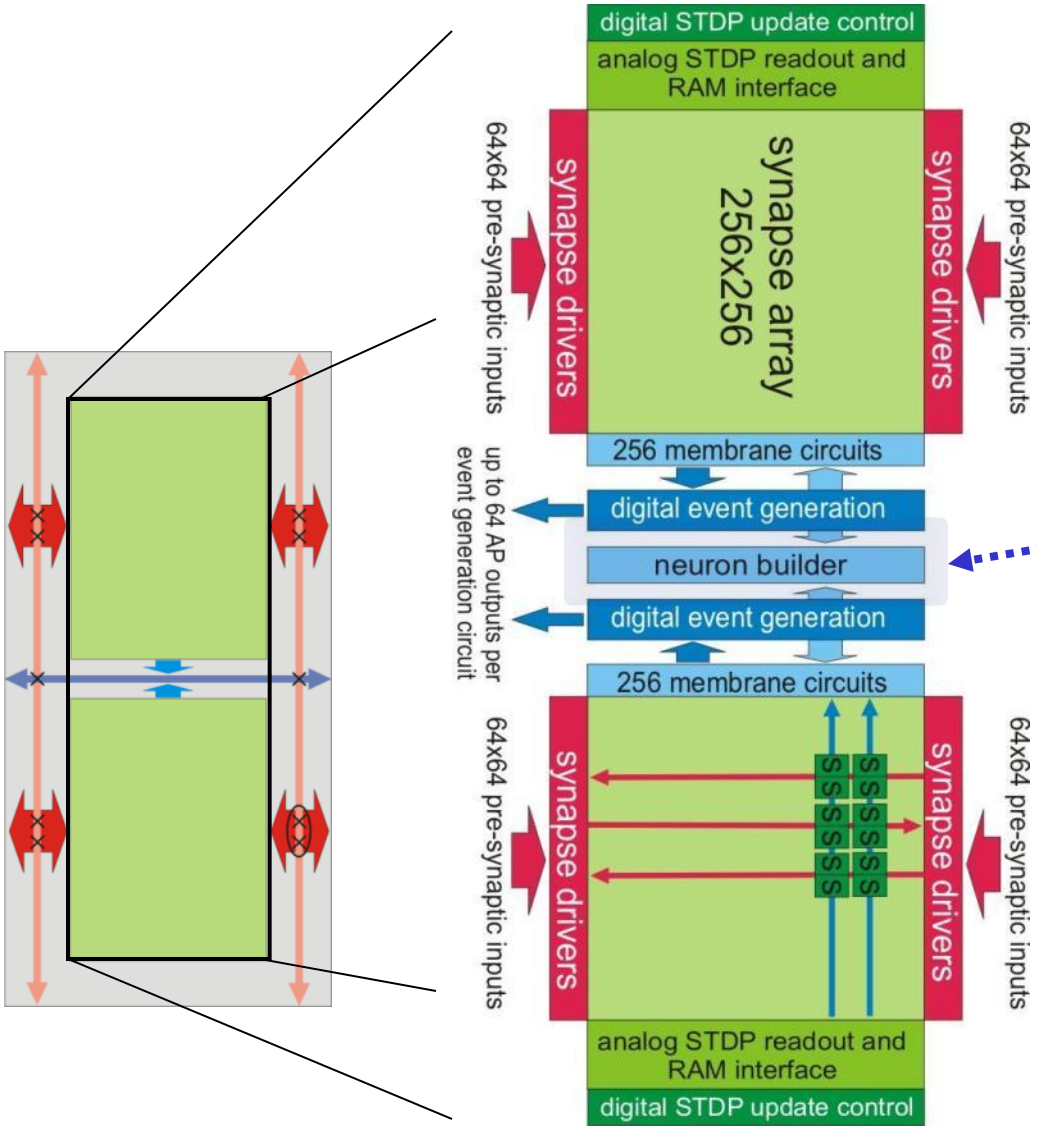
64 horizontal bus lanes

passive sparse switch matrix

connections to adjacent HICANN

unspecified part contains digital standard cell logic

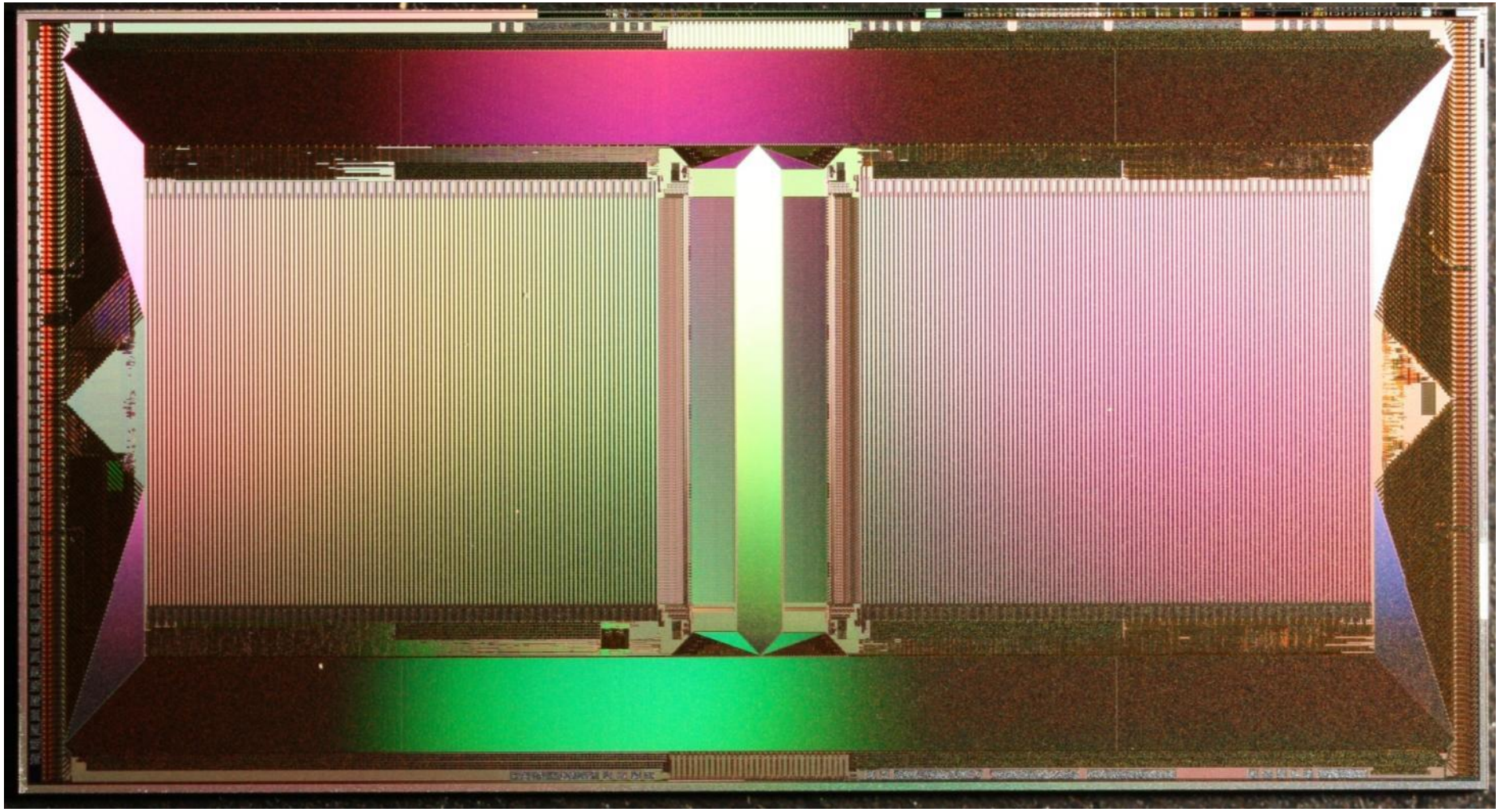
High Input Count Analog Neural Network (HICANN)



combining multiple membrane circuits with 256 synapses each allows neurons with up to 16k pre-synaptic inputs

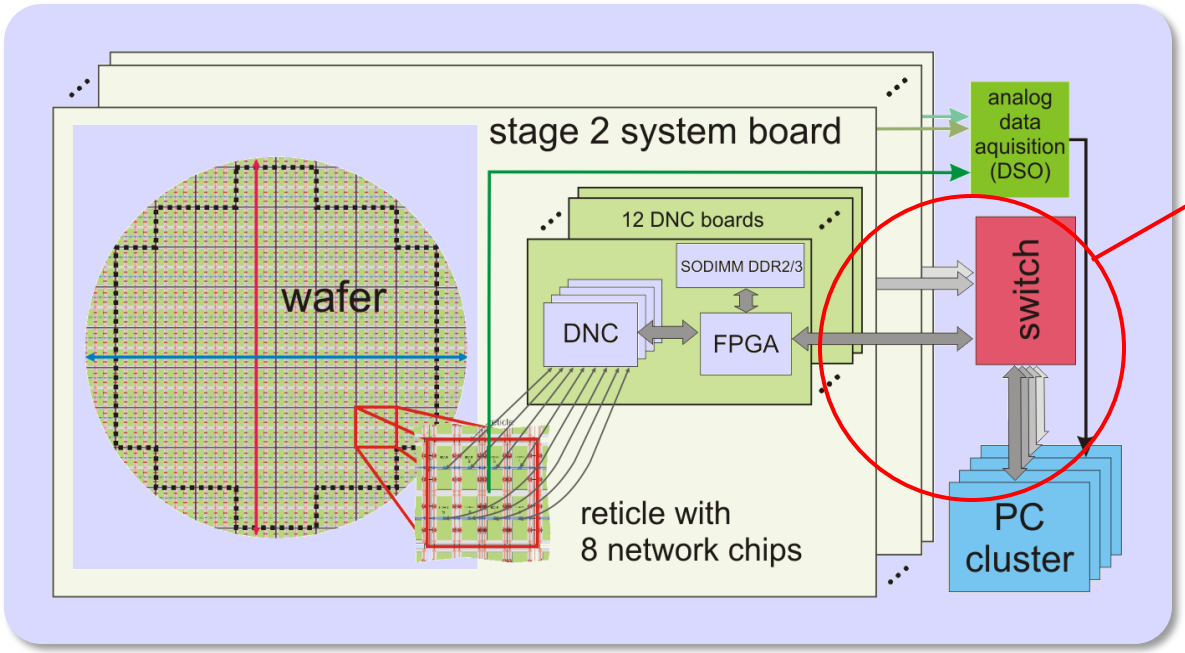
→ will be extended to real multi-compartment neurons with back-propagating action-potentials by current PhD

Microphotograph of the HICANN Die



10 mm

Adding Interaction to the System



latency of host communication can be low:

- O(µs) for GBit-Ethernet
- <100ns for specialized High-Speed serial links

possibilities:

- sensor-motor interactions: movement of a simulated retina
- robot in virtual environment simulated on PC cluster

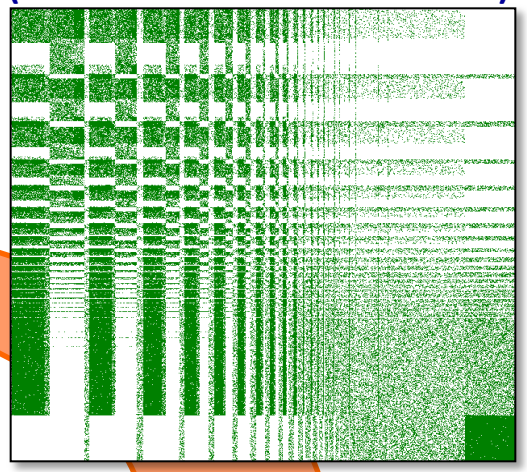
→ required latency 10 to 100ms: 1 to 10 µs

Software : From Networks to Experiments

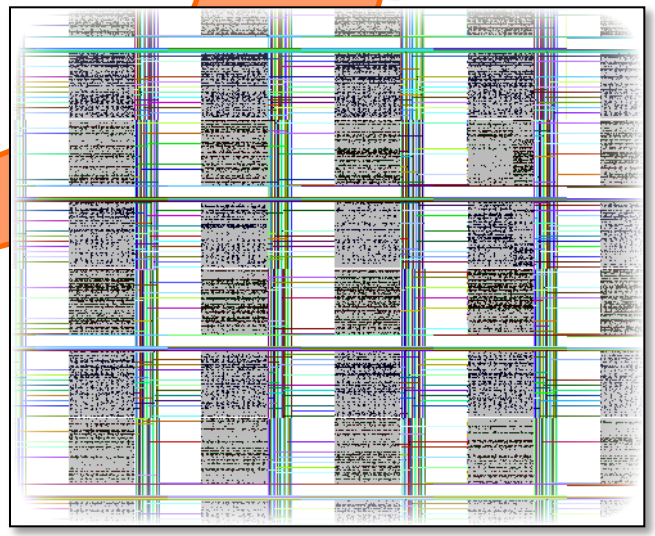
PyNN script

```
import pyNN.stage2 as pynn

pynn.setup()

neuronParams = {
    'v_init'    : -70.6,
    'w_init'    : 0.0,
[...]
}
pool0 = pynn.create(pynn.EIF_[...])
pool1 = pynn.create(pynn.EIF_[...])
[...]
pynn.connect(pool0, pool0, p=0.26, weight=0.5)
pynn.connect(pool1, pool0, p=0.16, weight=0.5)
[...]
pynn.run()
[...]
```

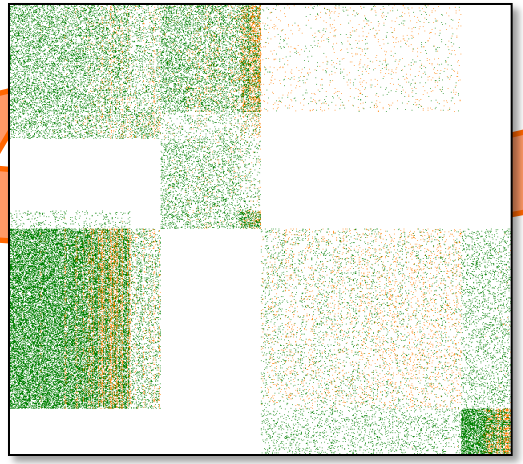
Mapping
(reordered connection matrix)



Routing



Configuration/Evaluation
(comparing connection matrix)



Summary

FACETS has developed solutions to enable large scale analog hardware as an alternative to numerical simulations:

- programmability of topology and model parameters
- flexible, biologically realistic neuron model
- low-power continuous-time communication
- scalable packet-based inter-wafer and host communication
  → possibility of interactive simulations (sensor-motor loop)
- a software framework for the translation of experiments from biology to hardware