VLSI Implementations of Very Large Scale Neuromorphic Circuits – Achievements, Challenges and Hopes

- Basics, Essentials and Motivations
- Approaches towards Very Large Systems
- We do need Software even for Neuromorphic Hardware
- Future Challenges and Plans

Workshop on Technology Maturity for Massively Parallel Adaptive Computing
Portland (OR), February 2009

Karlheinz Meier
Ruprecht-Karls-Universität Heidelberg

Contemporary IT systems

- Processor-memory based architectures with serial command execution (Turing)
- Predetermined algorithms define capabilies and performance (software)
- Based on well defined reproducible states and well defined reversible time evolution
- Electronics implementation of Boolean operators, high power consumption
- High yield requirements, little fault tolerance
- Limited by atomic distance scale in components (nm) : ultimately **component limited**

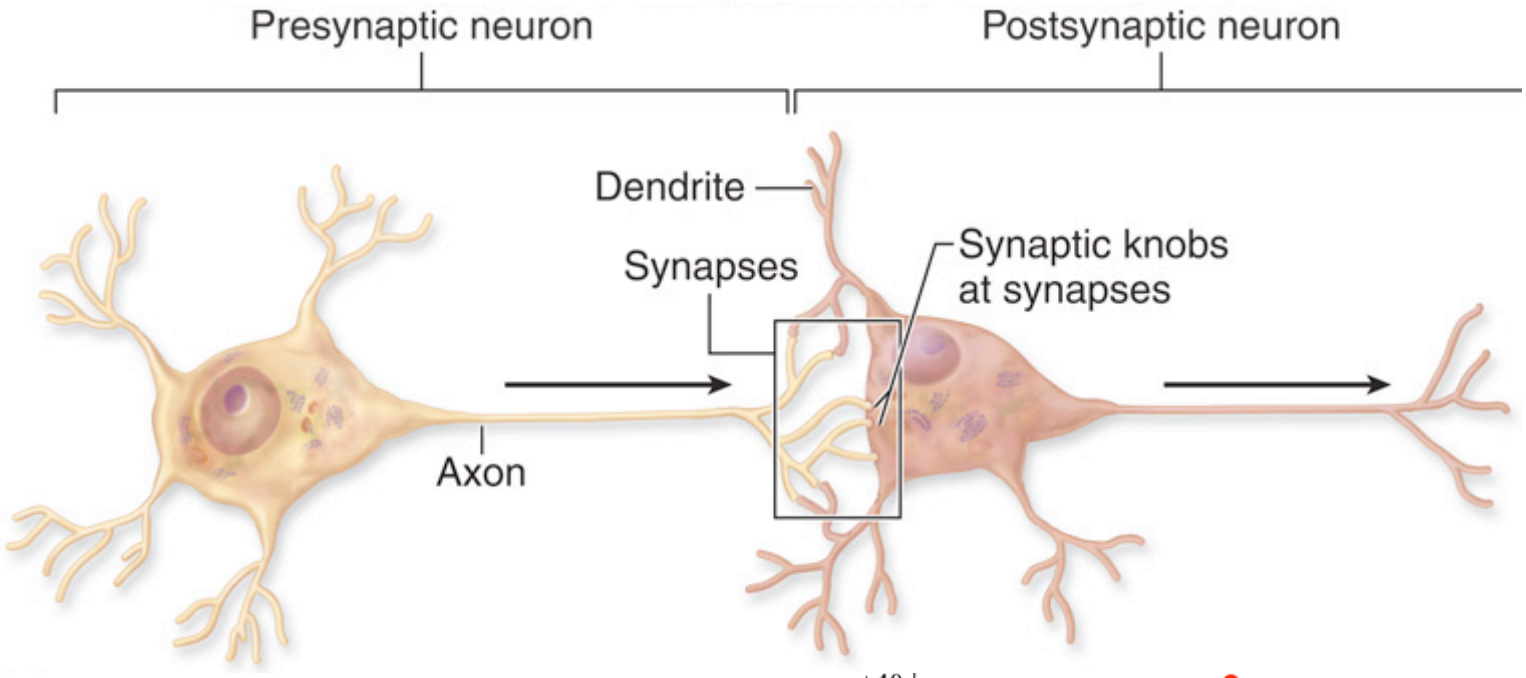
WELL UNDERSTOOD

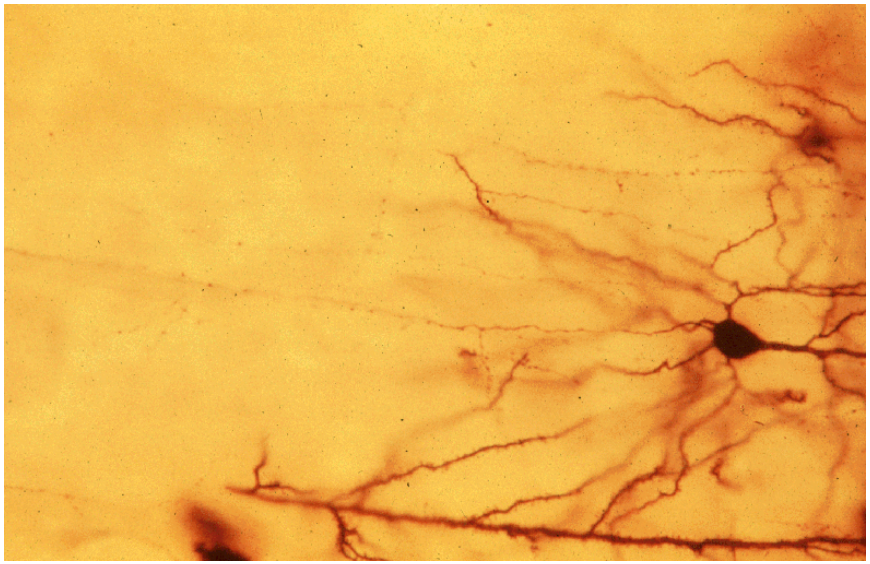
Neural computation

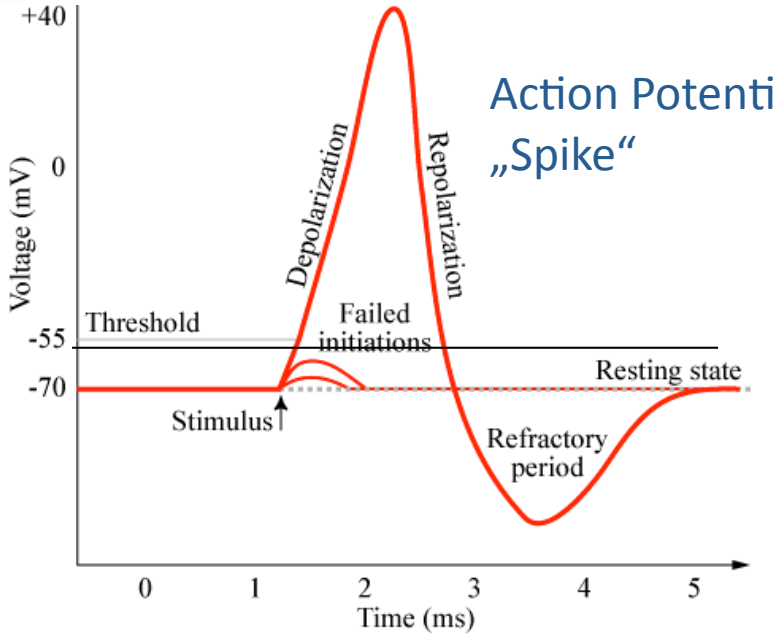
- Maximally parallel, non-linear compting elements with large diversity
- Time correlations drive the dynamics (e.g. STDP)
- Learning by internal self-organisation and by strong interaction with environment
- Low power consumption and high fault tolerance
- Limited by degree of complexity : **ultimately architecture/size limited**

NOT UNDERSTOOD (listed as a major challenge for 21st century science)

Biology Basics : Neurons - Synapses - Dendrites – "Spikes“







Action Potential = „Spike“

Essentials for 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

Essentials for 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

Modeling approaches

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”

Methodological approaches

- "Bottom-up" : Based on (simplified) reconstructed models of biological morphology and function : cells, synapses, connectivity, plasticity mechanisms ("atoms and their interactions") (e.g. FACETS-1)
- "Top-down" : Based on large scale functional blocks and their inter-relations (e.g. Jeff Hawkins)
- "First principles" : Design of computational paradigms, not necessarily biological plausible (e.g. FACETS-1, liquid computing)
- "Evolutionary" generation of computational structures based on biologically plausible elements

FACETS Facts and Figures

EU Integrated Project in Framework Programme 6
Information Society Technologies (IST) – Future Emergent Technologies (FET)
Bio-inspired Intelligent Information Systems (Bio-I3)

Funding Period :	9/2005 – 8/2010 (5 years)
EC financial contribution :	10.509.000 Euros
Participating Groups (Y3) :	15 (> 4 Disciplines)
Effort :	3807 Person Months (approx. 79 FTEs)
Graduate Students :	77
Refereed papers by Y3 :	124

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 : Emulation in analog and mixed-signal VLSI systems

```
# Cell parameters
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_refi
v_mean    = -60.   # (mV) 'mean'
tau_exc   = 5.     # (ms)
tau_inh   = 10.    # (ms)
```

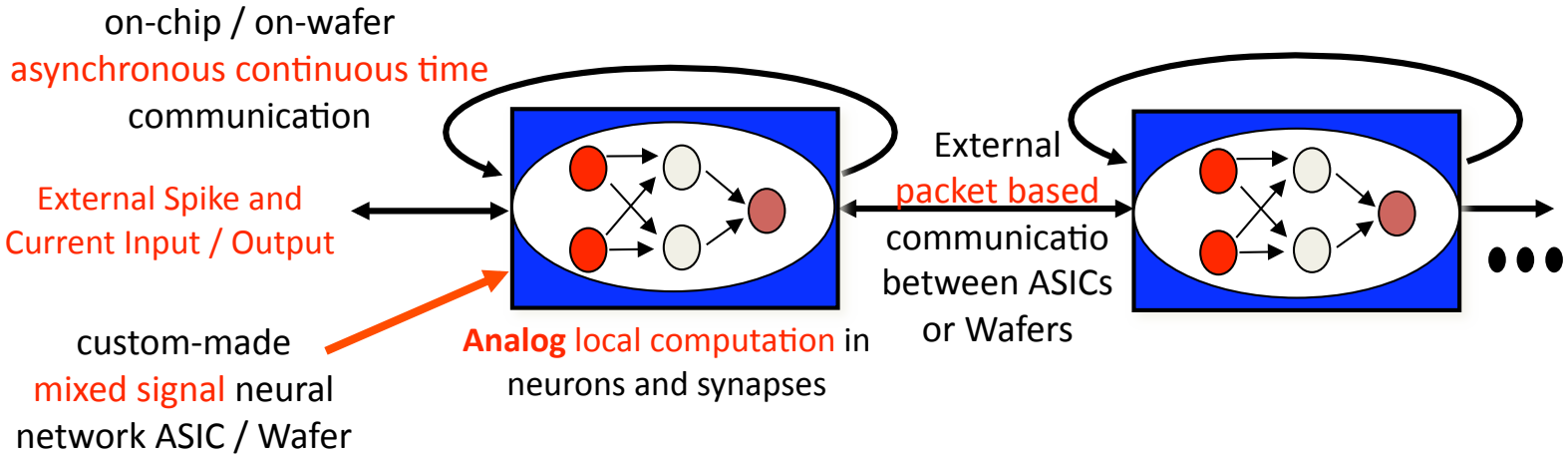
Methodology : Tool Development (Computing, VLSI)
Reduction of Biological Detail / Complexity

Common Goal : Study non-classical universal computing solutions

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

FACETS HW Concept : VLSI mixed-signal emulation

- hardware **mixed-signal approach:** local analog computation combined with high-speed state-of-the-art digital digital communication
- **Basically : Follow natures' example**



Stage 1

- Individual network modules used as building blocks, each module hosts one ANN ASIC and all main components to interface it (FPGA)
- High-speed links to connect the modules via a backplane

Stage 2

- Separate neural computation from setup / monitoring / control / readout
- Use Wafer Scale Integration for neural computation part

Stage 1 FACETS : Conductance-based Network Model

current source, no voltage dependence

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

membrane current

leakage current

sum over excitatory synapse currents k

sum over inhibitory synapse currents l

Voltage dependent part, changes membrane conductance

Synapses:

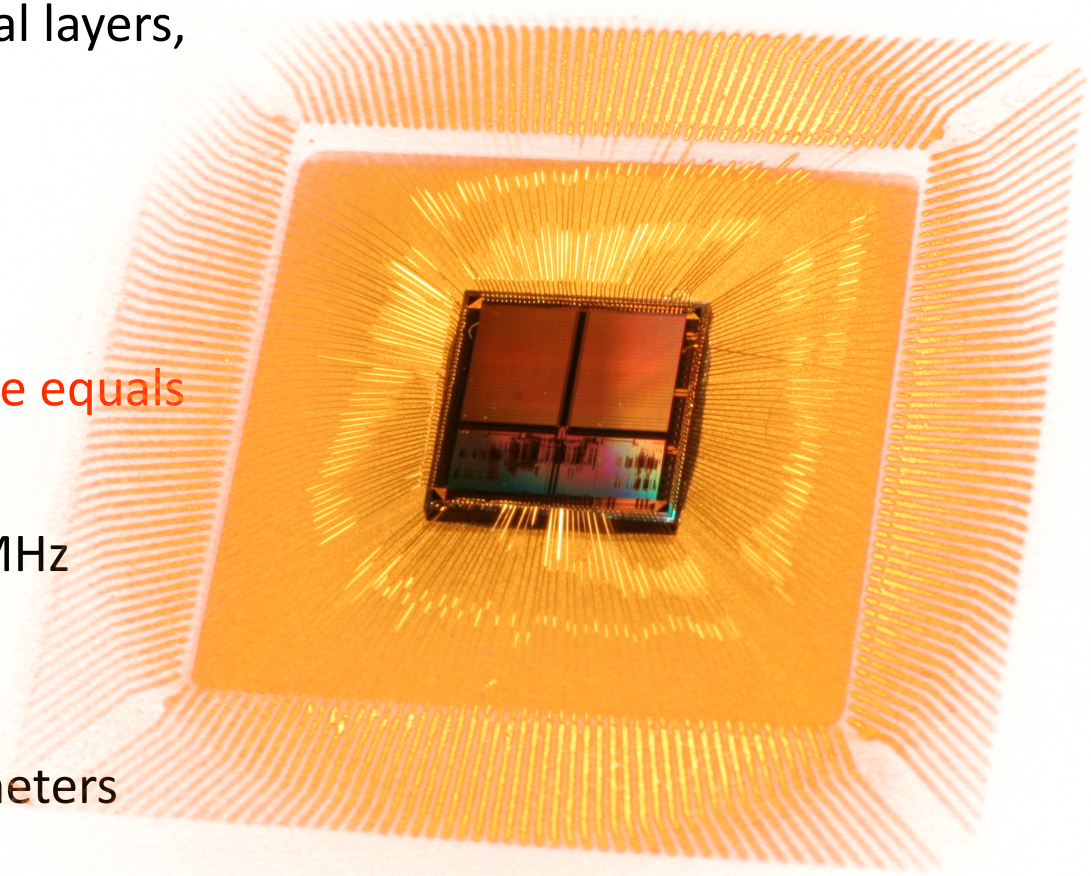
$p_{k,l}(t)$ exponential onset and decay (spike shape)

$g_{k,l}$ 0 to g_{max} with 4 bit (8 bit) resolution

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

Stage 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 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)



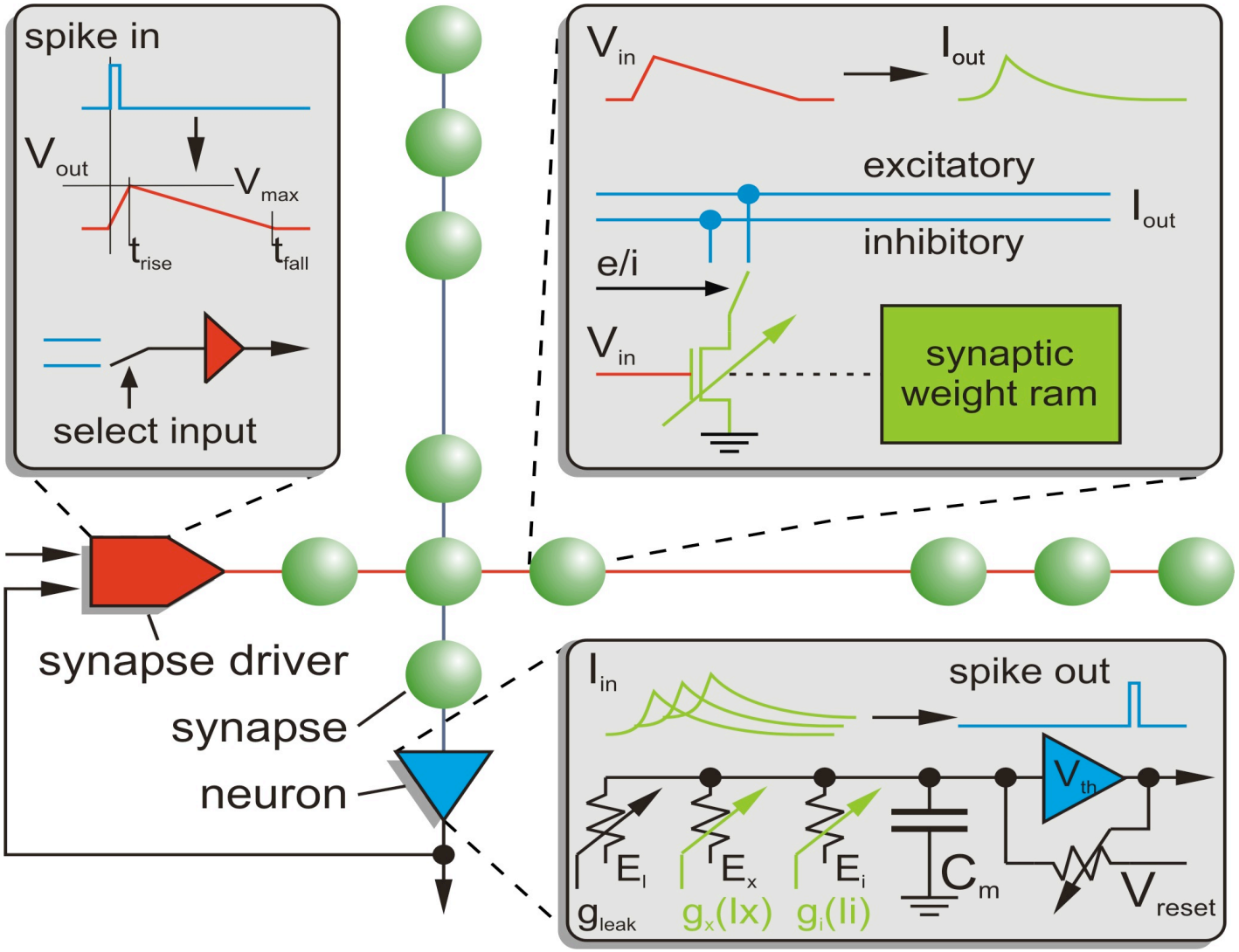
A New VLSI Model of Neural Microcircuits Including Spike Time Dependent Plasticity, Johannes Schemmel, Karlheinz Meier, Eilif Muller, Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04), IEEE Press, pp. 1711-1716, 2004

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}, $g_{k,l\,max}$)

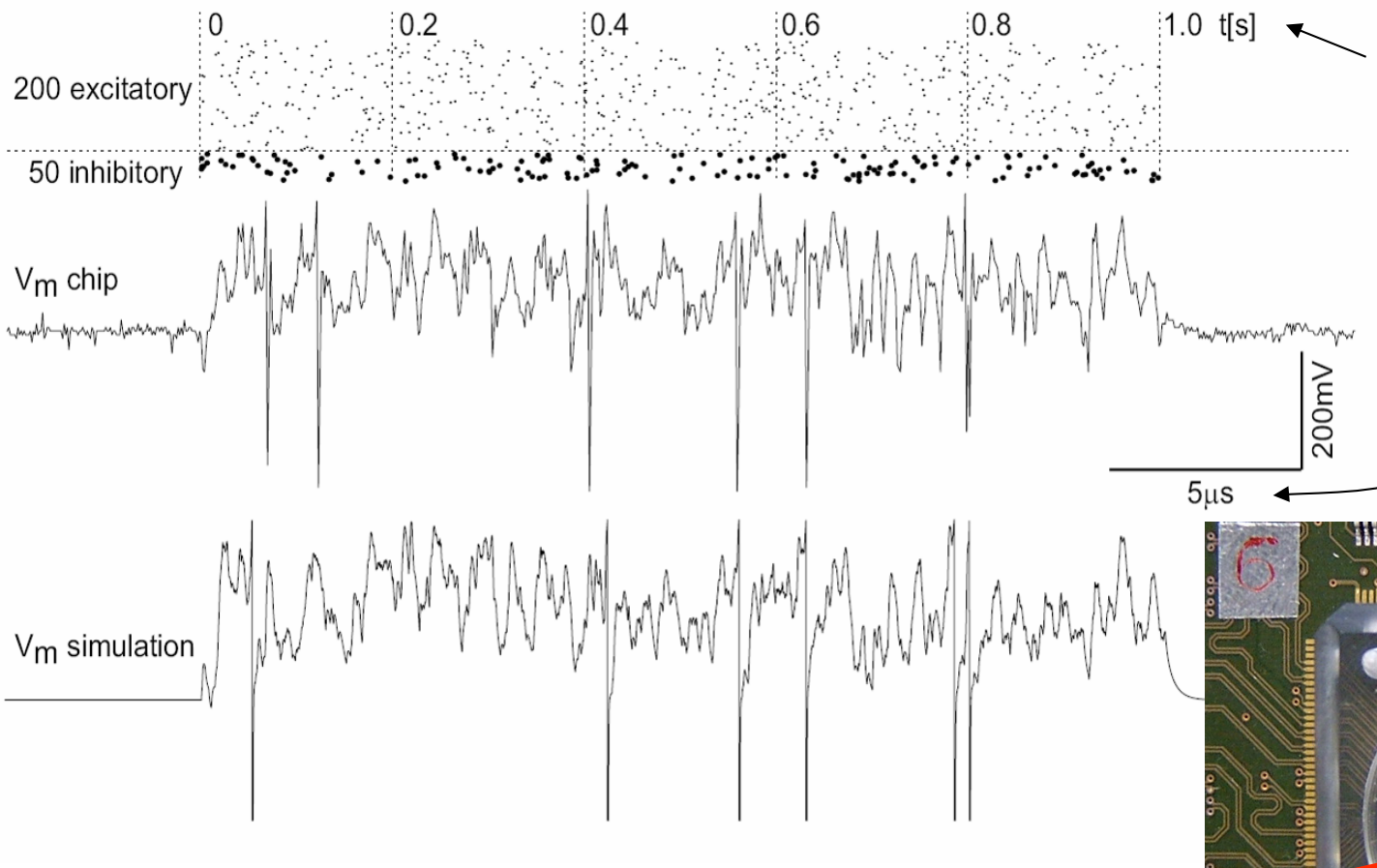
A New VLSI Model of Neural Microcircuits Including Spike Time Dependent Plasticity, Johannes Schemmel, Karlheinz Meier, Eilif Muller, Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04), IEEE Press, pp. 1711-1716, 2004

Overview of the Network Implementation



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

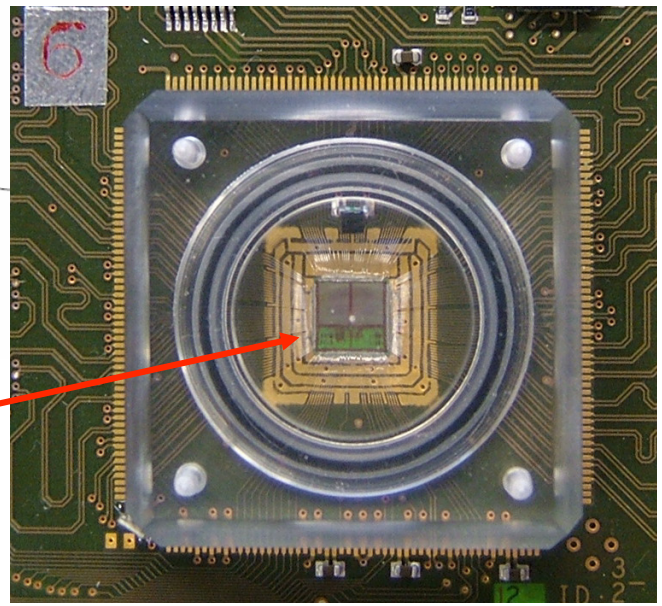
Response to Poisson-Distributed Input Spike Trains
Hardware vs. Simulation



biological time

chip time

1cm

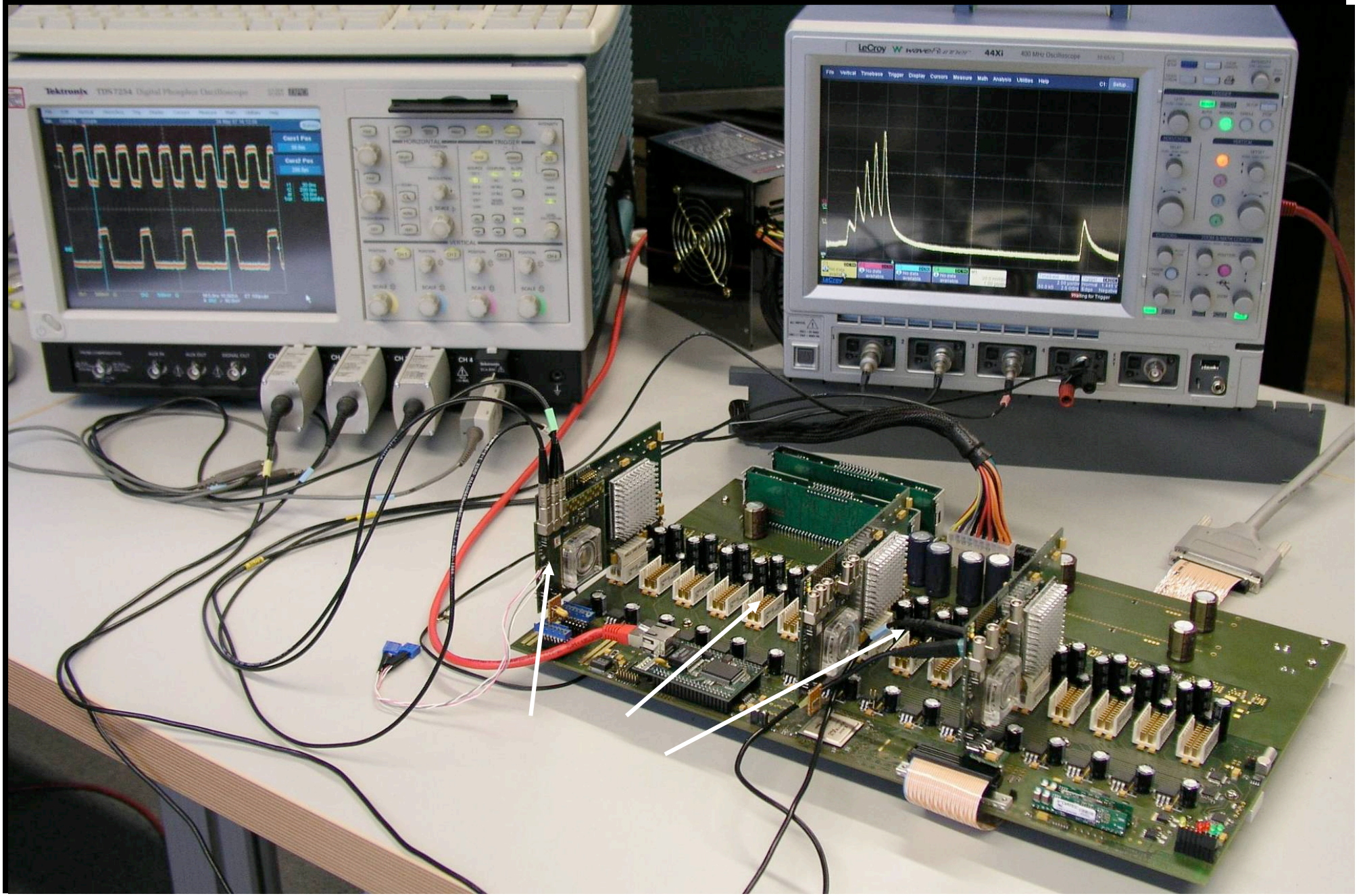


"Spikey" Chip

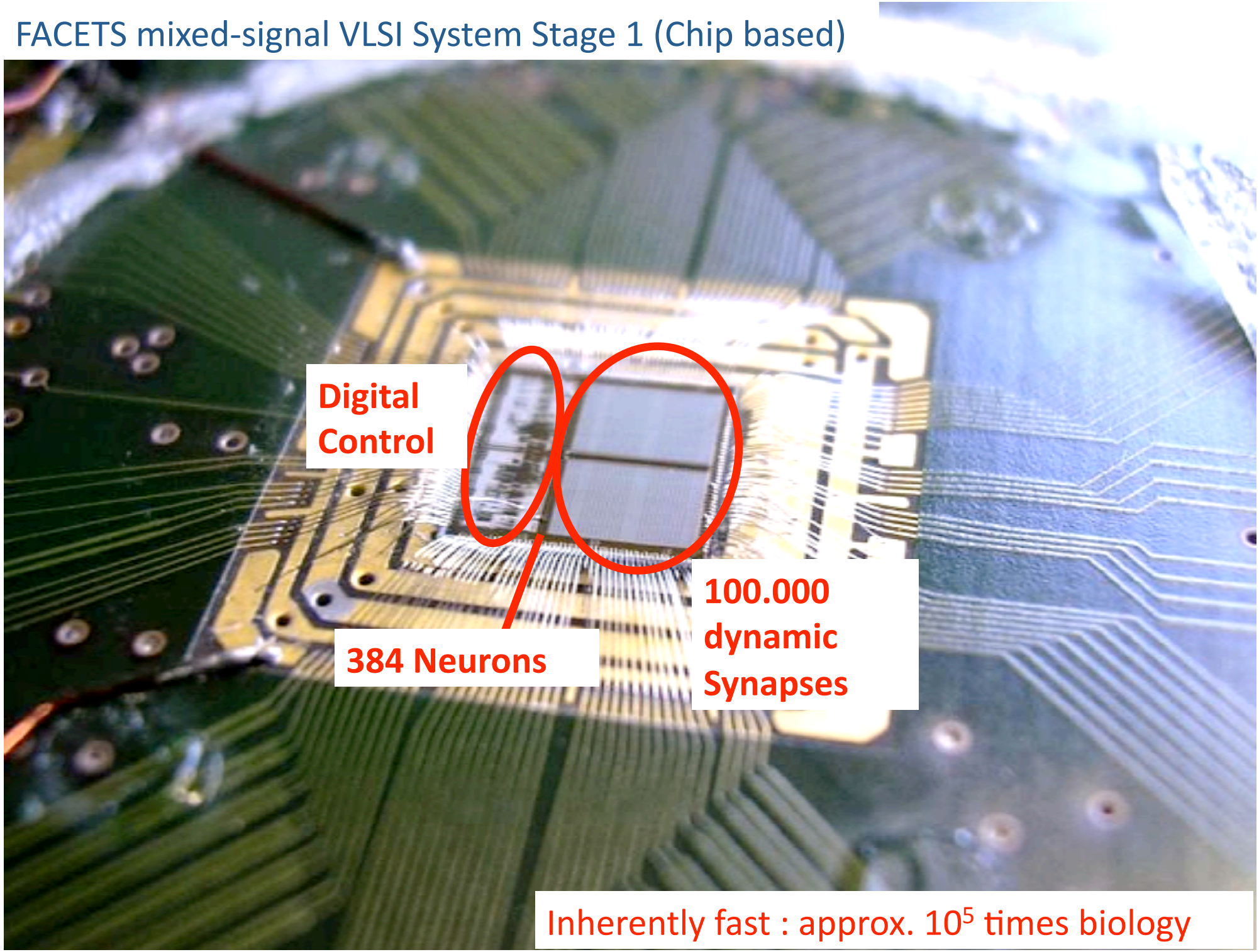
see Schemmel et. al., IJCNN 2006, ISCAS 2007

FACETS Stage 1 Backplane Set-Up

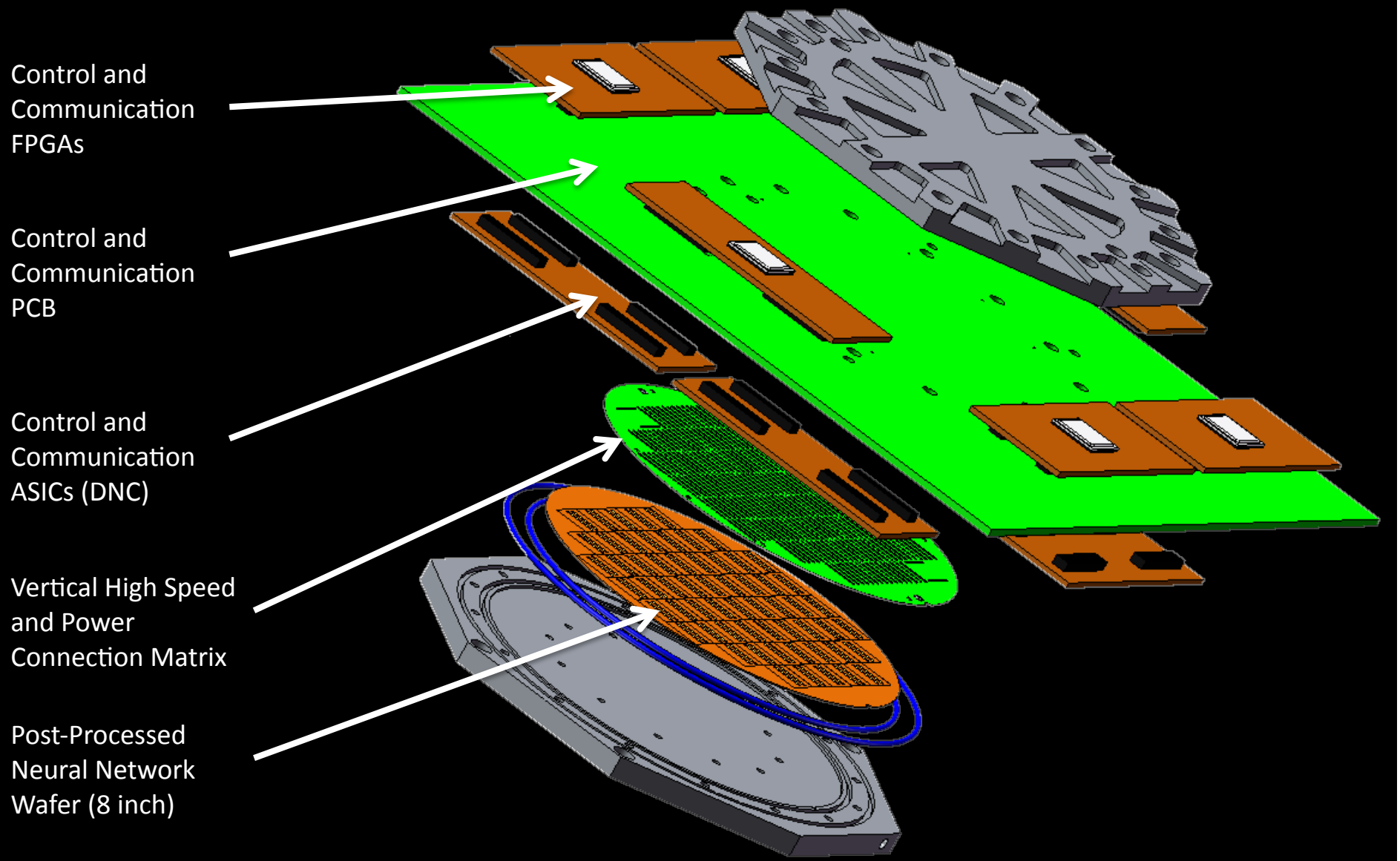
Up to 16 neural network boards can be combined to build a larger network



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



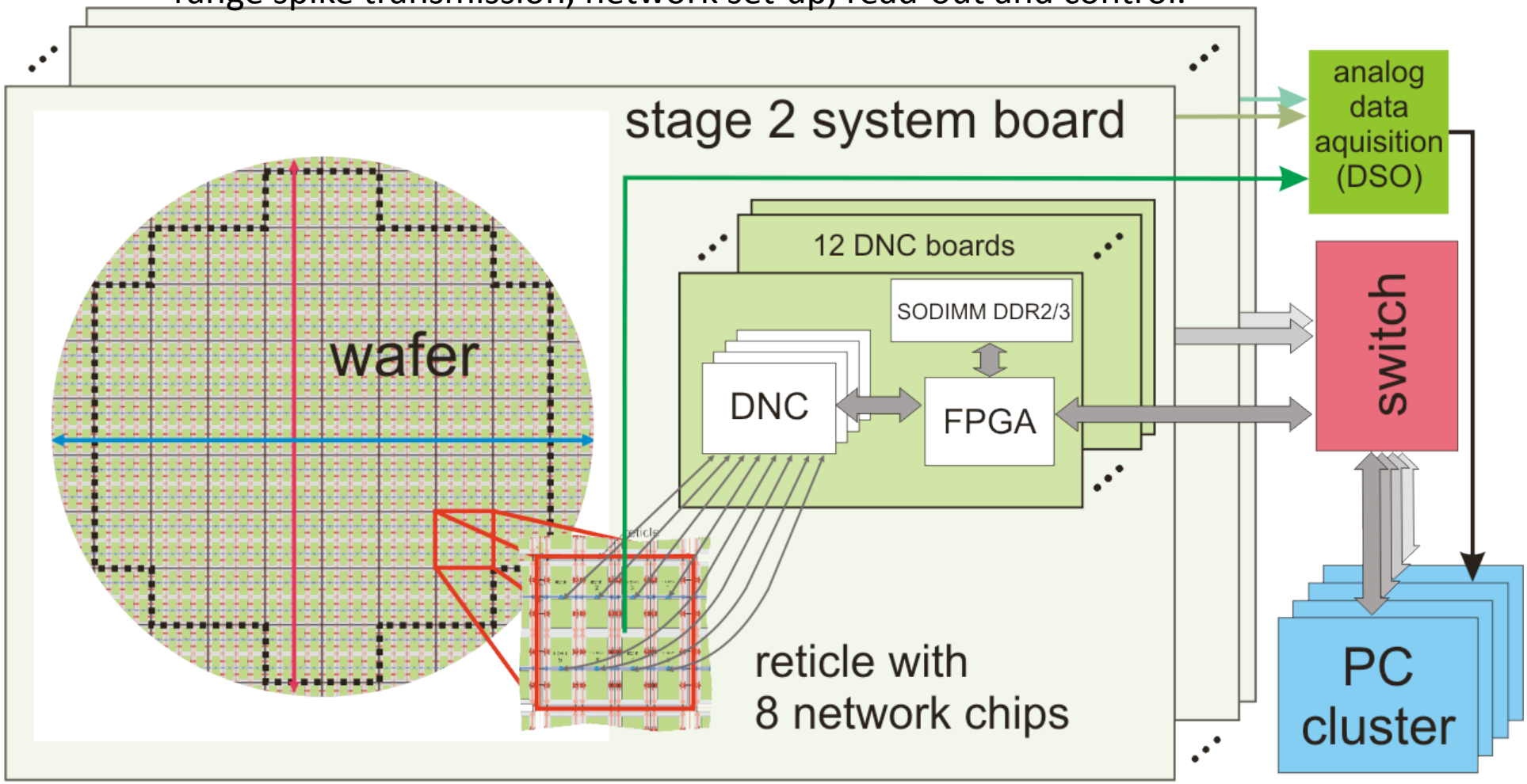
Stage 2 Technology : Neural Processing Unit, up to $2x10^5$ Neurons, $5x10^7$ Synapses
Idea : Separate Neural Circuits and Monitoring/Readout/Control



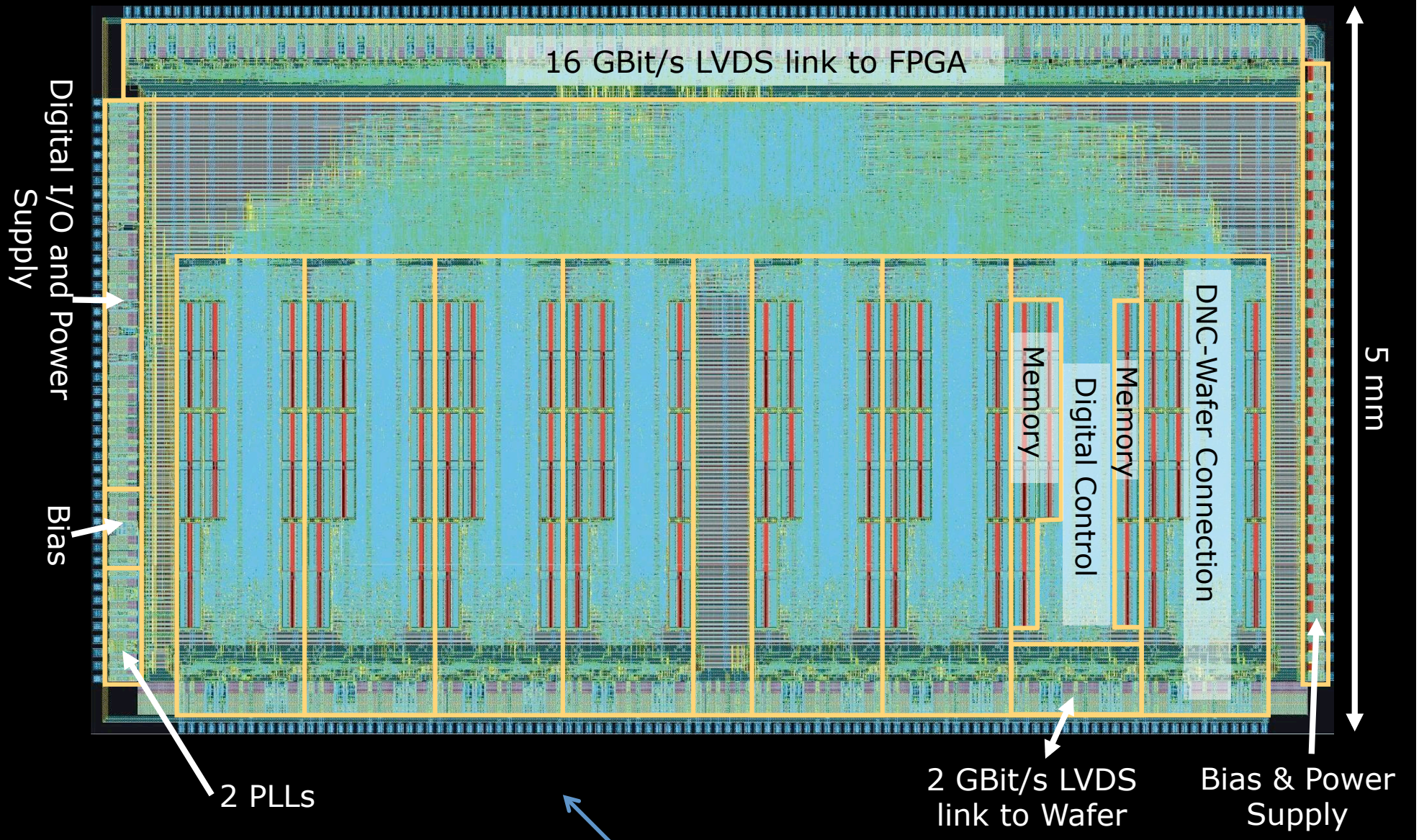
Stage 2 Architecture Overview

Hierarchical Communication Setup

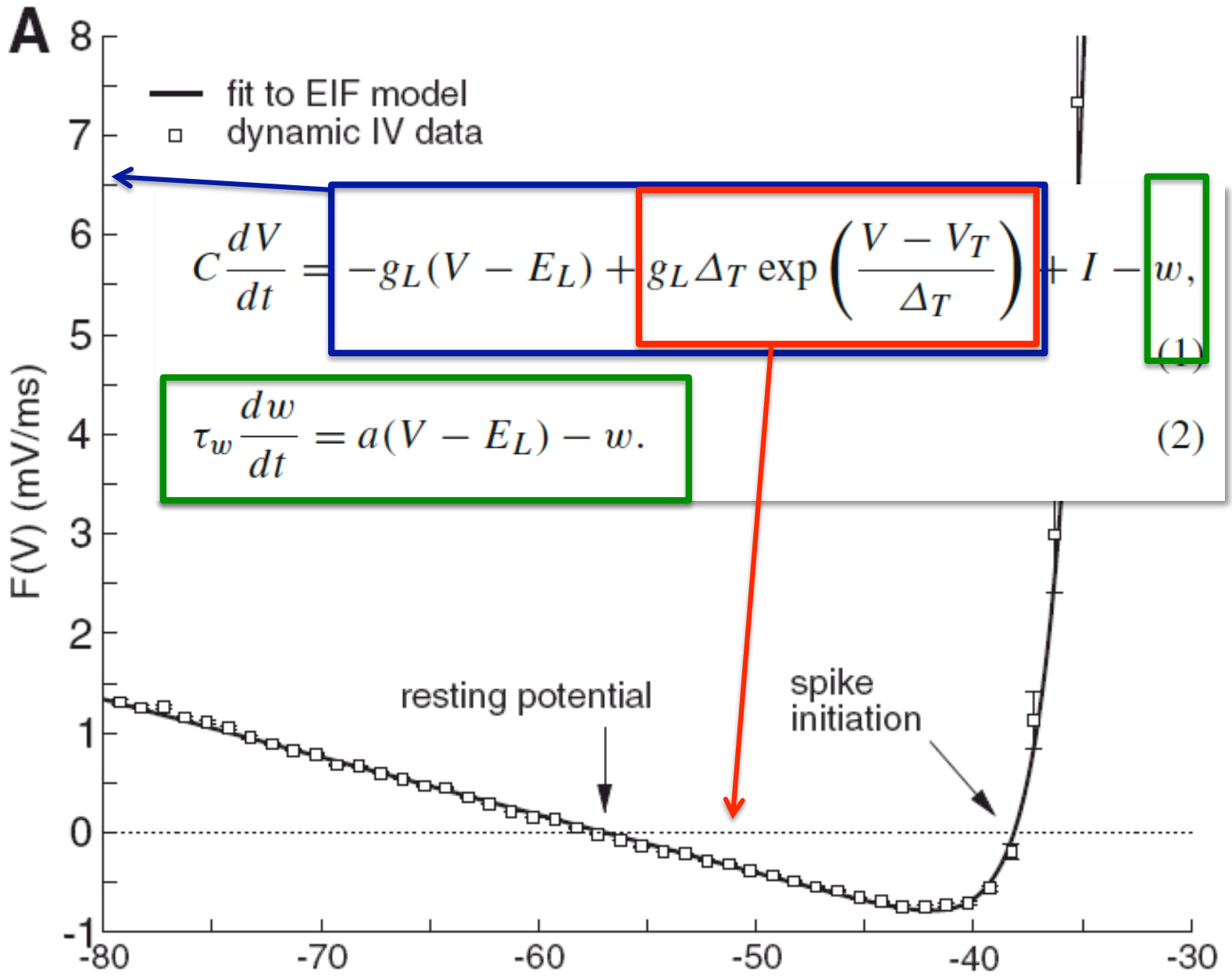
- Layer 1 : Use complete wafer to exploit inherent fault tolerance.
  On-Wafer continuous, asynchronous, fixed delay spike transmission
- Layer 2 : **Off-Wafer** Packet based digital communication for medium and long-range spike transmission, network set-up, read-out and control.



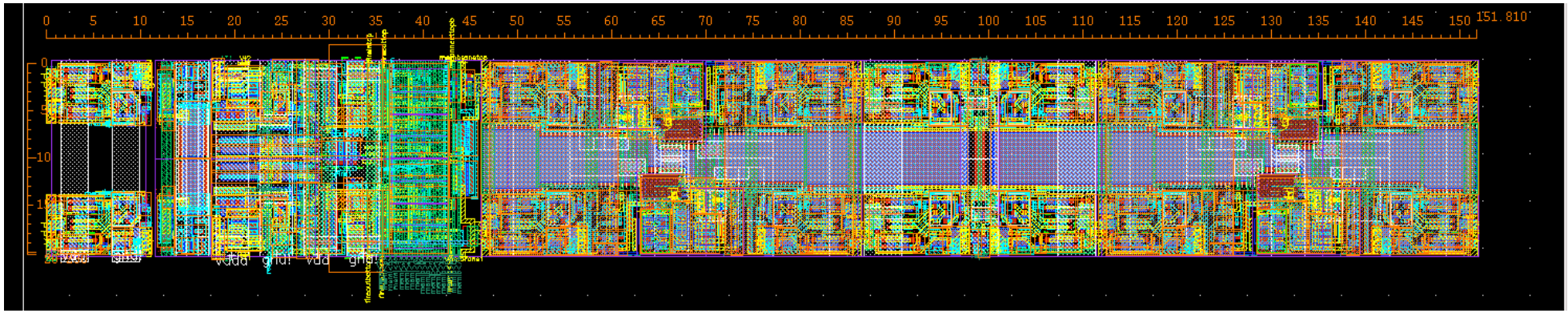
Digital Network Chips (DNC) : Managing packet based off-wafer routing of spike events (TU Dresden)

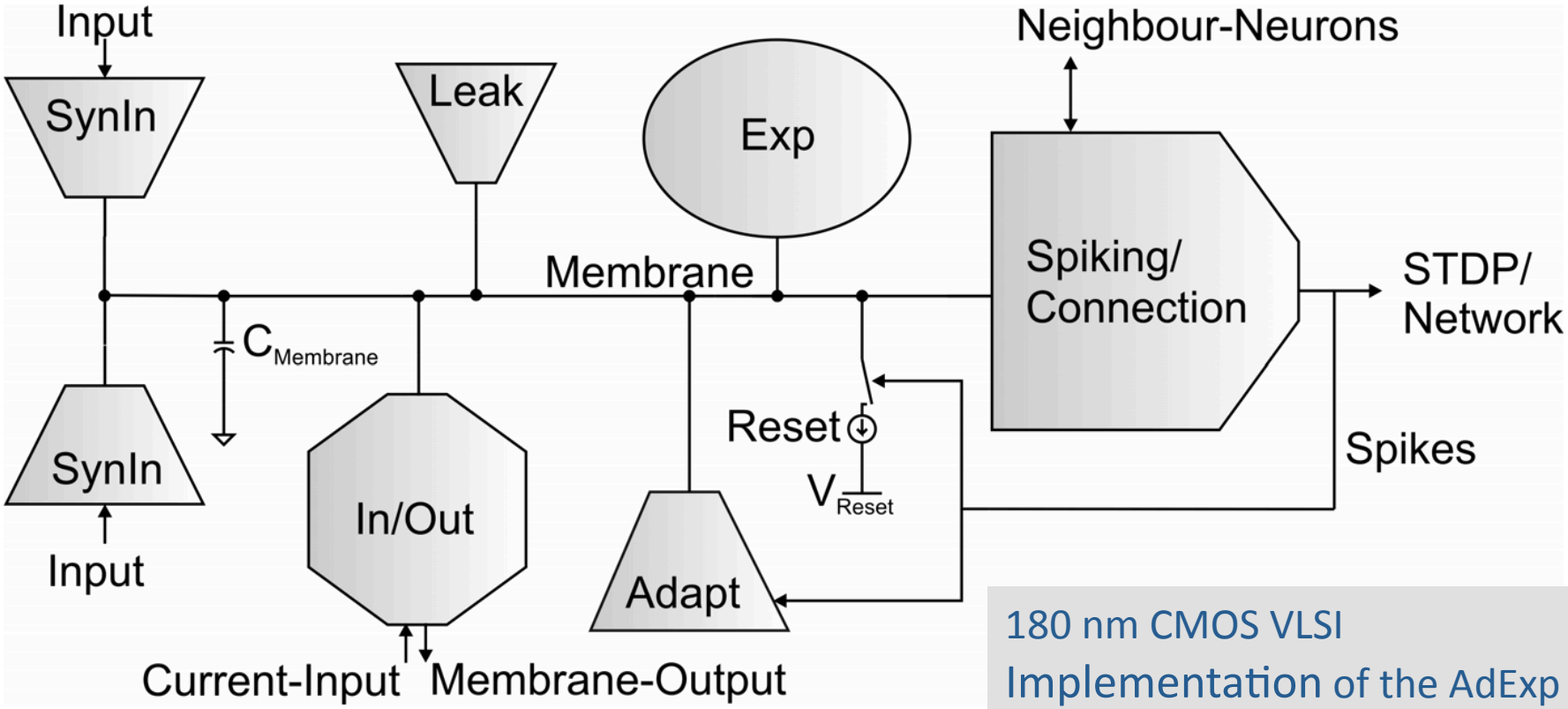


FACETS Adaptive-Exponential Integrate-and-Fire Model



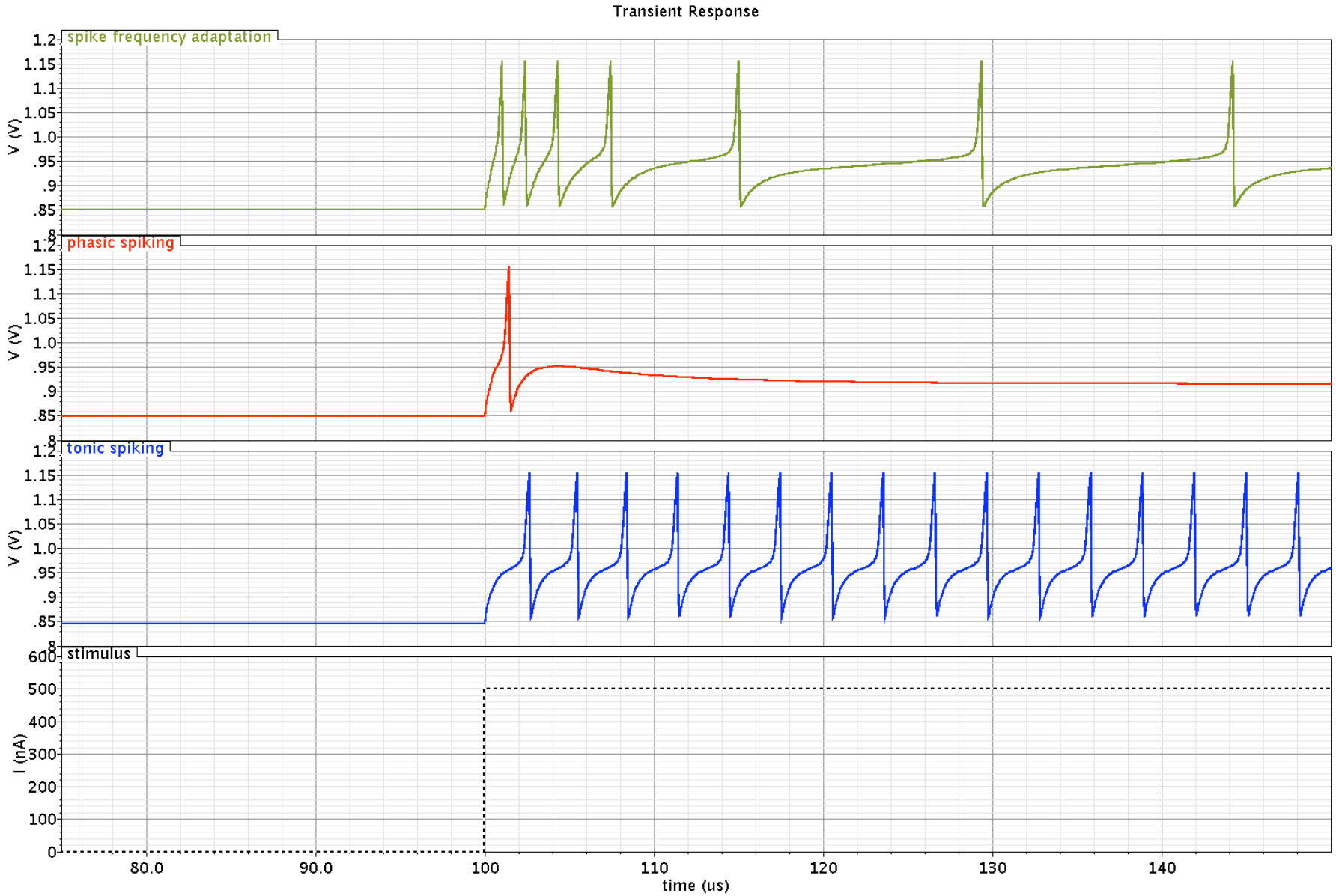
R. Naud et al.
Versatility and Rel
exponential integrate-and fire-model
Biol Cybern (2008) 99:335–347



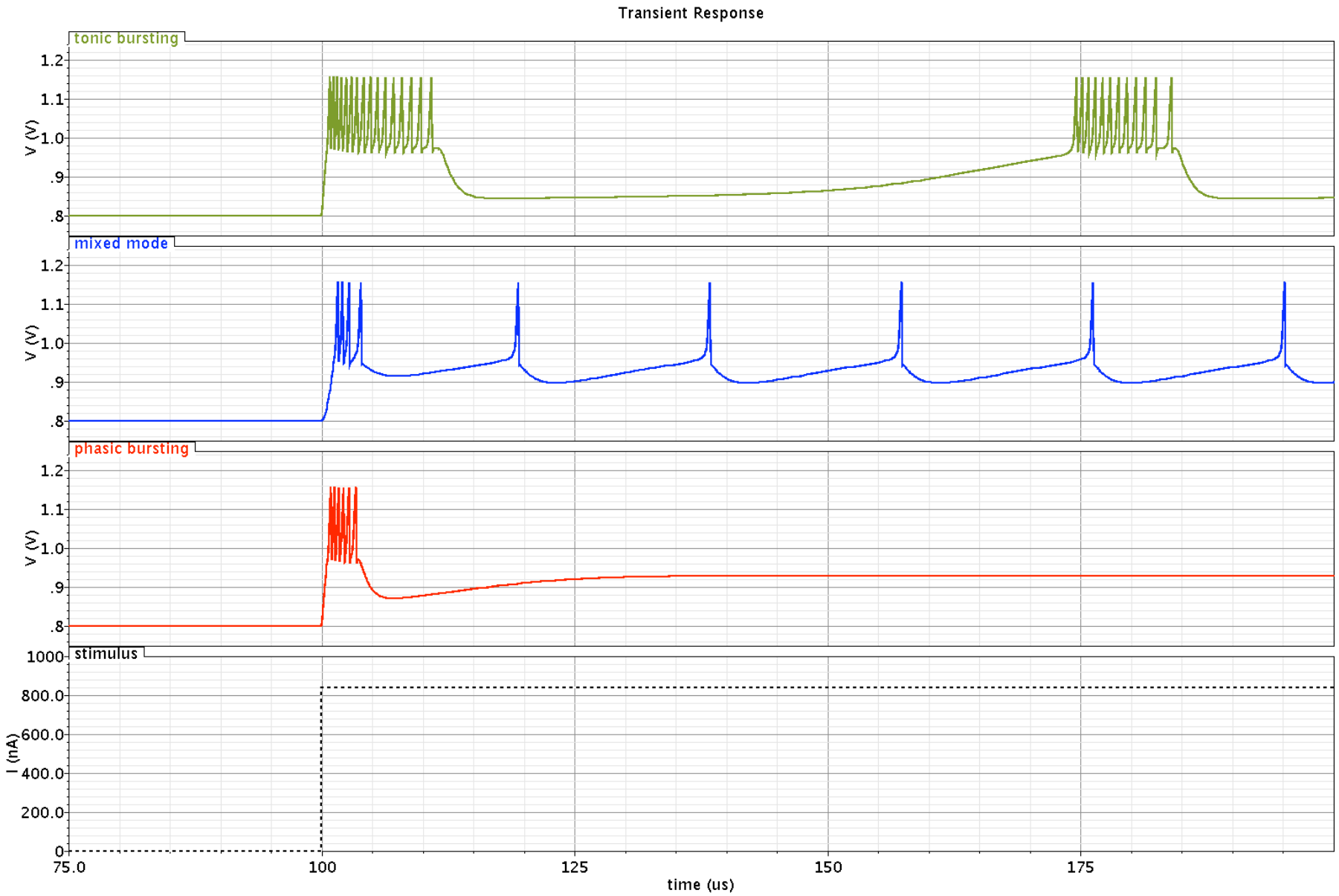


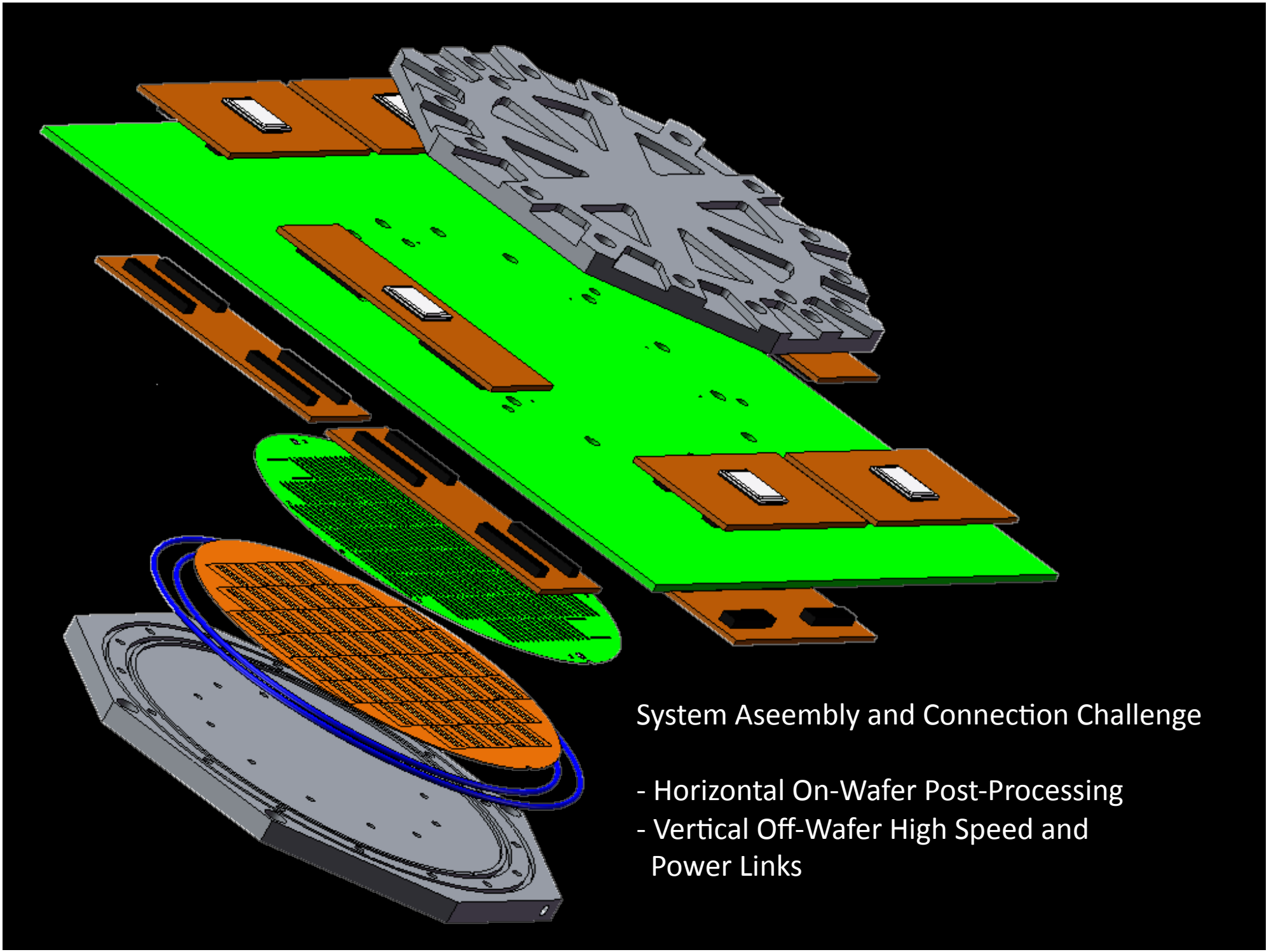
180 nm CMOS VLSI Implementation of the AdExp Integrate-and-Fire Neuron
Parameters stored on analog floating gates

Spike Firing Modes of the AdExp VLSI Neuron



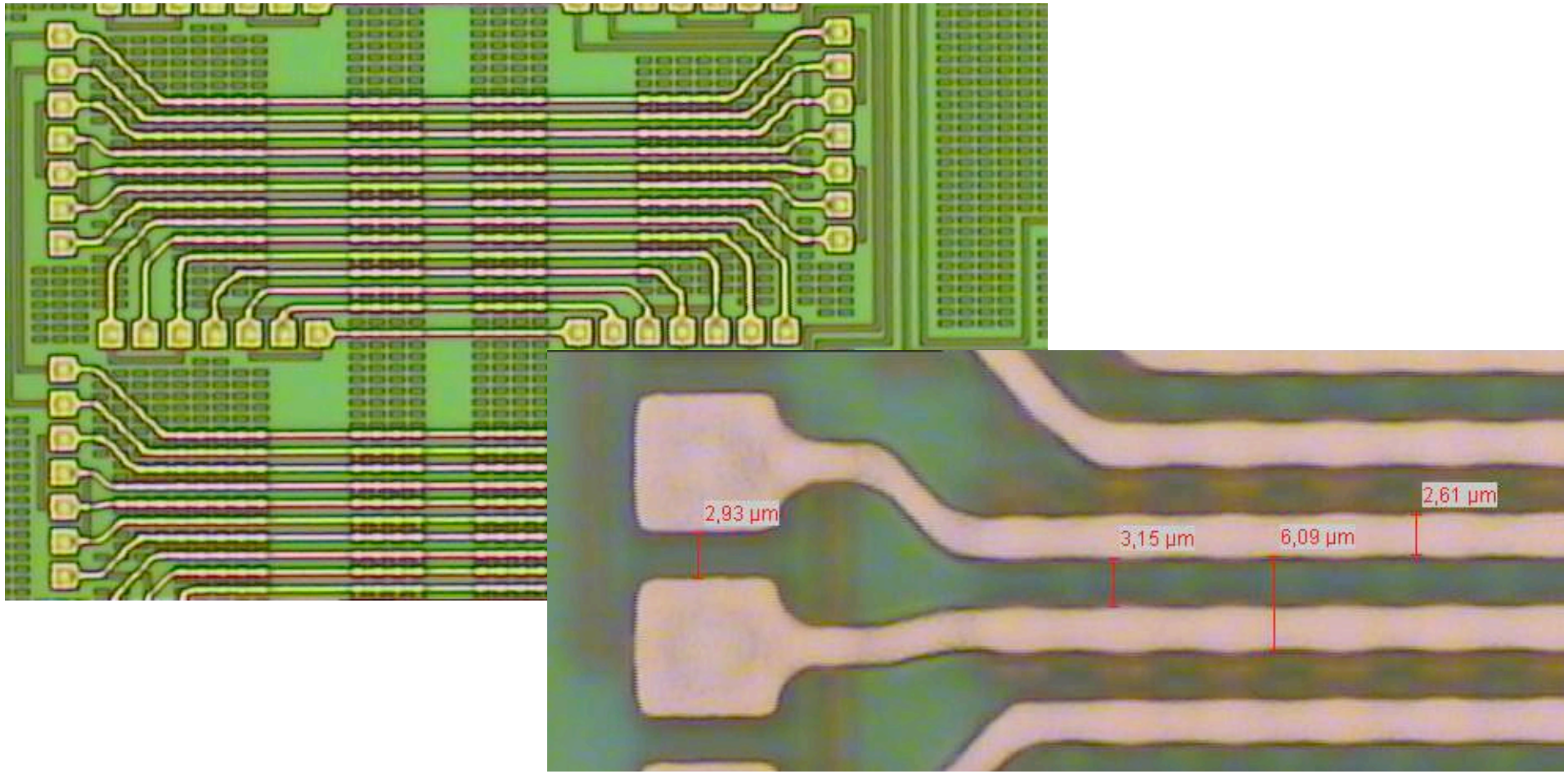
Burst Firing Modes of the AdExp VLSI Neuron




System Aseembly and Connection Challenge
- Horizontal On-Wafer Post-Processing
- Vertical Off-Wafer High Speed and
Power Links

Horizontal on-Wafer Post-Processing (Fraunhofer FhG Berlin)

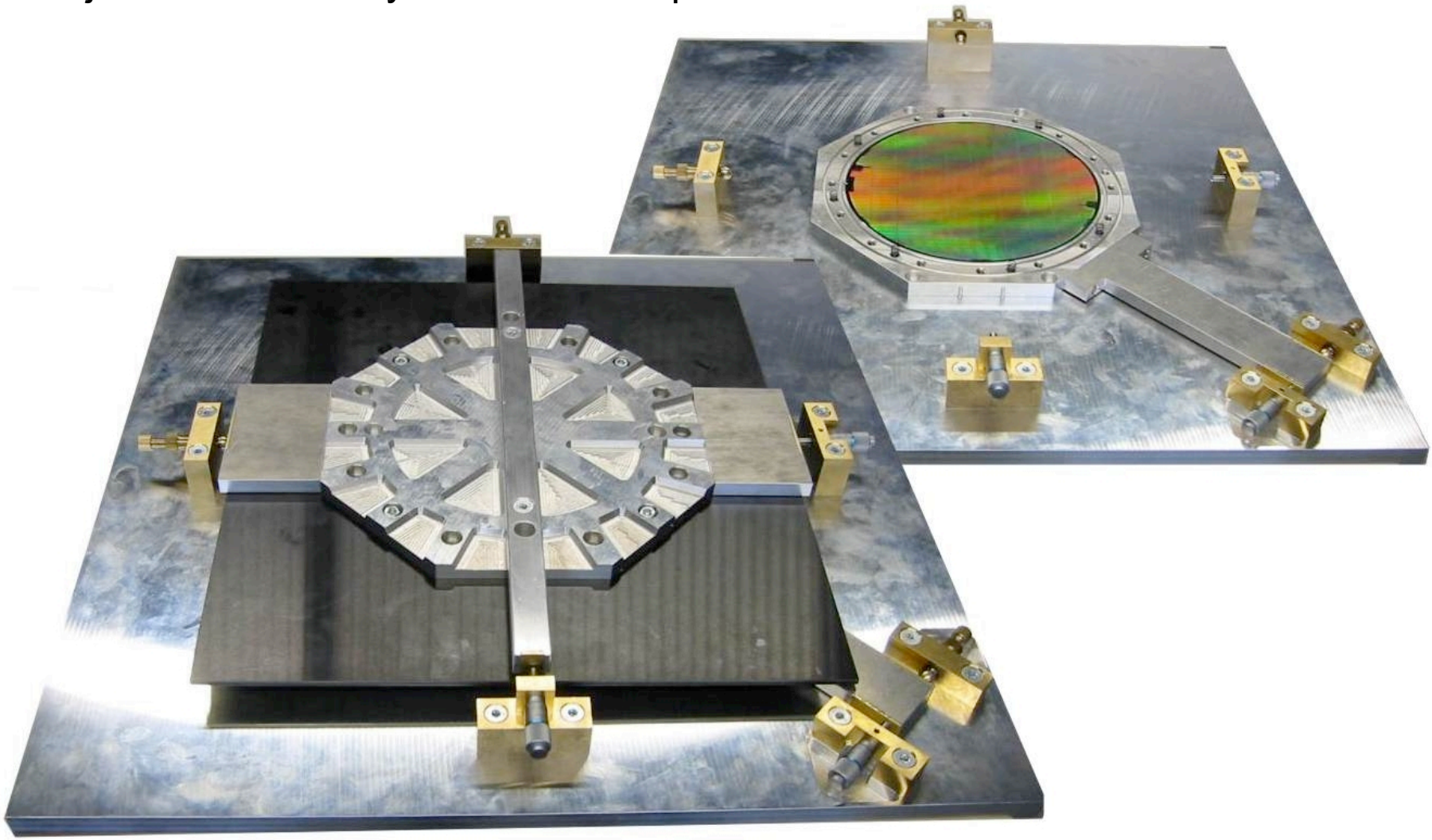
- Measured yield for 8 µm pitch test structures 100% on over 8 inch wafer
- Copper lines for spike transmission still susceptible to corrosion
- Version with gold lines und improved surface processing in production



Microscope views of an 8 µm pitch post processing structure

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
- Adjustment accuracy better than 50 µm over the 8 inch wafer



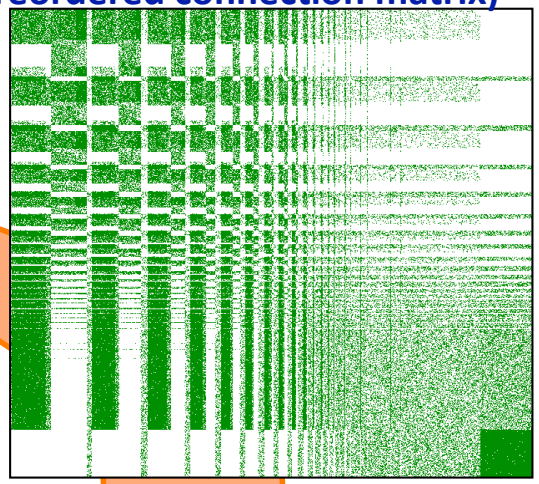
Software : From Networks to Experiments

PyNN script

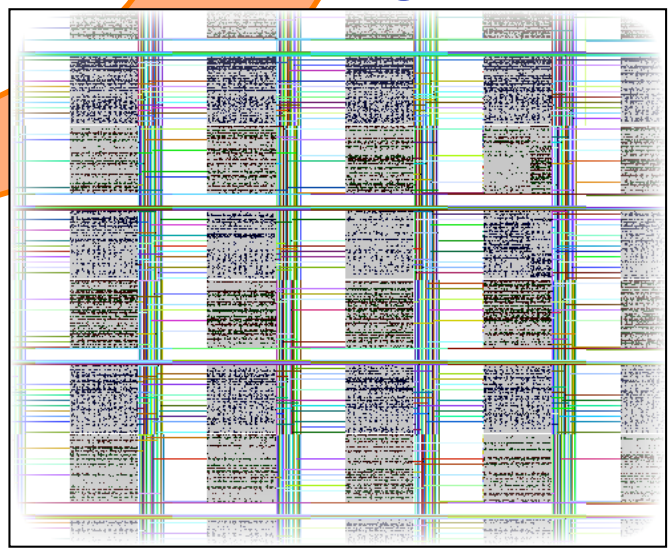
```
import pyNN.stage2 as pynn

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)

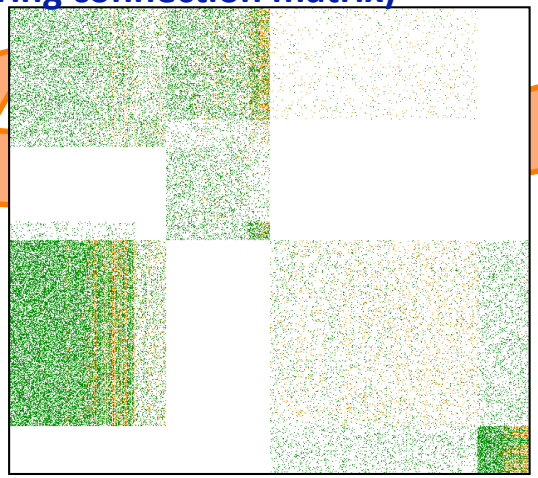


Routing

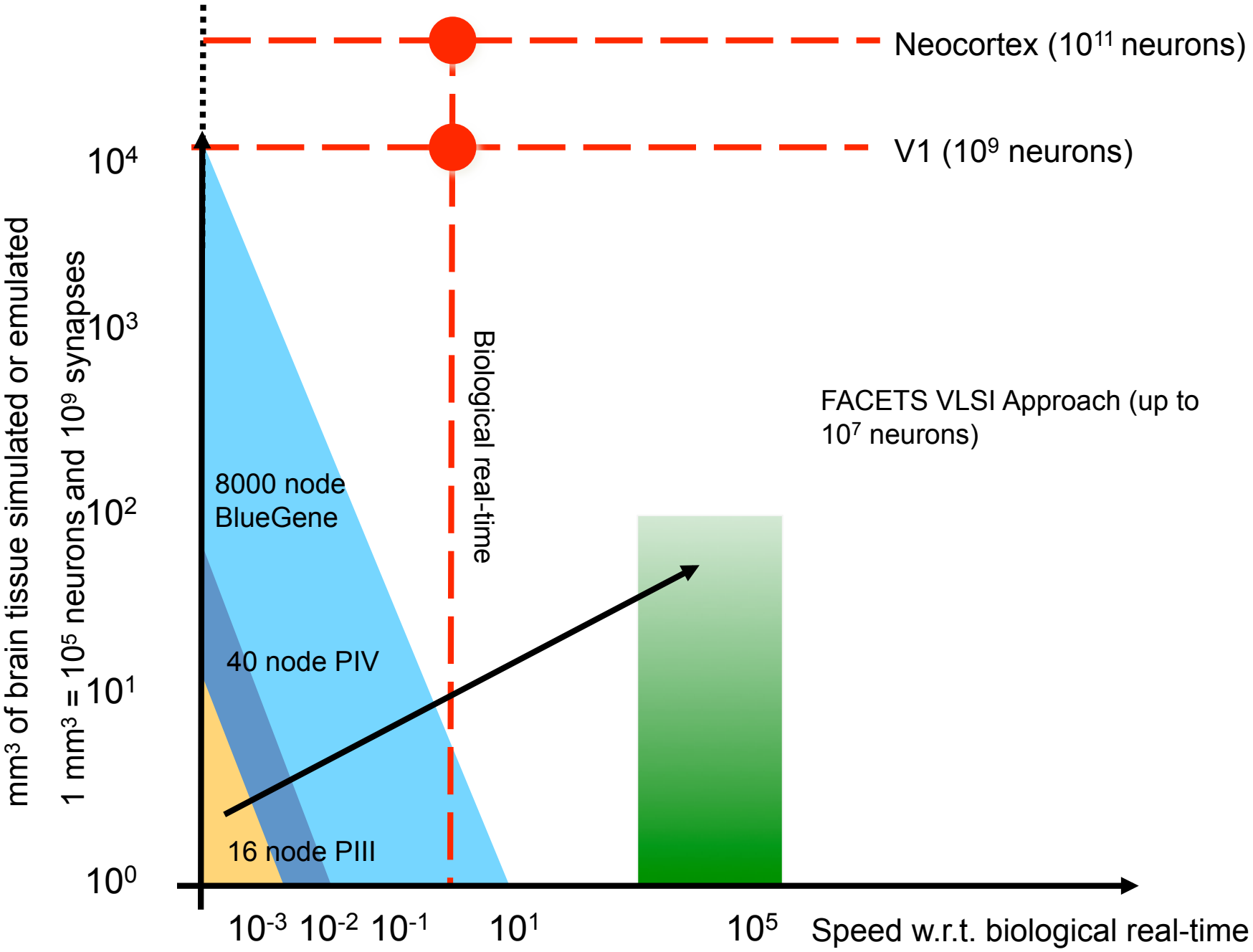


Configuration/Evaluation

(comparing connection matrix)



Complementarity Supercomputers vs. VLSI - Complexity vs. Speed



The Merits of fast (here : 10^5) neural VLSI

	Biology	Electronics
Precision of Spike based learning	10^{-04} s	10^{-09} s
Short Term Synaptic plasticity	10^{+00} s	10^{-05} s
Development	10^{+07} s	10^{+02} s (1.6 min)
Learning	10^{+09} s	10^{+04} s (2.8 h)
13 Orders of Magnitude		
Evolution	10^{+12} s	10^{+07} s (115 d)
19 Orders of Magnitude		

Access > 10 Orders of Magnitude in Time in an artificial System with a spatial complexity of >> 10^5 !?

Technological Challenges for CMOS NH

- Automated design and verification technologies for very large scale (50M synapses per wafer) massively parallel VLSI structures

- Distributed compact on-chip / on-wafer memory technologies for parameter storage and plasticity / learning / adaptation mechanisms (SRAM, current memories, floating gates), **NEW** : non-CMOS postprocessing : magnetic structures

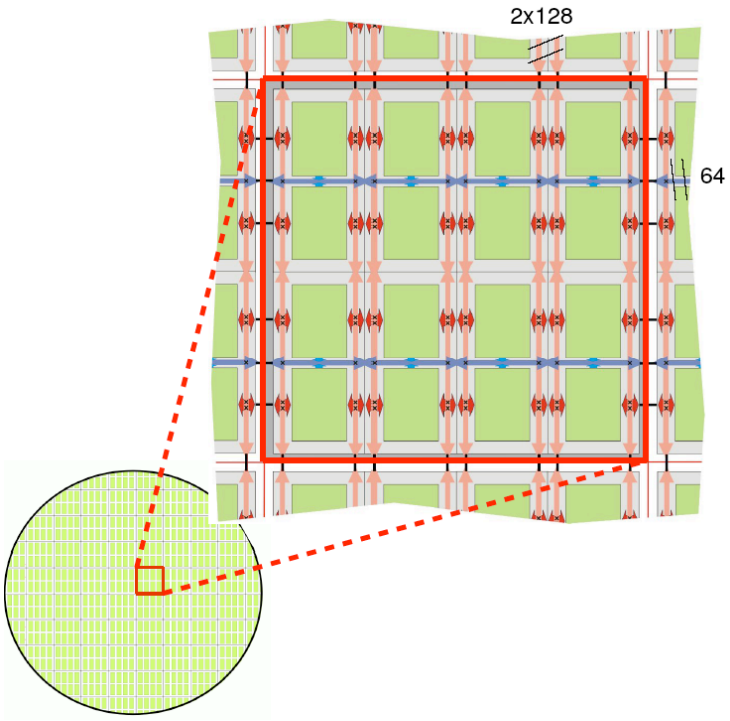
- On-wafer / inter-die horizontal high density connection technologies

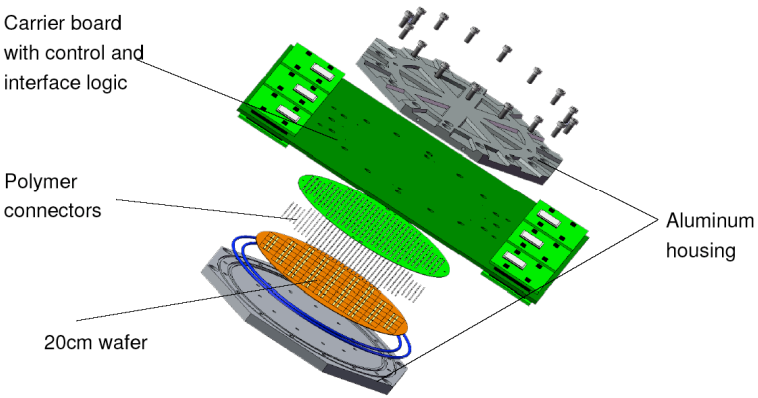
- Off-wafer vertical high density connection technologies

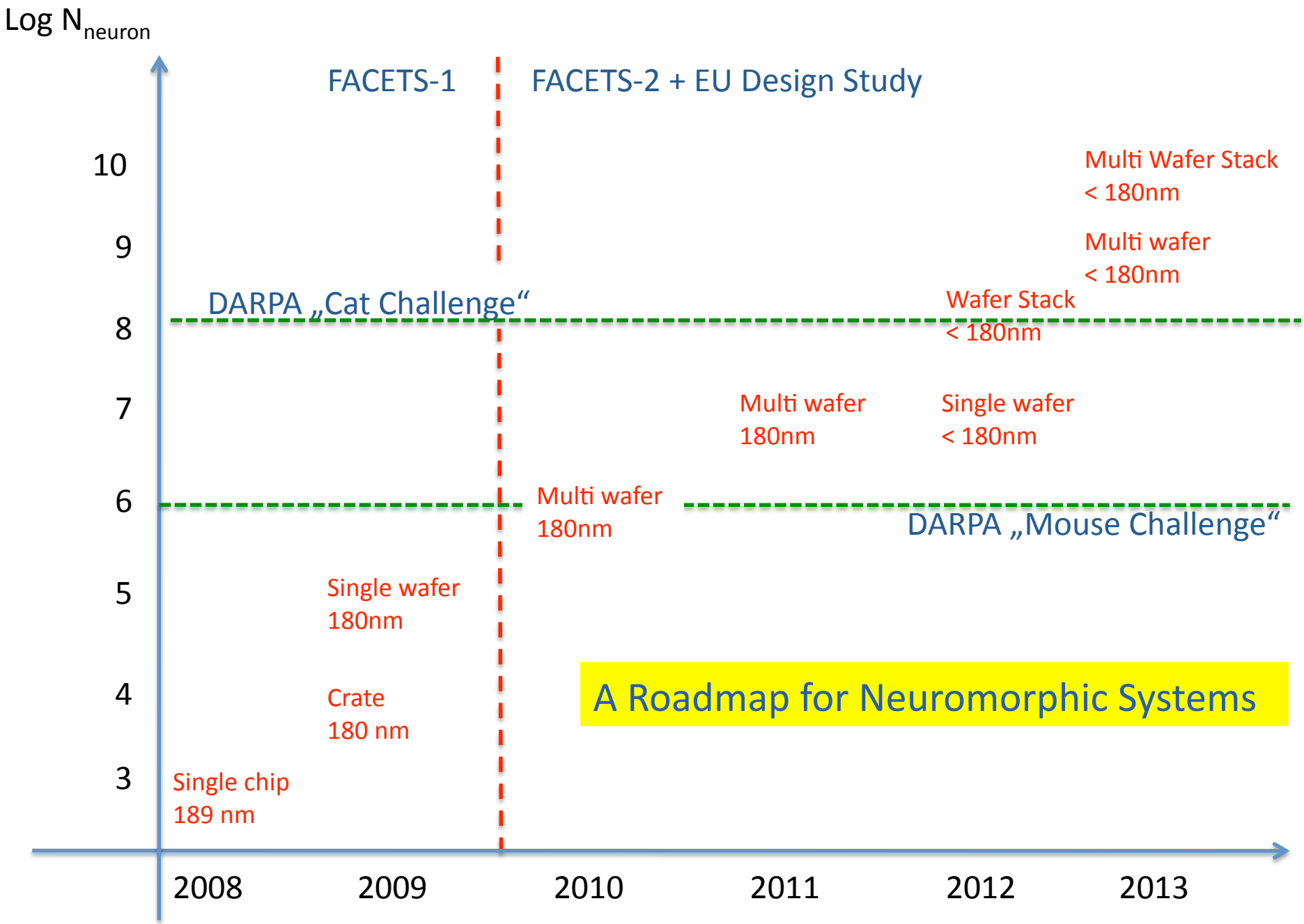
- **NEW :** Inter-wafer 3-dimensional connection technologies (wafer stacking) with low power analog design

- **NEW :** Access to deep sub-micron (< 100 nm) full wafer production for higher integration densities of neural cells

- Exploitation of the intrinsic fault- and mismatch-tolerance of neural circuits






Log N neuron
FACETS-1
FACETS-2 + EU Design Study
10
9
8
7
6
5
4
3
Multi Wafer Stack
< 180nm
Multi wafer
< 180nm
DARPA „Cat Challenge“
Wafer Stack
< 180nm
Multi wafer
180nm
Single wafer
< 180nm
Multi wafer
180nm
DARPA „Mouse Challenge“
Single wafer
180nm
Crate
180 nm
Single chip
189 nm
A Roadmap for Neuromorphic Systems
2008
2009
2010
2011
2012
2013

NH : Excellent opportunity to initiate the next HARDWARE revolution in information technology !?

Now is the time for more than toys : Build a LARGE SCALE NEUROMORPHIC DEMONSTRATOR as a joint (international) effort (follow examples from other scientific fields)“

Requires :

- Massively interdisciplinary approach (including graduate student training)
- Systematic effort towards brain mapping (morphology and function)
- Theory effort towards computational principles and physics of complex systems
- Infrastructure and capability for large scale hardware system development
- Access to deep-submicron technologies and cutting-edge connection technologies

Multi-Scale Funding approach (EU structure, to be discussed) :

Concepts :	Adressed in Integrated Projects (WE ARE HERE)
Technologies :	Adressed in Design Studies
Prototypes :	Adressed in Preparatory Phases
Systems :	Adressed in joint projects with industry

More ?

FACETS Project :
www.facets-project.org

FACETS Open Source Tool Group :
www.neuralensemble.org

Heidelberg Group :
www.kip.uni-heidelberg.de/visions