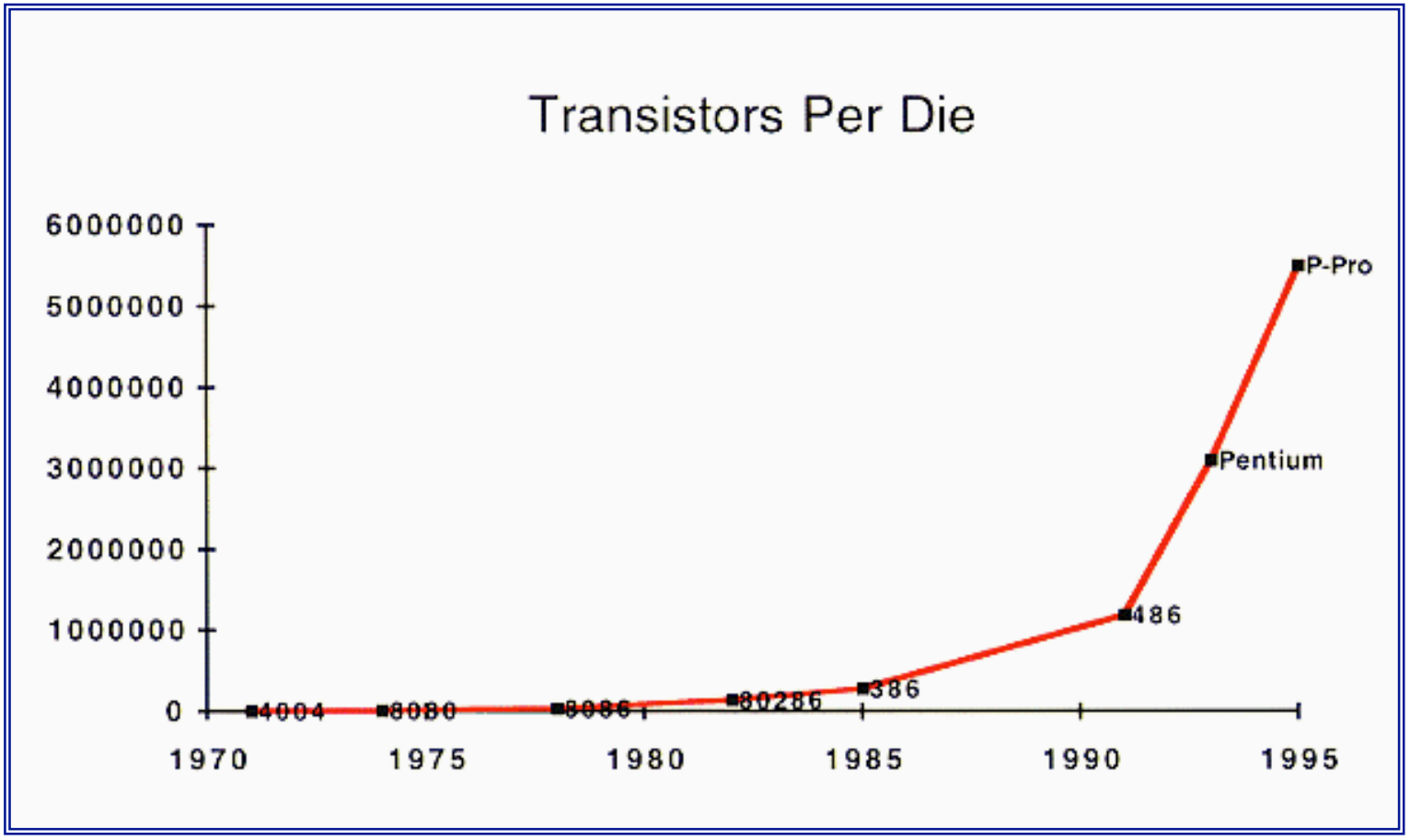
Neural circuits in mixed-signal VLSI
Towards new computing paradigms ?

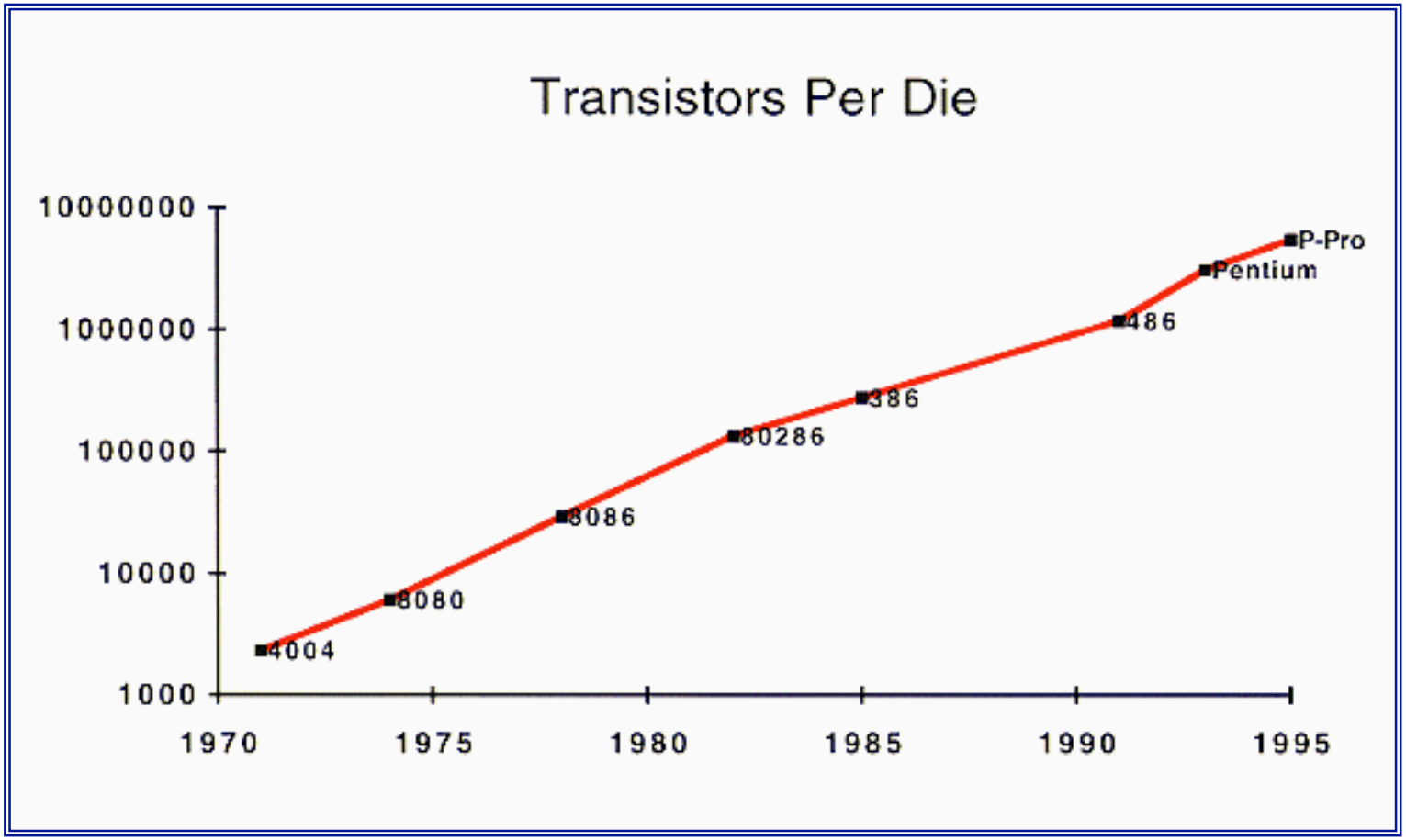
Seminar - Stockholm University - March 2007

Karlheinz Meier

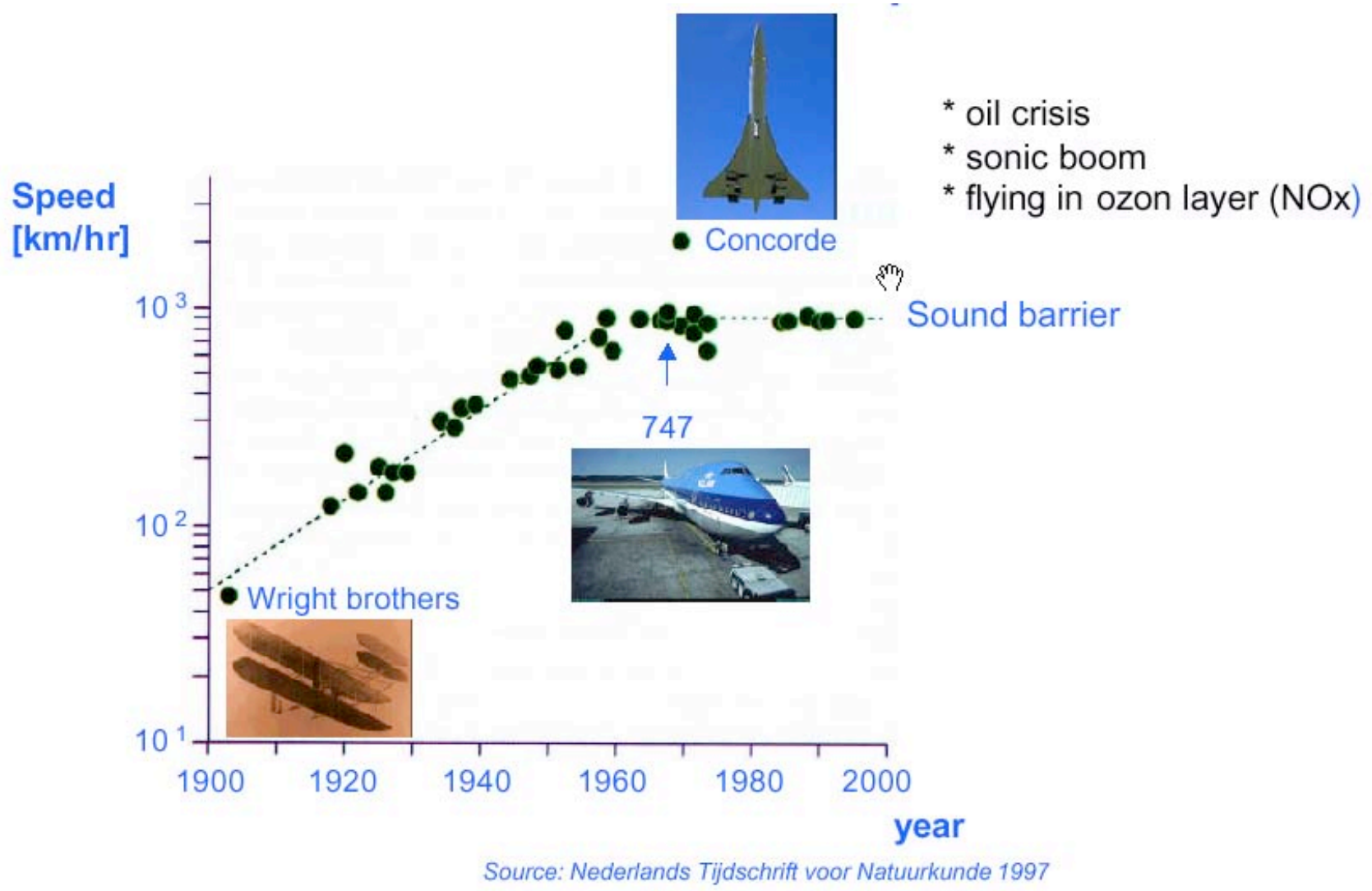
Kirchhoff-Institut für Physik

Ruprecht-Karls-Universität Heidelberg


Transistors Per Die
6000000
5000000
4000000
3000000
2000000
1000000
0
1970
1975
1980
1985
1990
1995
4004
8080
8086
80286
386
486
Pentium
P-Pro

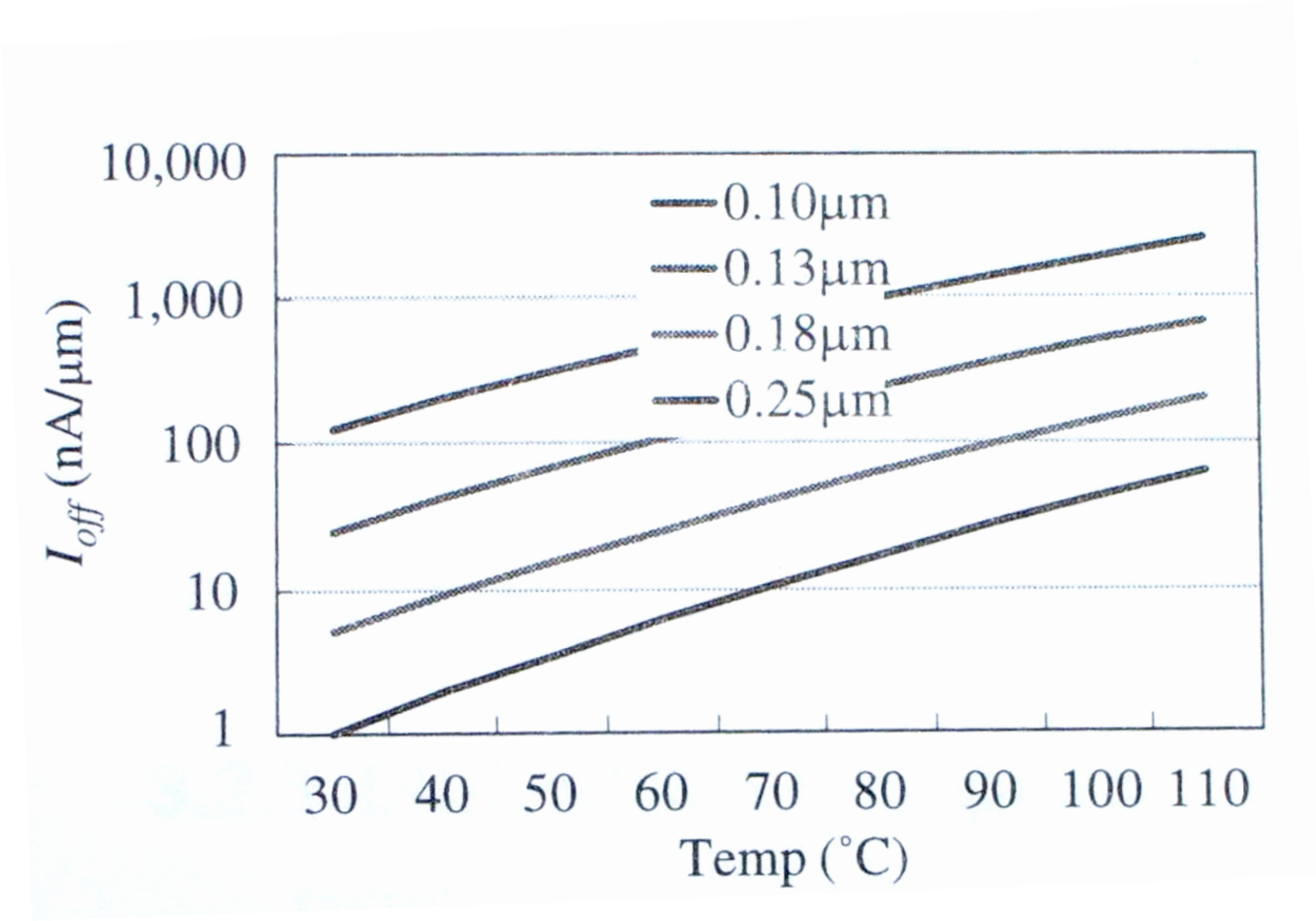

Transistors Per Die
10000000
1000000
100000
10000
1000
1970
1975
1980
1985
1990
1995
4004
8080
8086
80286
386
486
Pentium
P-Pro

A failing power law

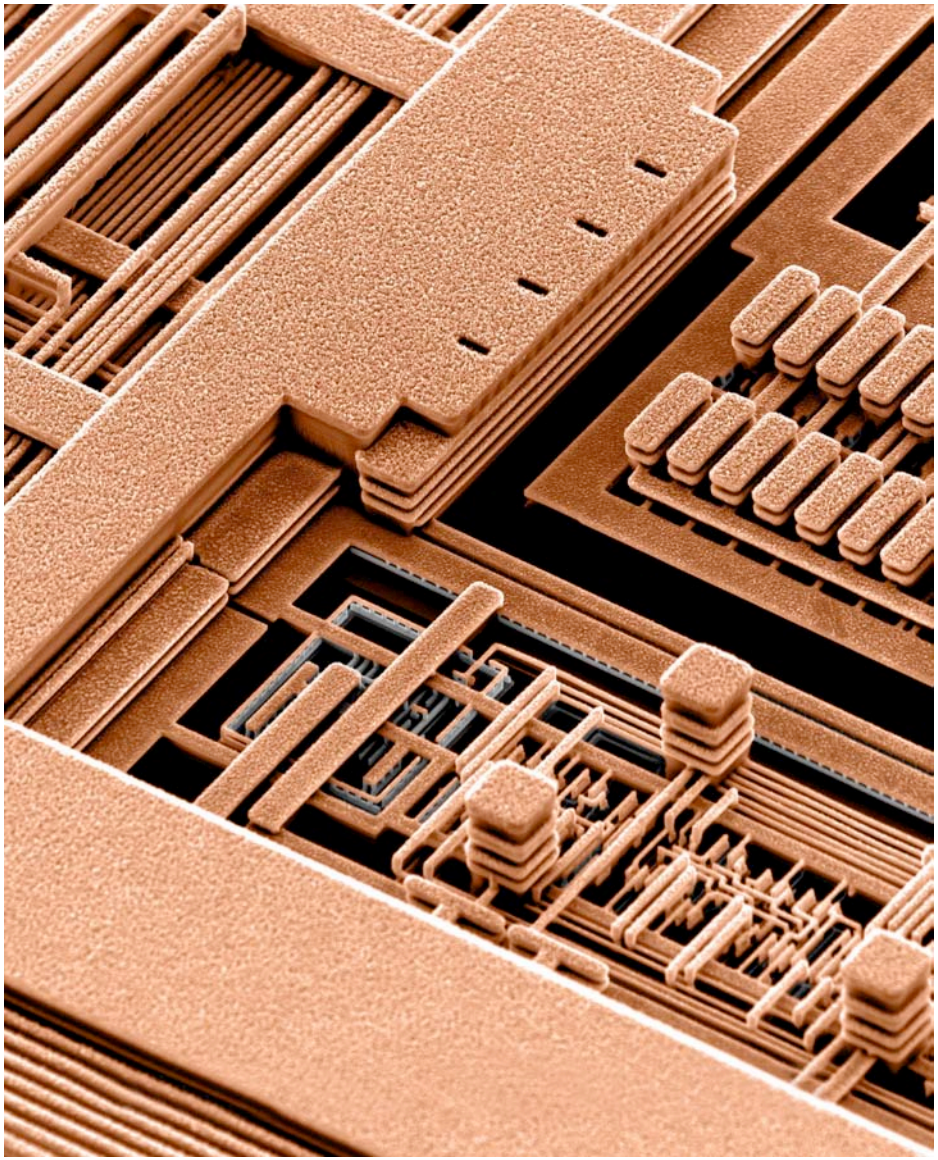


Source: Nederlands Tijdschrift voor Natuurkunde 1997

The local energy problem : Drastic increase of leakage currents



The structural energy problem



(almost) twodimensional system of „connecting wires"

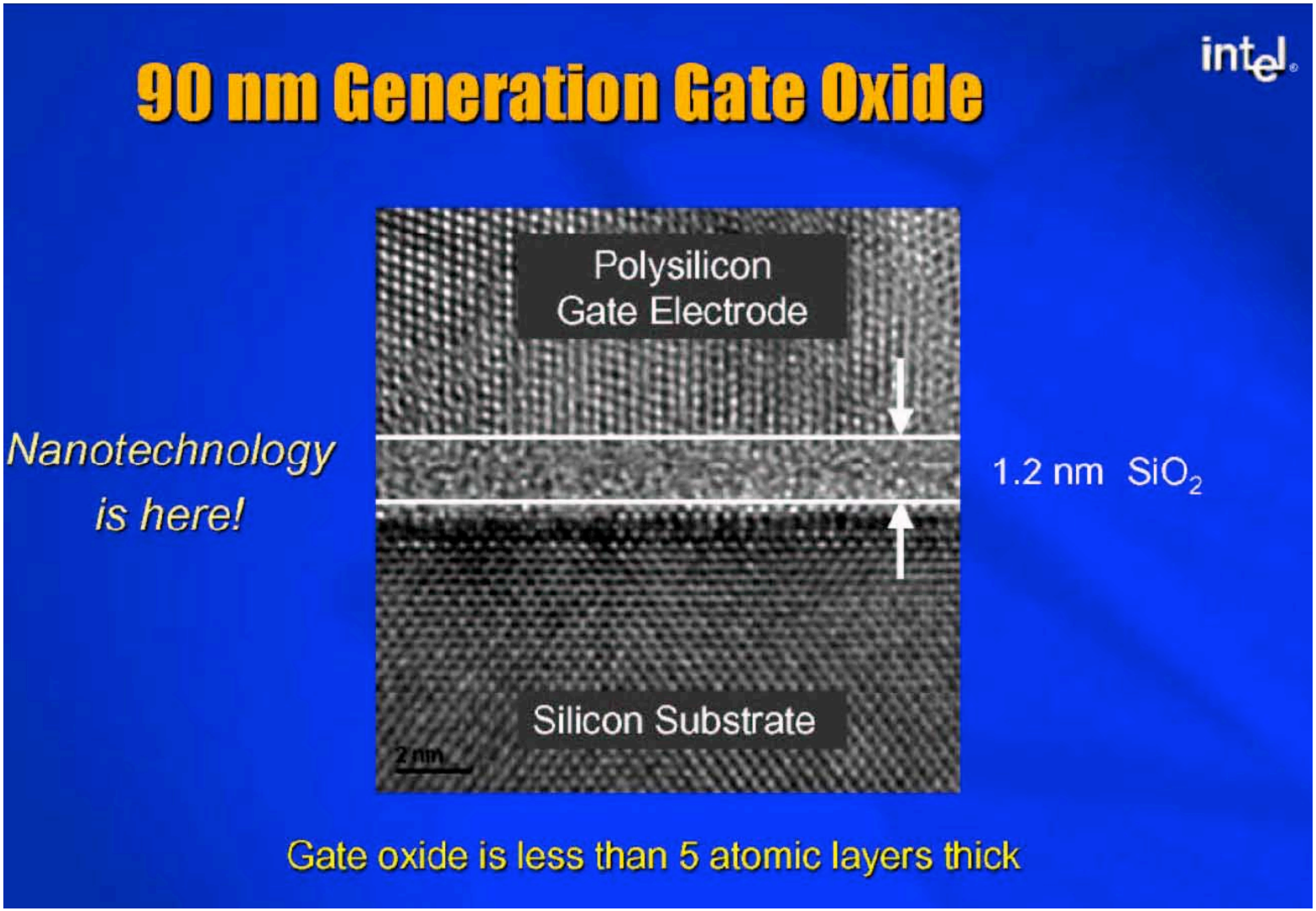
Spend typically 1000 times more enery in wires compared to transistors (as long as leakage currents are still small)

Energy Problem

Use this network to perform logic operations based on Boolean Algebra not well suited to calculated some real world probles (e.g. solve differential equations)

Architecture Problem

The Yield Problem (Lithography and Production)



INTEL Quote : „*Nanotechnology is here*" (90 nm Transistor)

4th NASA/DoD Workshop on Evolvable Hardware 2001

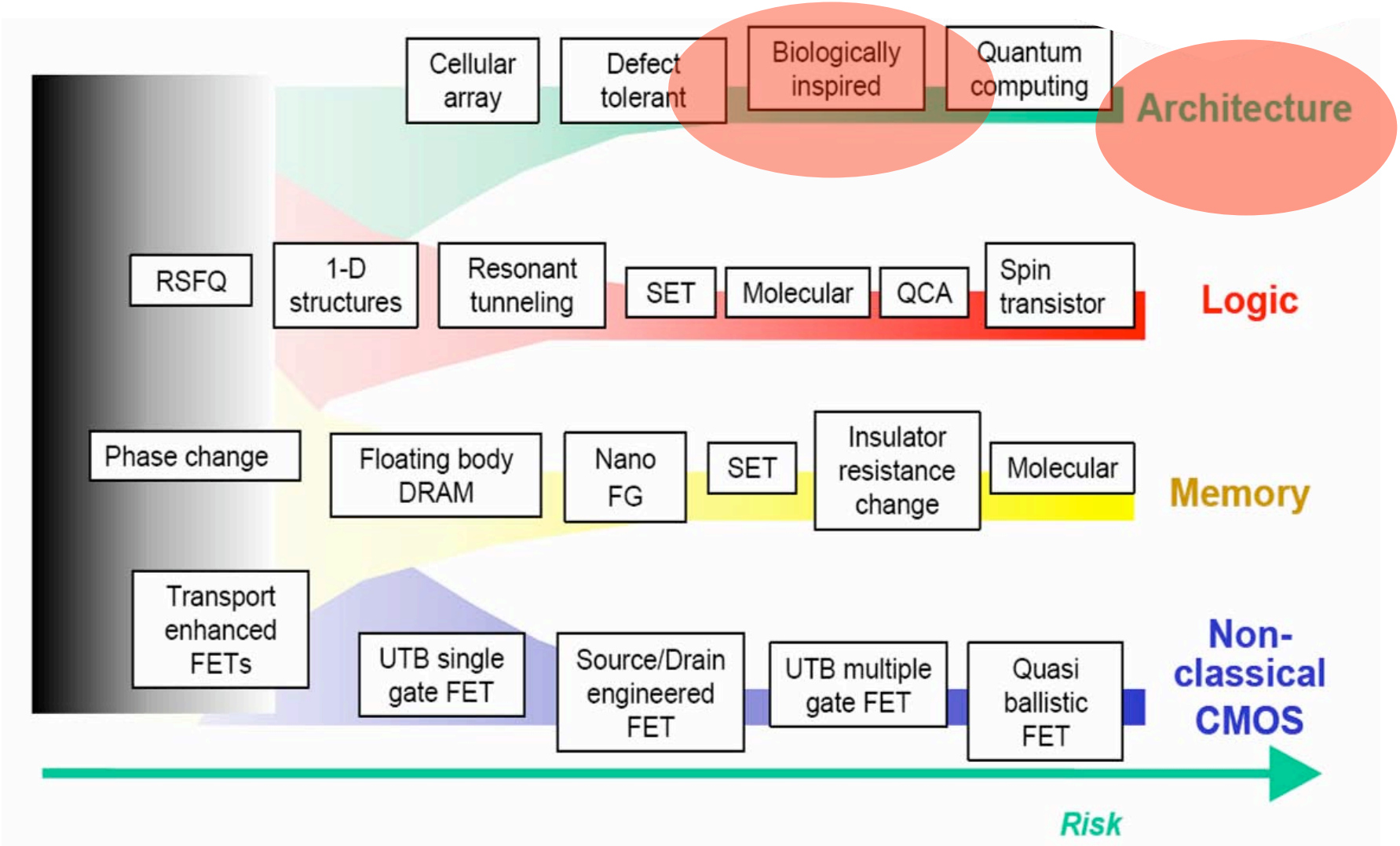
FPGA 2100

Feature Size	100 pm (Atomsize)
Wafer Diameter	2 m
Chipsize	5 cm x 5 cm
Transistors	$2.5 \cdot 10^{14}$ (0.1 x Synapses in Brain)
Logic Gates	1 Billion
Clock Speed	60 GHz
Power Dissipation	200 kW (5 families)
Foundry Investment	1 Tera Dollar
Costs for 1 set of lithographic masks	1 Giga Dollar
Application Development	5 Centuries

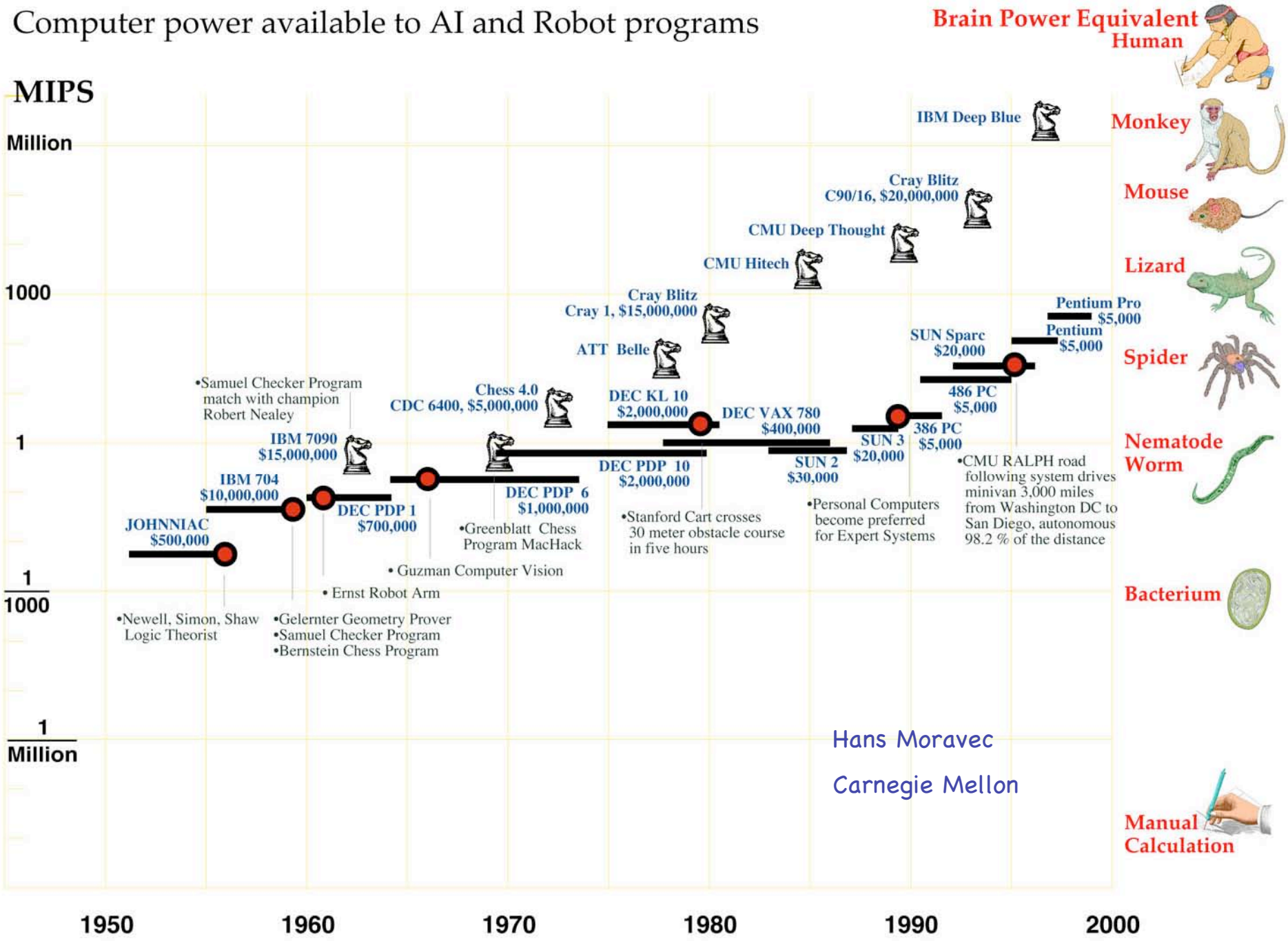
Problem Summary

➢ The local energy problem (leakage currents)	DEVICE
➢ The yield problem	DEVICE
➢ The structural energy problem (connections)	ARCHITECTURE
➢ The design problem (testability, simulatability)	ARCHITECTURE

International Technology Roadmap for Semiconductors (ITRS)

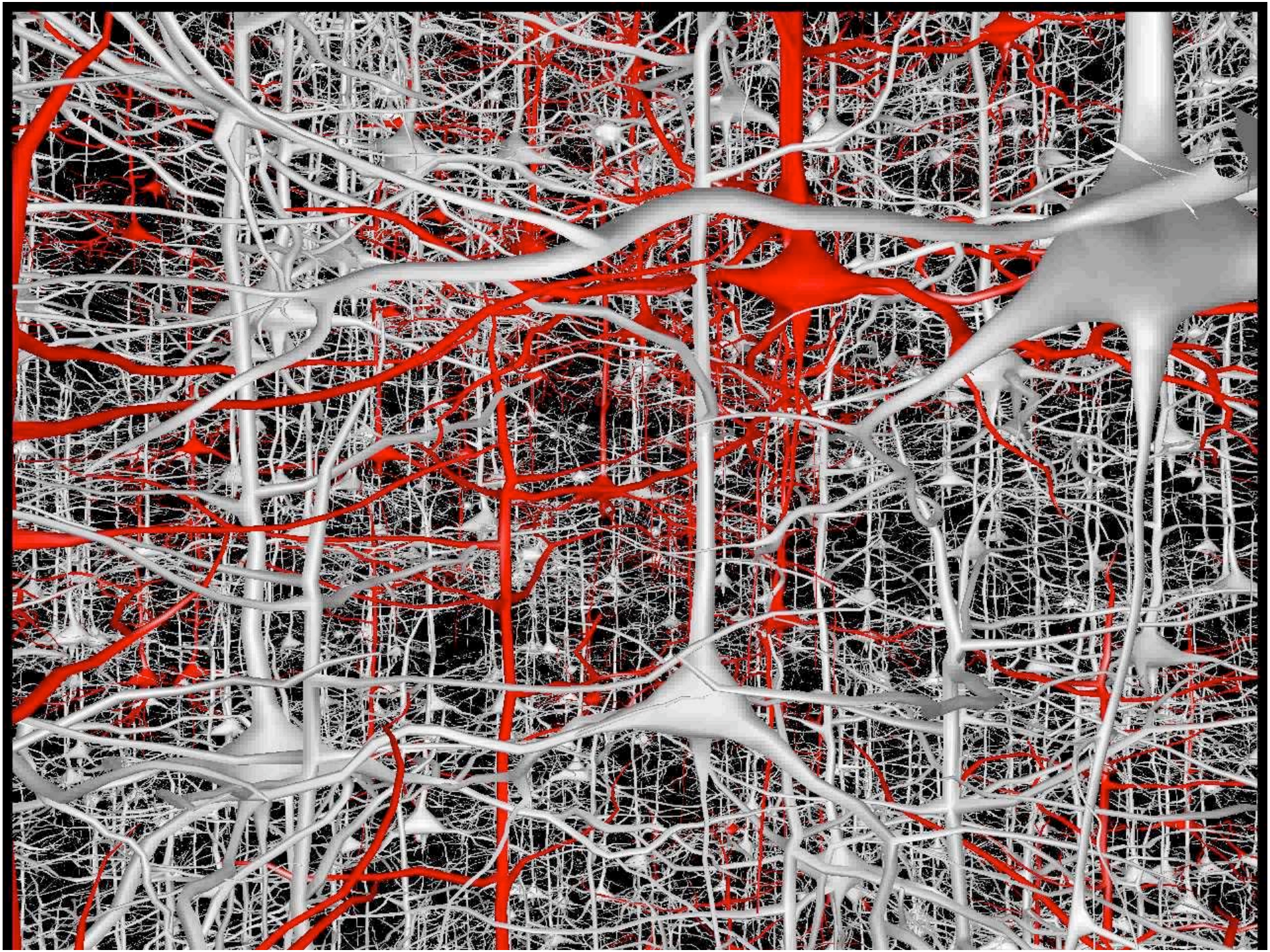


Computer power available to AI and Robot programs

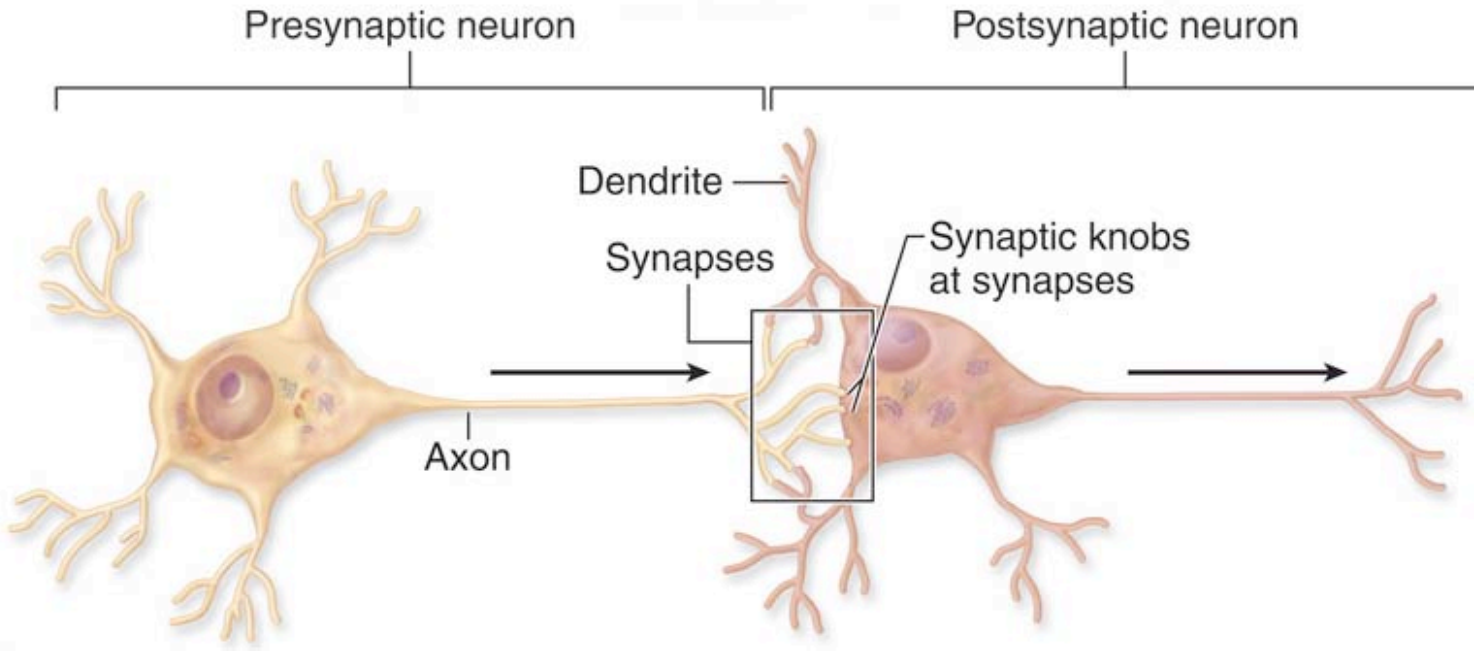


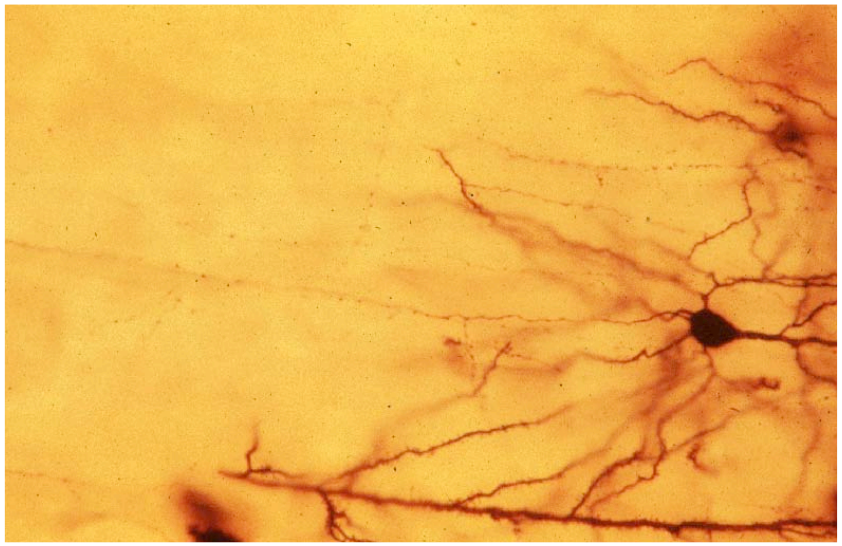
Hans Moravec

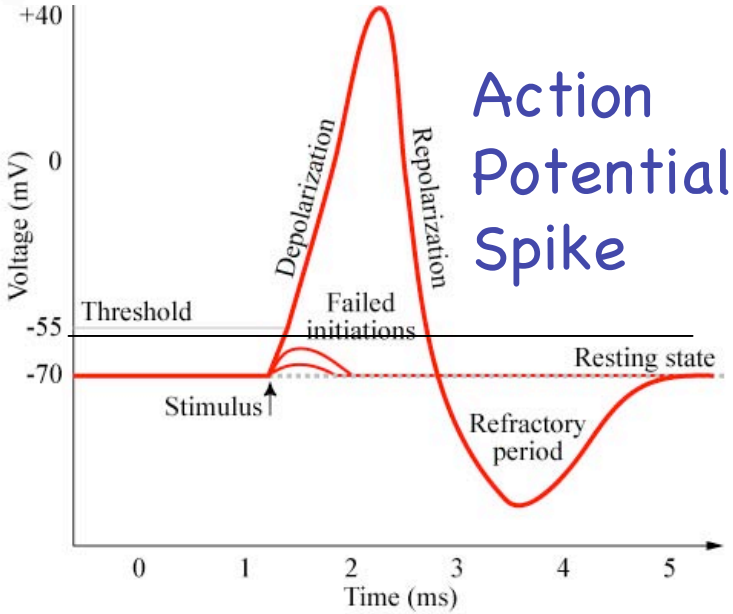
Carnegie Mellon



Real Biology : Neurons - Synapses - Dendrites - Spikes

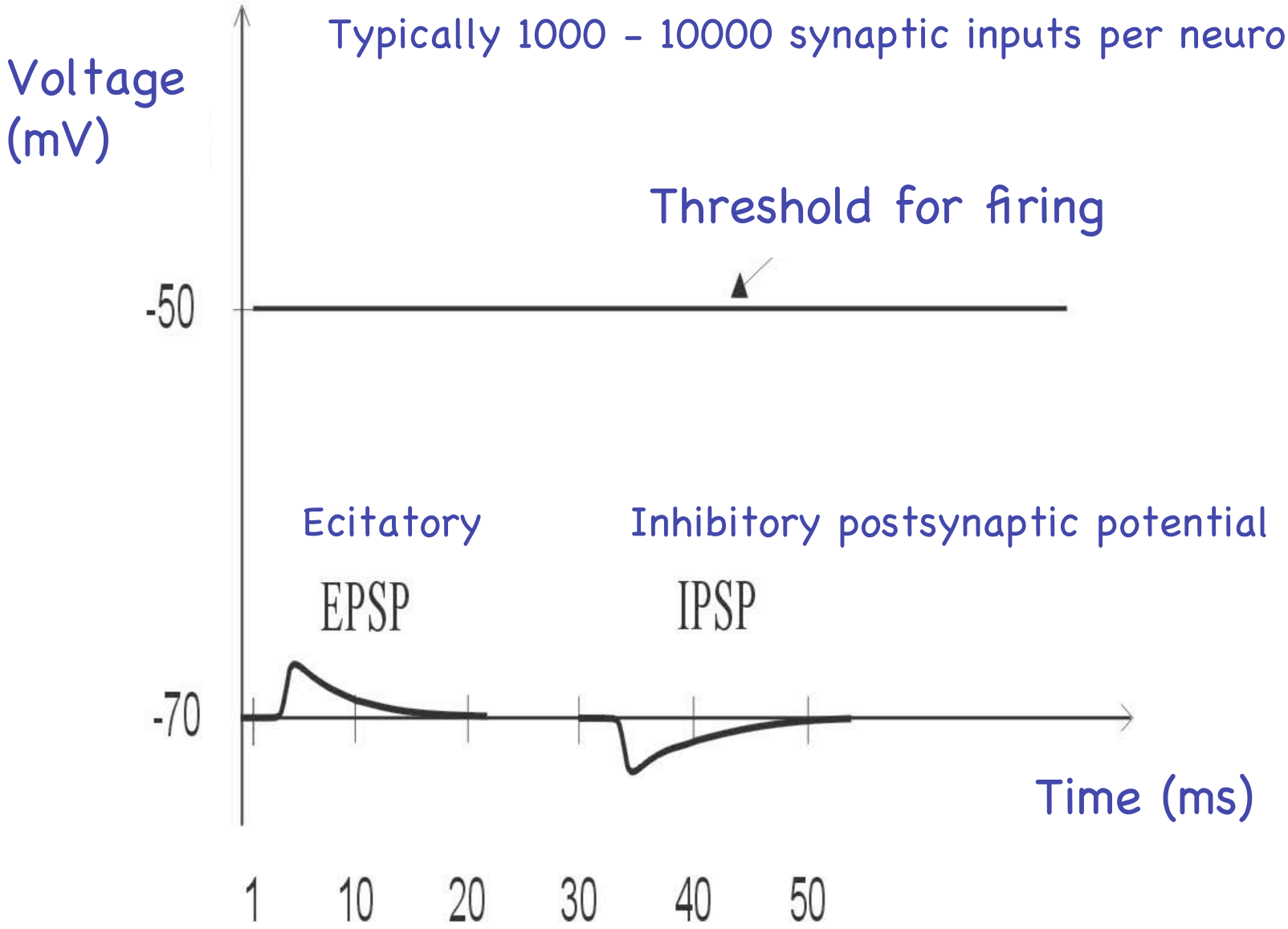






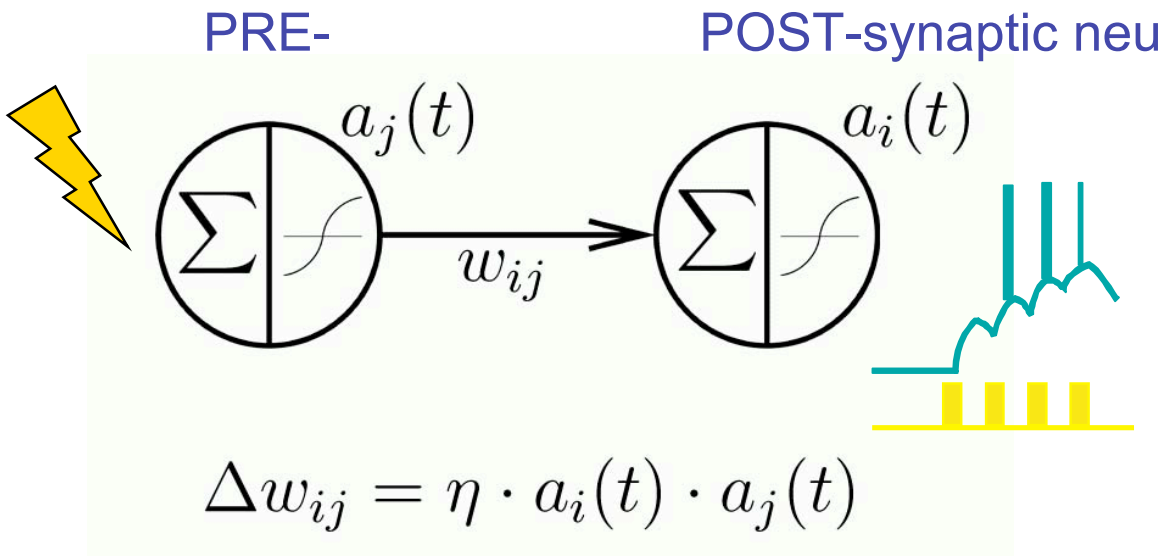
Action Potential = Spike

Excitation and Inhibition



More than „Spikes“ - „Plasticity“ and „local learning“

Hebbian Learning : The strength of a synaptic link changes, if pre- and post-synaptic timing are close (*un-supervised, local learning*)

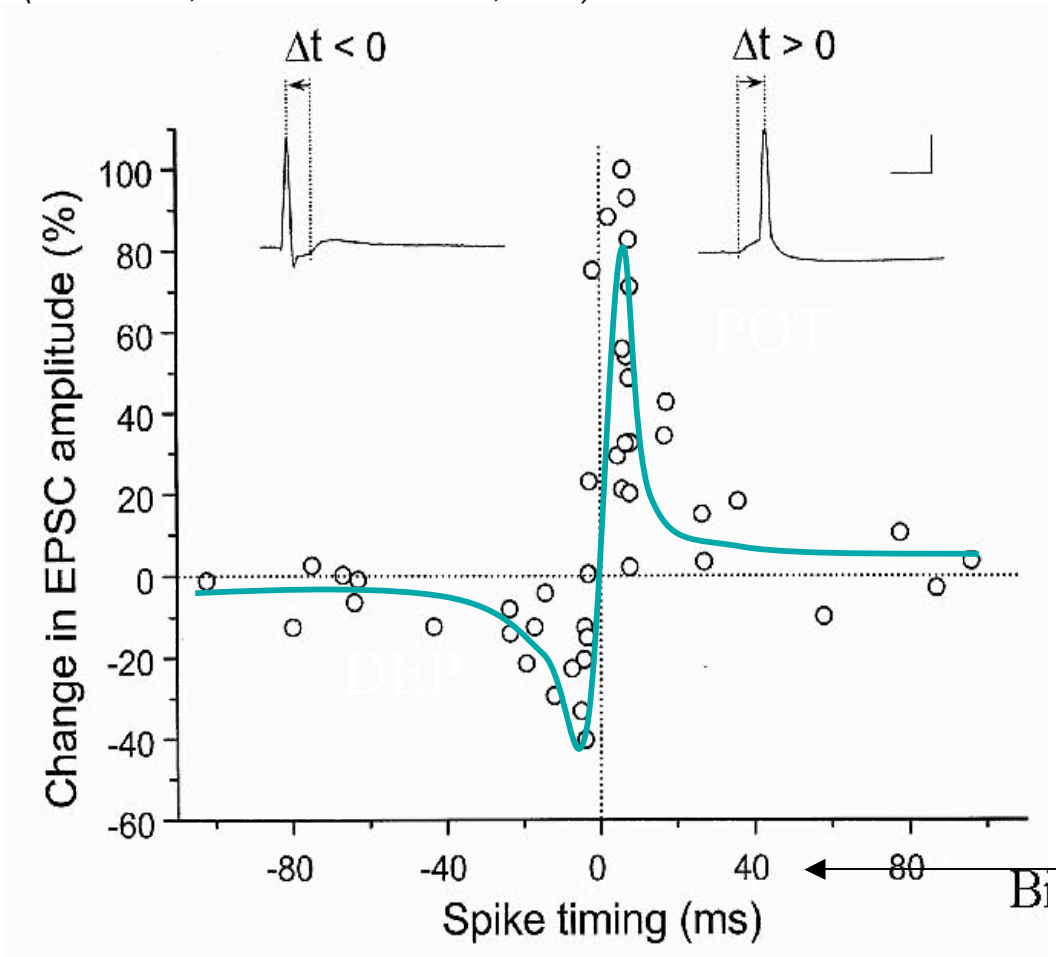


- **CLOSE** (µs, ms, s, ...) ?
- **WHAT** kind of change ?

Experimental Evidence for „Spike-Time-Dependent-Plasticity (STDP)“

In vivo intracellular recording (Adult Visual Cortex)

(Bi and Poo, Ann. Rev. Neurosci., 2001)



STDP

Extremly strong time dependence of facilitytion or depression of synaptic strength

Neural circuits require asynchronous MILLISECOND timing for long term learning !

AFTER - BEFORE

„synaptic spike“

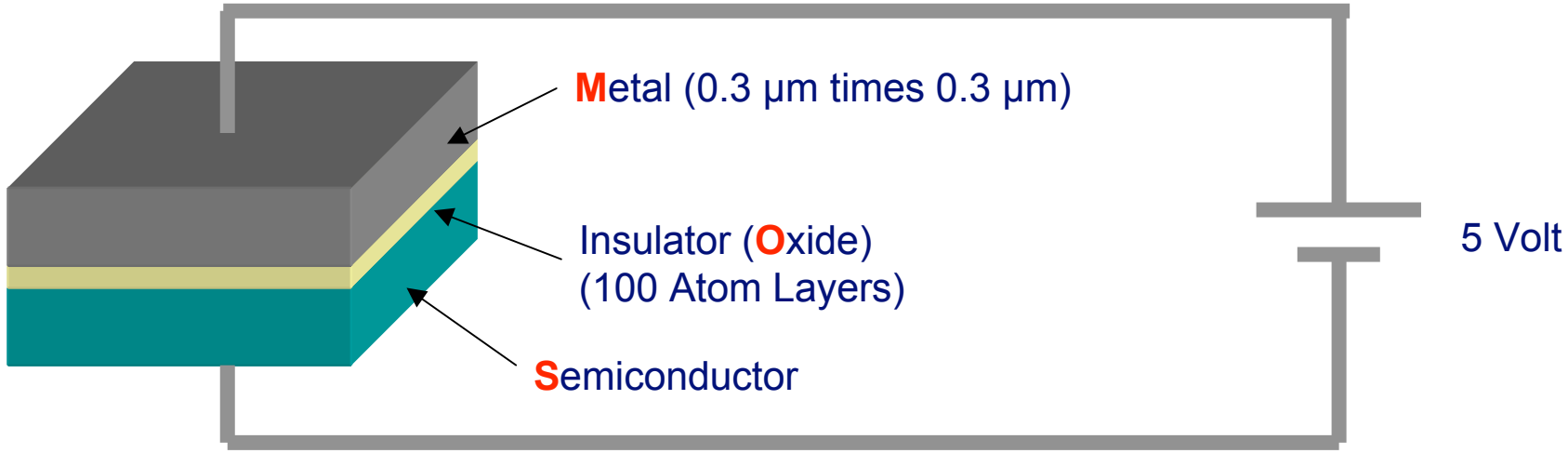
Modeling approaches

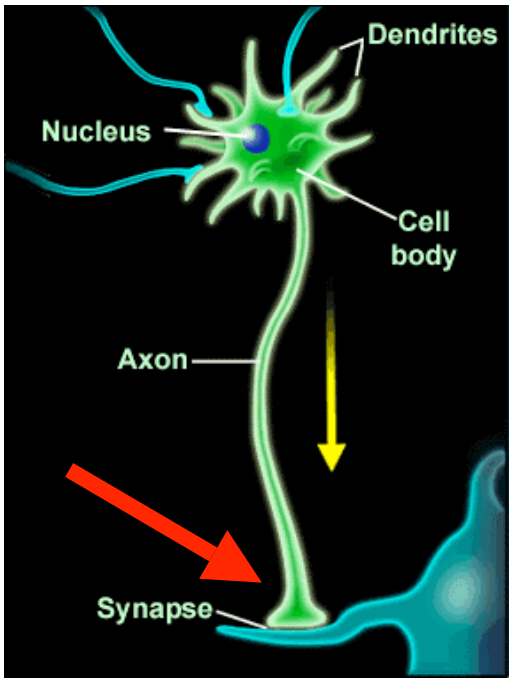
starting point: mathematical description

methods:

- **analytical treatment**
  proof of general properties and limits
- **numerical solution (general purpose or FPGA based simulation)**
  flexibility, parallel objects not obvious
- **physical model**
  artificial nervous system, artificial parallel object = biological objects
- *biological model*
  "custom-made biological nervous system"

Electronics vs. Biology on the device level - Not a big difference !





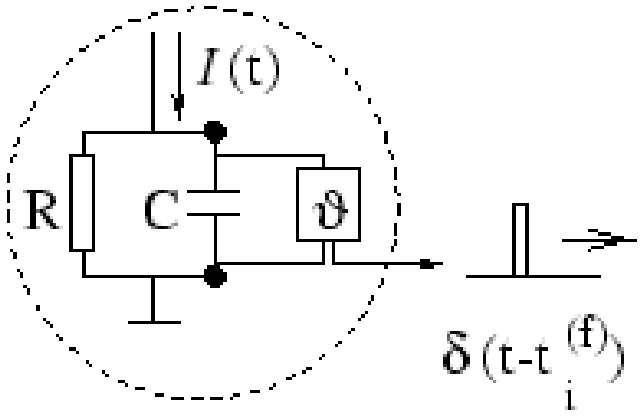
„Switching“ of a MOS element and a synapse :

Energy in BOTH cases approximately 1 fJ

„Energy“ ➔ Heat

Heat/Time ➔ Power Dissipation (kW)

A somewhat biological neuron model (integrate-and-fire, IF)



$I(t) = u(t)/R + C \cdot du/dt$

u(t) : membrane potential

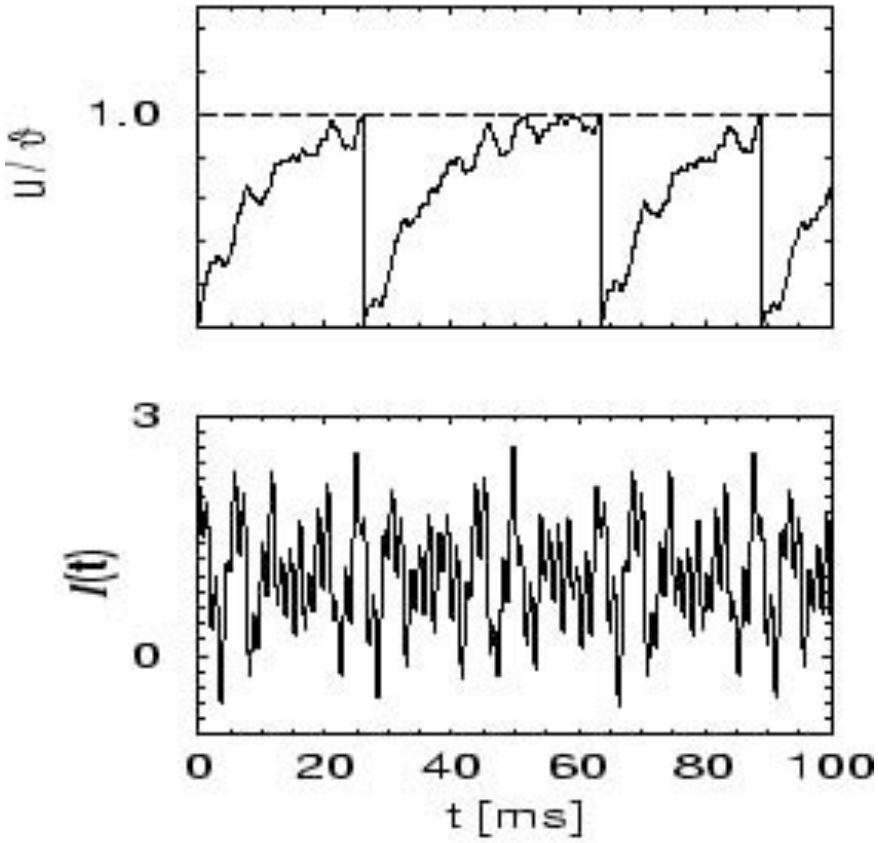
I(t) : input current

Using RC time-constant :

$\tau_m \cdot du/dt = - u(t) + R \cdot I(t)$

In addition :

„Spike-Generation“ and „Reset“ of u if u = ϑ



random current at input

The **FACETS** Consortium

Fast **A**nalog **C**omputing with **E**mergent **T**ransient **S**tates

U Bordeaux, CNRS (Gif-sur-Yvette and Marseille), U Debrecen, TU Dresden, U Freiburg, TU Graz, U Heidelberg *, EPFL Lausanne, Funetics S.a.r.l. Lausanne, U London, U Plymouth, INRIA Sophia-Antipolis, KTH Stockholm

*Coordinator

An Integrated Project in the 6th Framework Programme

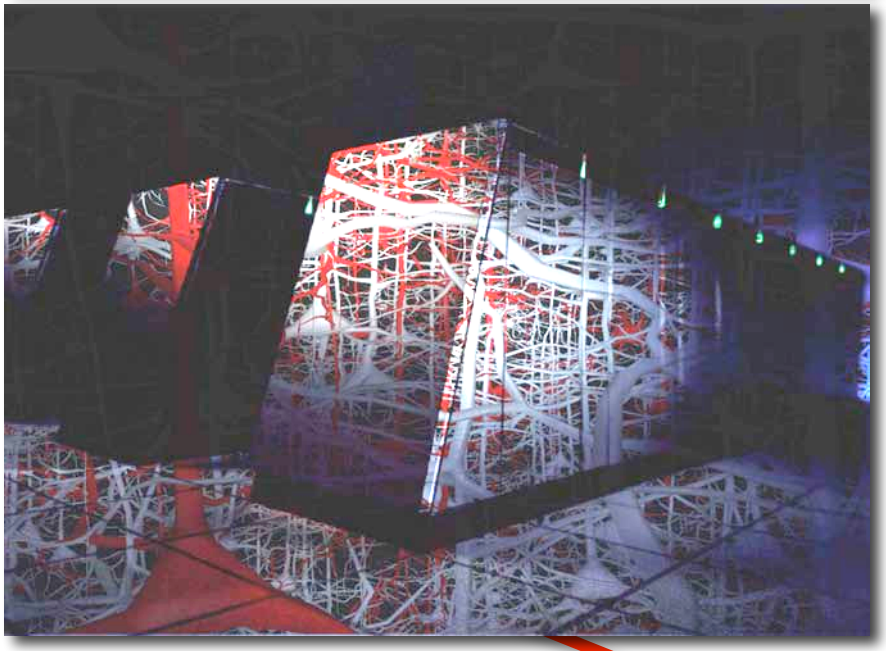
Information Society Technology - Future Emergent Technologies

FP6-2004-IST-FETPI Project Reference 15879

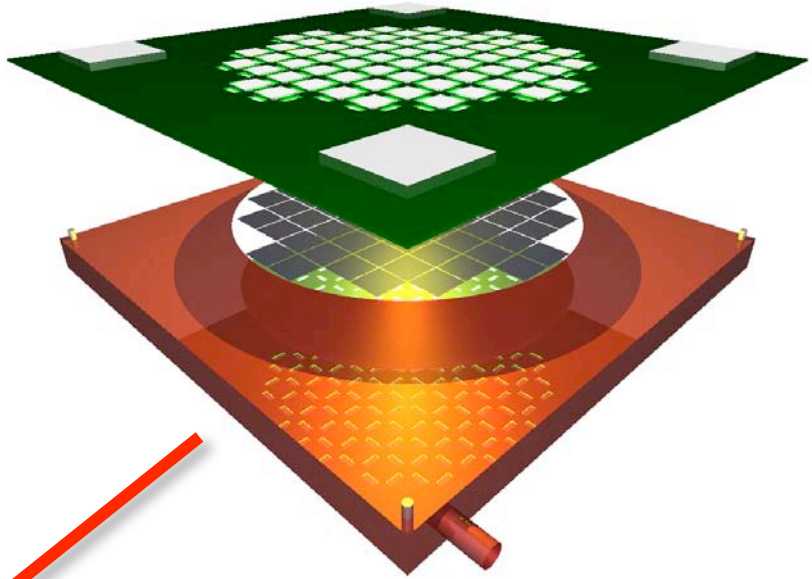
FACETS : Basic Idea, methodological approach and goals

Experimental Biology : Reverse Engineered Structural and Functional Blueprint of the Neocortical Microcircuit

Modelling : Simulation of Microcircuits with detailed cell modells



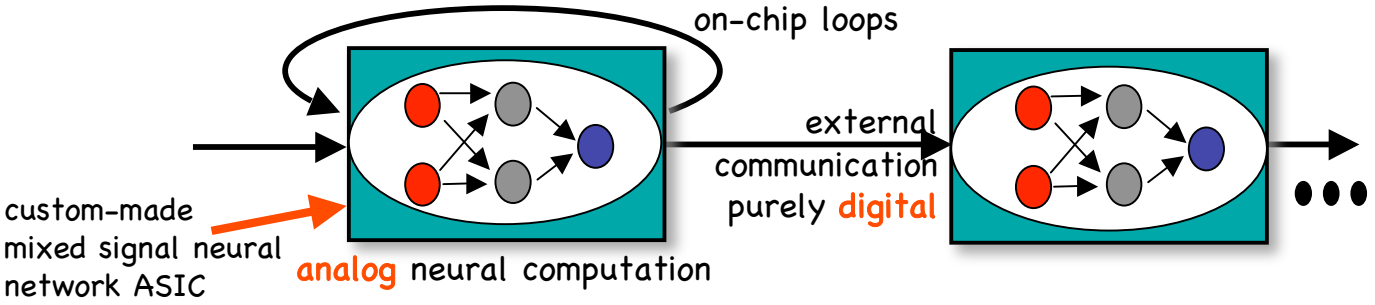
Circuits : Emulation in analog, fault-tolerant, scalable, high speed VLSI



Common Goal : Study non-classical universal computing solutions

Concept : VLSI mixed-signal emulation

- neural network implemented in custom-made neural network ASICs
- hardware mixed-signal approach: local analog computation combined with high-speed state-of-the-art 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
- use 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

A „scalable, very-large-scale, mixed-signal, massively-parallel, high-speed, flexible, biologically plausible" neural computation system

scalable : Clearly defined stackable components, no size or distance-dependent communication quality, digital long range communication

very-large-scale : Possibility to approach the numerical complexity of the visual cortex

mixed-signal : analog computational elements (neurons, synapses), digital (event-based) medium and long-range communication

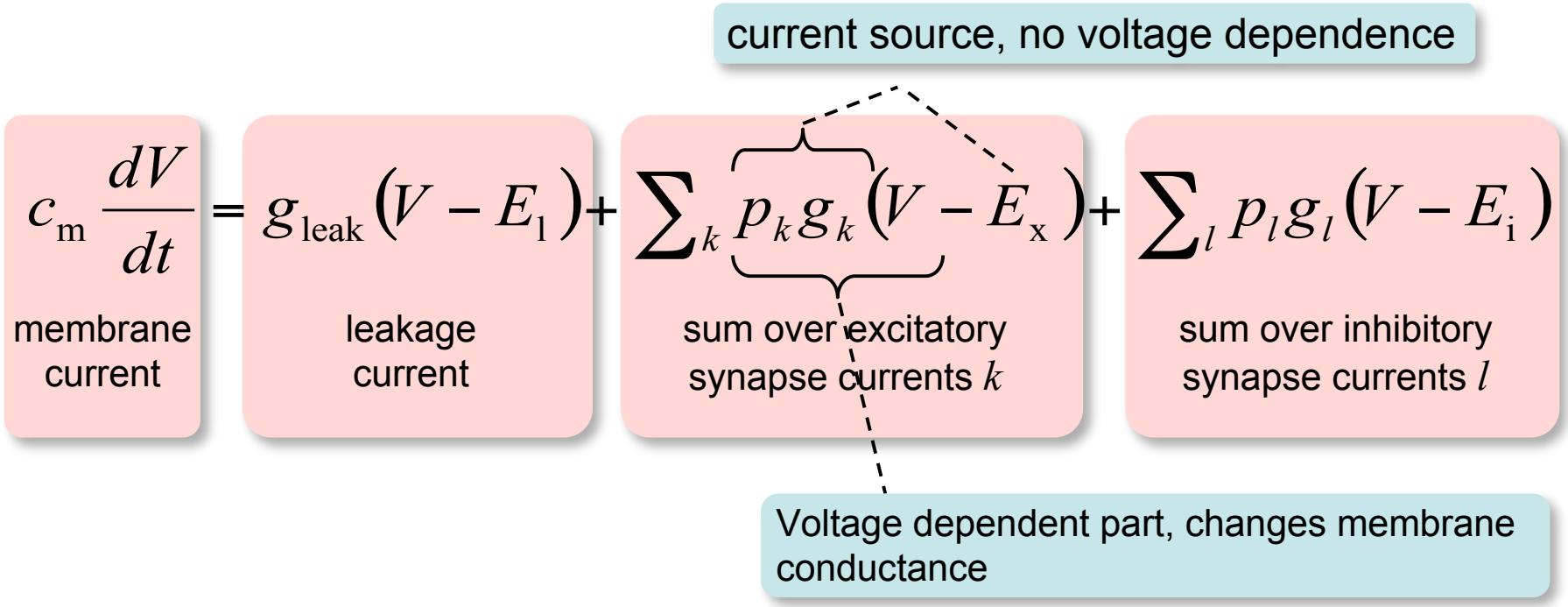
massively parallel : Obviously ....

high-speed : Time compression of factor 100.000 beyond biological real-time. Possibility to study all biologically relevant dynamics in convenient laboratory time scales. 1 ms becomes 10 ns, 1 year becomes 5 minutes.

flexible : user configuration of neuron and synapse parameters, implementation of diversity, programmable medium and long range connectivity, on chip-storage and update of synaptic weights.

biologically plausible : based on inputs from biological measurements

Stage 1 FACETS Model : Conductance-based Network Model



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_m / g_{total} is time-dependent

Specifications of the Stage 1 FACETS VLSI Model

- 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

Chip Specifications

- technology: UMC 180 nm, 6 metal layers, 1 polysilicon layer
- chip size: 5 x 5 mm^2 (Europractice constraints)
- 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

Digital Network Model

Event based communication between different model neurons

Two network models transport events from neuron *x* to neuron *y*:

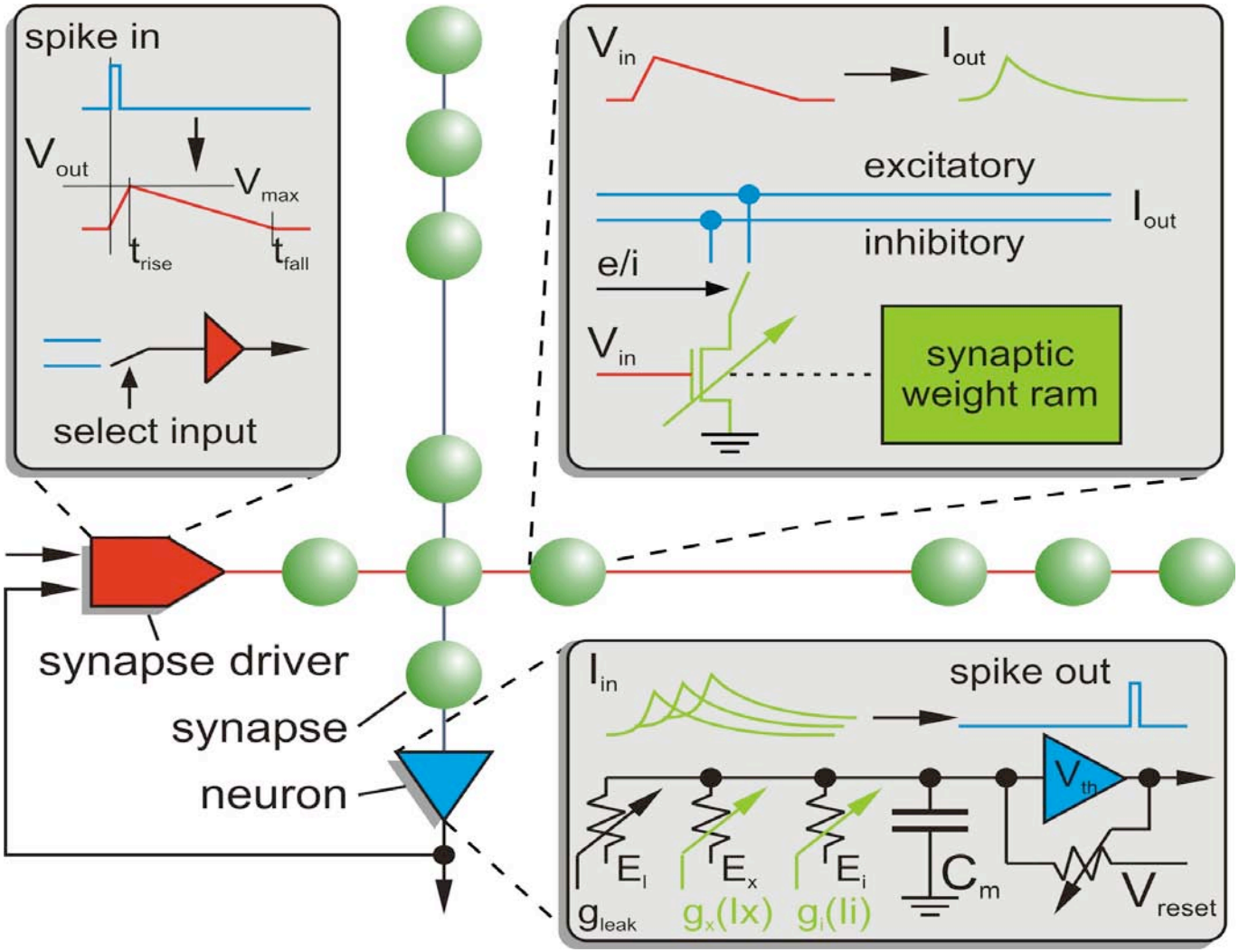
1. on-chip:

- dedicated electrical connections transport the output from neuron *x* to the input of neuron *y*
- continuous time
- constant delay

2. off-chip

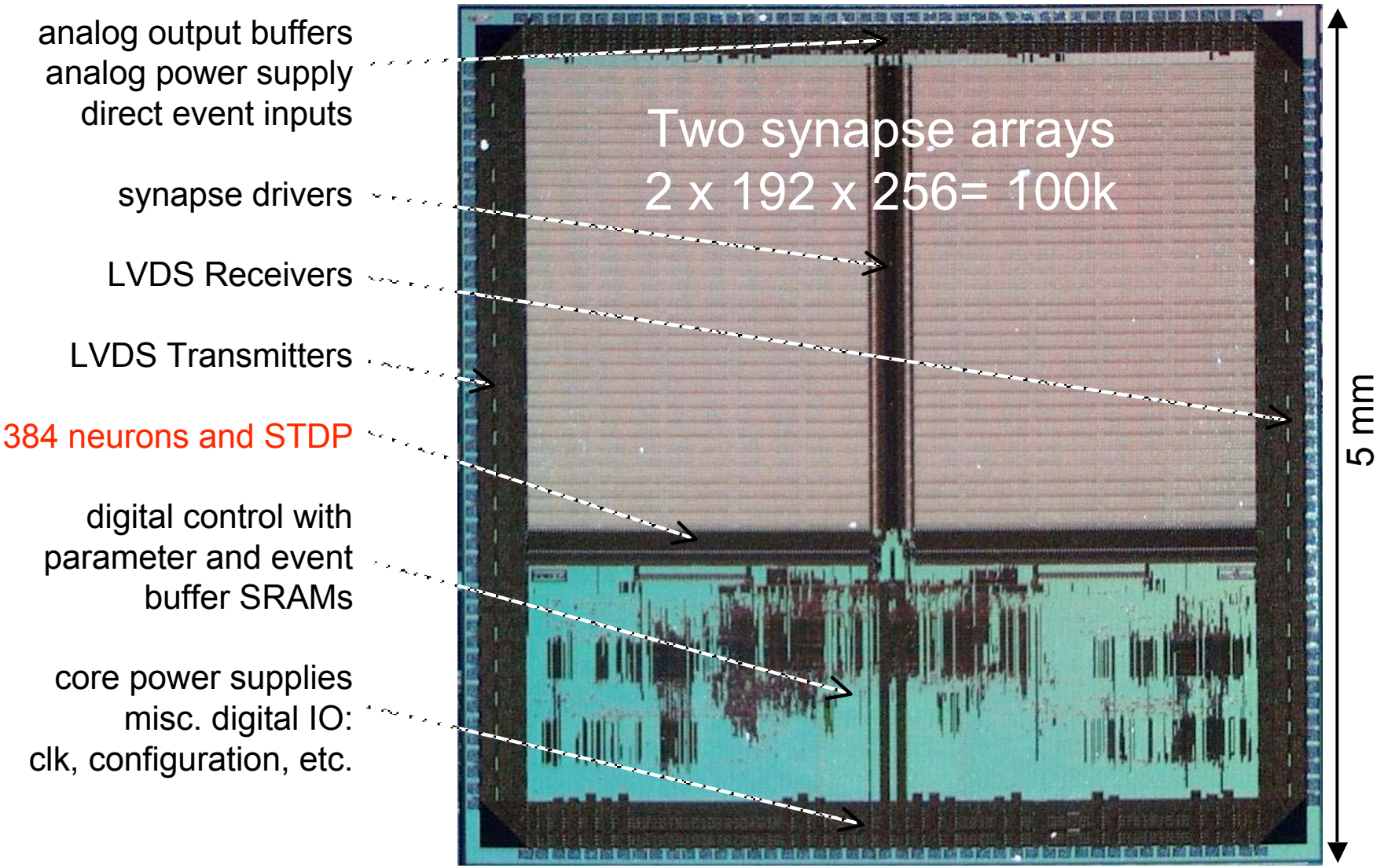
- event based external interface
- digitized event time (150ps resolution)
- variable delay, can be compensated by external routing logic
- two bi-directional 800 MByte/s links
- Hypertransport physical layer specification
- transport protocol allows for daisy-chaining of multiple chips

Overview of the Network Implementation

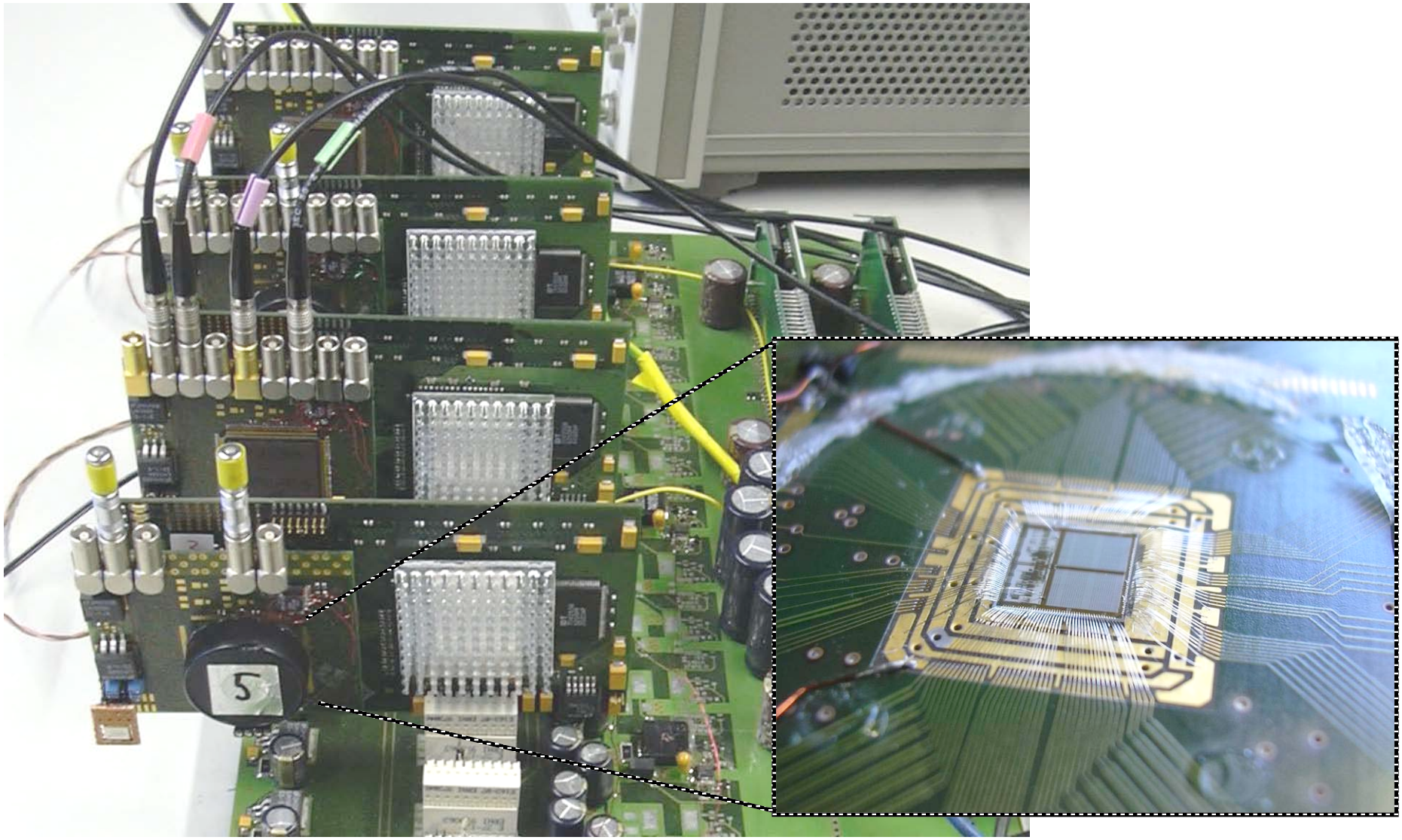


A New VLSI Model of Neural Microcircuits Including Spike Time Dependent Plasticity, Johannes Schemmel, Karlheinz Meier, Eilif Muller, Proceedings of the 2004 International Joint Conference on Neural Networks (IJCNN'04), IEEE Press, pp. 1711-1716, 2004

Stage I Analog Neural Network Chip



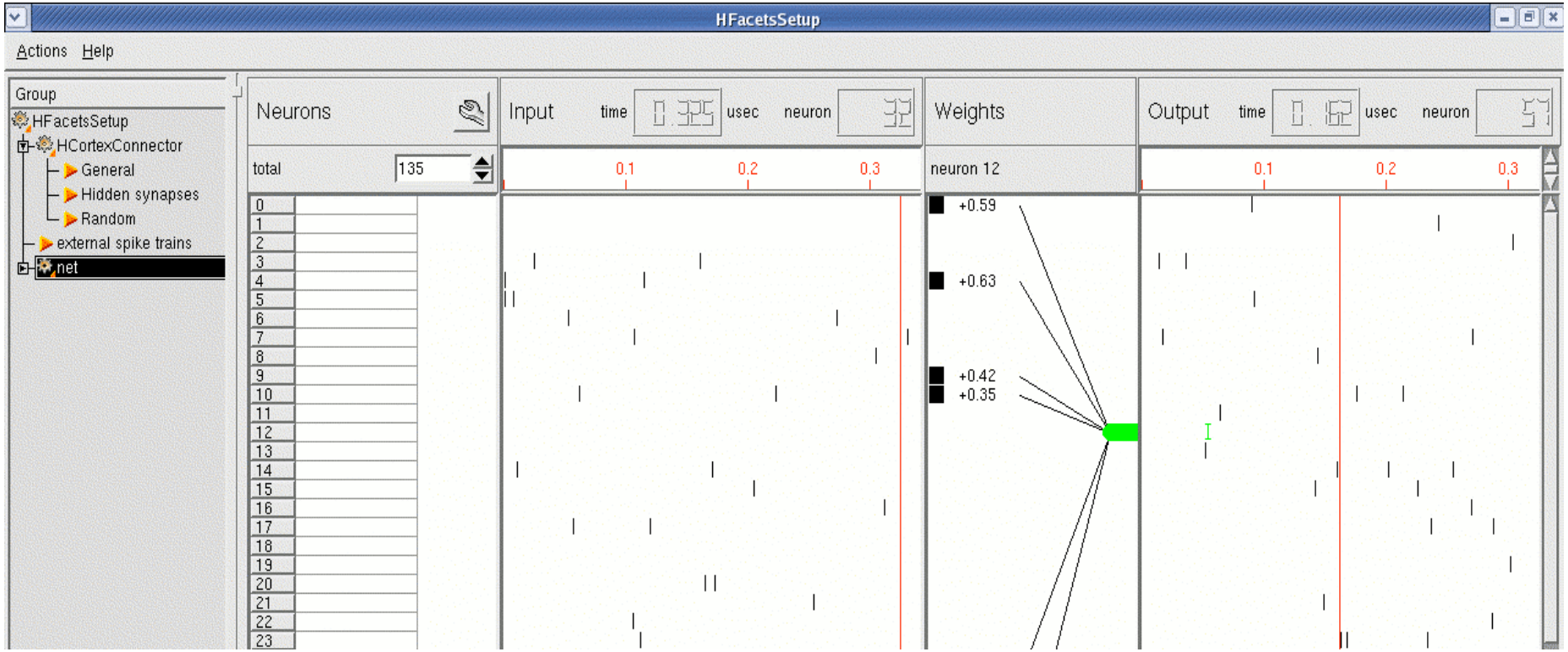
Experimental Setup



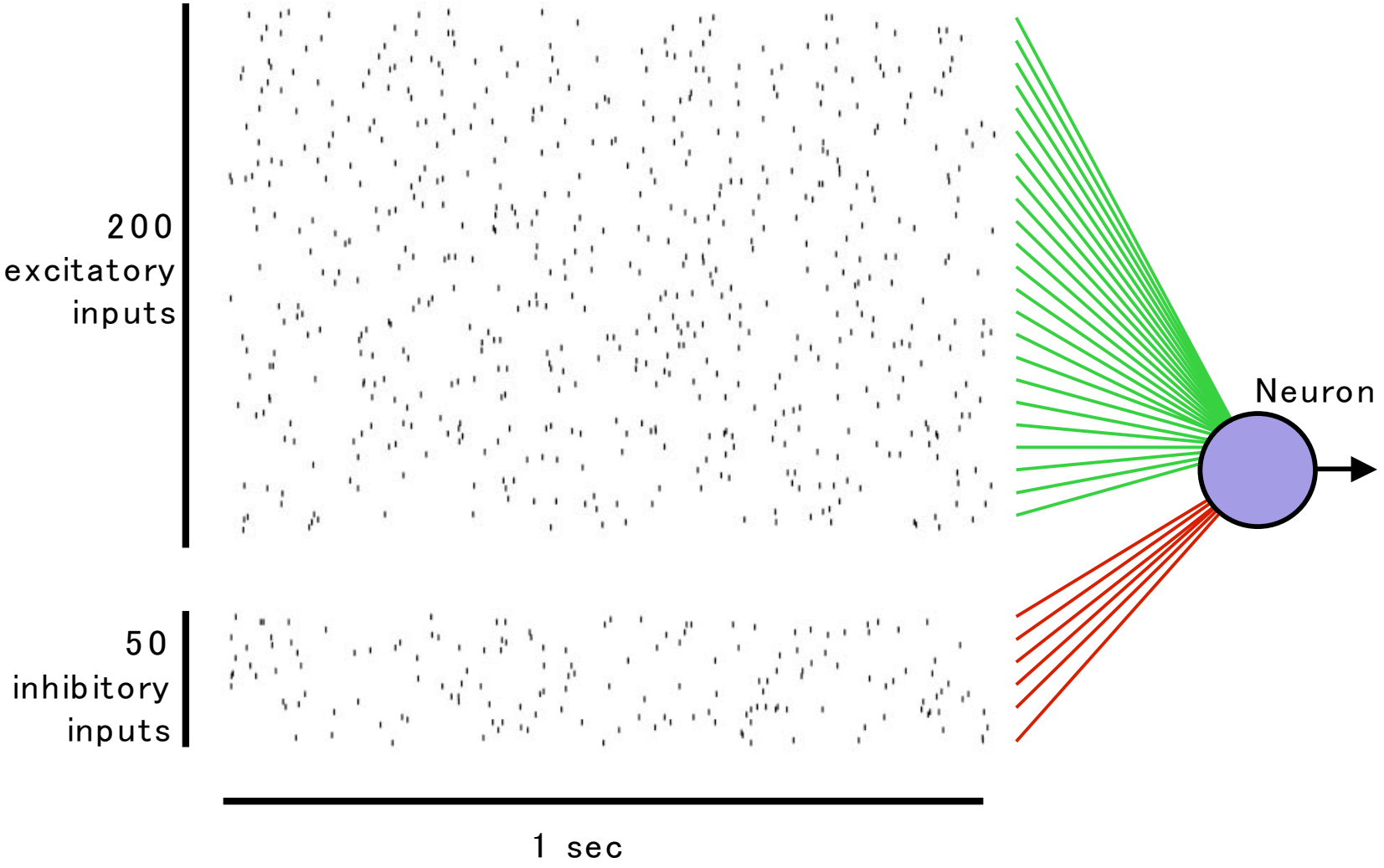
High-level software interfaces

Different user demands:

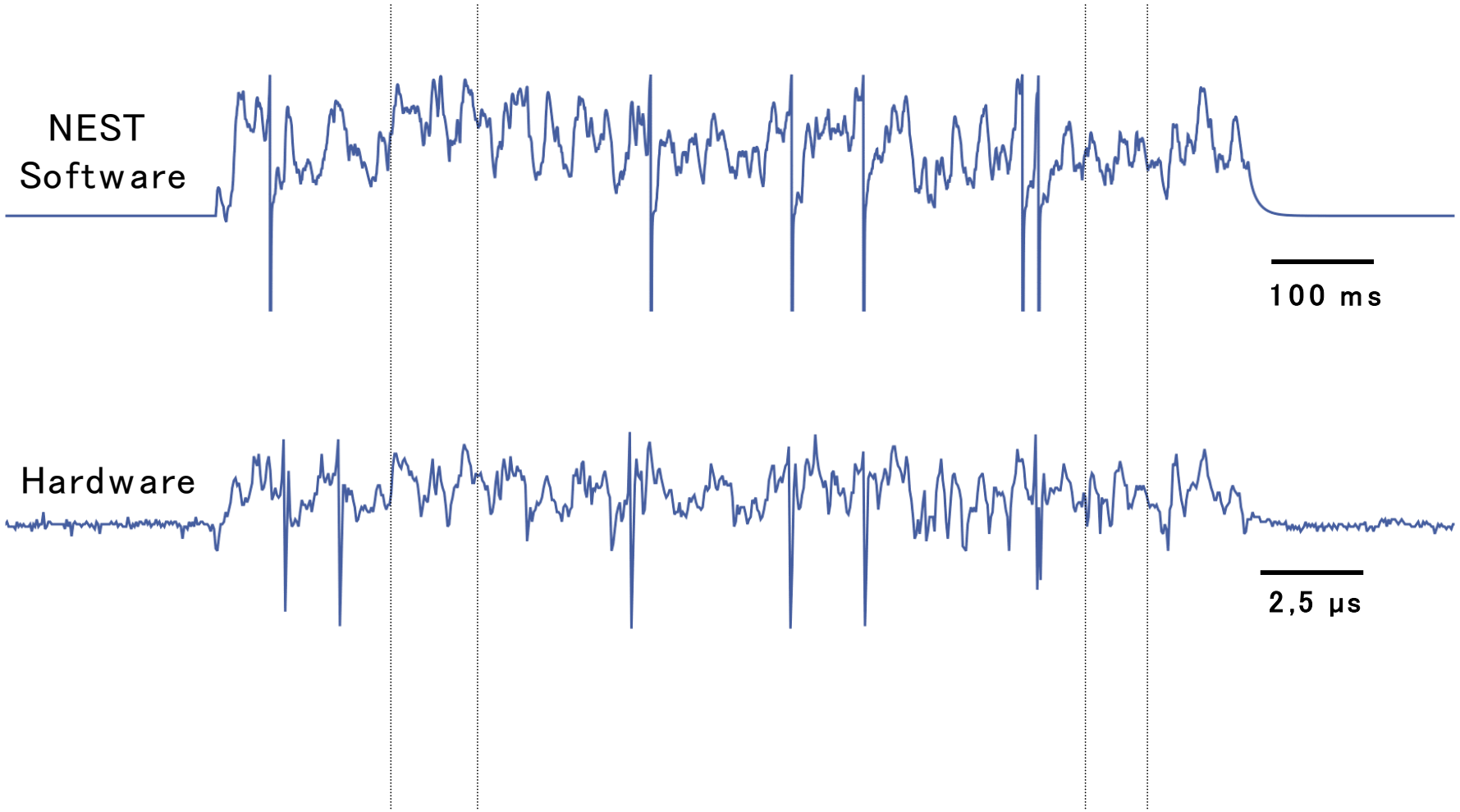
- An interpreter-based interface with huge scripting power for efficient generation and operation of large experimental setups (Python)
- A convenient graphical user interface (C++)



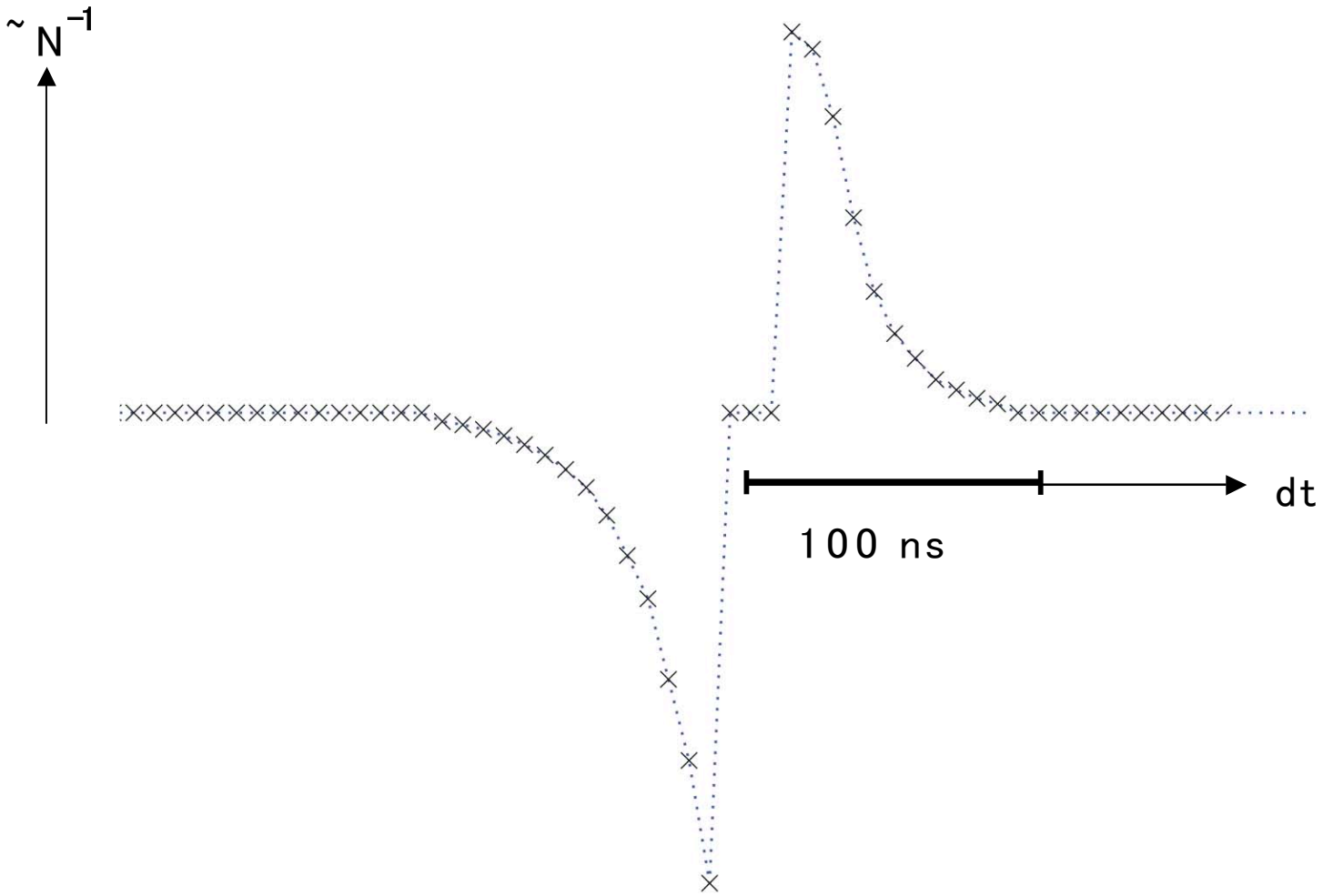
Single Neuron Bombardment Setup



Equivalence between Hardware and Software



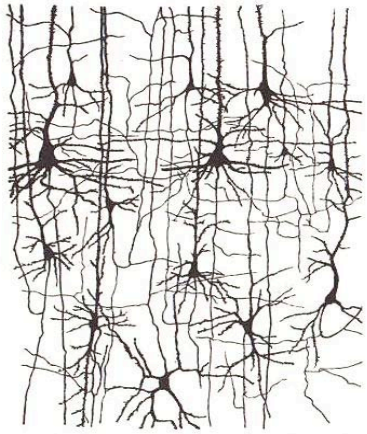
Measured STDP Modification Function



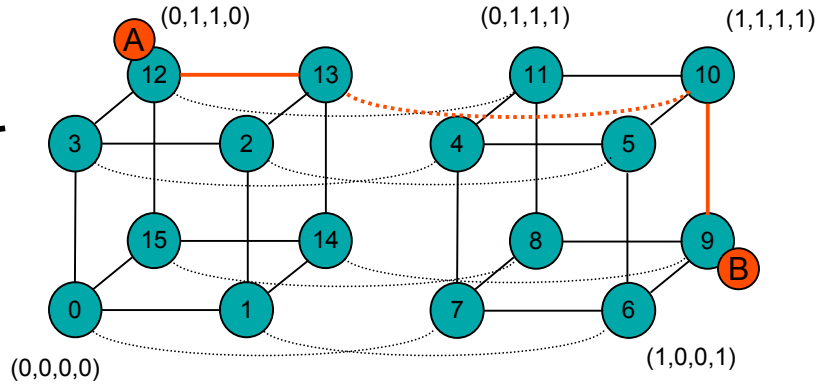
N = number of correlated spike pairs necessary to trigger a weight change
dt = time between pre- and postsynaptic spike

FACETS Stage 1 Network Overview

netlist of model neural network



FACETS stage-1 network hardware



mapping software

A (12) 13 10 B (9)

routing / switching logic

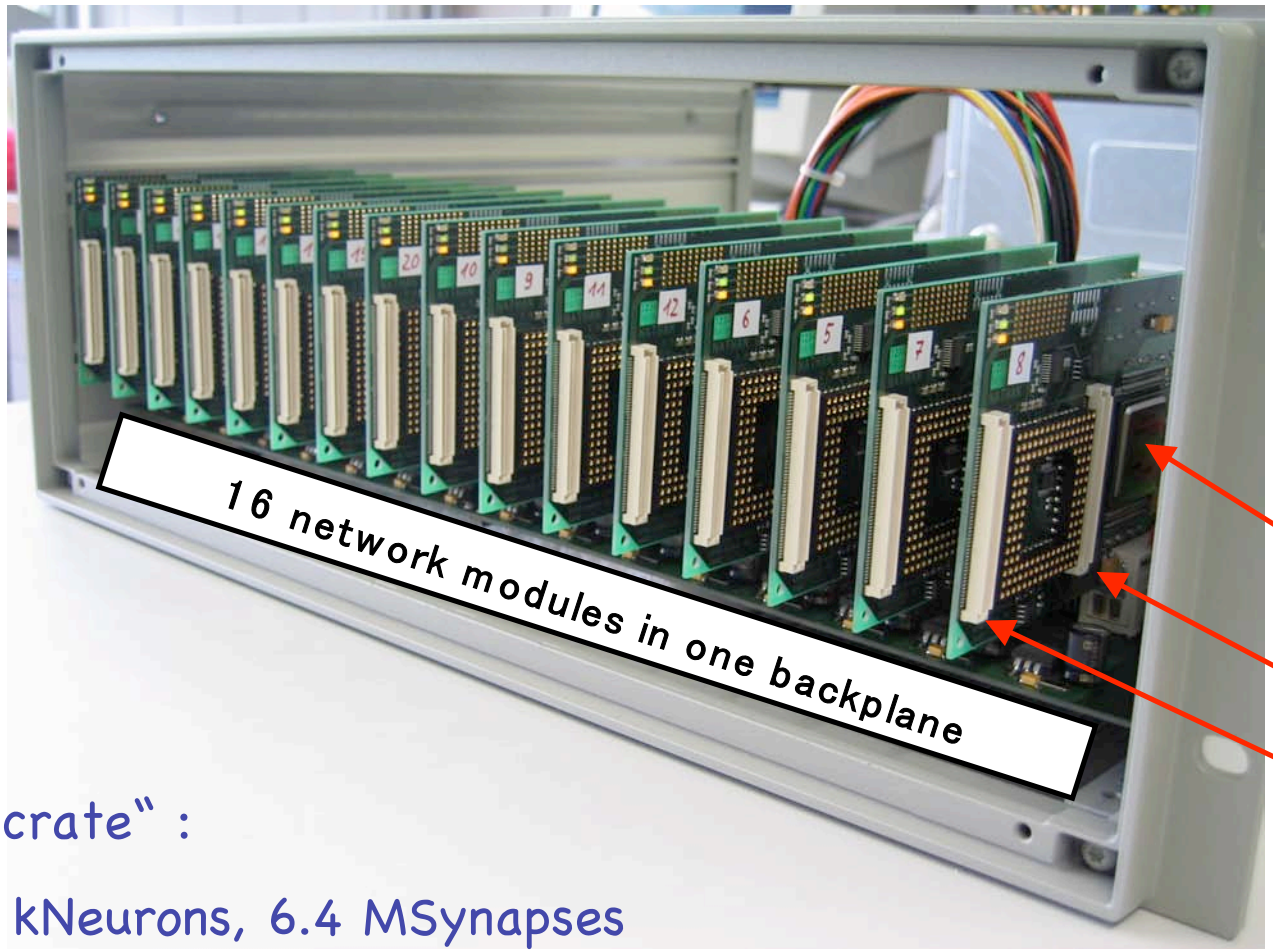
routing / switching logic

routing / switching logic

routing / switching logic

physical layer, isochronous network of point-to-point connections at 3.125 GBit/s

Stage 1 : Crate System



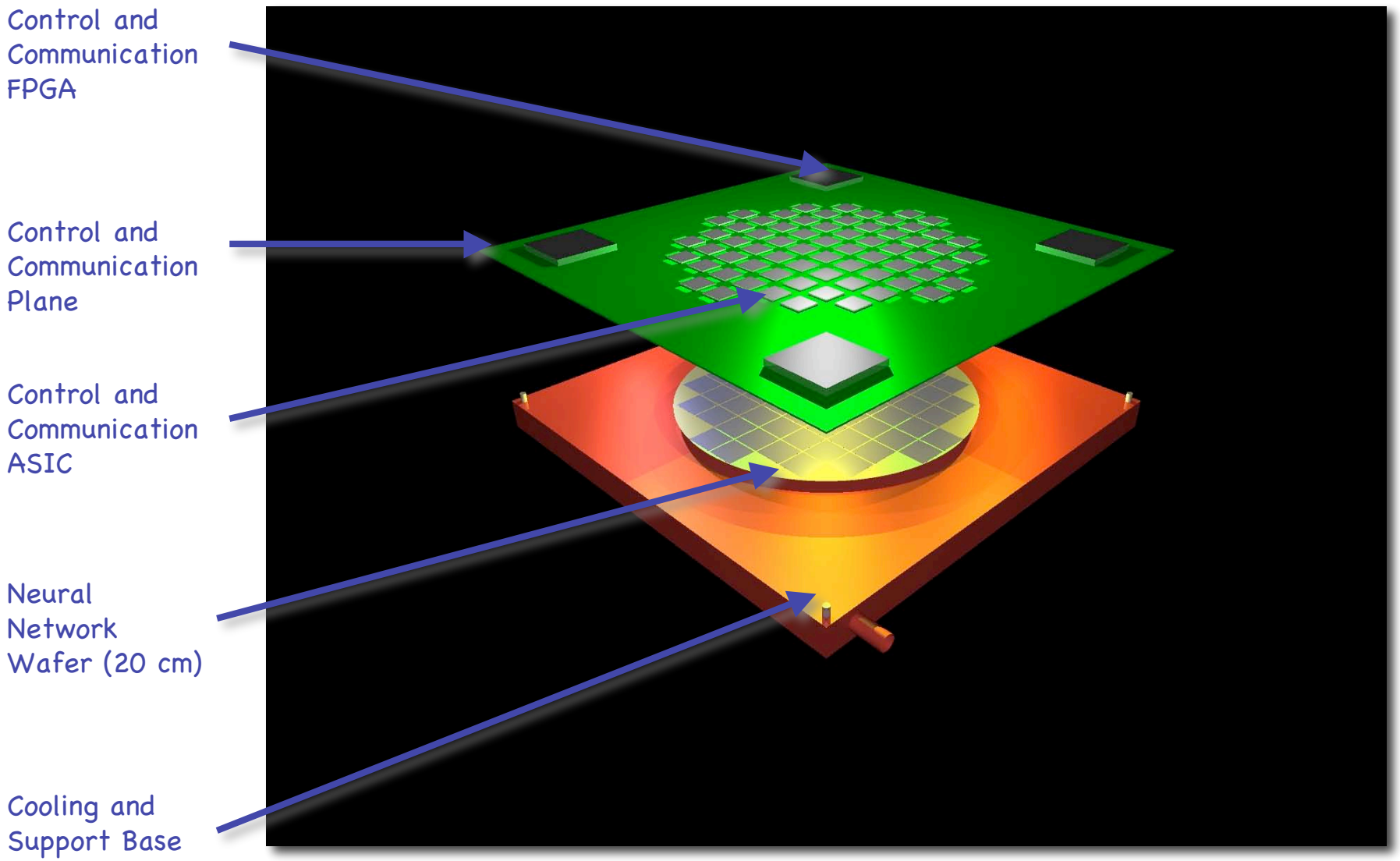
1 „crate" :

25 kNeurons, 6.4 MSynapses

0.25 mm^3 cortex

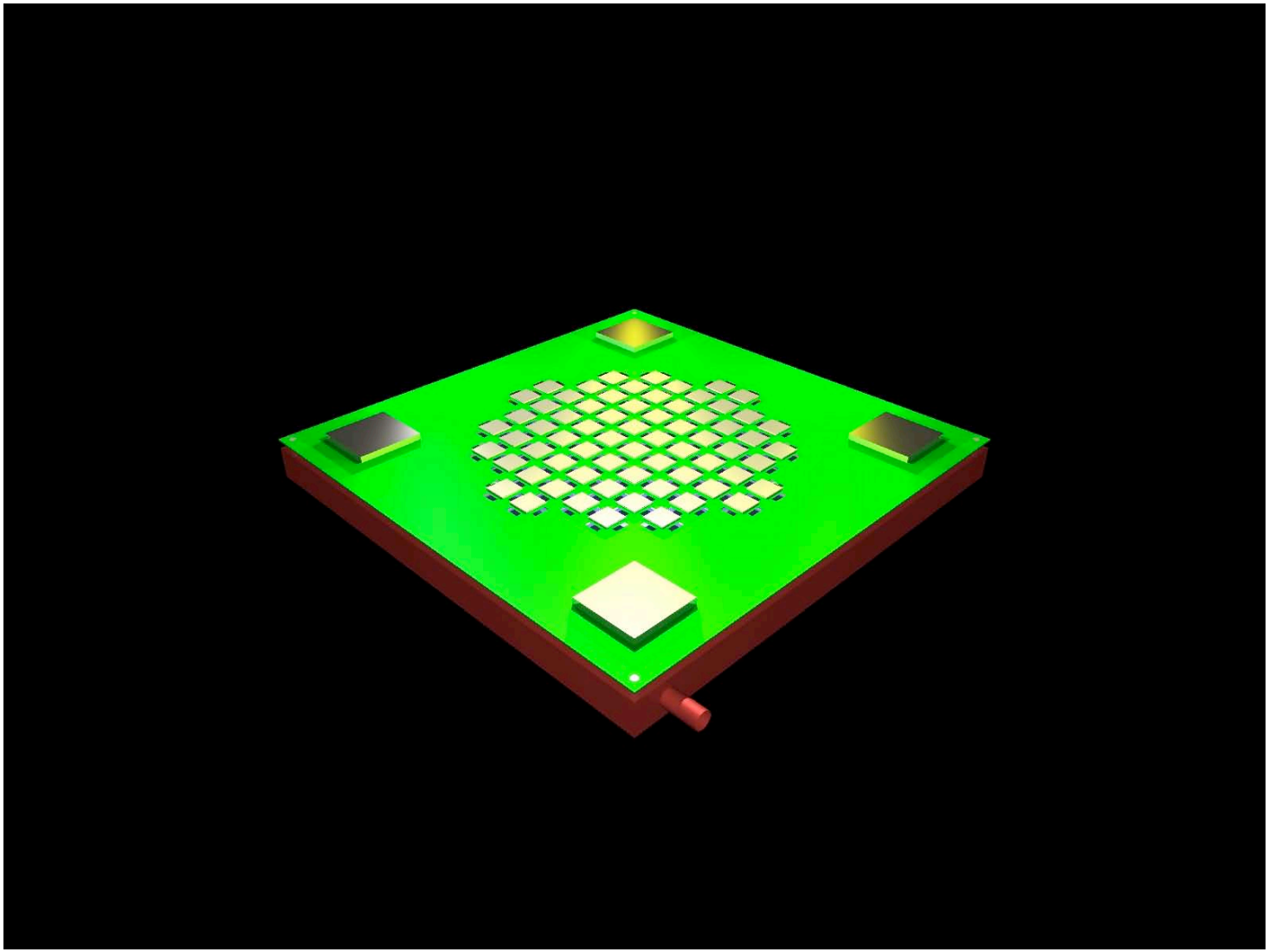
Not much scaling beyond this is economically plausible (50 k€/crate)

Stage 2 Technology Step : Neural Processing Unit, 5×10^5 Neurons, 10^9 Synapses



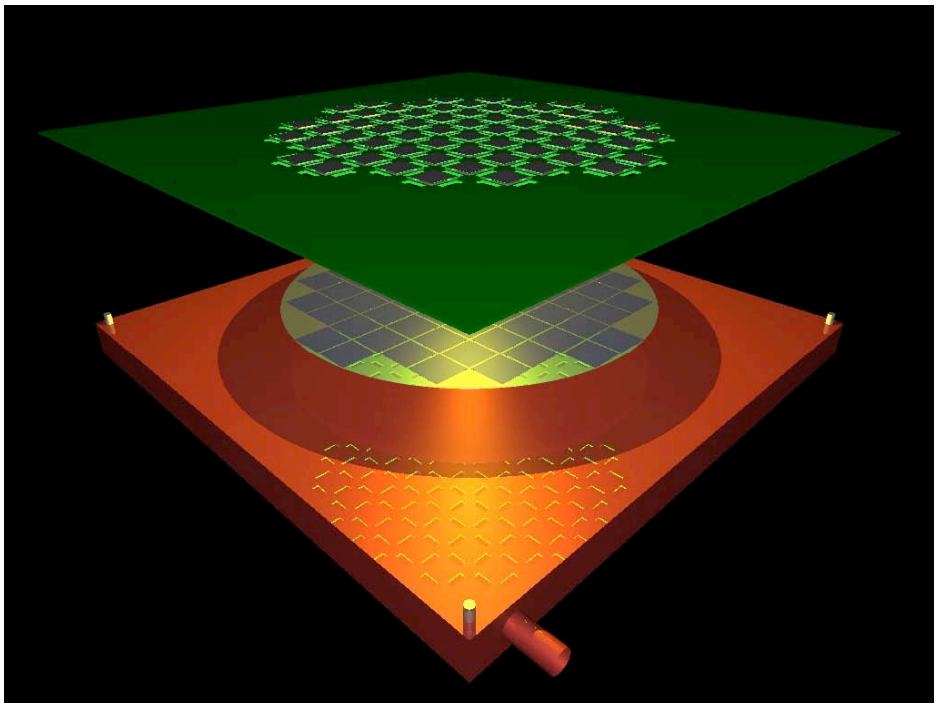
Idea : **Separate** Neural Circuits and Monitoring/Readout/Control

Closed Unit



Specifications

Process Technology	UMC 180 nm CMOS
Wafer Size	20 cm
Synapse Size	10 μm^2
Synapses per Wafer	10^9
Synapse-to-Neuron Ratio	1000-2000
Neurons per Wafer	$5.0\ 10^5$
Power of single neuron/synapse system	1 – 50 µW
Power of single NPU incl. digital overhead	about 100 W



Challenge :

FAULT TOLERANCE (a biological feature !)

Neuron model :

To be defined

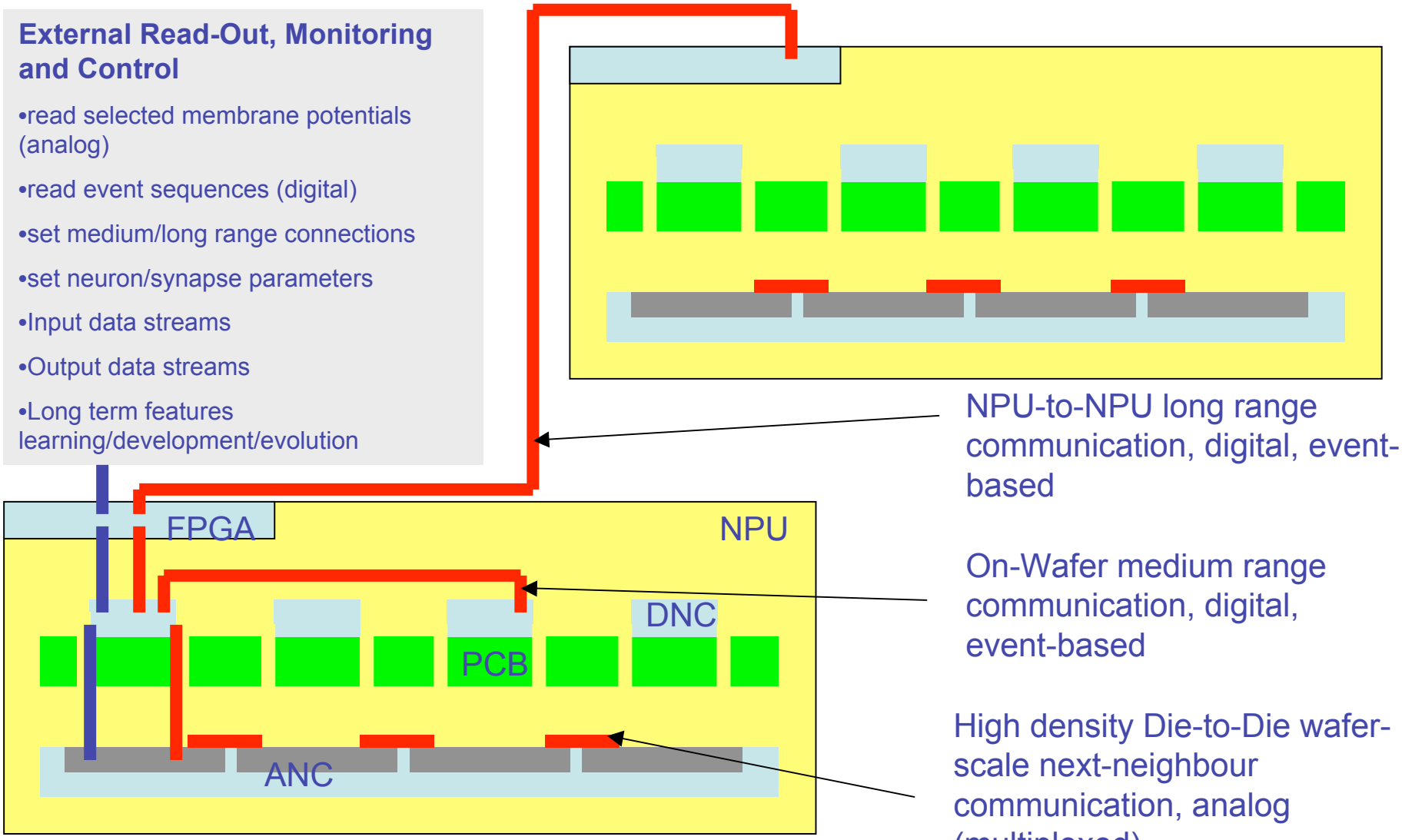
Adaptation, failing synapses, ...

FACETS project is needed !

Connectivity, Control, Communication

External Read-Out, Monitoring and Control

- •read selected membrane potentials (analog)
- •read event sequences (digital)
- •set medium/long range connections
- •set neuron/synapse parameters
- •Input data streams
- •Output data streams
- •Long term features learning/development/evolution



NPU-to-NPU long range communication, digital, event-based

On-Wafer medium range communication, digital, event-based

High density Die-to-Die wafer-scale next-neighbour communication, analog (multiplexed)

Wafer-Scale-Integration : The Communication Challenge

Synapse area in 180 nm CMOS : 10 μm x 10 μm, synapse density : 10.000/mm^2

10 Hz biological rate -> on-chip : 1 MHz/synapse, 1 THz/cm^2

Information Flow on an 8 inch wafer (16 bits per event) : 2.5 Petabytes/s

Reticle size : 25 mm x 25 mm -> no connections beyond this from UMC

Fraction of events crossing reticle border : 0.2 (assumption)

Events passing through a given reticle border : $1.25 \cdot 10^{10}$ events/mm/s Events must be routed from synapse A to synapse B, need connections :

- constant propagation delay
- slowly changing routing topology (long term plasticity)
- use programmable topoloy for different circuits

Solution : One connection per synapse : connection based routing :

$1.25 \cdot 10^{10}$/mm/s$\cdot 10^{-6}$ = $1.25 \cdot 10^4$ connections/mm

NOT possible with conventional wire bonding !

Almost possible with additional process step : wafer scale integration

Post-processing (metallization) on top of CMOS process

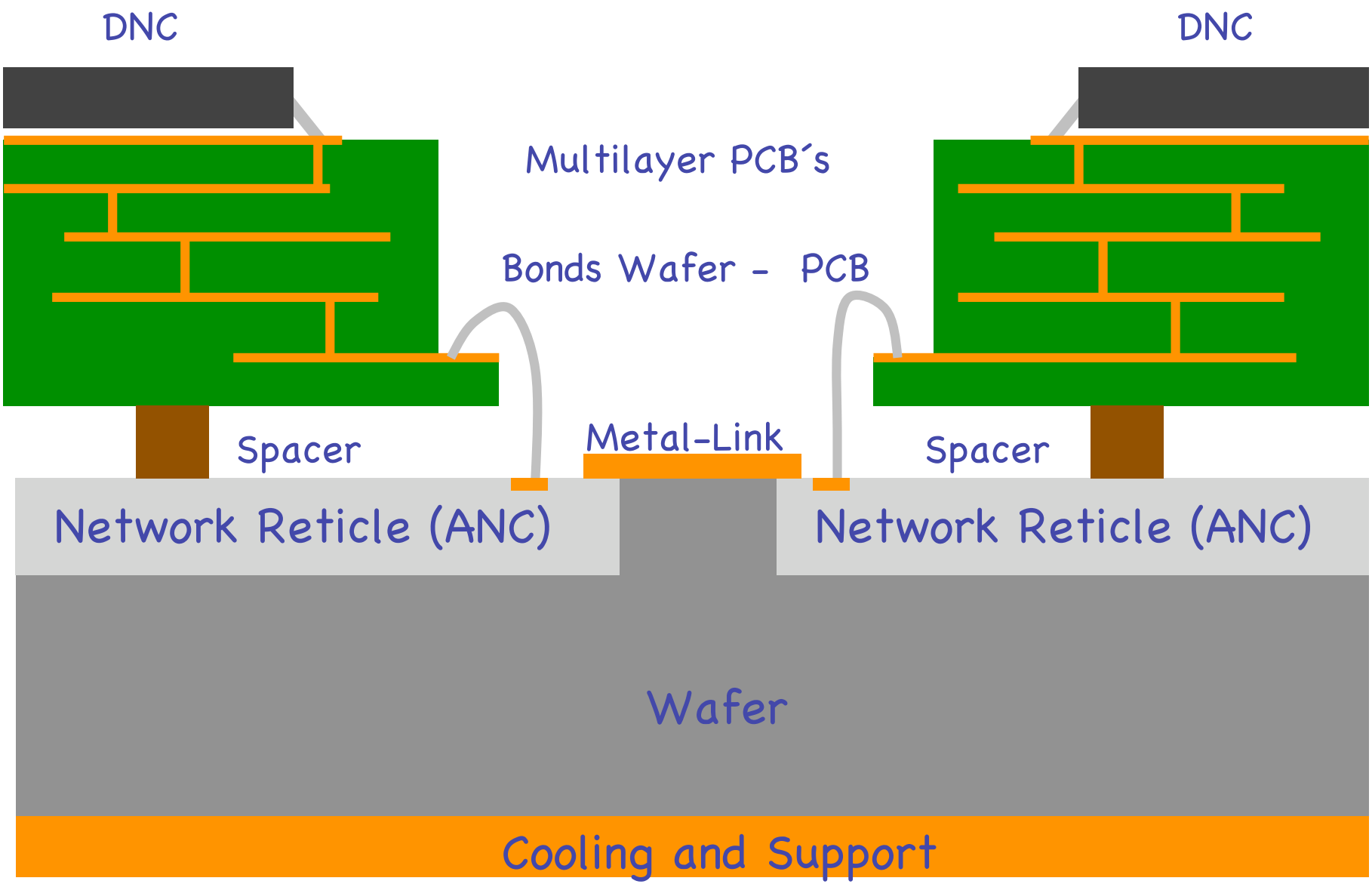
Digital PCB (Motherboard) Functionality

- Low latency event network for digital event routing, non-local connections
- maximum delay for inter-area communication : 50 ns (5 ms biology)
- delay between distant cortical areas : 500 ns maximum (50 ms biology)

Pathway for 500 ns . Source neuron - TDC - PCB - routing on source PCB - routing to target PCB - routing on target PCB - DTC - Target synapse

- Digitization of selected analog membrane potentials
- Trigger functionality for event analysis (e.g. detection of correlated firing, selected cortical area readout)
- Statistical analysis for on board data reduction
- Power monitoring and defect management
- Interface for superstructures

Mechanical Structure and Connectivity



FACETS Long Term Goal : Building Superstructures

Example for 5 x 5 x 5 Superstructure with 10^8 Neurons and 10^{11} Synapses

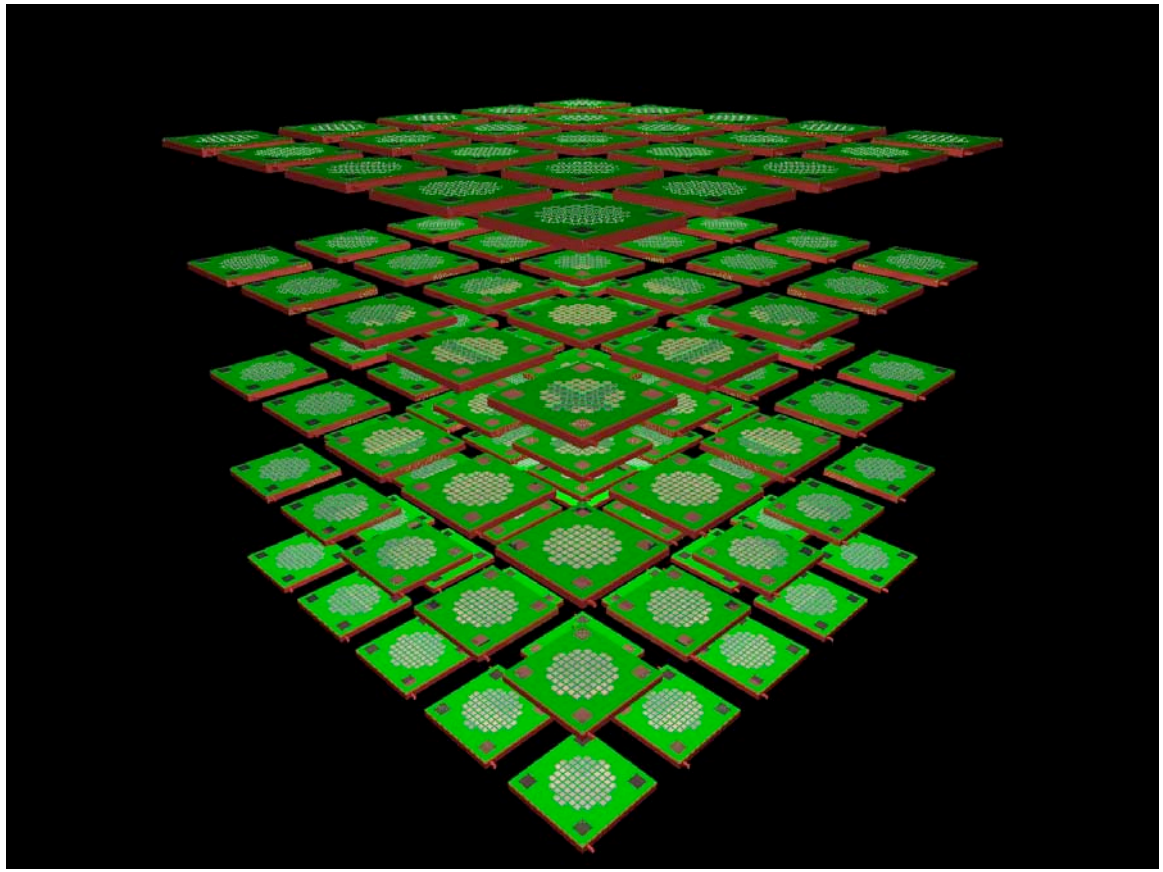
Major Engineering Effort
(Power, Mechanical Structure)

Major Software Effort
(Monitoring, Set-up and Control)

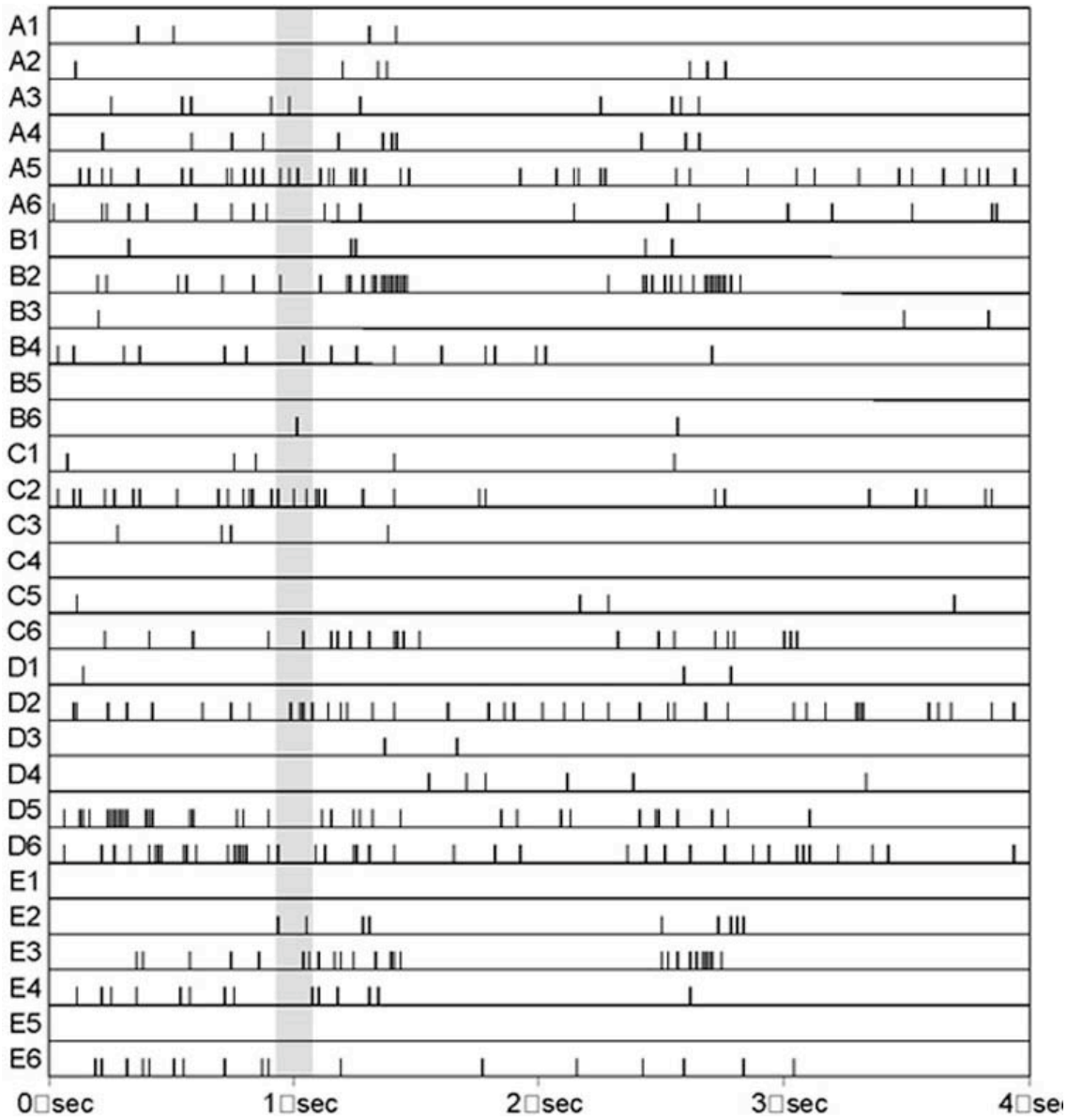
Major Effort in Model Building and Concepts for Experiments

1000 mm^3 of Cortex

10% of VI

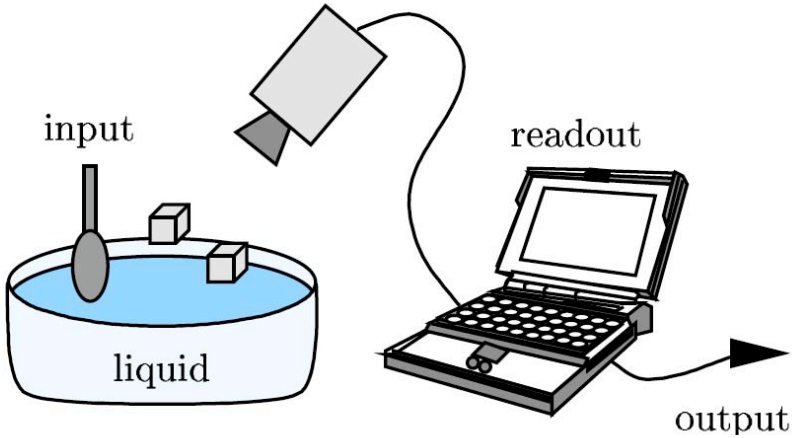


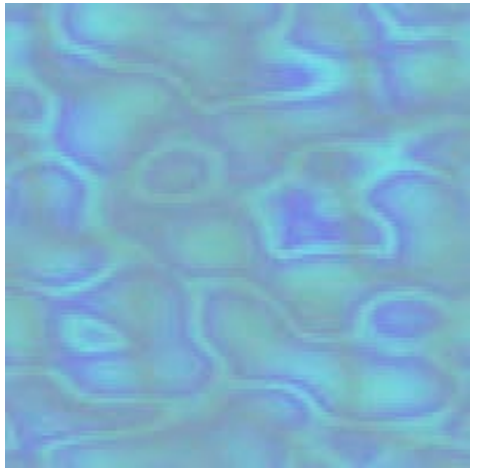
4 seconds of 30 neurons in a monkey brain (Krüger,Aiple 1988)

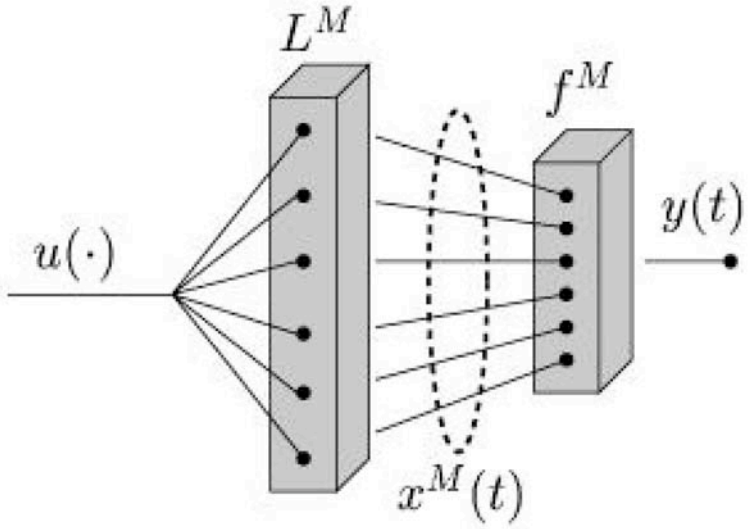


Liquid Computing - Liquid State Machine (LSM)

W. Maass, T. Natschläger, and H. Markram. Real-time computing without stable states : A new framework for neural computation based on perturbations. Neural Computation, 14(11):2531-2560, 2002.







1. „Liquid“

$$x^M(t) = (L^M u)(t)$$

2. „Readout“

$$y(t) = f^M(x^M(t))$$

Operation of a Liquid State Machine (LSM) (W. Maass)

135 (15x3x3) **randomly** connected IF neurons, 20% inhibitory

Randomly chosen synaptic strengths

2 Poisson-distributed pulsetrains u und v

0.5 seconds

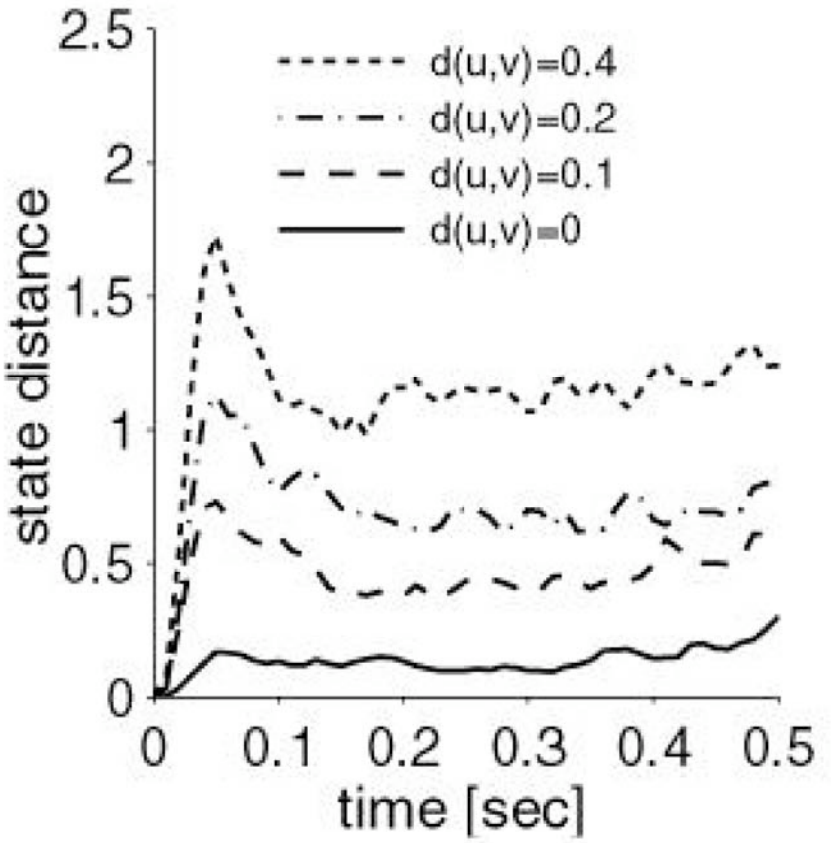
Compare distances d of input and output pulse trains

Output distance > Input distance

Separation without dedictaed set-up

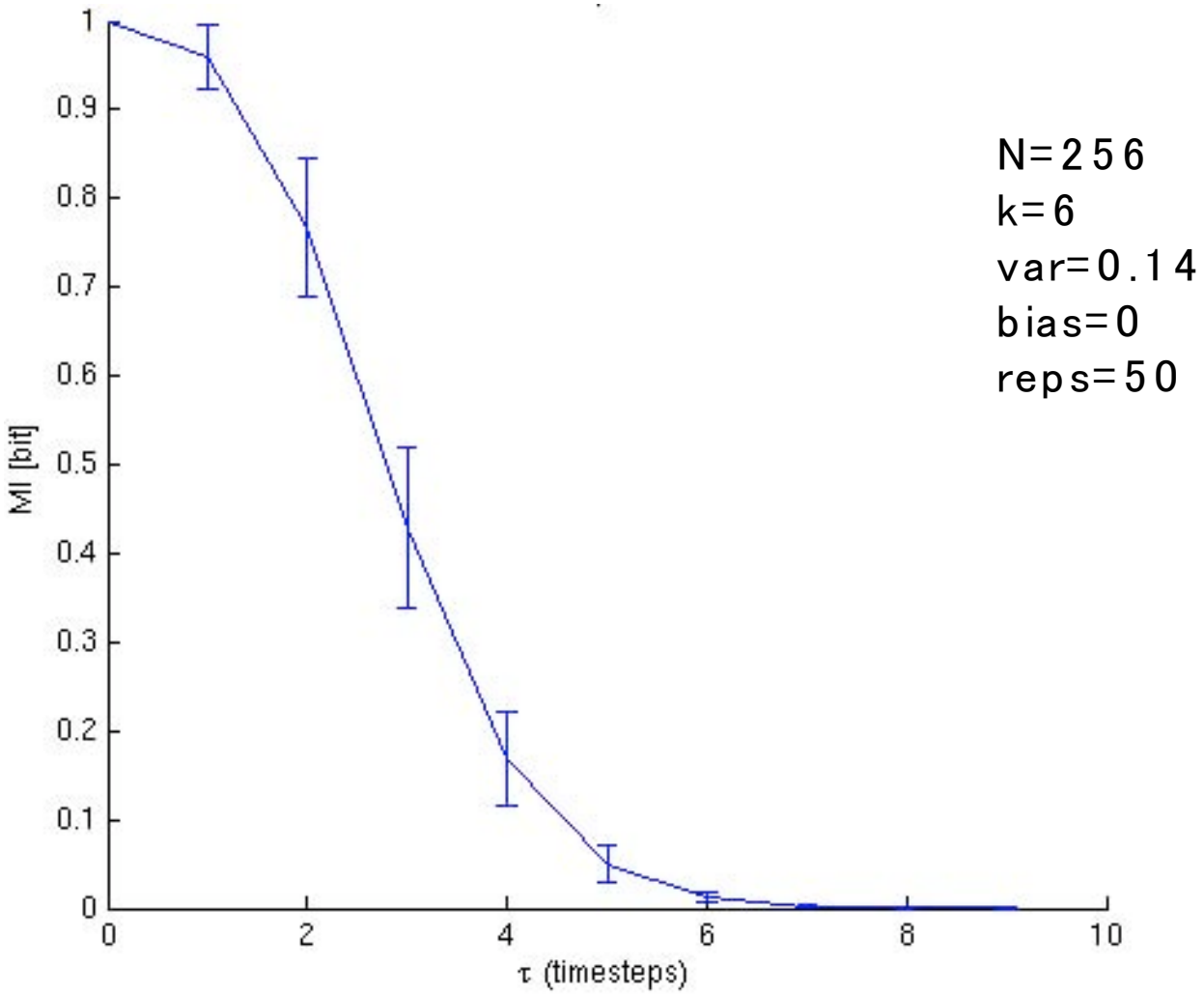
Universality

$$\left\| x_u^M(t) - x_v^M(t) \right\|$$



quasi-instantaneous availability of results : „Any-Time-Computing“

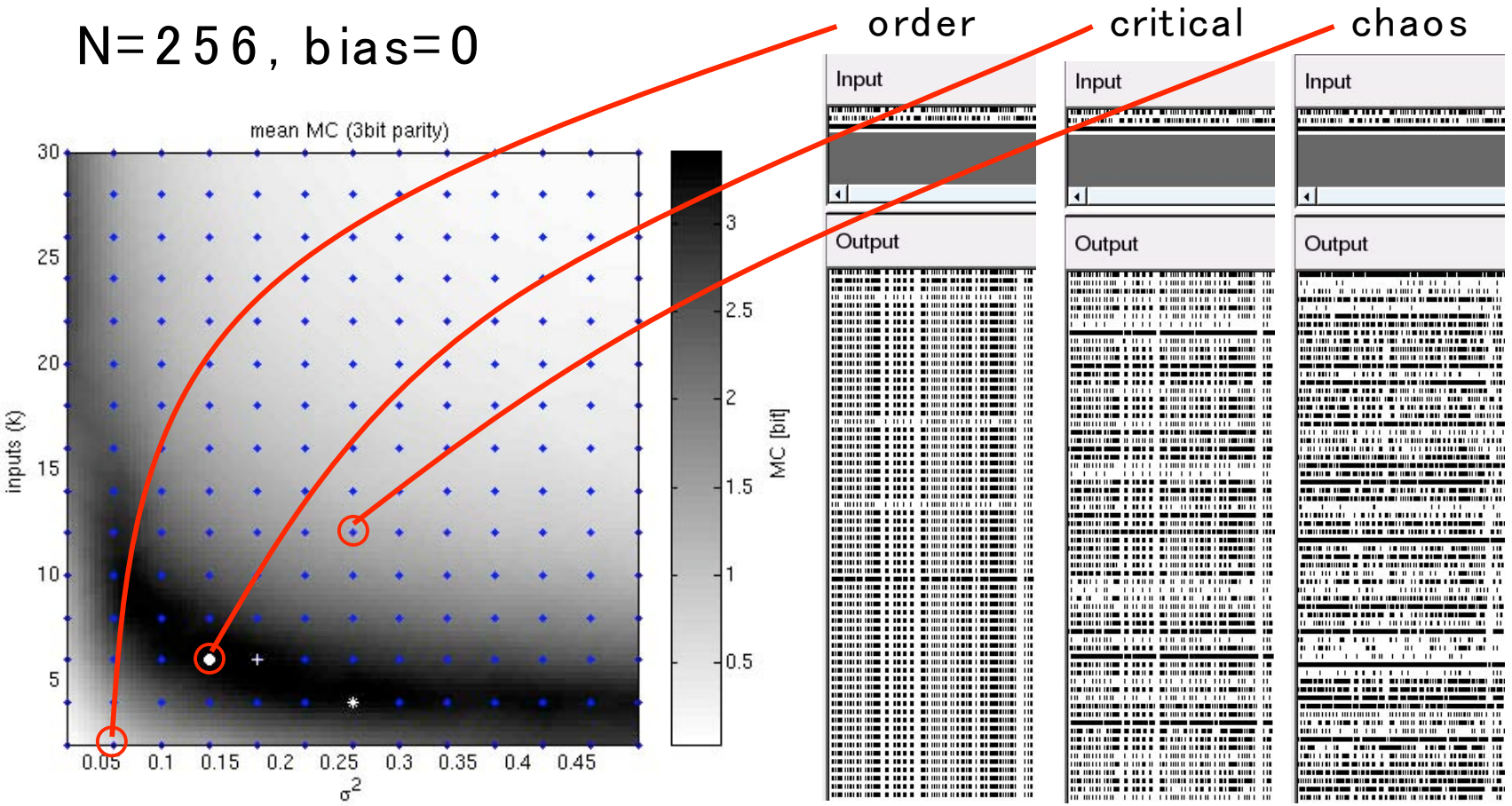
Memory Curve (3bit time-delayed parity)



Edge of Chaos Computation in Mixed-Mode VLSI - "A Hard Liquid", Felix Scürmann, Karlheinz Meier, Johannes Schemmel

In Lawrence K. Saul and Yair Weiss and Leon Bottou, editors, Proc. of NIPS 2004, Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, 2005.

Mean MC (3-bit time delayed parity)

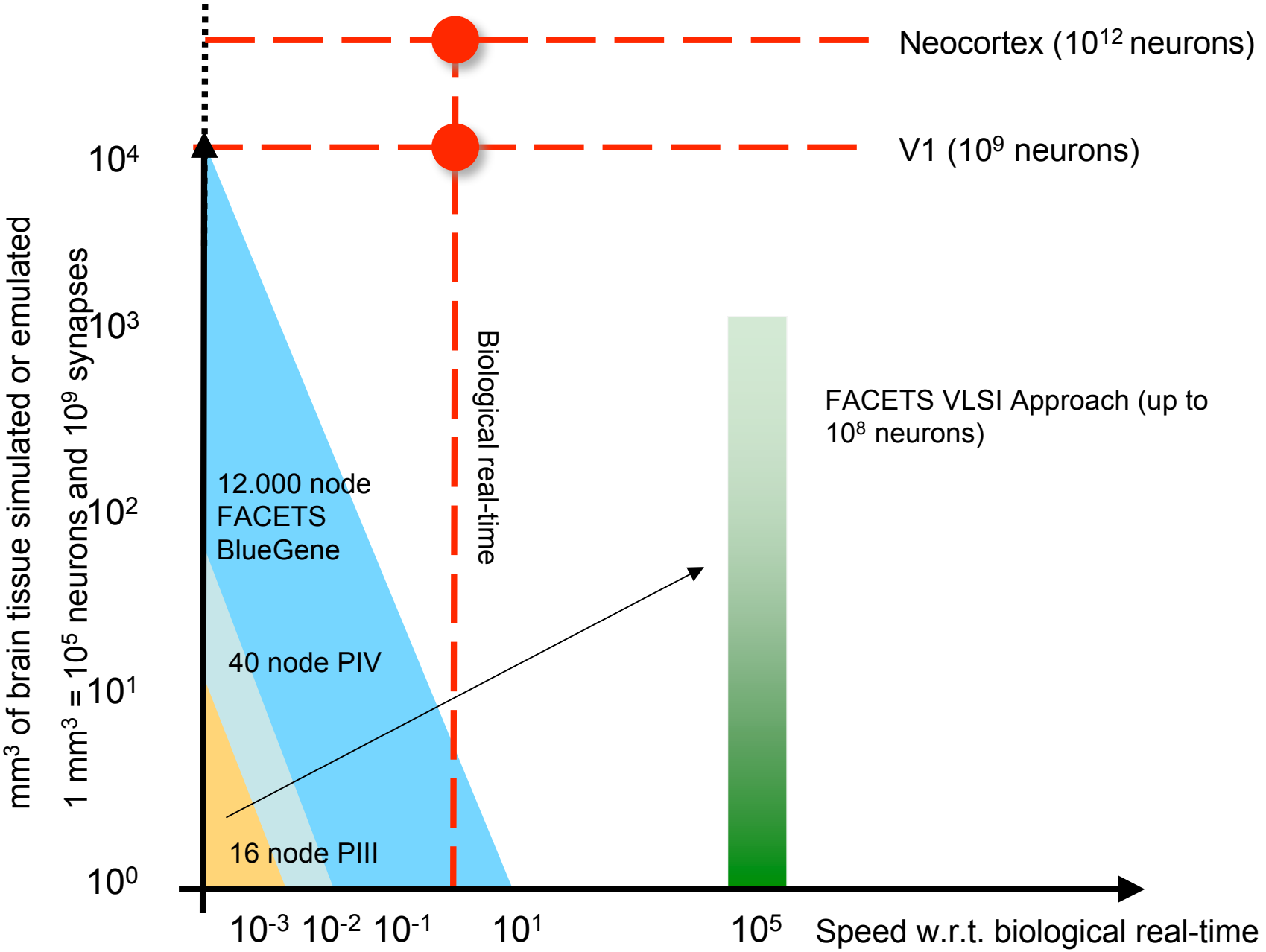


50 networks per data point, sigma < 0.35 bit

Edge of Chaos Computation in Mixed-Mode VLSI - "A Hard Liquid", Felix Schürmann, Karlheinz Meier, Johannes Schemmel

In Lawrence K. Saul and Yair Weiss and Leon Bottou, editors, Proc. of NIPS 2004, Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, 2005.

FACETS : Complementarity Supercomputers and VLSI - Complexity vs. Speed



Why FACETS Hardware ?

Two Answers :

I. Research tool for neuroscience : Bridge the gap in timescales from milliseconds to years

II. New type of information processing : Use low yield, low power, make use of self-organisation (learning)

www.facets-project.org

www.kip.uni-heidelberg.de/vision