Universität Bielefeld Cognitive Interaction Technology Center of Excellence

CITEC

Cognitronics:

Resource-efficient Architectures for Cognitive Systems

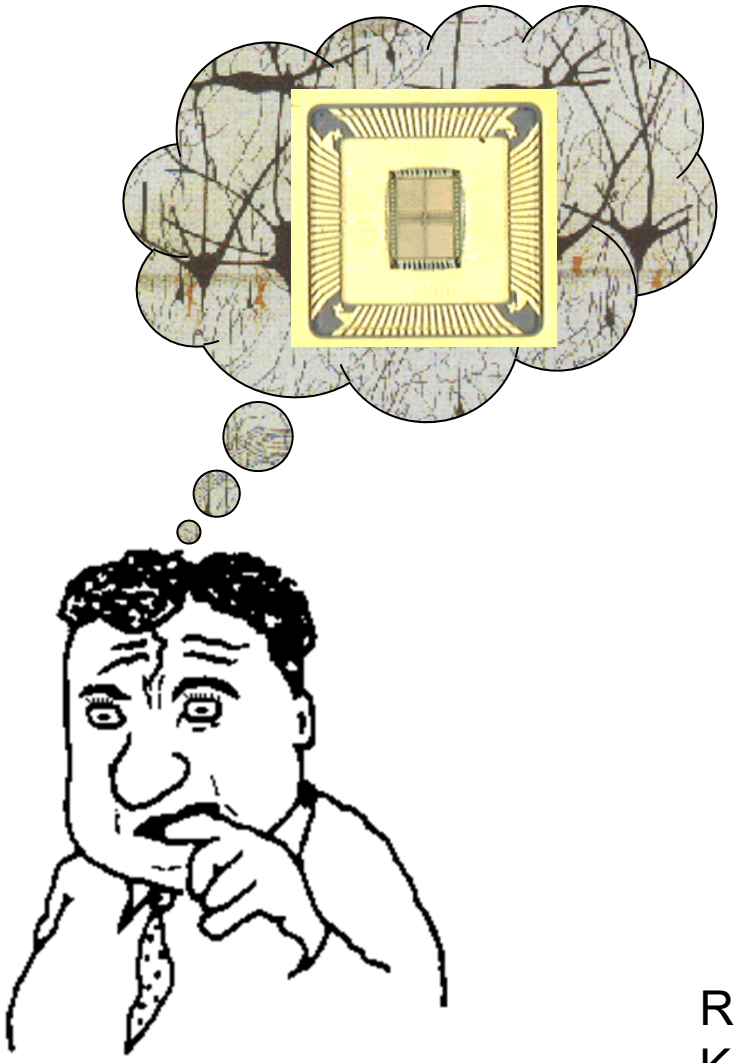
Ulrich Rückert

Cognitronics and Sensor Systems

14th IWANN, 2017

Cadiz, 14. June 2017

rueckert@cit-ec.uni-bielefeld.de
www.ks.cit-ec.uni-bielefeld.de



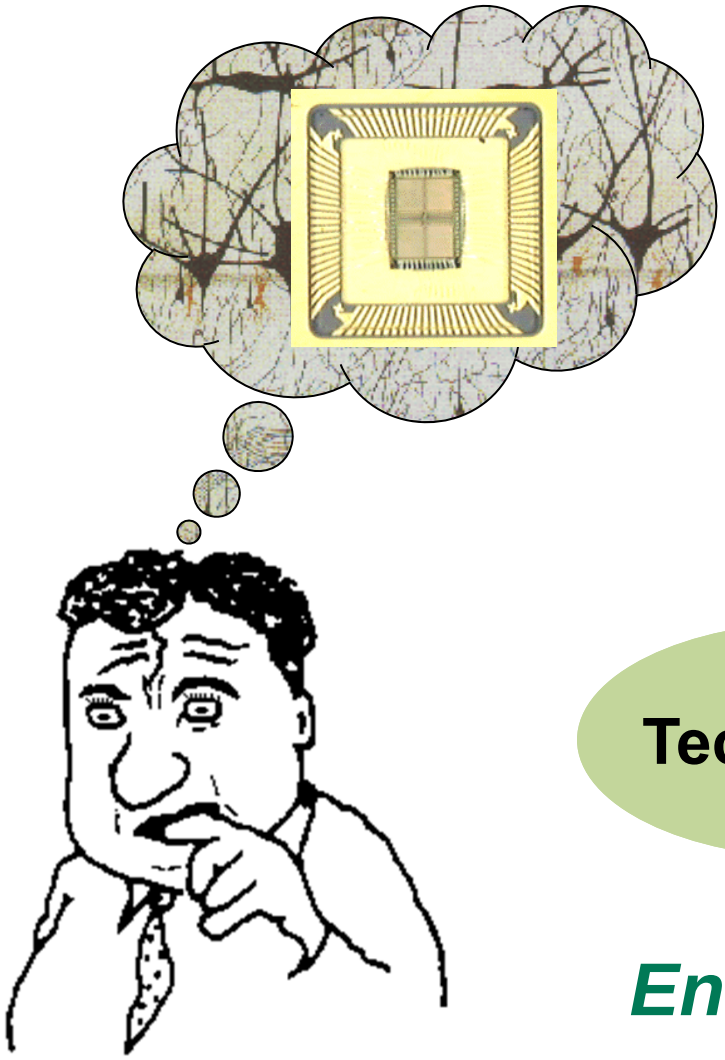
Is there a Silicon Way to Neural Networks ?

(1921-1993)

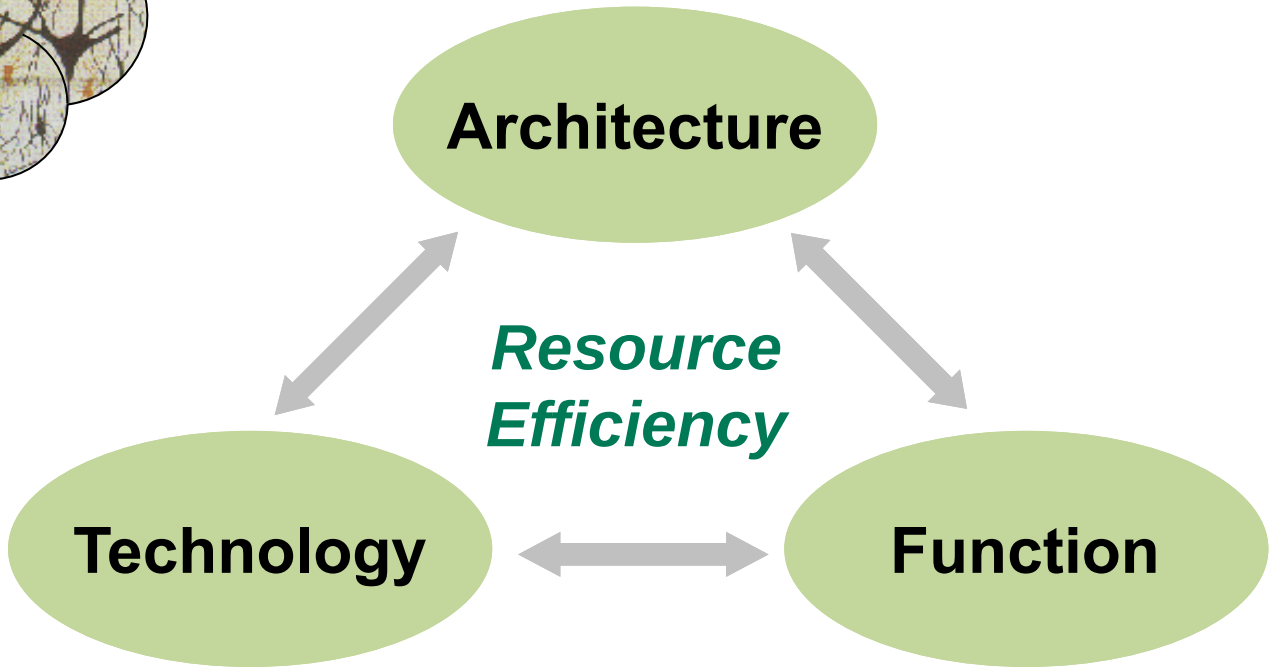
R. Caianiello, WOPPLOT 1986, Munich
K. Goser / U. Rückert, IEEE Micro Vol.9, No.6 1989

Overwiew

- **Introduction** — Background
- **Technology** — Femtoelectronics for ANN
- **Architectures** — Design Alternatives
- **Applications** — Cognitive Robotics
- **Discussion** — Questions

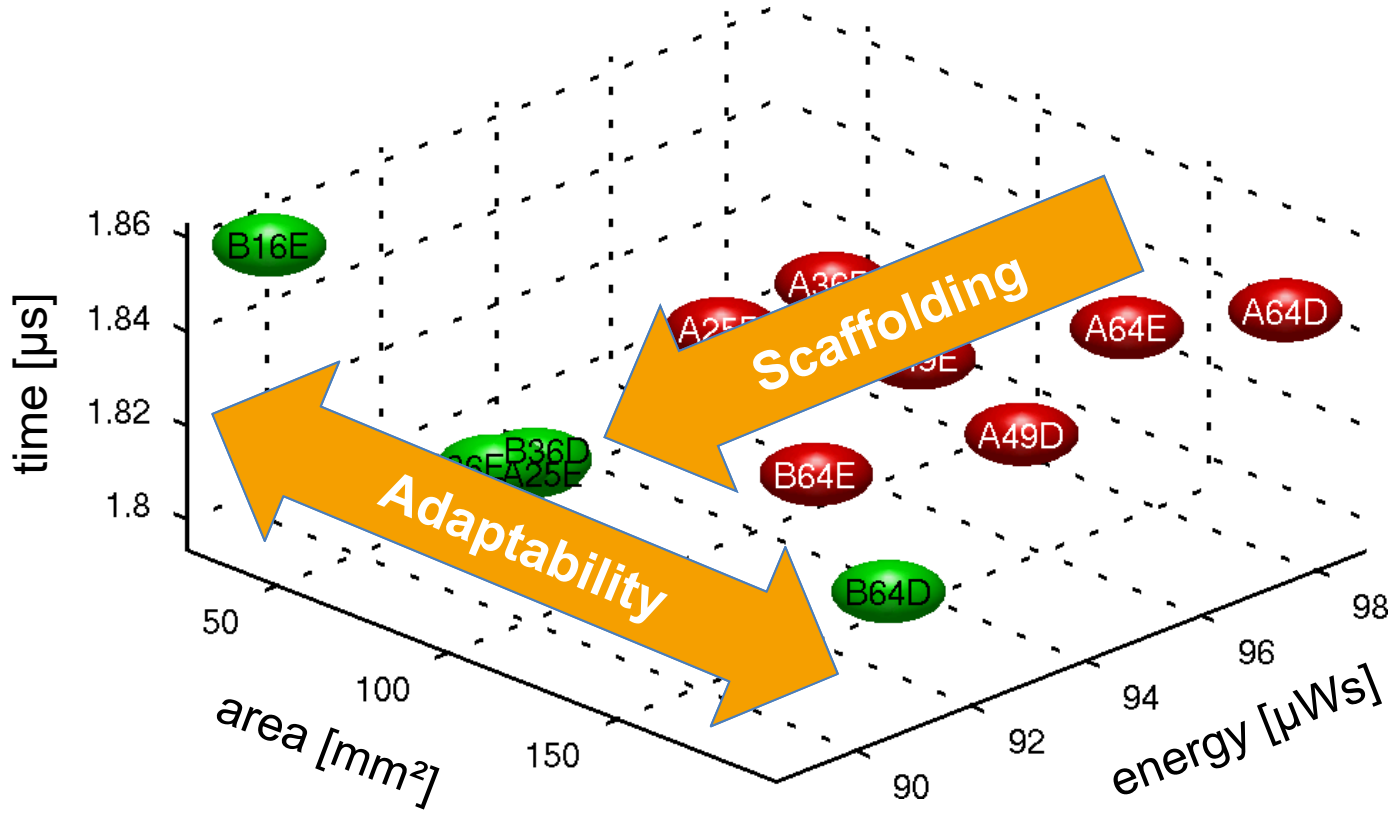


How to use 10^{10} (unreliable) devices/cm^2 efficiently?



Energy – Mass (Volume) – Time

Design Space Exploration

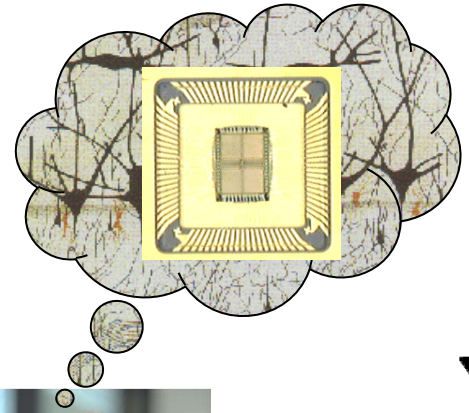


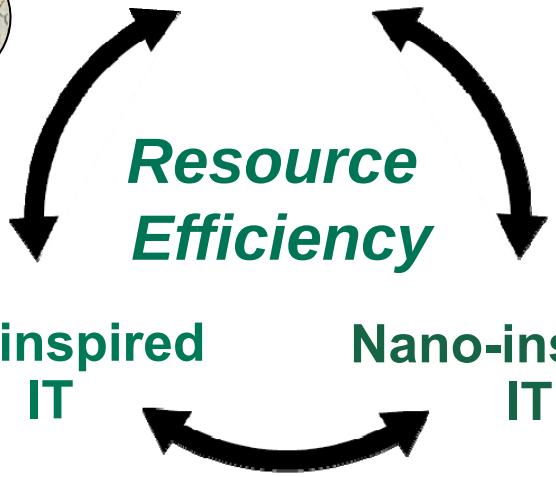
pareto-optimal

not pareto-optimal

pareto-optimal design in respect to area, energy, and time

Cognitronics and Sensor Systems





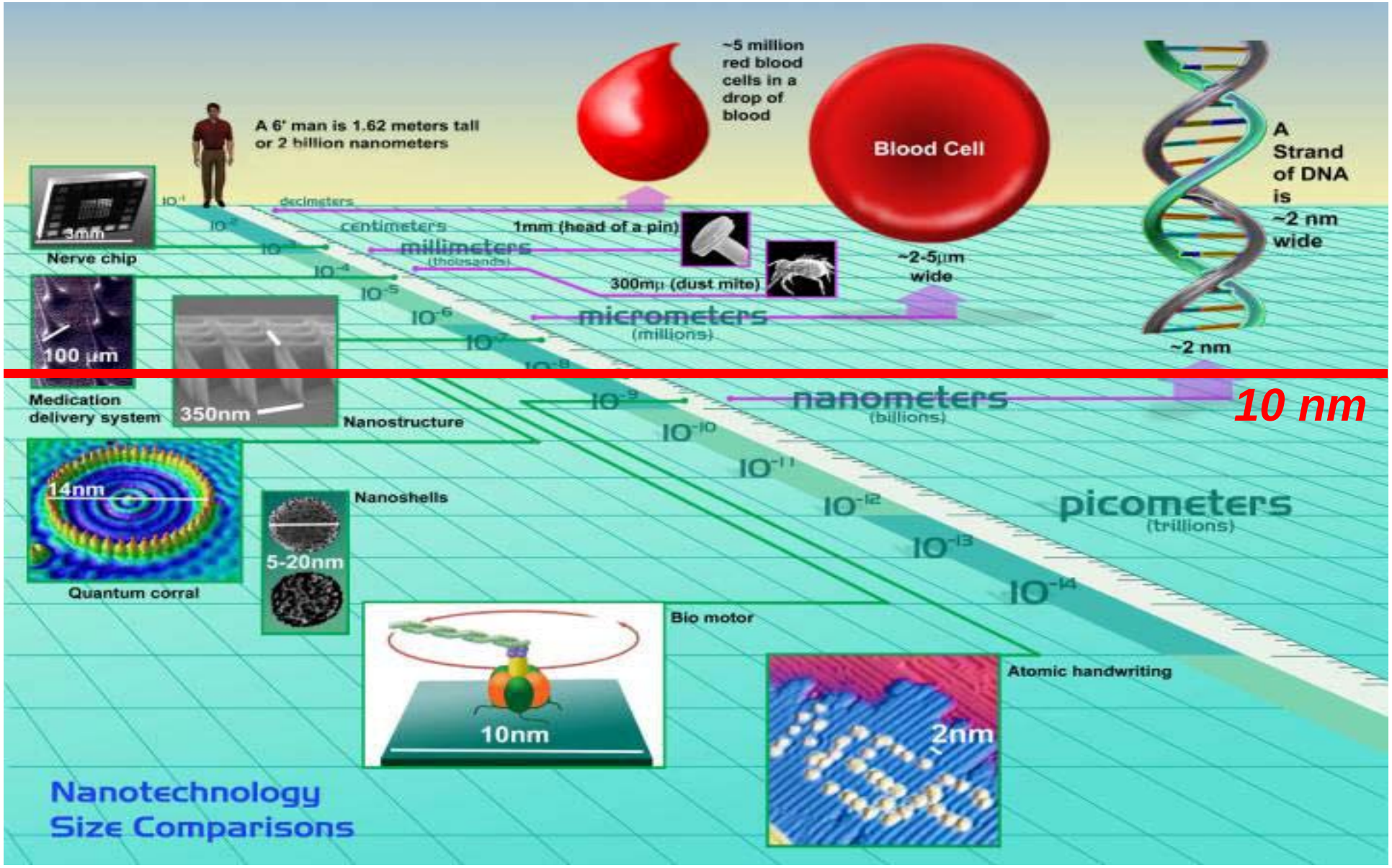
Bio-inspired Nano-Architectures for Cognitive Interaction Technology

How to generate complex behaviour in the limits of restricted resources?

Overwiew

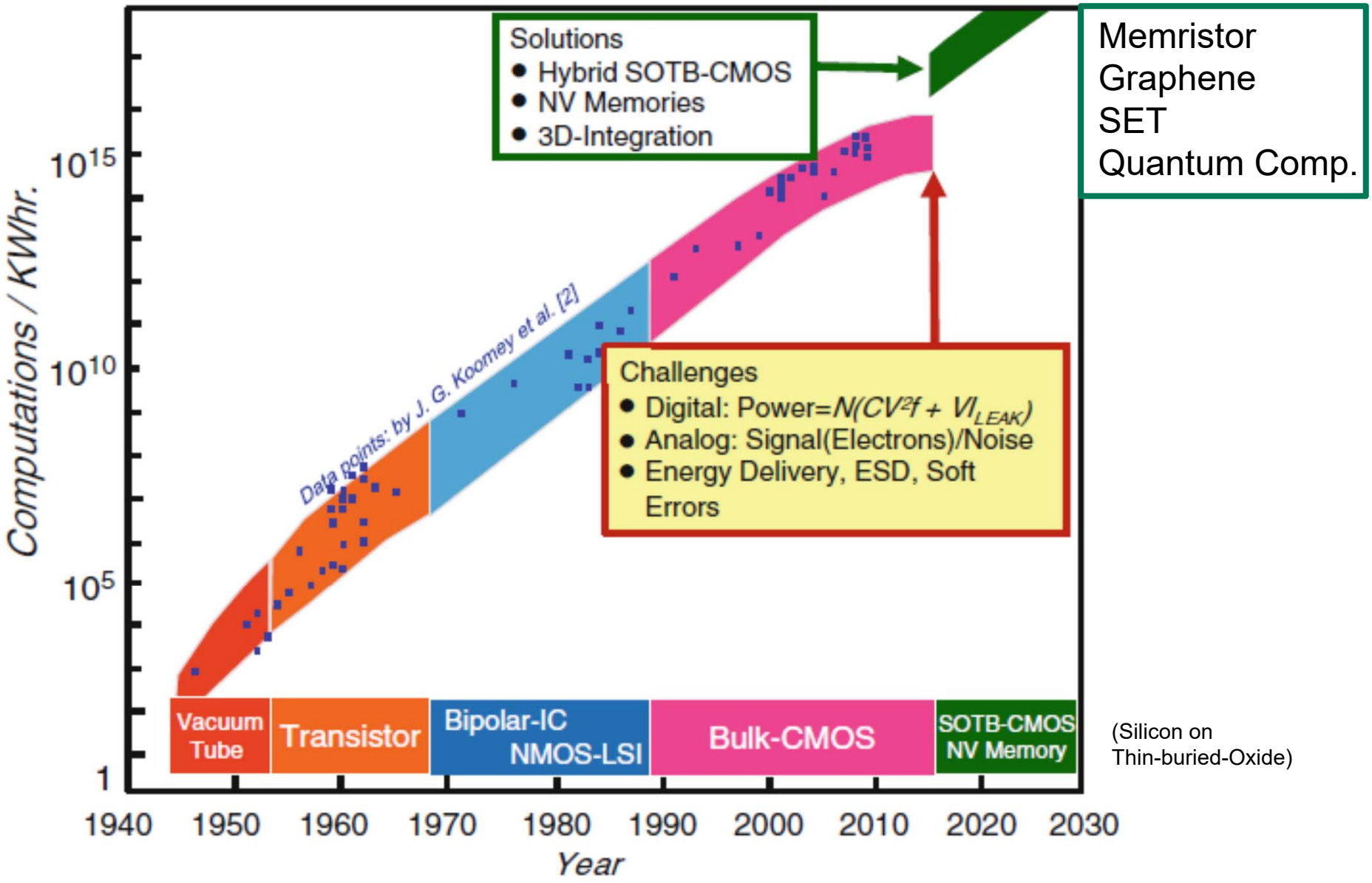
- Introduction — Background
-  Technology — Femtoelectronics for ANN
- Architectures — Design Alternatives
- Applications — Cognitive Robotics
- Discussion — Questions

Nanotechnologie



J. Bordogna, NSF, USA, 2001: https://www.nsf.gov/news/speeches/bordogna/rochester/sld006.htm

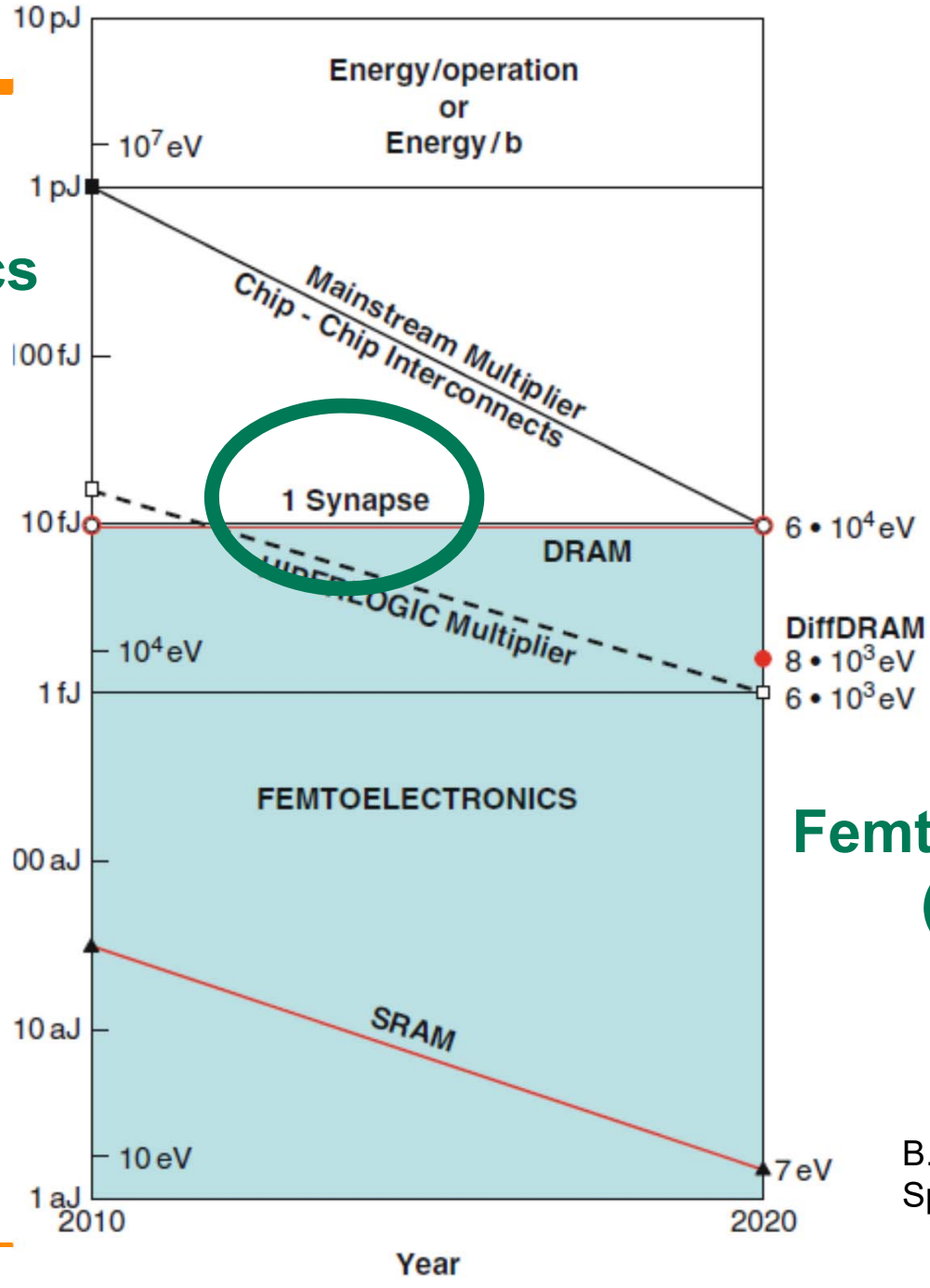
Evolution of Computations per KWh



(Silicon on Thin-buried-Oxide)

J.G. Koomey et al. 2009, IEEE Annals of the History of Computing

Nanoelectronics (structure)



Femtoelectronics (energy)

B. Höfflinger: CHIPS 2020
Springer 2012

Nanotechnology - Challenges

How to use 10^{10} (unreliable) devices/cm^2 efficiently?

Need for novel architectures:

regular & modular structure	➔	reduced design complexity
fault-tolerance, redundancy	➔	higher yield
parallelism	➔	reduced power consumption!
scalability	➔	faster and simpler mapping to next generation technologies

Neural architectures are very good candidates!

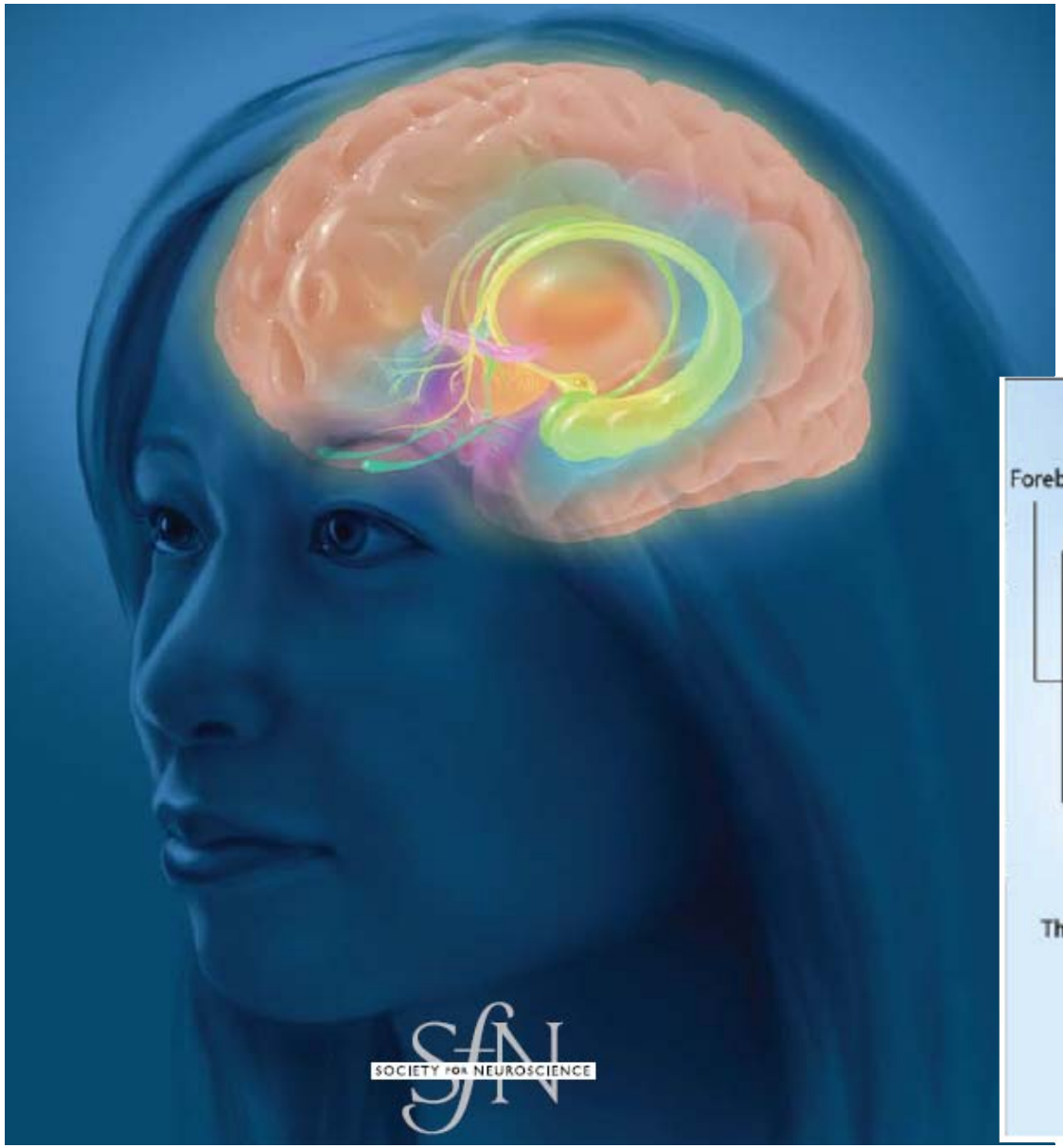
Technology Push!

(1990)

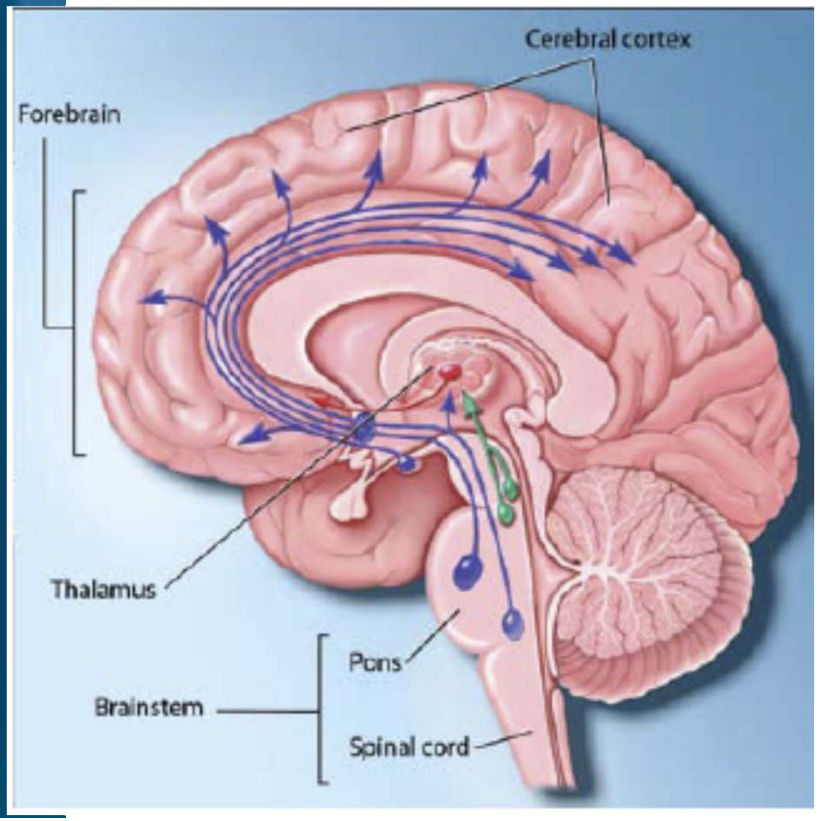
Overwiew

- Introduction — Background
- Technology — Femtoelectronics for ANN
-  Architectures — Design Alternatives
- Applications — Cognitive Robotics
- Discussion — Questions

Brain Architecture ?

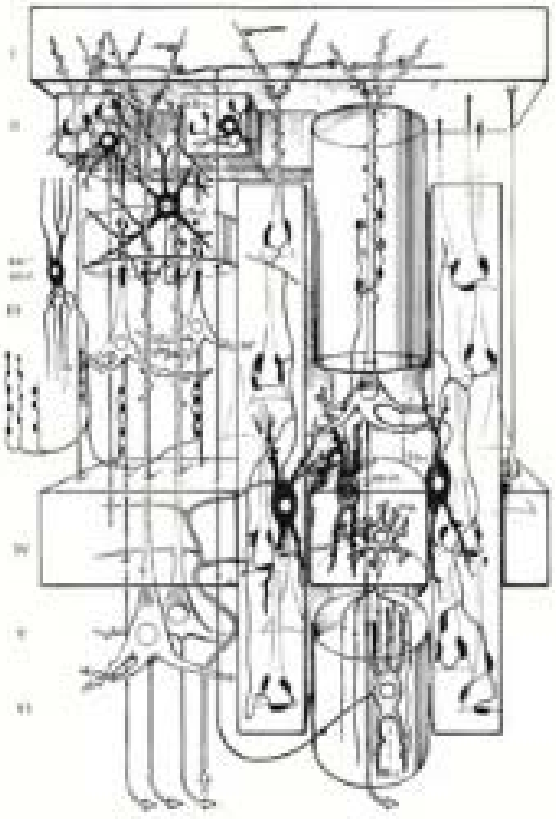


~ dm^3
~ 1,5 kg
~ 10^{11} neurons
~ 30 Watt



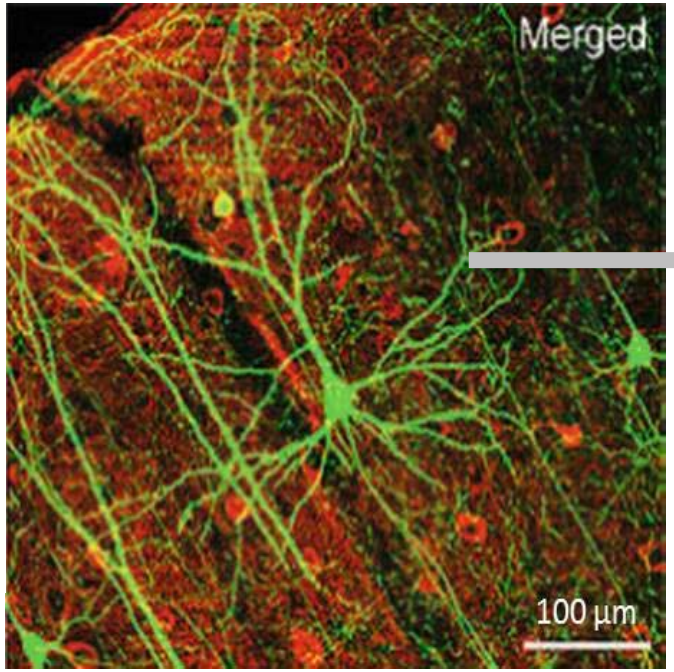
Brain Facts, Society For Neuroscience, 2008, www.sfn.org

Brain Architecture ?



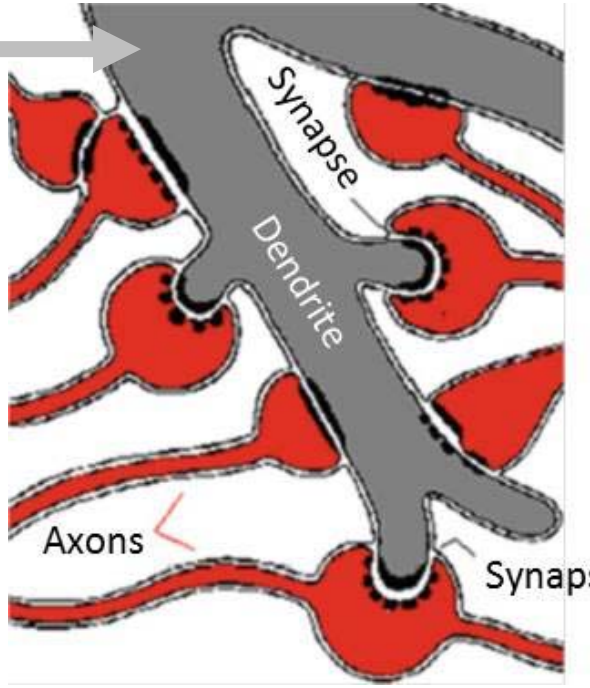
János Szentágothai, *Proc. Roy. Soc. London Ser. B* 201: 219-248)

Cortical column
~ 1 mm³
~ 100.000 neurons



Neuron
< 100.000 µm³
~ 10 Hz (1-100 ms)
~ pJ

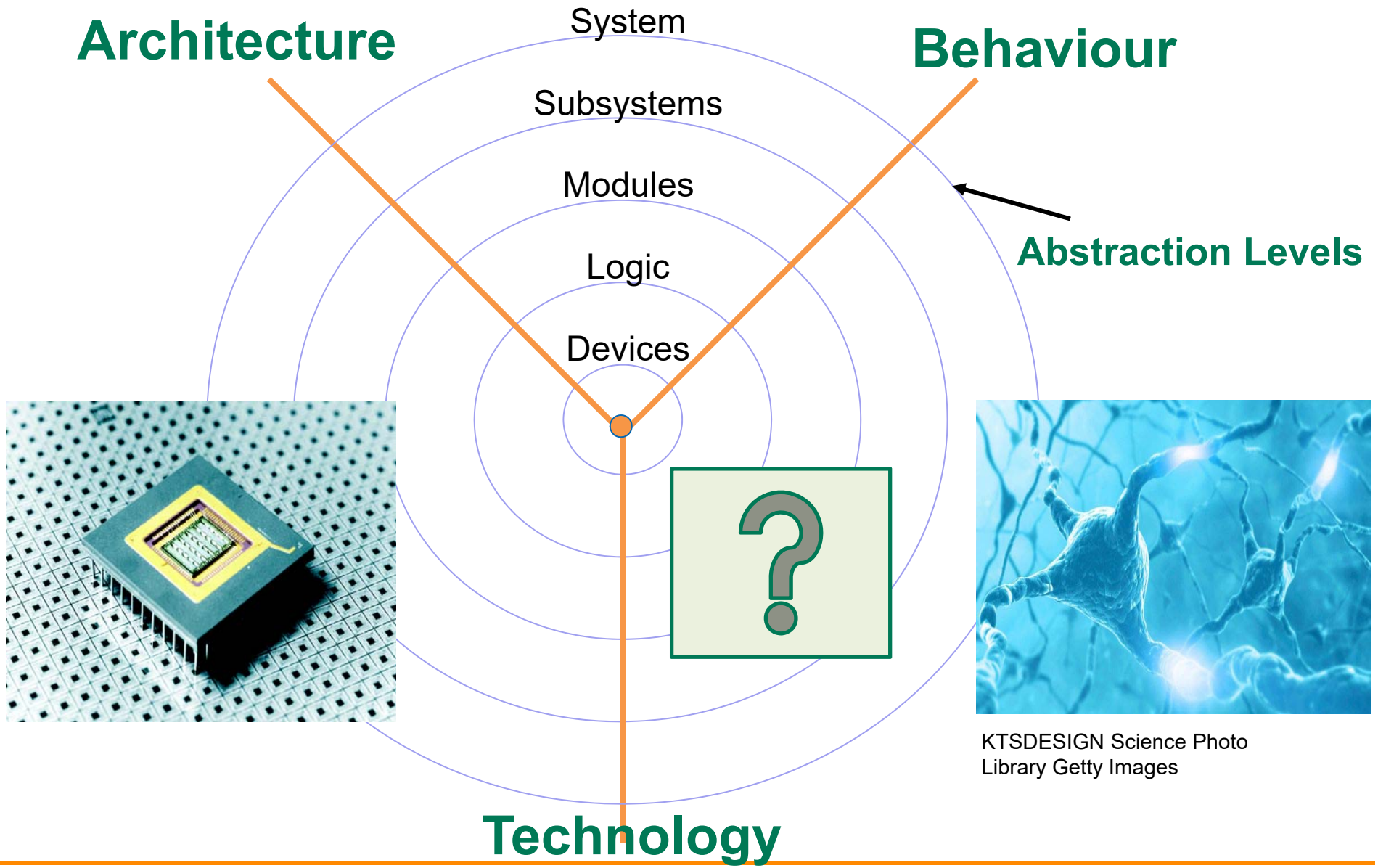
Synapse
~ 0,01 µm³
~ 1ms
~ fJ



http://faculty.washington.edu/chudler/synapse.html

https://en.wikipedia.org/wiki/File:PLoSBiol4.e126.Fig6fNeuron.jpg

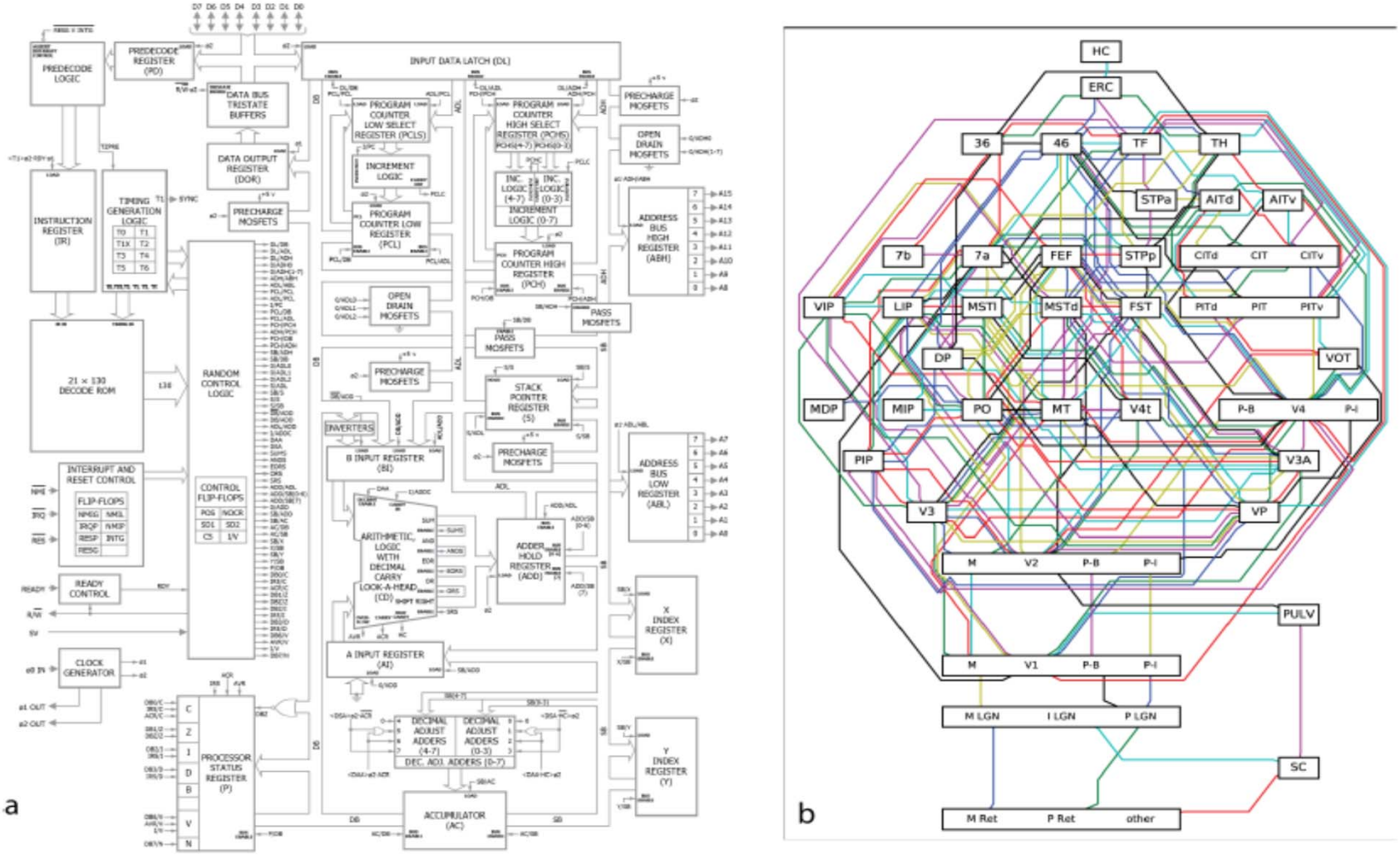
Modular System Design



KTSDESIGN Science Photo Library Getty Images

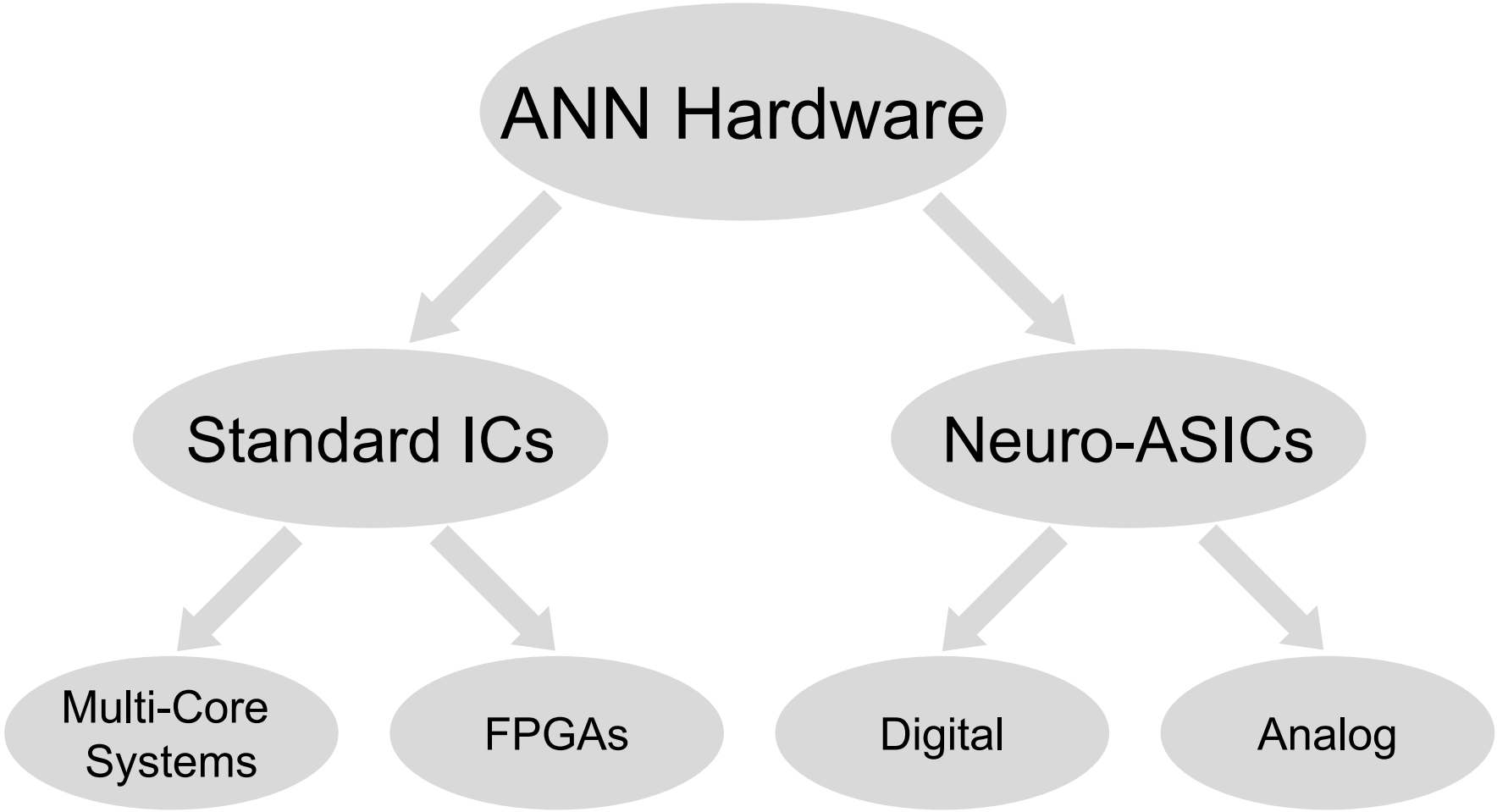
Could a Neuroscientist Understand a Microprocessor?

E. Jonas, K.P. Kording, PLOS May, 2016



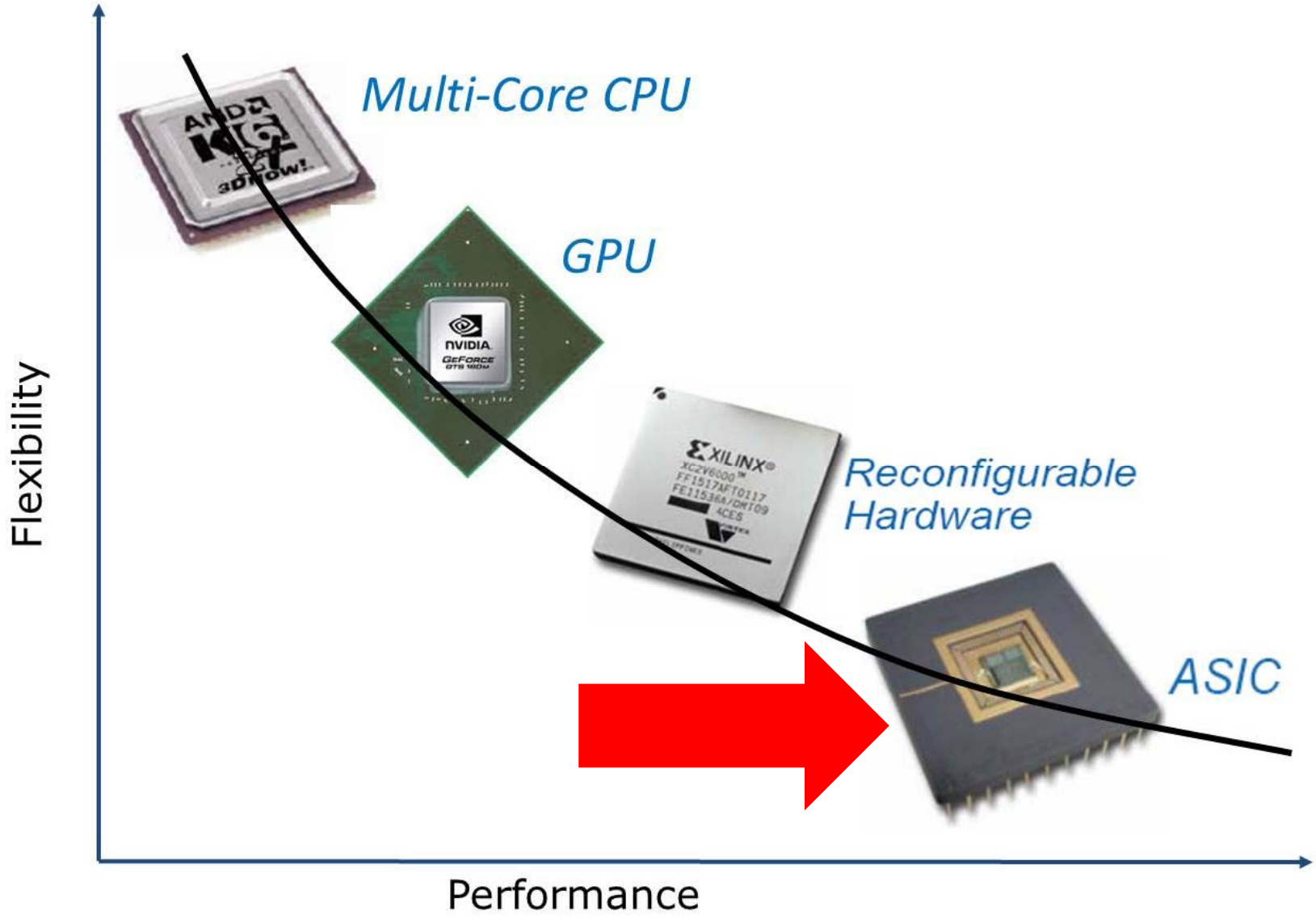
http://dx.doi.org/10.1101/055624

ANN Hardware Approaches

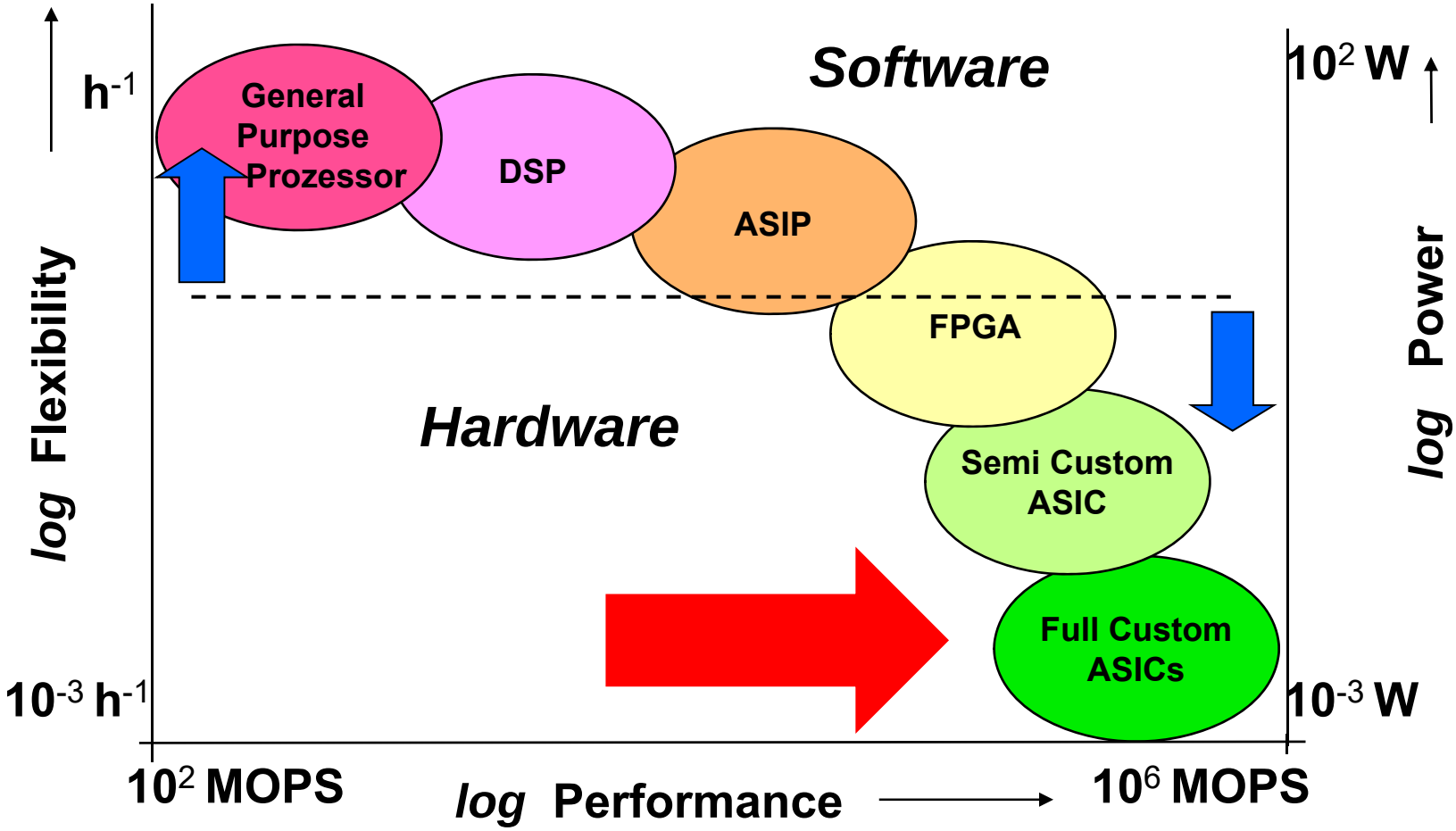


Rückert: Brain-Inspired Architectures for Nanoelectronics, Chap. 18 in Chips 2020, Vol. 2, Springer 2015

Flexibility / Performance Trade-Off



Architectures

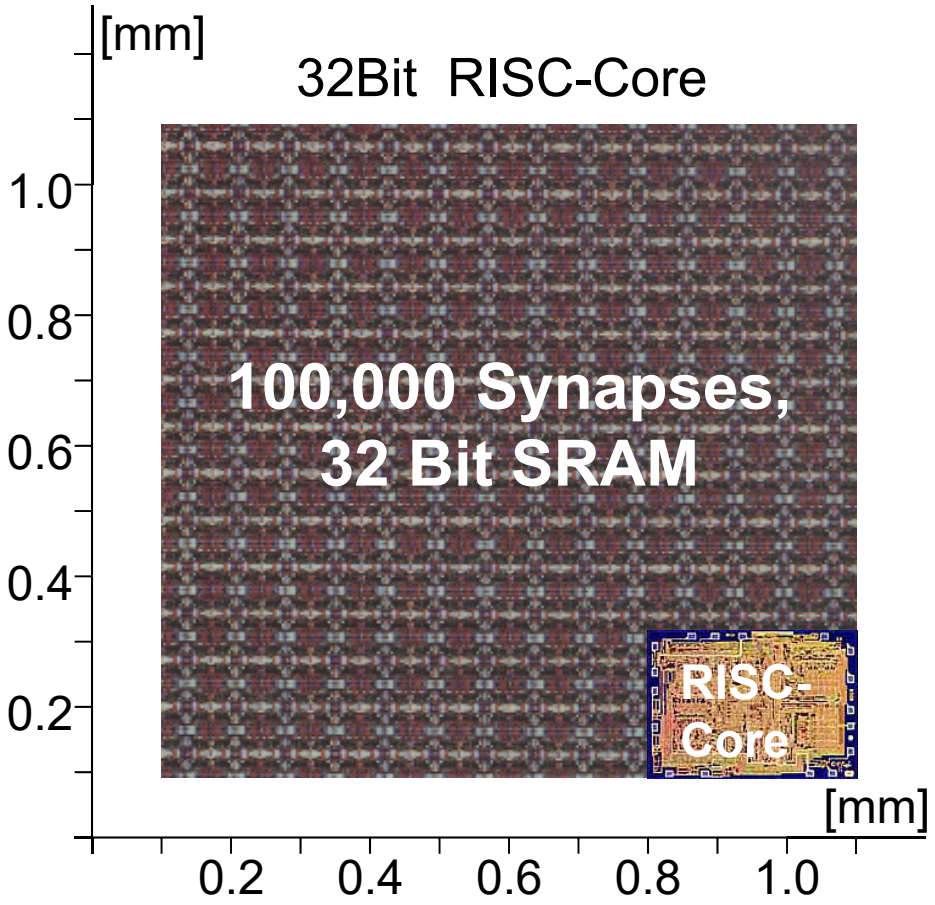


DSP	Digital Signal Processor	FPGA	Field Programmable Gate Array
ASIP	Application Specific Instruction Set Processor	ASIC	Application Specific Integrated Circuit

[Source: EECS, RWTH Aachen]

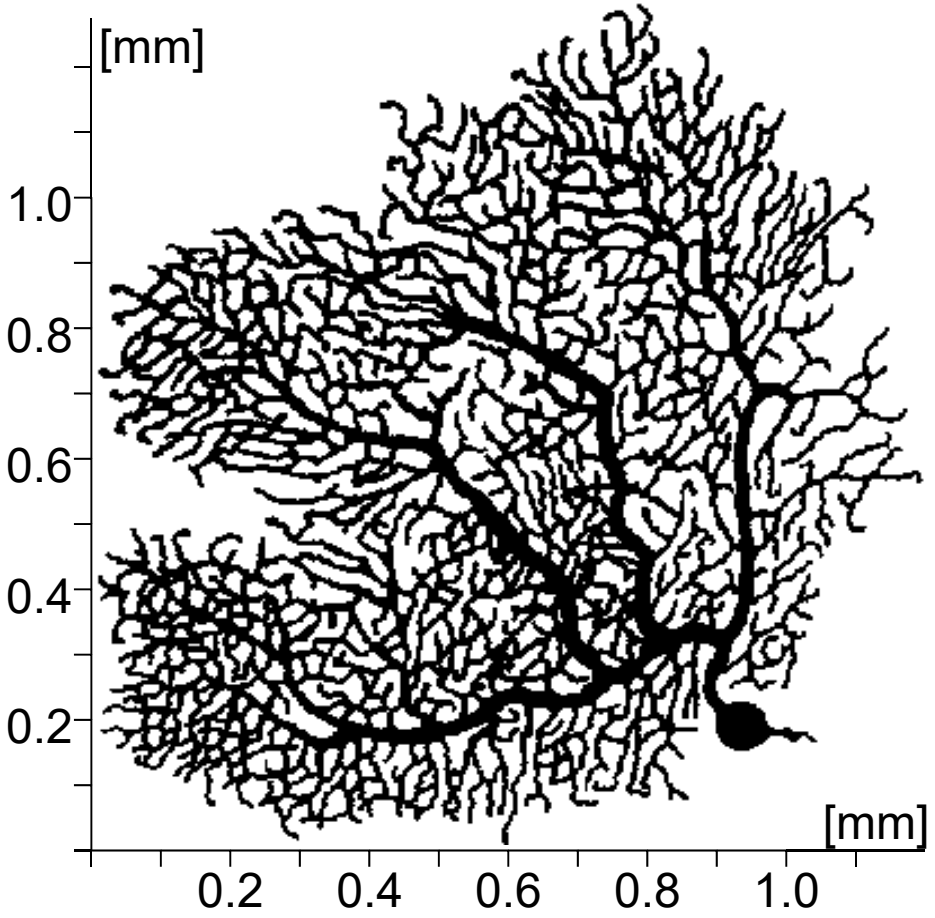
Digital Silicon Neuron

Digital-Neuron (µW)



32 nm CMOS

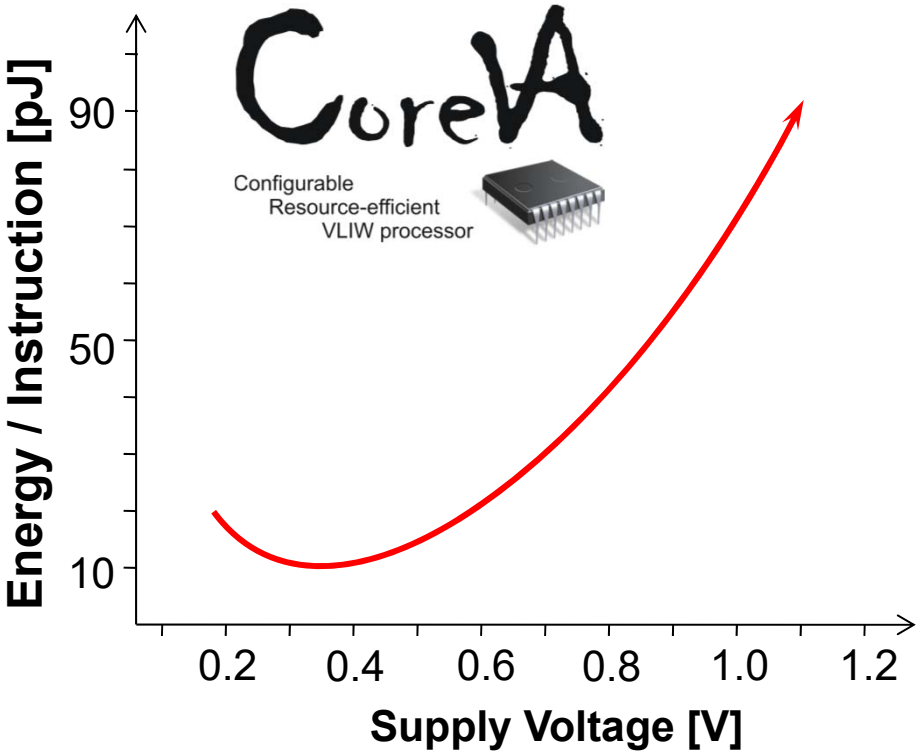
SRAM cell area ~ 0,1µm²



Purkinje cell of cerebellar cortex (nW)

Ultra-Low-Power Processor

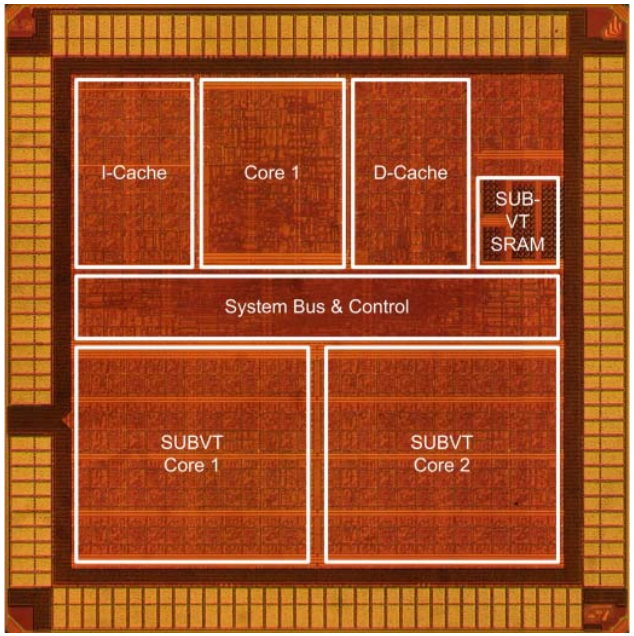
Adaptable supply voltage & operating frequency



Power consumption

1.35 µW – 133 kHz – 0.325 V

10,4 mW – 95 MHz – 1.2 V

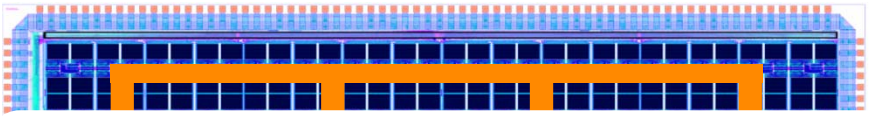


Sub-threshold CMOS

0.36 mm² area, 65 nm

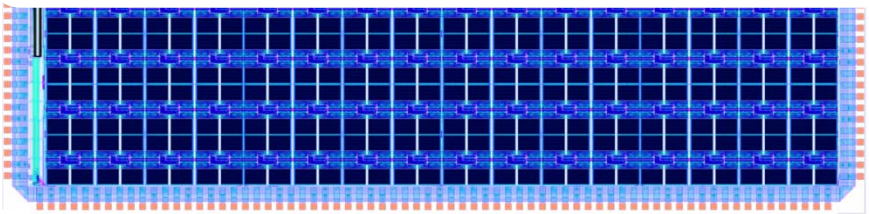
ISSCC2012

From RISC to MPSoC



Power Challenge:

1000W versus 1W!



Scalable Network-on-Chip Architecture

- First Chip Realization
  - 2 Cluster / 8 Cores
- 65 nm CMOS Technology
  - 16 Cluster / 64 Cores
- 32 nm CMOS Technology
  - 64 Cluster / 256 Cores
- 16 nm CMOS Technology
  - 256 Cluster / 1024 Cores

1 cm²

Amdahl´s Law

It roughly estimates the attainable speedup of a parallel algorithm.

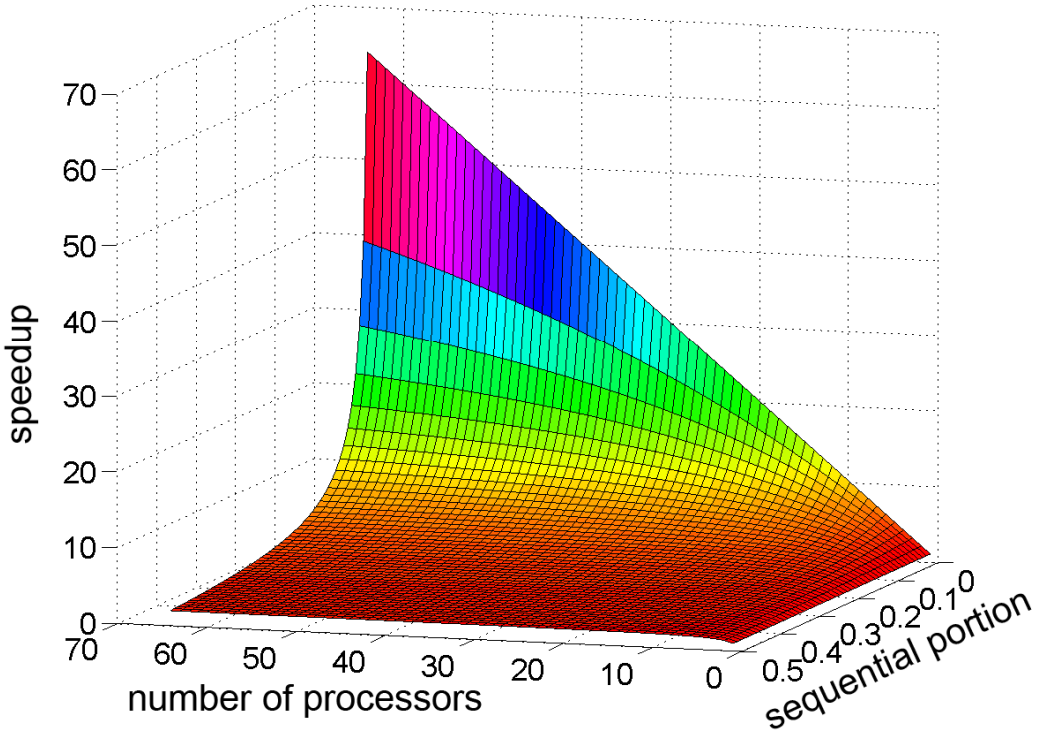
Each algorithm has a parallel and a sequential portion.

- $\tau_{par}+\tau_{ser}=1$
- Runtime with p processors:

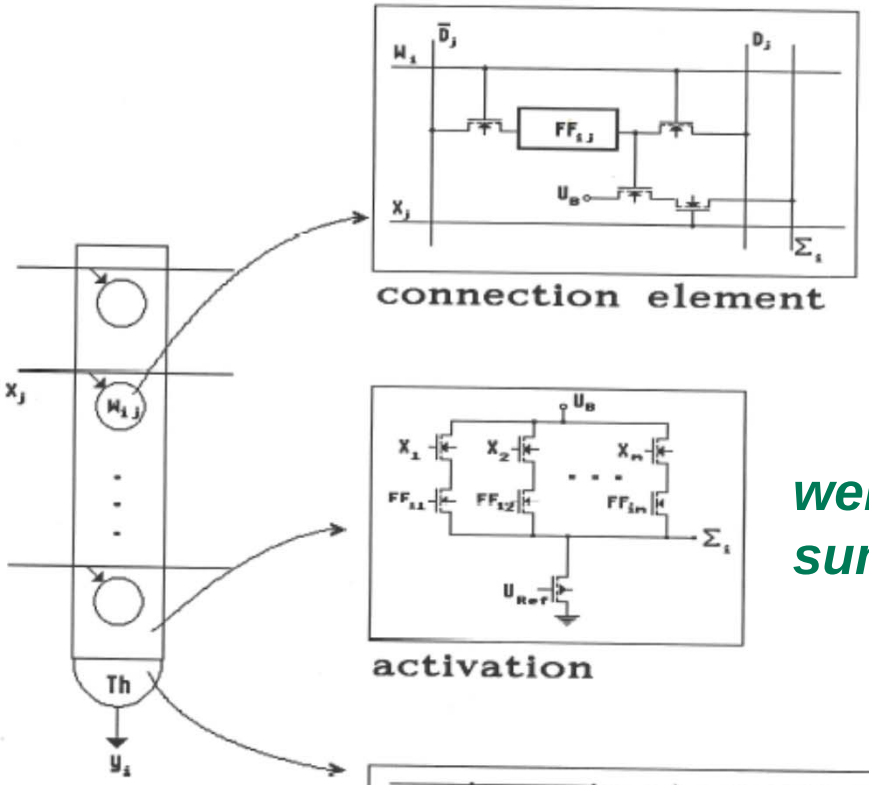
$$t(p) = \frac{\tau_{par}}{p} + \tau_{ser}$$

- Speedup

$$s(p) = \frac{1}{\frac{\tau_{par}}{p}+\tau_{ser}}$$

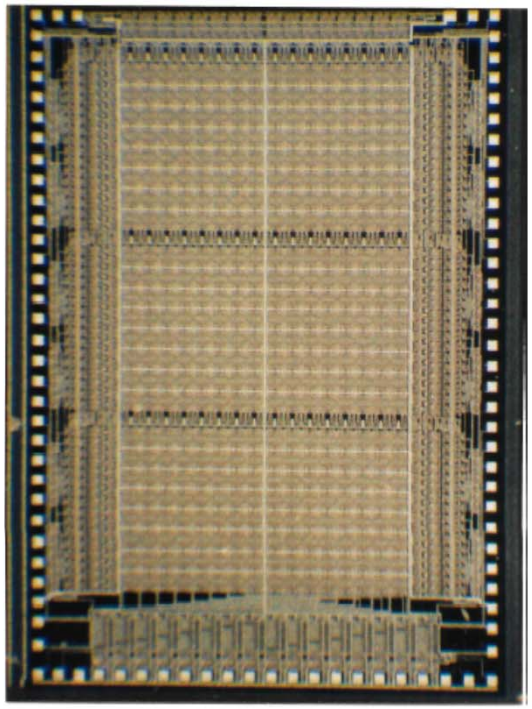


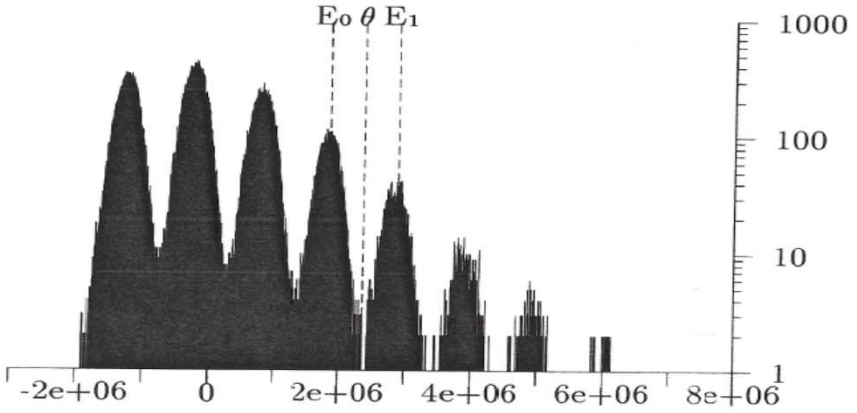
Analog Design Studies



weighted sum

decision





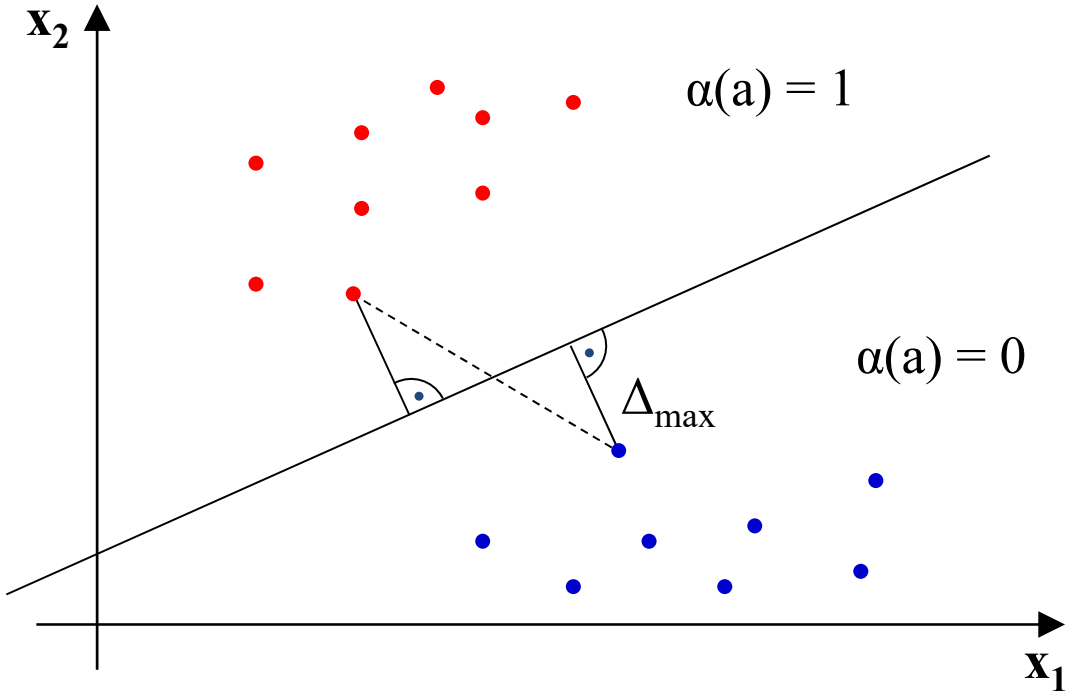
Calculation Precision

Binary Threshold Neuron:

weights w binary
Input vector x binary

m = length of x
l = number of active inputs in x

$$\Delta_{max} < \min \left\{ \frac{\left| \sum_{i=1}^{m} w_{ij} \cdot x_i^t - Th \right|}{\sum_{i=1}^{m} |w_{ij}| \cdot x_i^t + |Th|} \right\}$$



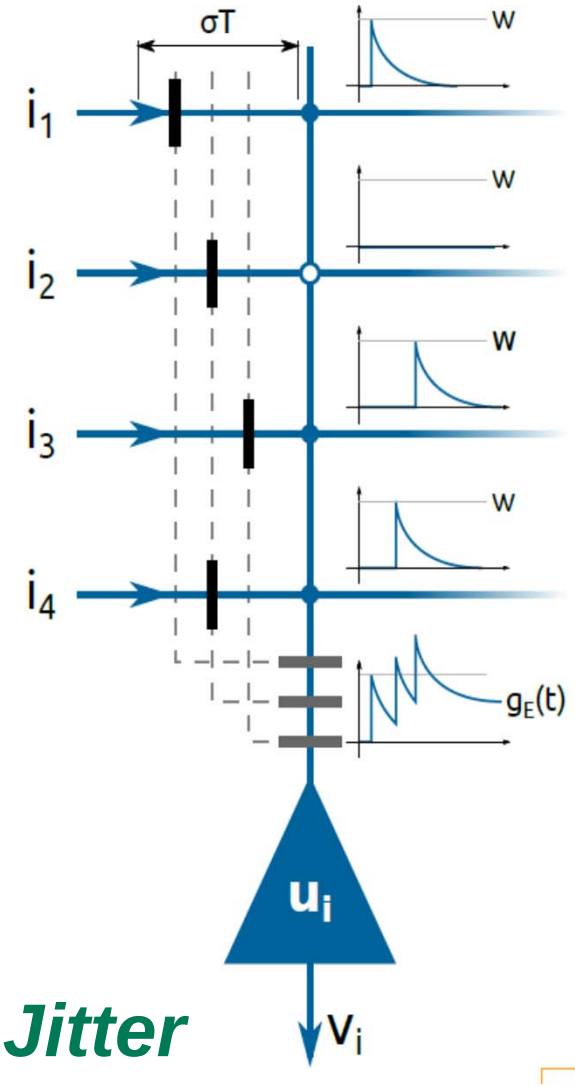
$$\Delta_{max} < \frac{0.5}{2l - 0.5} \approx \frac{1}{4l}$$

Sparse Coding !

Integrate-and Fire-Neuron

Spiking Neurons

- **Given:**
  - Presynaptic excitatory spikes, jitter σT
  - Binary synapse weights:
    $w_0 = 0$ (not connected), $w_1 = w$
- **Goal:** ***(Optimization Problem)***
  Find parameters $\mathcal{P}$ with
  - N input spikes $\Rightarrow P(\text{spike})$ large
  - $N - 1$ input spikes $\Rightarrow P(\text{spike})$ small



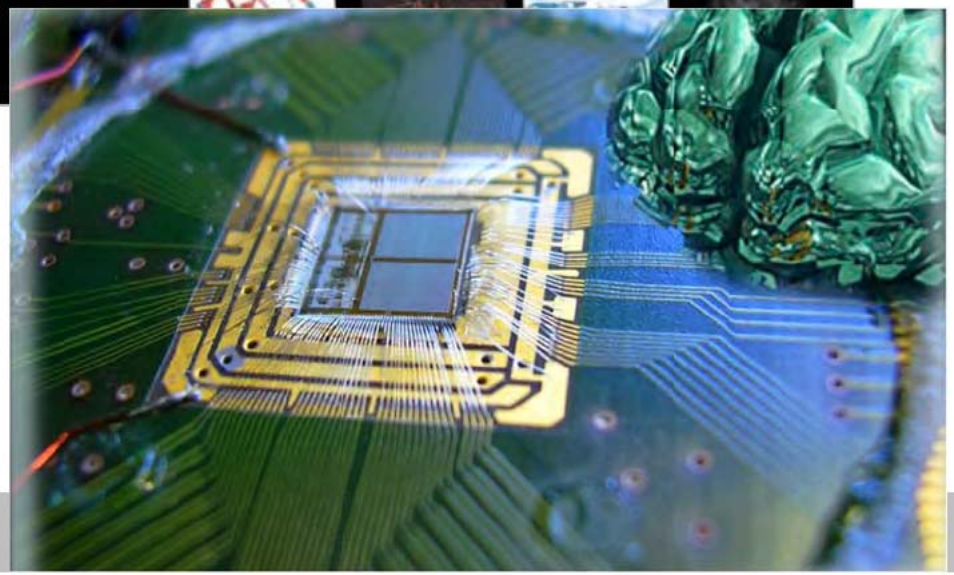
Asynchronous Behaviour ➔ Spike Jitter

Human Brain Project

HBP
The Human Brain Project

Two major goals of the Human Brain Project

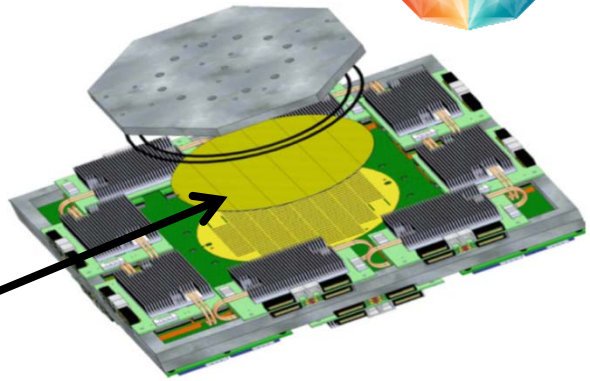
2. Developing powerful simulations, both in conventional computers and in specialized hardware



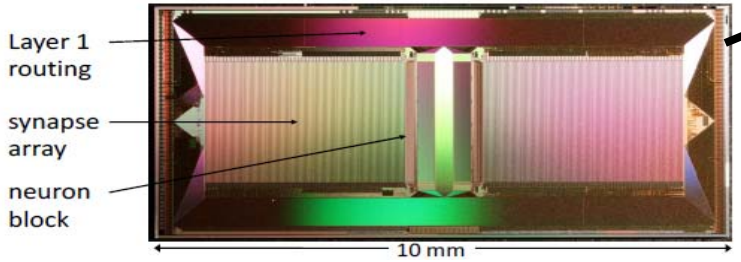
Human Brain Project

Three (neuromorphic) computing platforms:

- High Performance Computers (HPC)
- Wafer-Scale-System (Heidelberg)

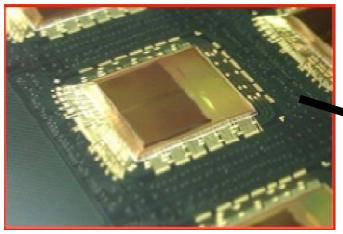


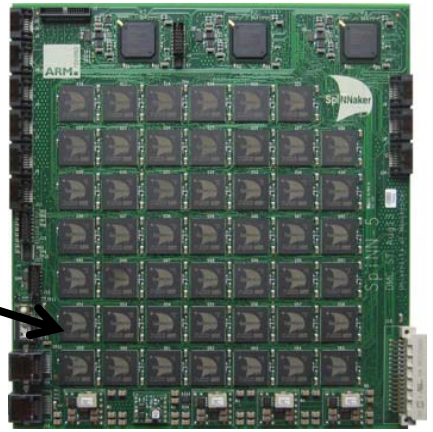
analog Neurons



- Multi-Core-System (Manchester)

digitally emulated Neurons





HBP Neuromorphic Computing Concepts

MANY-CORE NUMERICAL MODEL SYSTEM

0.5 – 1 Million ARM processors – address-based, small packet, asynchronous communication – running at real-time

Location : Manchester (UK)

PHYSICAL MODEL SYSTEM

Local analog computing with 4 Million neurons and 50 Million synapses – binary, asynchronous communication – emulation speed is x 10 000 real-time

Location : Heidelberg (Germany)

Benchmarking of Neuro-ASICs

Neuro-ASICS	Feature Size	Die Size	Neurons	Synapses	Bit/Syn.	ESE
SpiNNaker	130nm	1,02cm²	1,600	*128x10^6	8	10^{-8}J
TrueNorth	28nm	4,30cm²	10^6	256x10^6	1	10^{-11}J
HICANN	180nm	0,50cm²	8-512	114,688	4-8	10^{-10}J
Neurogrid	180nm	1,68cm²	65,536	*16x10^6	#13	10^{-10}J
numbers per chip				*off-chip	#shared	ESE: Energy/synaptic event

Human Brain: about 10^{11} Neurons, 10^{15} Synapses, 30W average power

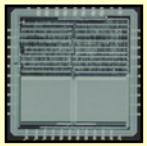
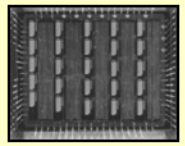
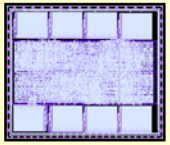
Neurocomputer Projects

- CNAPS, Adaptive Solutions, USA, 1990 (Digital, ASIC)
- Synapse, Siemens, Germany, 1990 (Digital, ASIC)
- ETANN, INTEL, USA, 1990 (EEPROM, Analog, ASIC)
- MoNA, (mixed signal), Porrmann/Rückert
- ….
- SpiNNaker, Univ. of Manchester, U.K., 2007 (Digital, ASIC)
- HICANN, Universität Heidelberg, 2010 (Analog/Digital, ASIC)
- Neurogrid, Standford University, 2007 (Analog/Digital, ASIC)
- Neurosynaptic Core TrueNorth, IBM, USA, 2010 (Digital, ASIC)
- BlueHive, Cambridge, U.K., 2012 (FPGA)
- Cadence, Tensilica, Google, ARM, ….
- …

Overwiew

- Introduction — Background
- Technology — Femtoelectronics for ANN
- Architectures — Design Alternatives
-  Applications — Cognitive Robotics
- Discussion — Questions

Basic Cognitive Functions

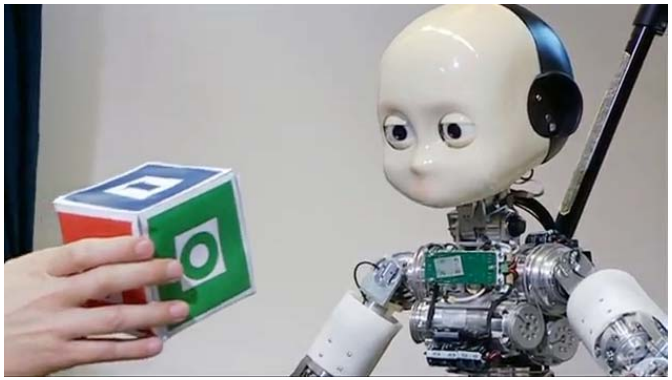
IN → ANN → Out		Task	Evaluation	Model	Realization
discrete	discrete	association	storage efficiency	NAM	
continuous	discrete	classification	error probability	SOM	
continuous	continuous	controlling, approximation	distance measure	LCNN	

- ANN = Artificial Neural Networks
- SOM = Selforganizing Maps

NAM = Neural Associative Memory
LCNN = Local Cluster Neural Nets

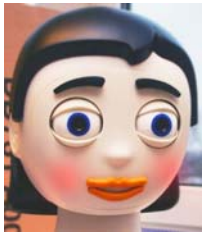
CITEC Research Topics

Motion Intelligence



Attentive Systems

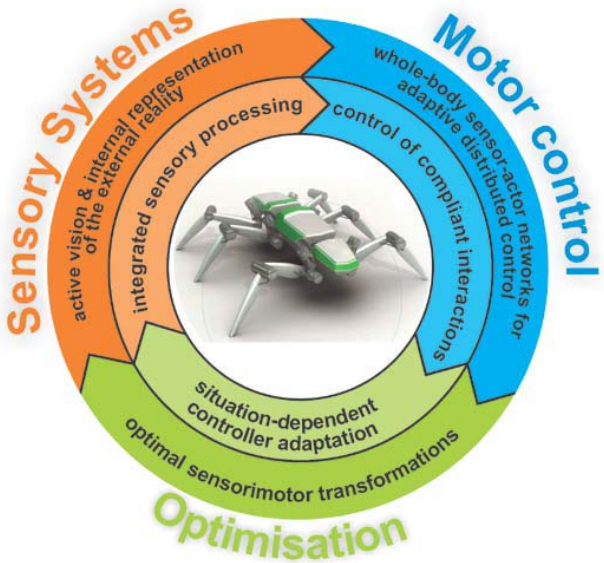
Ressource-efficient architectures for cognitive systeme: Cognitronics



Memory and learning

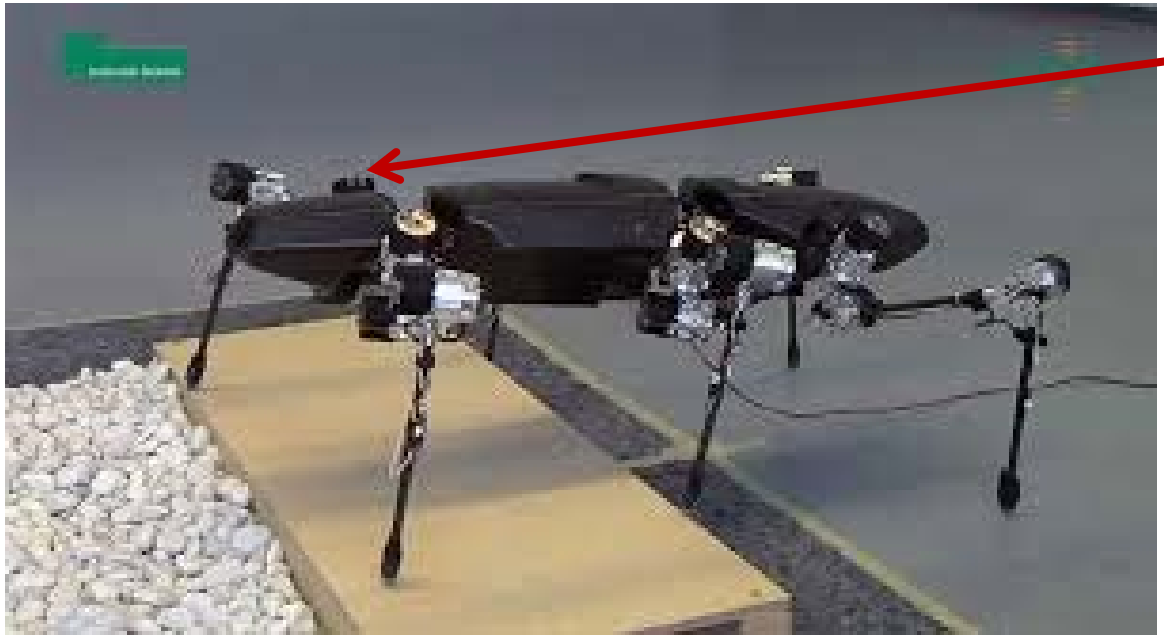
Situated Communication

The CITEC Hardware in the loop



How to generate complex behaviour in the limits of restricted energy resources?

HECTOR - The six-legged walking robot
see YouTube



Axel Schneider
Martin Egelhaaf
Volker Dürr
Marc Ernst
Ulrich Rückert
Elisabetta Chicca

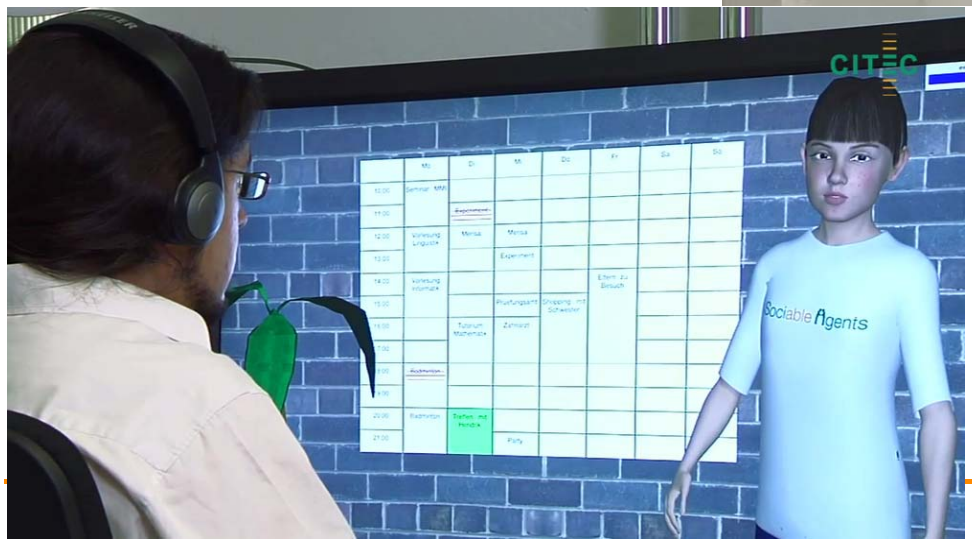
Cognitive Interaction Technology

Long-term experiments in real world scenarios

- 1000 Processors
- TeraByte Memory
- Embedded Sensors

~1dm³

< 10W



Art Exhibition Hall, Bielefeld

www.cit-ec.de

CITEC Smart Home Apartment

Miele

esa

Overwiew

- Introduction — Background
- Technology — Femtoelectronics for ANN
- Architectures — Design Alternatives
- Applications — Cognitive Robotics

- Discussion — Questions

Bio-inspired Principles

- ***Sparse codes and acticity (Cell Assemblies)***

  reduced power consumption - simple communication

- ***Asynchronous information processing***

  reduced power consumption - higher robustness

- ***Massiv Parallelism and Redundancy***

  reduced power consumption - higher robustness

- ***Continuous Self-organization***

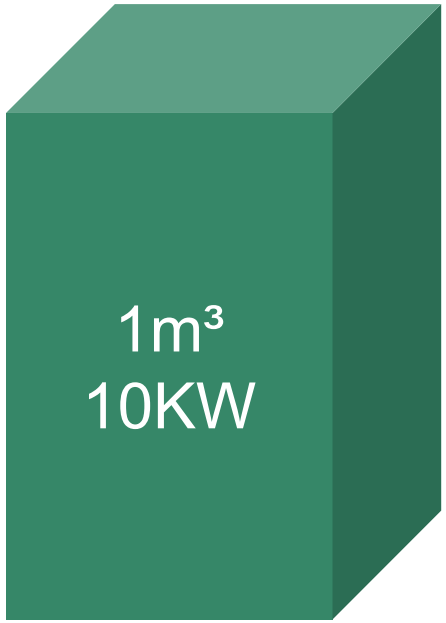
  fault-tolerance - optimal use of resources

- ***Stability in large, dynamic networks***

  local rules for stable and robust global behaviour

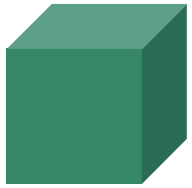
System Integration Roadmap

1 Billion Cores
(Peta Byte)

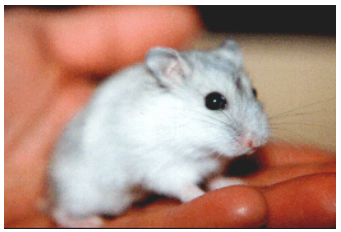


System in a Rack

1 Mill. Cores
(Tera Byte)



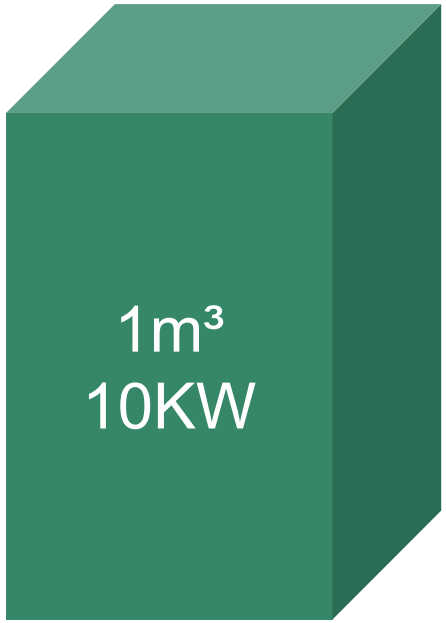
System in a Box



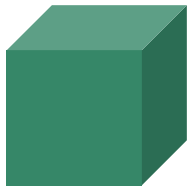
1.000 Cores
(Giga Byte)

System in a Dice

System Integration Roadmap

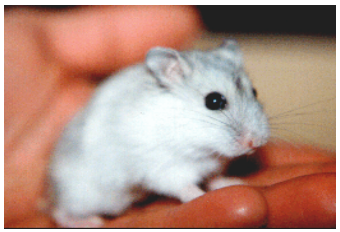


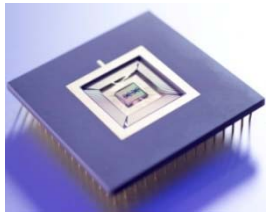
System in a Rack



1dm³
100W

System in a Box





1cm³ ; 1W

System in a Dice

Conclusion

At least in the next decade neural network implementations are dominated by nanoelectronics.

Until fundamental concepts become better understood, the main advantages of silicon should not blind us to alternative neural network design.

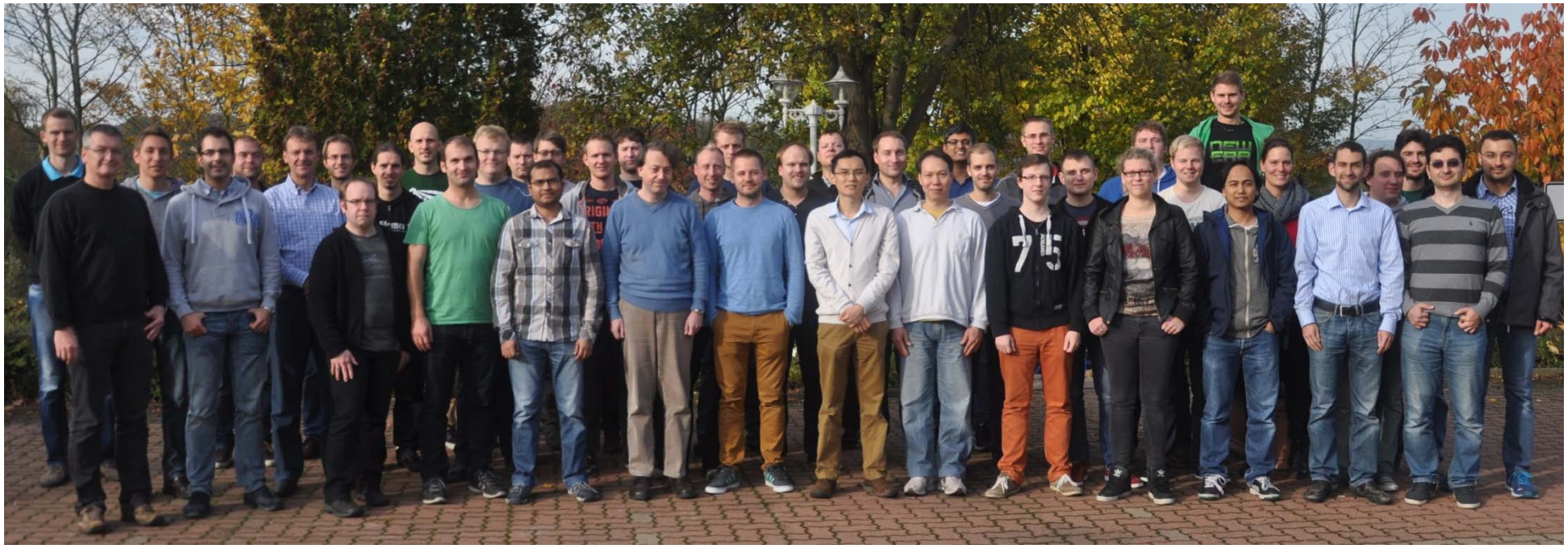
E. R. Caianiello

Universität Bielefeld Cognitive Interaction Technology Center of Excellence

CITEC

Thank your for your attention!

Ulrich Rückert



The Cognitronics and Sensor Systems Research Group