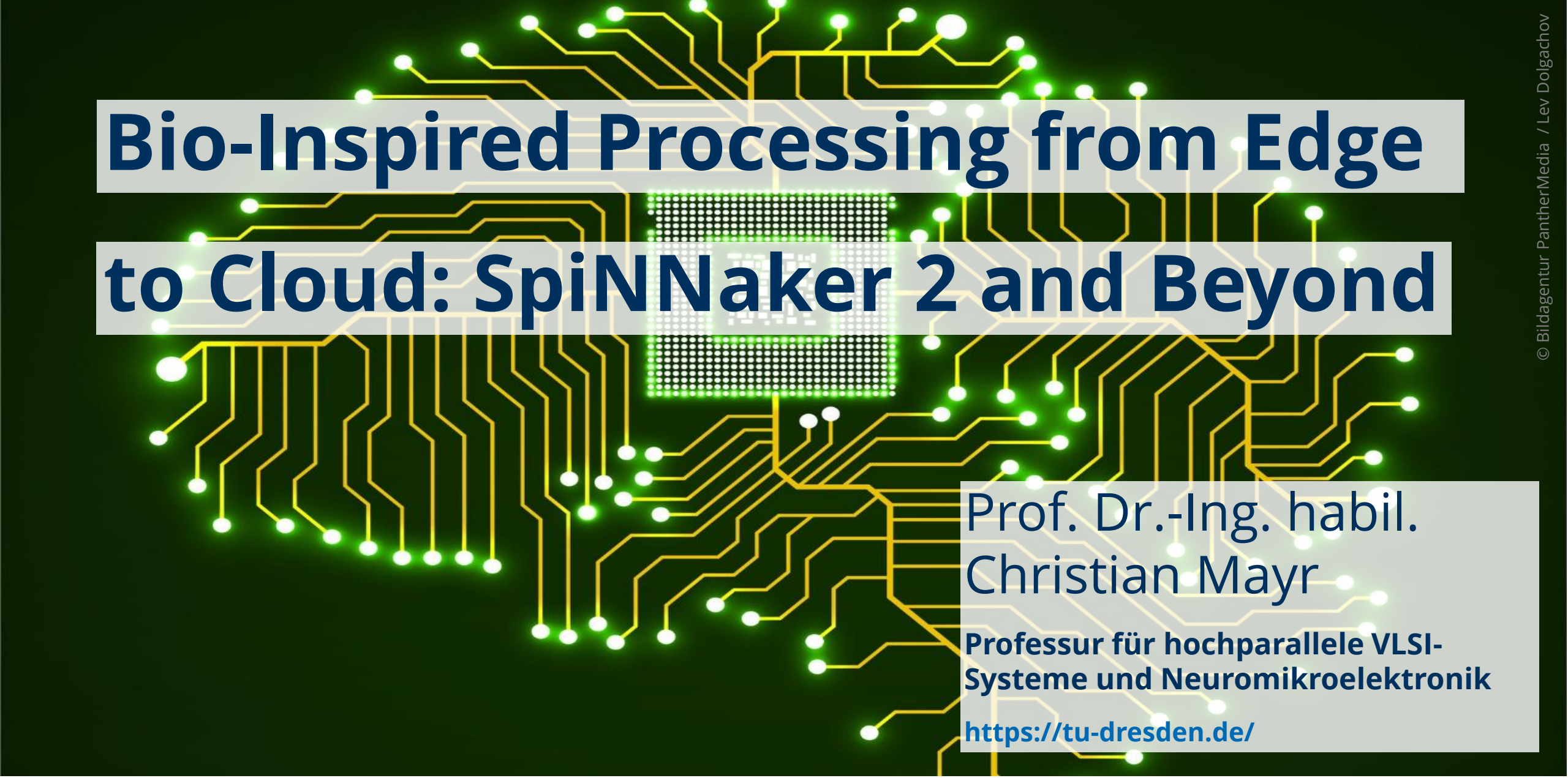


Bio-Inspired Processing from Edge to Cloud: SpiNNaker 2 and Beyond

Prof. Dr.-Ing. habil. Christian Mayr

Professur für hochparallele VLSI-Systeme und Neuromikroelektronik

https://tu-dresden.de/

Chair of Highly Parallel VLSI Systems, TU Dresden

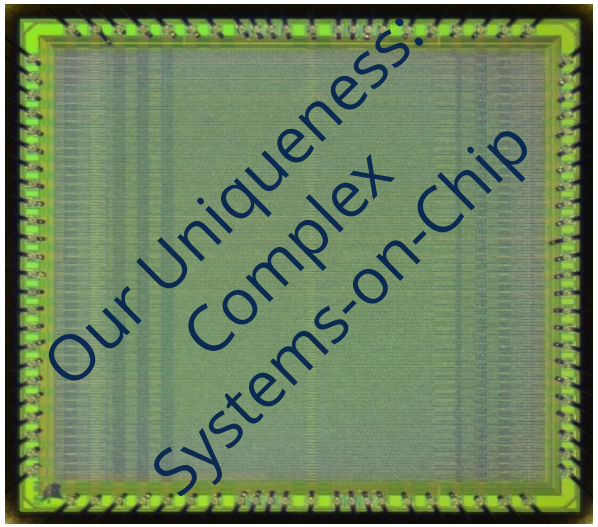
- Head of Chair: **Prof. Christian Mayr**
- 40+ group size: Analog/Digital designers, AI algorithm people, medical/biotechnologists, roboticists …
- Our "biotope" spinoffs: Racyics GmbH, Siliconally GmbH, Coinbau GmbH, SpiNNcloud Systems GmbH, Silicon Matter u.G. **>250 people**
- Local environment: Silicon Saxony is **Europes largest microelectronics- / ICT-location** and the fifth-largest worldwide. More than every third chip produced in Europe has the label „Made in Saxony".

Sensor interfaces, communication

Multi-processor system design

Advanced CMOS design (Adaptive body biasing in 22nm FDX)

Neuromorphic Circuits, Neural Interfaces and AI accelerators

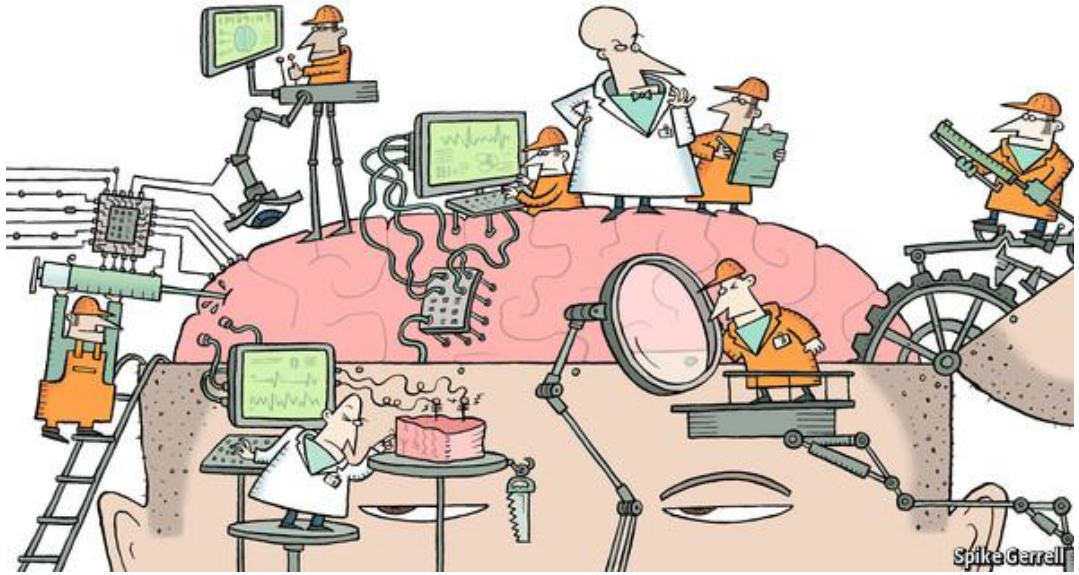


2012 to 2022: 26 multi-component SoC tapeouts from 90nm to 22nm CMOS

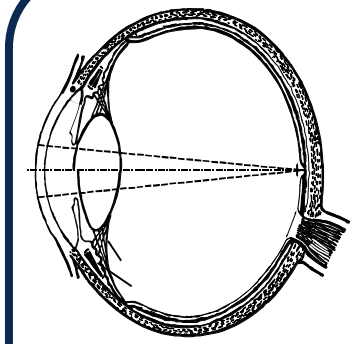
Chips/Demonstrators used by ca. 30 external groups

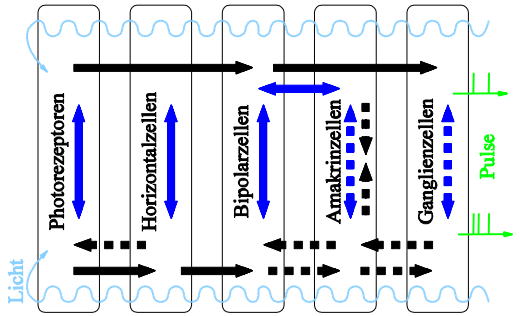
From single transistor to 70k chip system: We cover 14 orders of magnitude!

- SpiNNaker 2
- Bio-Inspired AI: Algorithms & Use Cases from Edge to Cloud
- Domain-Specific SpiNNedge Chips
- Outlooks
  - Merging DNN, SNN and Symbolic AI
  - Use Case Quantum
  - (Use Case Drug Discovery)
  - (SpiNNaker3)



The Eye

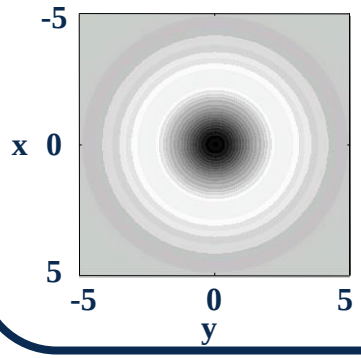


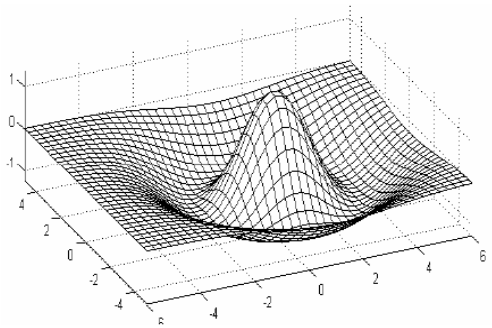


Biological Processing Chain

100:1 compression via sparse event coding

Equivalent Image Convolution

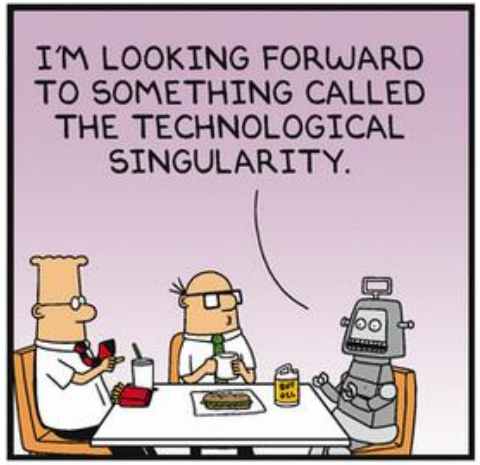




Additional mechanisms besides feed-forward compression:

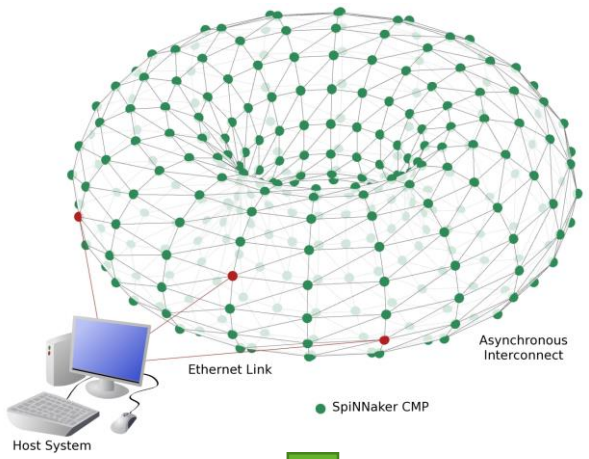
- Feedback/gating in LGN
- Attentional feedback via saccades
- Sensory prediction
- Delta-Encoding/Adaptation
- Etc

Neuromorphic Principles:



- The Brain removes redundancy and non-relevant information at every step: Sparsity
- Computation and communication in the brain scale with activity, they are energy-proportional
- The brain constantly adapts and predicts, thus it's very robust
- The brain is highly parallel&asynchronous, i.e. no Amdahl limit
- Various higher-level concepts to be taken from biology:
  - Attention/Gating layer/Region of interest ->sparsity at network level
  - Internal physics/reasonableness model
  - Decision confidence
  - Online, low-resource learning

SpiNNaker2



10 Mio ARM M4F cores in 22FDX, each with...

Dynamic Power Management
- DVFS and PSO

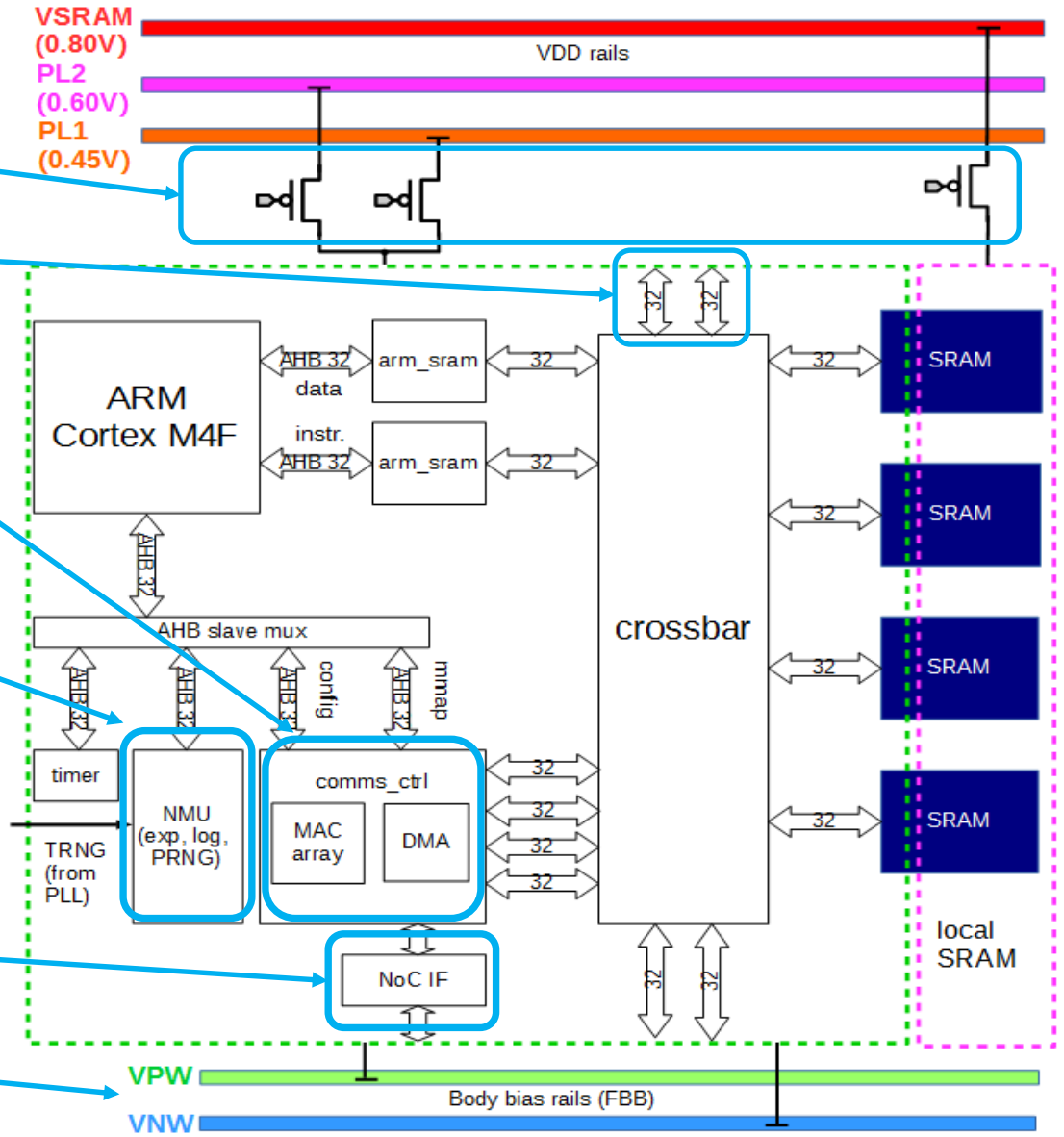
Memory sharing
- Synchronous access to neighbor PEs

Multiply-Accumulate accelerator
- MAC array with DMA

Neuromorphic accelerators
- Exp/log
- Random numbers (PRNG, TRNG from ADPLL noise)

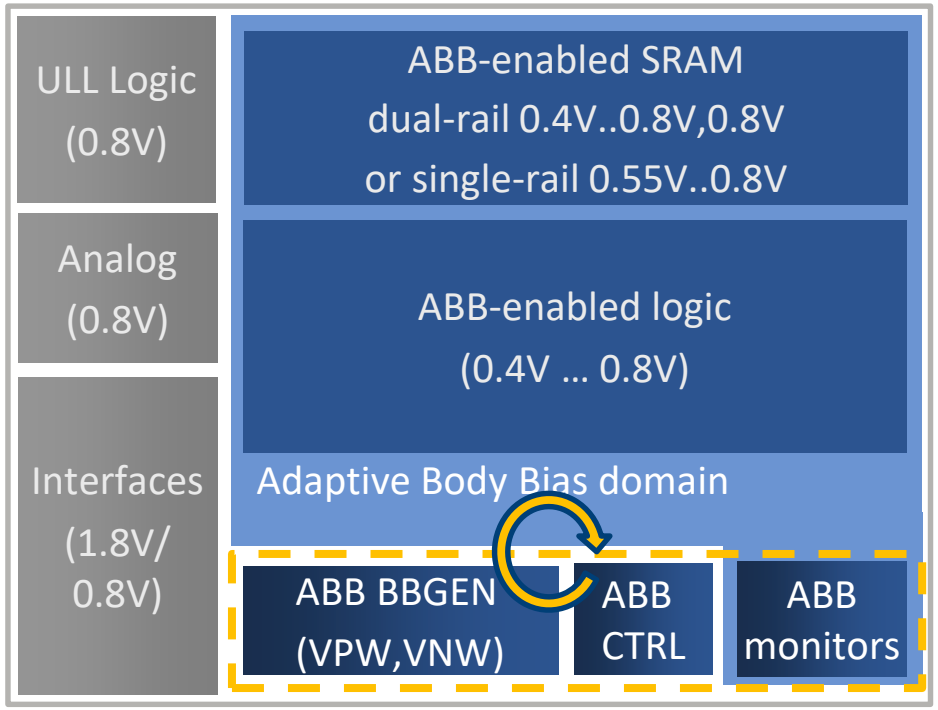
Network-on-Chip
- On- and off-chip memory access
- SpiNNaker packet (spike) handling

Adaptive Body Biasing

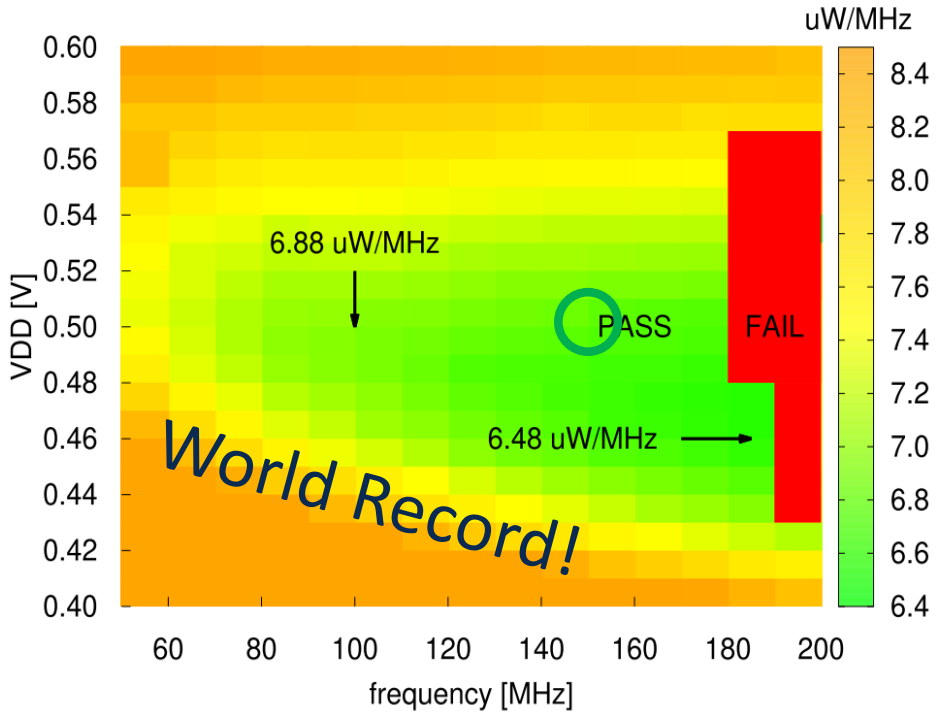


Adaptive Body Biasing in 22nm FDSOI

- Racyics ABX® body bias generator IP
- Racyics ABX® implementation methodology*
  → Improved PPA
  + standard cells + SRAM



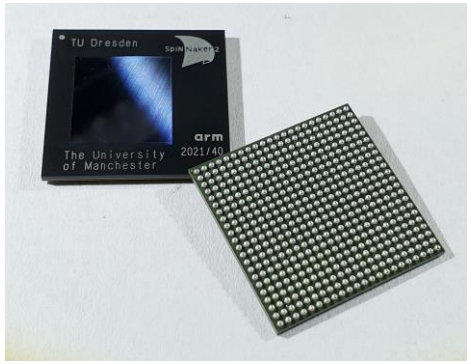
Arm Cortex-M4 Testchip (MPW2213)
with FBB [2]



[2] S. Höppner *et al.*, "How to Achieve World-Leading Energy Efficiency using 22FDX with Adaptive Body Biasing on an Arm Cortex-M4 IoT SoC," *ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC)*

- Optimal leakage
- Improved dynamic power
- In analog: adapt speed/leakage (e.g. in SC circuits) at runtime

SpiNNaker2

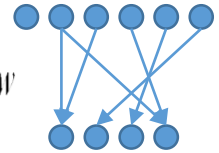


SpiNNaker2 Chip:

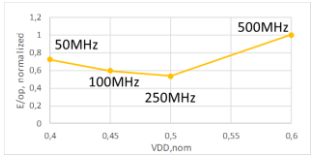
- 153 ARM cores
- >100 person design team
- 22FDX Global Foundries
- Developed in EU flagship Human Brain Project
- Development cost: >20Mio
- Deployment cost: >10Mio

Hybrid design for deep neural networks, spiking neural networks and symbolic AI

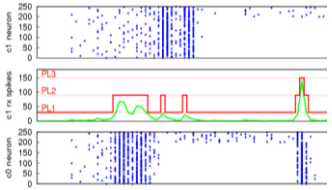
$$\tau_w \dot{w} = a(V_{Mem} - V_{rest}) - w$$



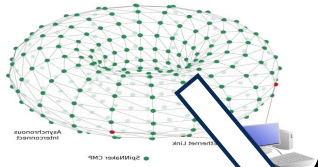
Outperforming Intel, Nvidia, Google on real time AI



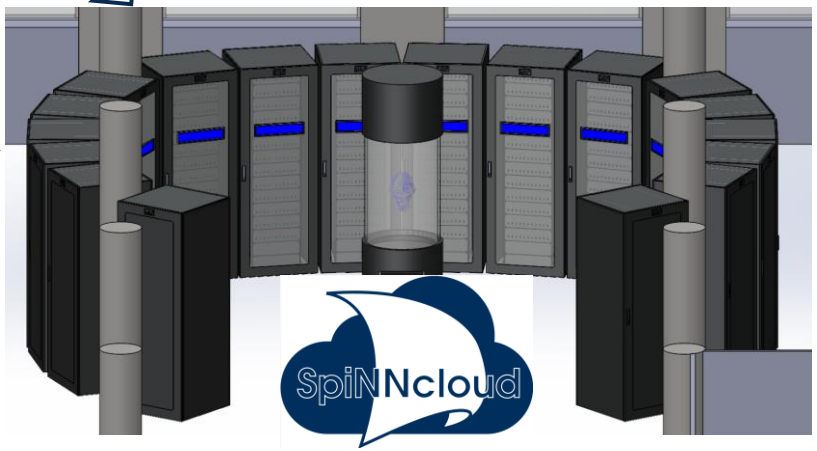
Brain-inspired **dynamic data sparsity**, i.e. ultra-efficient highly-parallel operation of AI algorithms on streaming data



Largest real-time AI platform worldwide, **10^14 parameters**
3 PFLOPS CPU
0.4 ExaOPS in AI accelerator



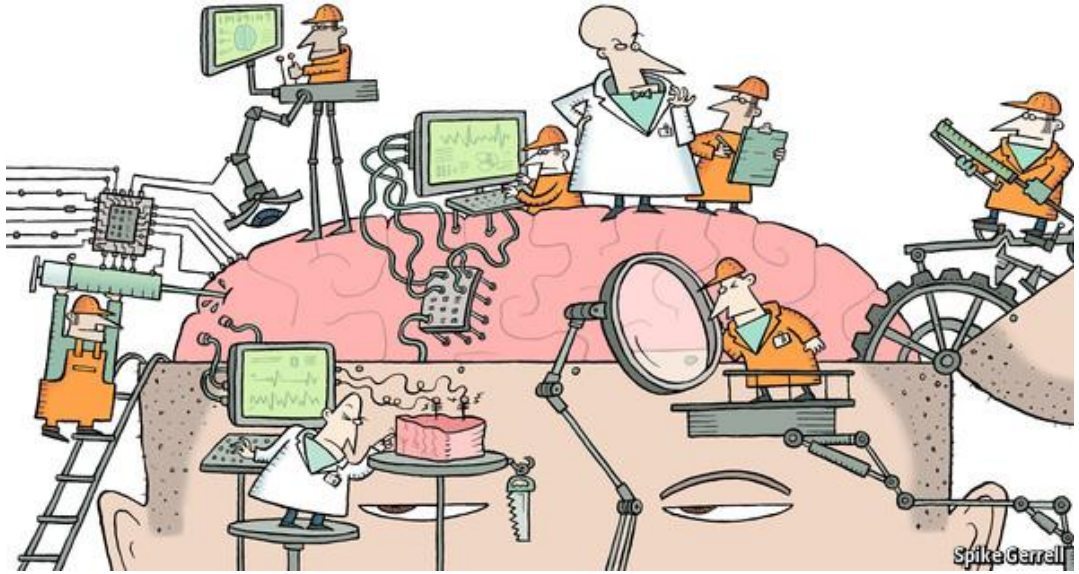
Physical: 10^7 processors, 70.000 chips, 14 racks, 100.000.000.000.000 transistors



Brain-like capabilities for autonomous systems

Drug Screening, fast medical data analysis

- SpiNNaker 2
- Bio-Inspired AI: Algorithms & Use Cases from Edge to Cloud
- Domain-Specific SpiNNedge Chips
- Outlooks
  - Merging DNN, SNN and Symbolic AI
  - Use Cases Quantum

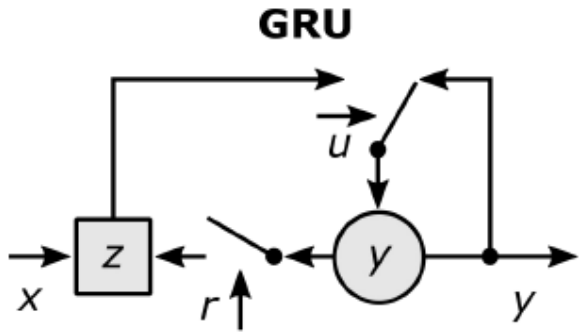


DVS Demo

Event-based Gated Recurrent Unit (EGRU)

Based on GRU, a very performant recurrent architecture for deep learning.

GRU Equations:



Update gate

Reset gate

$$\mathbf{u}^{\langle t \rangle} = \sigma\left(\mathbf{W}_u \left[\mathbf{x}^{\langle t \rangle}, \mathbf{y}^{\langle t-1 \rangle}\right] + \mathbf{b}_u\right), \quad \mathbf{r}^{\langle t \rangle} = \sigma\left(\mathbf{W}_r \left[\mathbf{x}^{\langle t \rangle}, \mathbf{y}^{\langle t-1 \rangle}\right] + \mathbf{b}_r\right),$$

$$\mathbf{z}^{\langle t \rangle} = g\left(\mathbf{W}_z \left[\right.\right.$$

Add event-ge

$$y_i^{\langle t \rangle} = c_i^{\langle t \rangle} H\left($$

Heaviside step functi

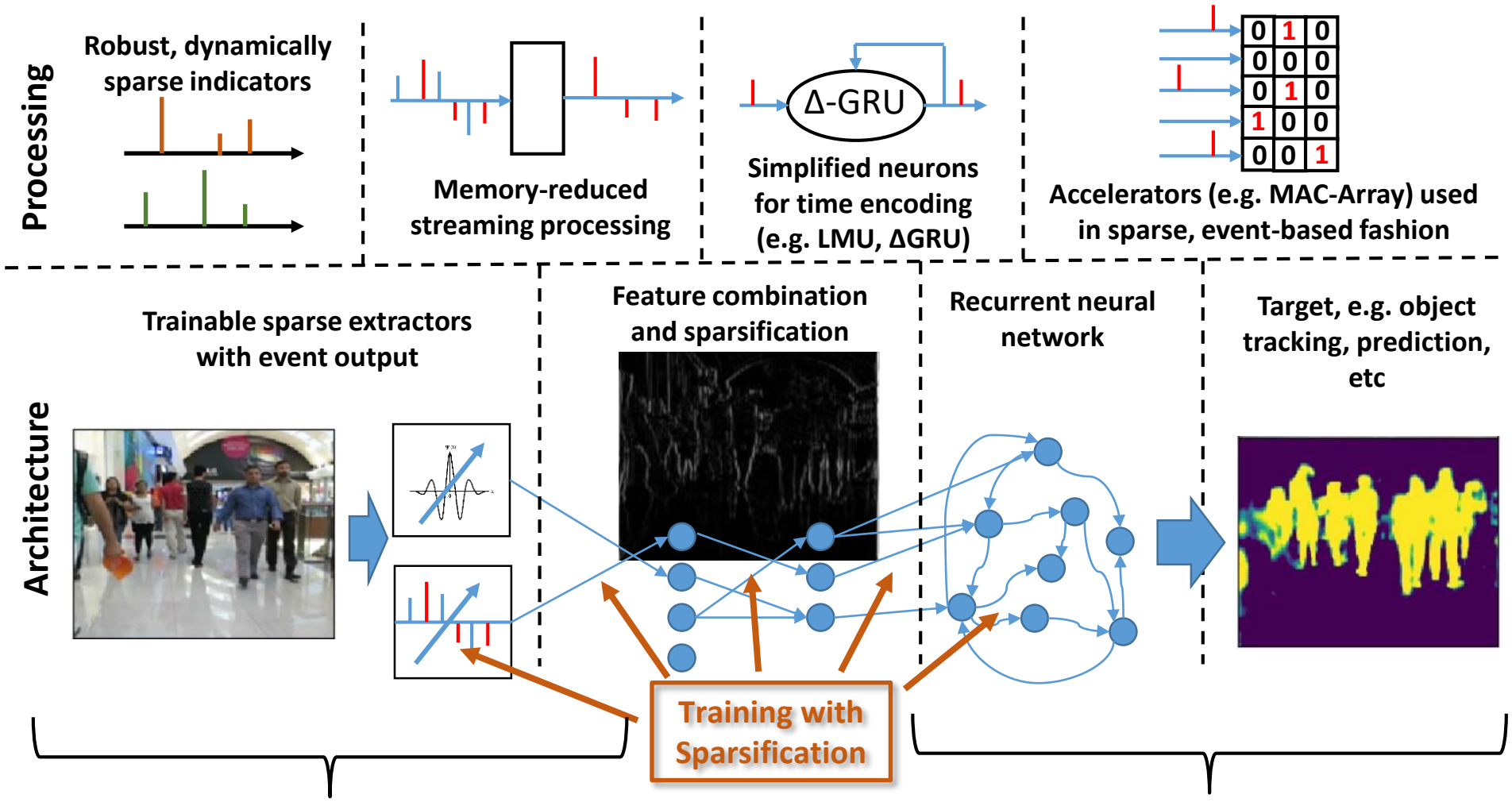
DVS 128 gesture recognition:

reference	architecture (# units)	para-meters	effective MAC	accu-racy	activity sparsity	backward sparsity
He et al. [23]	LSTM (512)	7.35M	7.34M	86.81%	-	-
Innocenti et al. [31]	AlexNet+LSTM+DA	9.99M	638.25M	97.73%	-	-
ours	GRU (1024)	15.75M	15.73M	88.07%	0%	-
ours	**EGRU** (512)	5.51M	4.19M	88.02%	83.79%	53.55%
ours	**EGRU** (1024)	15.75M	10.54M	90.22%	82.53%	56.63%
ours	**EGRU**+DA (1024)	15.75M	10.77M	97.13%	78.77%	58.20%

- SpiNNaker 2
- Bio-Inspired AI: Algorithms & Use Cases from Edge to Cloud
- Domain-Specific SpiNNedge Chips
- Outlooks
  - Merging DNN, SNN and Symbolic AI
  - Use Case Quantum

3uW always on gesture recognition

Event-Based DNN/RNN with Domain-Specific Preprocessing



SpiNNedge ASICs:
Sparse Preprocessing and Neural Network Acceleration for Edge Applications (radar, video, audio, robotic, biomedical)

EventProp/EVNN/EGRU:
Event-based RNN&DNN for distributed/ large-scale sparse AI applications

Sparse Preprocessing and Event-Based Sparse Data Path:

- Reduce signal data rate/features to minimum required by task
- Individual sparse numerical accelerator functions
- Tie together flexibly and efficiently via sparse data path

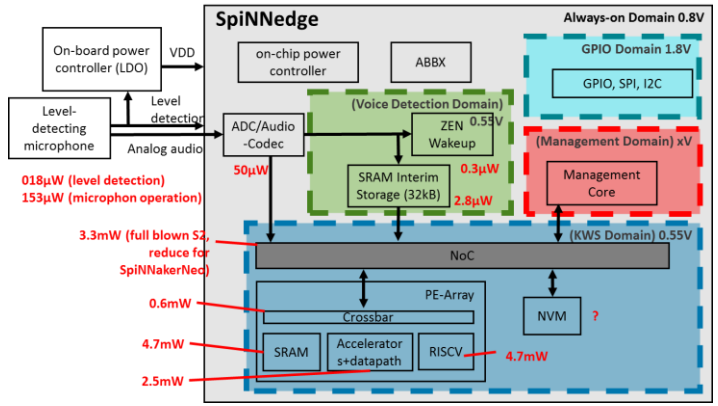
Memory Power Management/NVM:

- Small-scale NVM arrays as SRAM replacement
- With our sparse access pattern: Memory power management possible, i.e. power-cycling of memory

Multi-Stage Wakeup/Attention, e.g. Audio:

- 3uW always-on intelligent wakeup (speech/no speech)
- 5mW keyword spotting
- 1W Full language processing

SpiNNEdge Concept



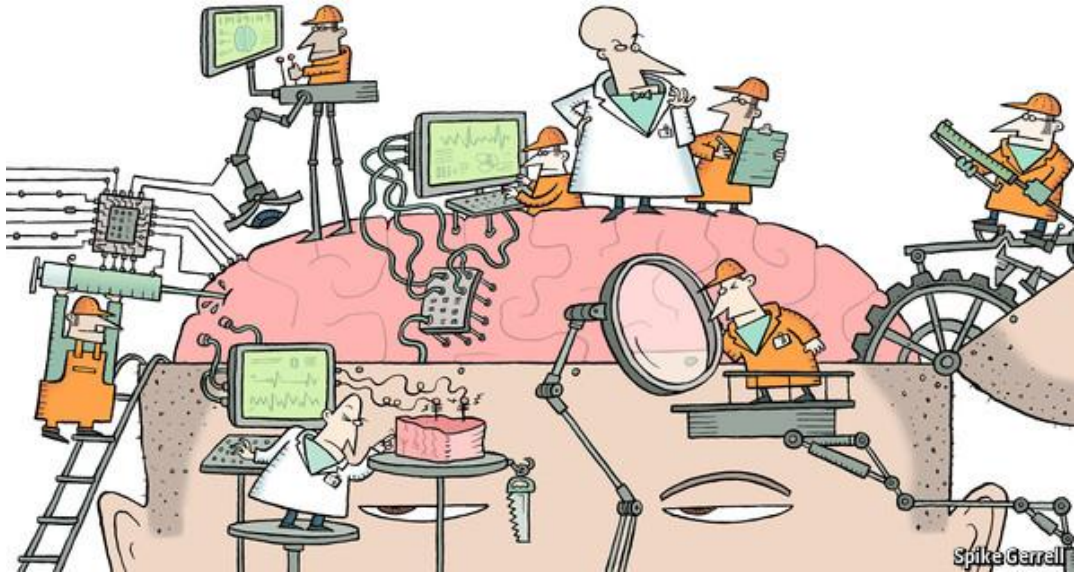
Adaptive Mixed-Signal Data Acquisition:

- Backbiasing for adaptive analog speed/leakage
- Pervasive digital assist in the loop for world-record area and power efficiency
- Signal-adaptive, attention-driven resolution and sampling rate

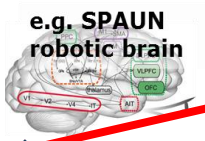
Distributed AI from Edge to Cloud:

- Use SpiNNaker2 as a powerful, scalable robot brain
- Combine with edge chips for distributed processing
- 1Mio BMBF/SprinD award winning SpiNNedge AI chips now available for radar, tactile, biomedical signals, soon also: audio&video

- SpiNNaker 2
- Bio-Inspired AI: Algorithms & Use Cases from Edge to Cloud
- Domain-Specific SpiNNedge Chips
- Outlooks
  - Merging DNN, SNN and Symbolic AI
  - Use Case Quantum



DNN/SNN/Symbolic AI Merger: Humanoid Robot



ng real-world,

transmission
ency

ed decisions

ntations
the

es the
eraction.

e gap
ting
e AI
wave of AI.

HPC vs SpiNNaker2 vs Quantum

	High Performance Computing	SpiNNaker2	Quantum Computing
Framework:			
Parallelism:	Cores: 100k	Cores: 10M MAC: 640M Pseudo-Parallelism: Synaptic Updates in 1 ms (computing and communication): 10^{14}	30 ideal Qubits (number of Quantum entaglements): $>10^{30}$

Further SpiNNaker2 Features:

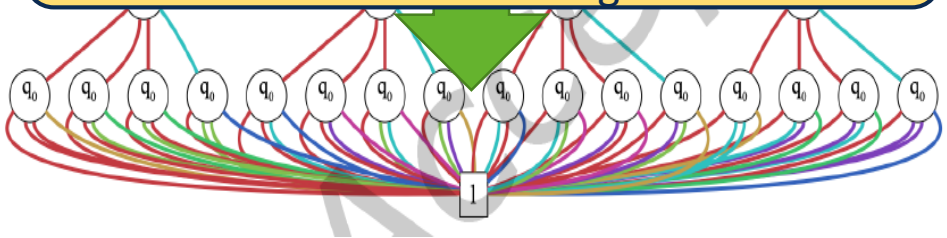
- High-frequency flexible random generators (10^{15} random numbers per second)
- Entire memory contiguously adressable via same NoC: 10^{14} 16 Bit parameters
- Small, fast packet sizes for coherence updates

IQM & SpiNNcloud

Idea: Neuromorphic QC-Emulator via Decision Diagram

Enabled By: Highly-parallel MC-transversal of decision tree: simple, ridiculously parallel computation

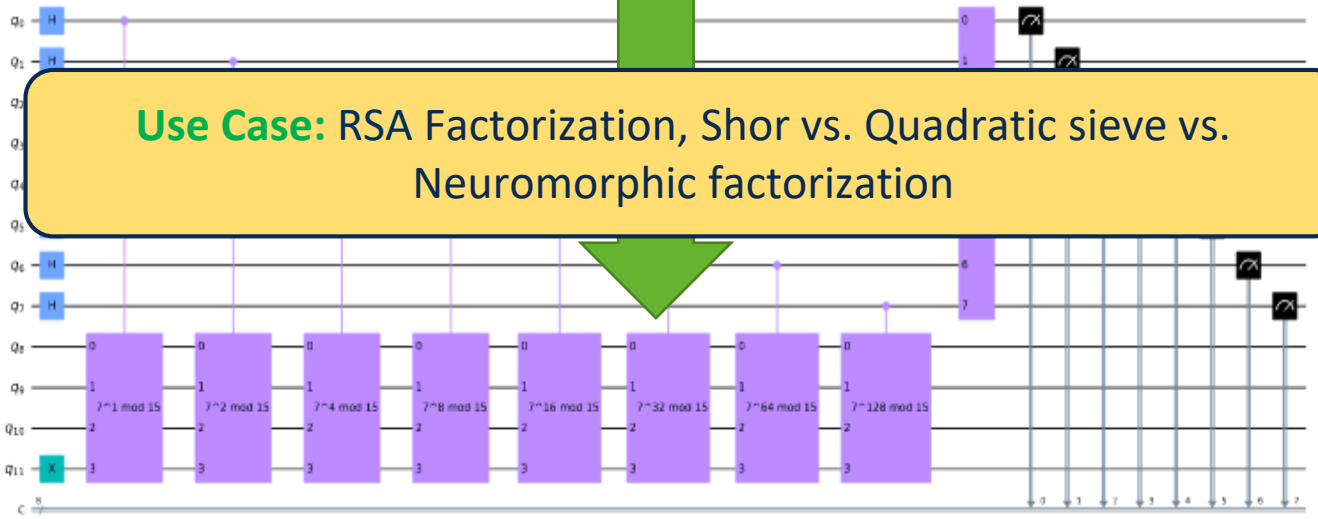
Use case: digital twin of real IQM QC system, for algorithm development and QC hardware design



Idea: „Smart“ QC concepts

Enabled by: Hybridization of algorithms: HPC smart searches, not very parallel. QC: brute force, ridiculously parallel

Use Case: RSA Factorization, Shor vs. Quadratic sieve vs. Neuromorphic factorization



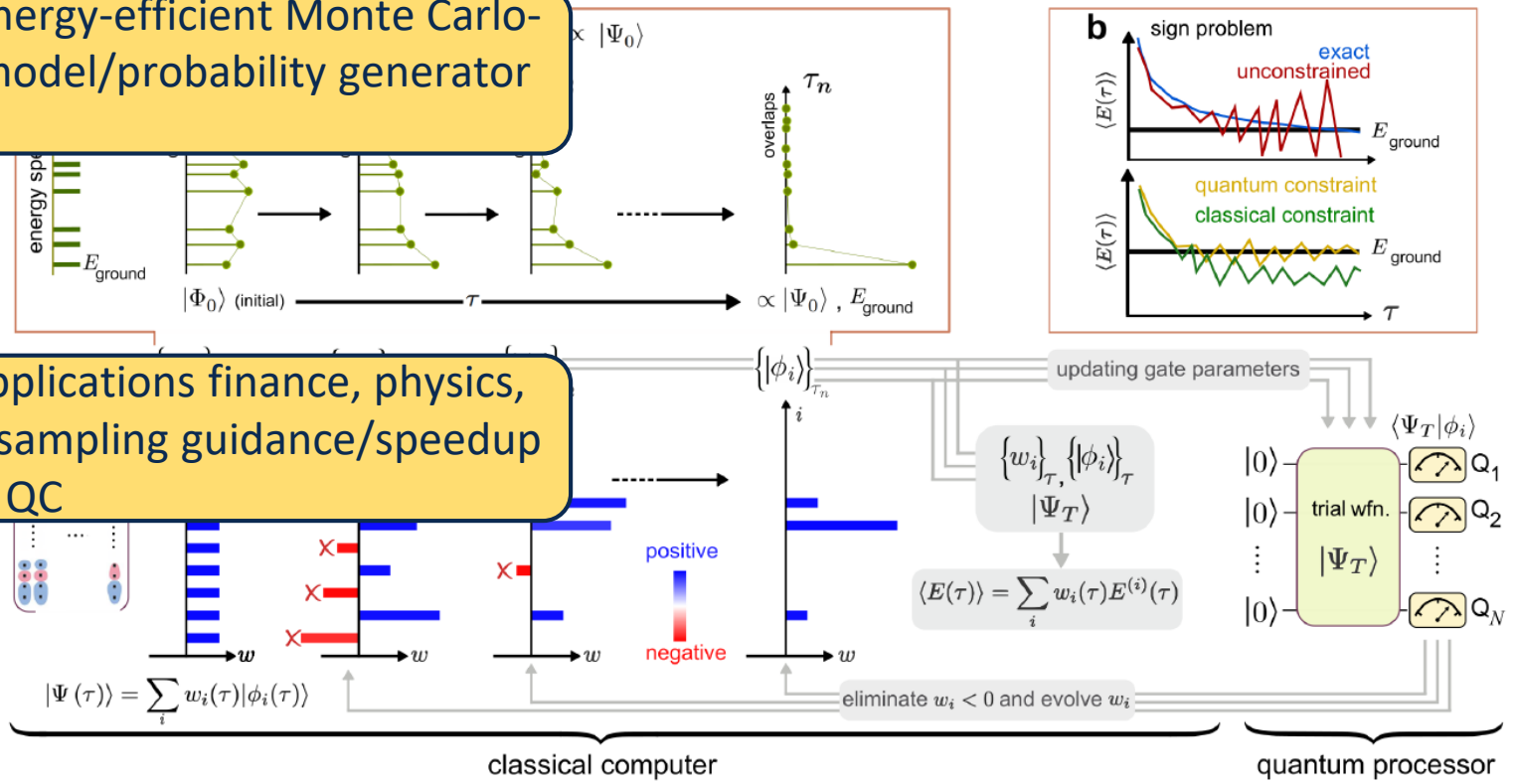
TECHNISCHE UNIVERSITÄT DRESDEN

IQM & SpiNNcloud

Idea: Hybrid QC-NM-HPC concepts, achieve synergies that no single system could do, reach early Quantum Advantage

Enabled By: Neuromorphic Highly-parallel, energy-efficient Monte Carlo-sampling. On the fly adjustable MC system model/probability generator in loop with QC

Use case: Hybrid Monte Carlo algorithms (applications finance, physics, etc): Explicit sampling in neuromorphic HPC, sampling guidance/speedup via implicit model in QC



TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

Meiler Lab
(Vanderbilt/Leipzig)

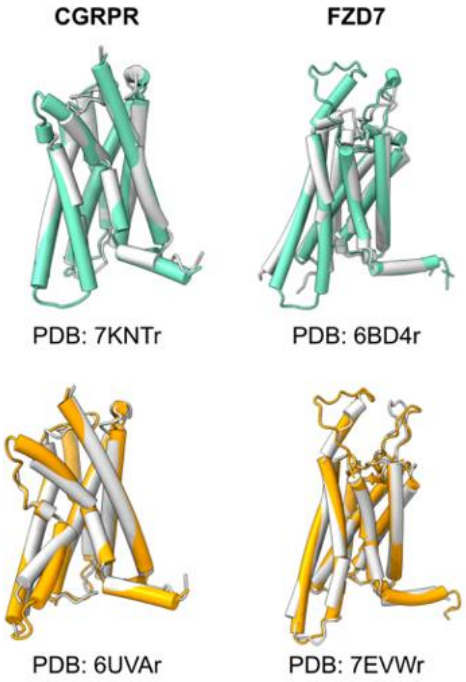
&

MERCK

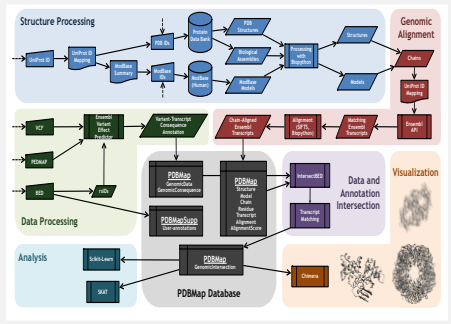
&

SpiNNcloud

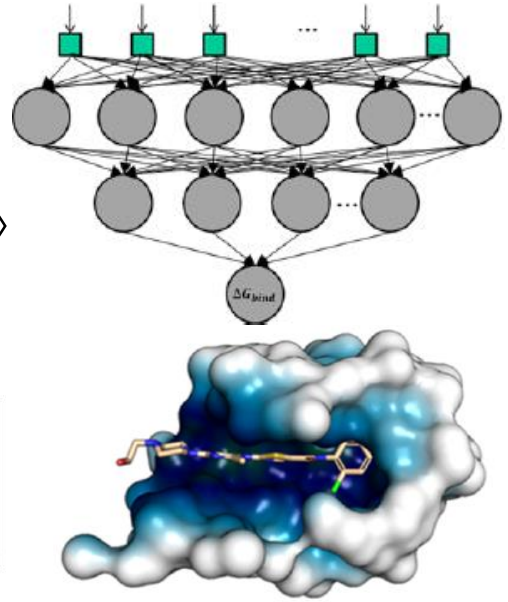
KI-predicted structures of protein receptors in different conformations



Personalisation via illness mutation in Genom of patient



KI Ultra-High throughput Screening of 21,000,000,000 drug candidates



D. Del Alamo, D. Sala, H. S. McHaourab and **J. Meiler**; "Sampling alternative conformational states of transporters and receptors with AlphaFold2"; *Elife*; **2022**; Vol. 11 p. || B. P. Brown, J. Mendenhall, A. R. Geanes and **J. Meiler**; "General Purpose Structure-Based Drug Discovery Neural Network Score Functions with Human-Interpretable Pharmacophore Maps"; *J Chem Inf Model*; **2021**; Vol. 61 (2): p. 603-620. || C. Schuss, O. Vu, M. Schubert, Y. Du, N. M. Mishra, I. R. Tough, J. Stichel, C. D. Weaver, K. A. Emmitte, H. M. Cox, **J. Meiler** and A. G. Beck-Sickinger; "Highly Selective Y4 Receptor Antagonist Binds in an Allosteric Binding Pocket"; J Med Chem; 2021; Vol. 64 (5): p. 2801-2814. || J. N. Gallant, J. H. Sheehan, T. M. Shaver, M. Bailey, D. Lipson, R. Chandramohan, M. Red Brewer, S. J. York, M. G. Kris, J. A. Pietenpol, M. Ladanyi, V. A. Miller, S. M. Ali, **J. Meiler** and C. M. Lovly; "EGFR Kinase Domain Duplication (EGFR-KDD) Is a Novel Oncogenic Driver in Lung Cancer That Is Clinically Responsive to Afatinib"; *Cancer Discov*; **2015**; Vol. 5 (11): p. 1155-63

TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

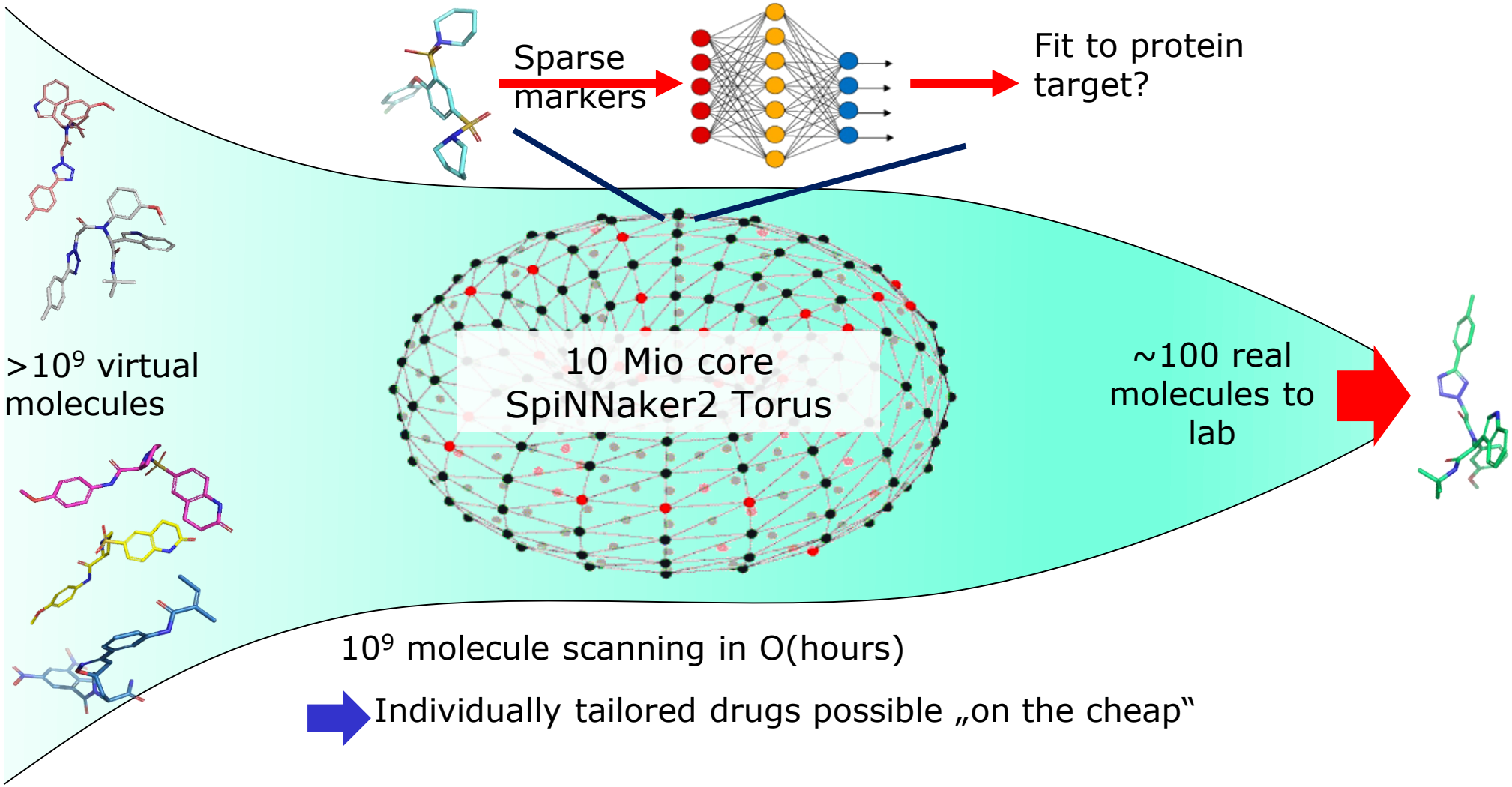
Meiler Lab
(Vanderbilt/Leipzig)

&

MERCK

&

SpiNNcloud



TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

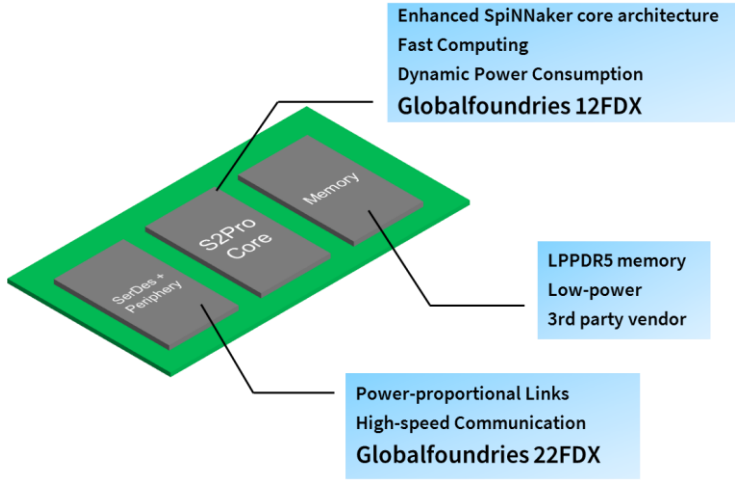
6 Year Perspective: SpiNNaker3

2015: SpiNNaker1

130 nm, mostly standard-ARM-IP-based, event-based comm, brain simulation and neuromorphic use cases

2022: SpiNNaker2

22nm FDX, 10x core scaling, >50x capacity scaling, ARM cores plus dedicated accelerators for both neuromorphic and DNN processing



2028: SpiNNaker3

12nm FDX or smaller, 100x capacity scaling, event-based communication and processing paradigm throughout whole architecture
-> „Brain-inspired general purpose computing“

- Take numerical simplicity and large-scale applicability from DNNs, merge with brain-inspired efficiency
- Sparse, condensed real time AI processing chain on all levels (inference, learning, architecture, down to accelerators)

Overall SpiNNaker3:

- Ultra-High Density AI system, geared towards real-time AI processing, processing model supports a very high, partially asynchronous parallelism, supports load balancing among cores and a large-scale grey silicon approach
- Full merger between DNN and brain inspiration, >50x less traffic/processing load, i.e. factor 50 better latency and efficiency
- Downscaling current SpiNNaker2 by >100x form factor at same capability
- **Breaking the barrier that now prohibits >100 Trillion parameter AI models (esp for learning)**
- **This time: Tight coupling with feedback from user community. Yes. That is you!**

Processing Element Level:

- Everything proposed in other slides, i.e. Memory power management, fine-grained NVM
- All the refinements in processing (e.g. Accelerators for sensor preprocessing, for learning, ABB-assisted SRAM, etc)
- Ultra-configurable sparse data flow architecture for accelerator interconnect

Chip Stack Level:

- 2.5D/3D stacking (e.g. Not TVM, more like BEOL-based stacking),
- algorithm-enabled voltage stacking (as bitcoin mining does it) of chips in 3D-stack, with fine-grained, highly dynamic load balancing for voltage drop regulation across 3D stack
- DC-voltage-free intra-stack communication (capacitively coupled or optical)
- **Chiplet integration**

System Level:

- Direct optical off-chip and intra-board links (i.e. **Integrate optical transmitters in chip** or package)
- Various derived systems possible, from Cloud scale via desktop size ‚Brain in a box' down to ultra-efficient edge systems

Connect with SpiNNaker2 and our other systems:

- Workshops (Capo Caccia, Telluride)
- Commercial: SpiNNcloud Systems
- 10Mio core academic Machine in Dresden (from 03/2023)
- Edge systems available on limited basis
- SpiNNaker2-awards

Plus our spinoffs and Steve's APT group at University Manchester:

SILICONALLY

Racyics

MANCHESTER 1824
The University of Manchester

COINBAU®

SpiNNcloud

Near future: Multiple large-scale SpiNNaker2 machines, in hybrid HPC/GPU/QC setups

- International: Sandia, Oak Ridge, Johns Hopkins, Yale, Manchester, China, Middle East
- Germany: CMI, DZA, LAIQUA, LEAM, Scads.AI, Infineon AI center

From 07/2023: SpiNNaker3 design on 7Mio grant....

Kelber, Florian, et al. "Mapping deep neural networks on spinnaker2." *Proceedings of the neuro-inspired computational elements workshop*. 2020.

Yan, Yexin, et al. "Comparing Loihi with a SpiNNaker 2 prototype on low-latency keyword spotting and adaptive robotic control." *Neuromorphic Computing and Engineering* 1.1 (2021): 014002.

Yan, Yexin, et al. "Efficient reward-based structural plasticity on a SpiNNaker 2 prototype." *IEEE transactions on biomedical circuits and systems* 13.3 (2019): 579-591.

Höppner, Sebastian, et al. "Dynamic power management for neuromorphic many-core systems." *IEEE Transactions on Circuits and Systems I: Regular Papers* 66.8 (2019): 2973-2986.

López-Randulfe, Javier, et al. "Time-Coded Spiking Fourier Transform in Neuromorphic Hardware." *IEEE Transactions on Computers* (2022).

Vogginger, Bernhard, et al. "Automotive Radar Processing With Spiking Neural Networks: Concepts and Challenges." *Frontiers in neuroscience* 16 (2022).

Höppner, Sebastian, et al. "The SpiNNaker 2 processing element architecture for hybrid digital neuromorphic computing." *arXiv preprint arXiv:2103.08392* (2021).

Liu, Chen, et al. "Memory-efficient deep learning on a SpiNNaker 2 prototype." *Frontiers in neuroscience* 12 (2018): 840.

SpiNNaker2 is inherently disruptive

SpiNNaker2 is a paradigm shift for AI. It brings data processing close to sensors and redefines the traditional central server paradigm, which makes machine learning faster, better and way more efficient. As machine learning is just beginning to enter our daily lives in medical technology, autonomous driving and robotics, SPRIND is seeing huge potential that SpiNNaker2 will disrupt the way we use AI.

Rafael Laguna de la Vera - *Director of the German Federal Agency for Disruptive Innovation (SPRIN-D)*