TECHNISCHE UNIVERSITÄT DRESDEN

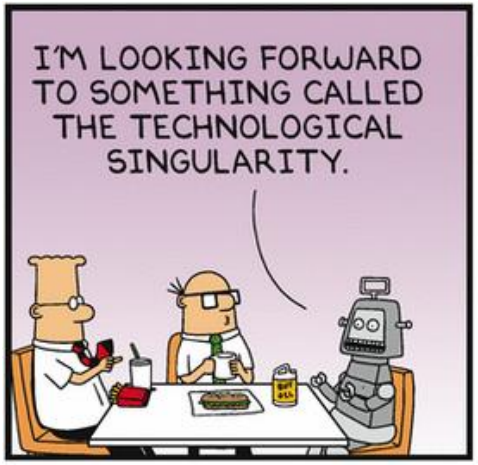
Elektrotechnik und Informationstechnik, Professur hochparallele VLSI Systeme und Neuromikroelektronik

SpiNNaker 2: A Platform for Real-Time Bio-Inspired AI and Cognition

Christian Mayr

Neuromorphic Processing: The Abstract View

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 and efficient
- 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

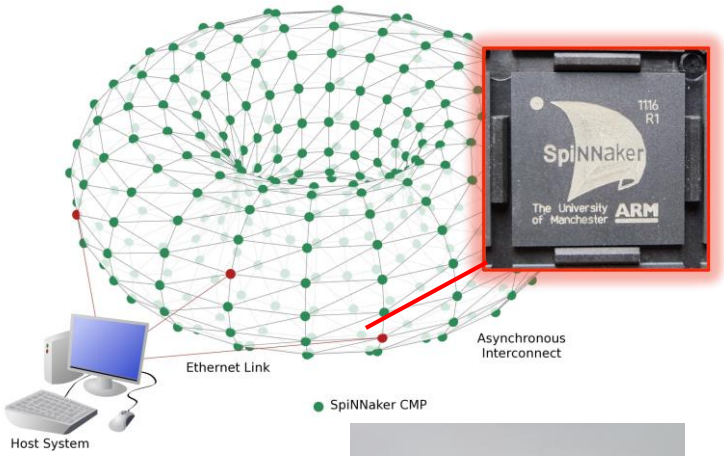
Human Brain Project: SpiNNaker 1

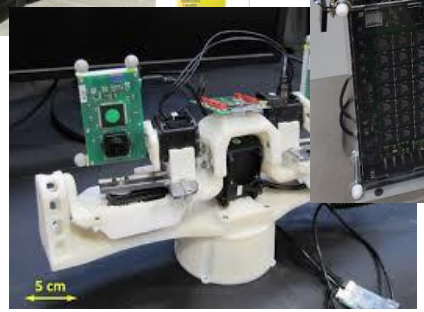
SpiNNaker1:

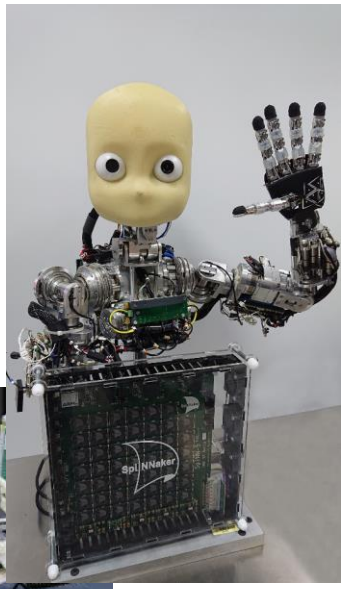
- ~40 systems in use around the world
- Popular robotics compute system
- Inbuilt millisecond time scale: Inherently stays real-time compared to regular AI or supercomputing machines

SpiNNaker2:

- Enhance capacity for brain size network simulation in real time at >10x better efficiency
- Keep programmability: flexible neurons, complex plasticity (e.g. three factor), graded events, etc
- Add various accelerators for deep and spiking neural networks and sparse information processing
- **Merge brain inspiration and deep neural networks at all levels**: processors, network, algorithms







Dynamic Power Management
- DVFS and PSO

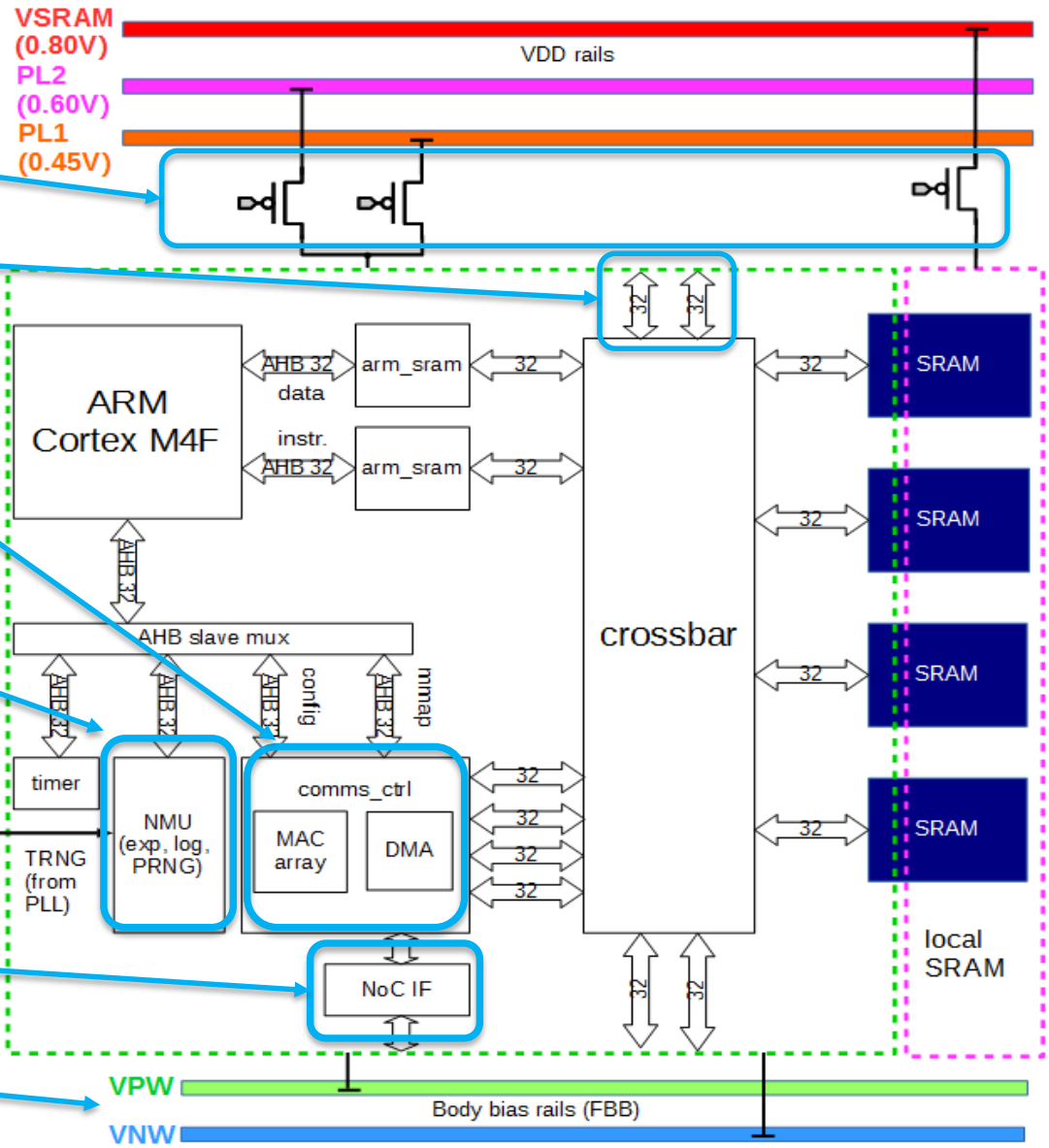
Memory sharing
- Synchronous access to neighbor PEs

Multiply-Accumulate accelerator
- MAC array with DMA

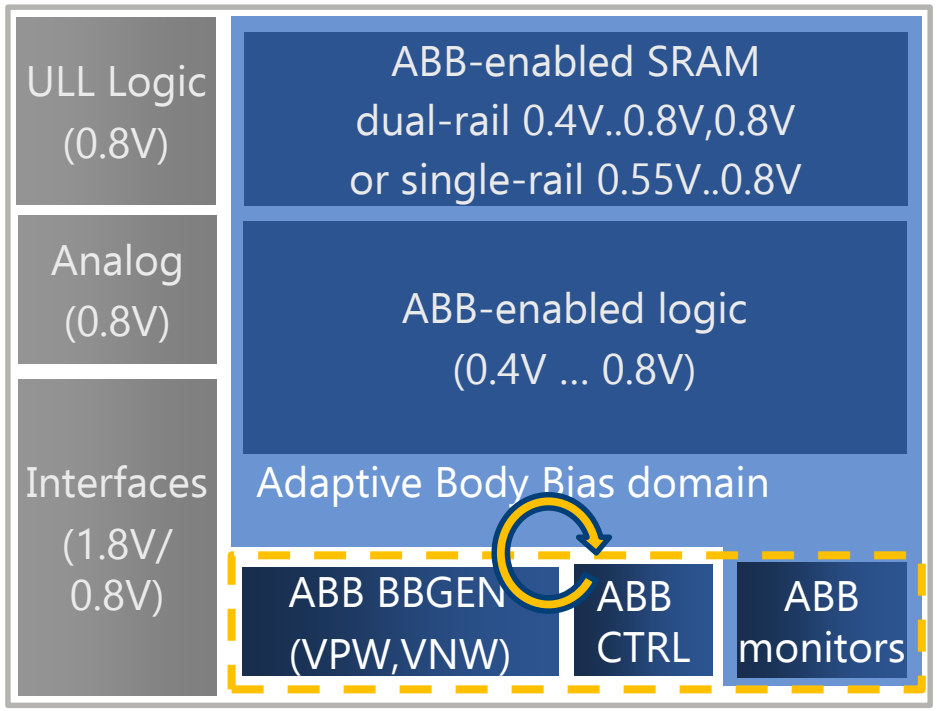
Neuromorphic accelerators
- Exp/log
- Random numbers (PRNG, TRNG from ADPLL noise)

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

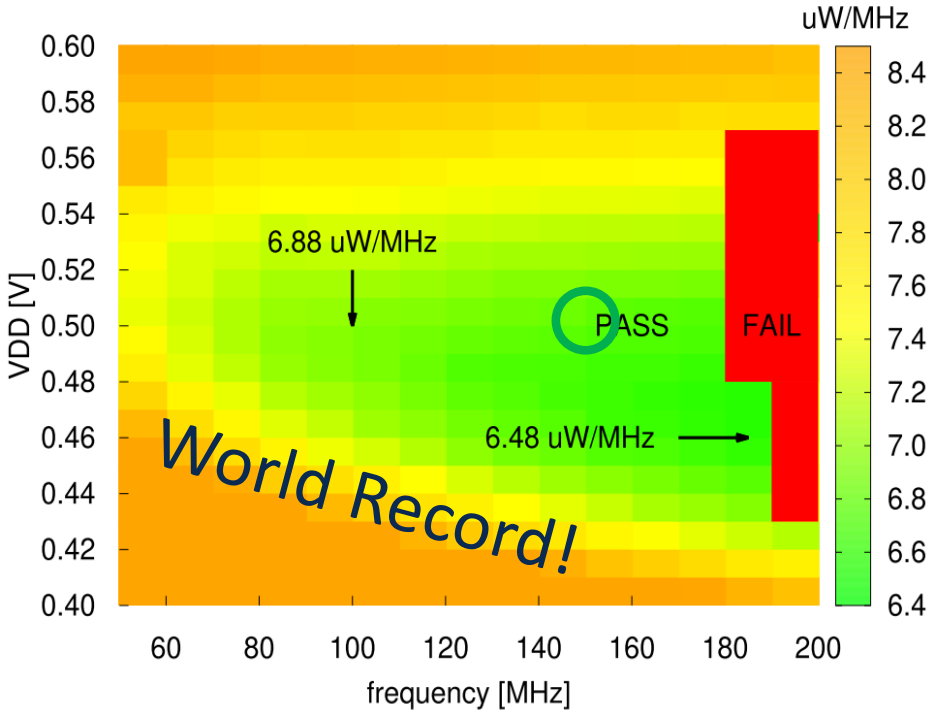
Adaptive Body Biasing



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



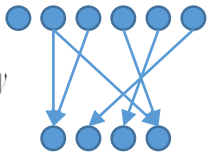
Arm Cortex-M4 Testchip
(MPW2213) with FBB [2]



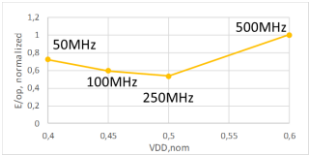
[2] S. Höppner *et al.*, "How to Achieve World-Leading Energy Efficiency using 22FDX with Adaptive Body Biasing on an Arm Cortex-M4 IoT SoC," *ESSDERC 2019 - 49th European Solid-State Device Research Conference (ESSDERC)*

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

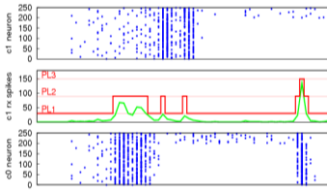
$$\tau_w \dot{w} = a(V_{Mem} - V_{rest}) - w$$



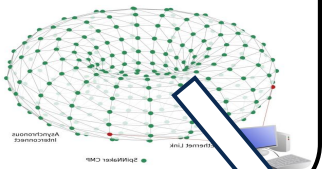
Outperforming Commercial Systems on **real-time AI**



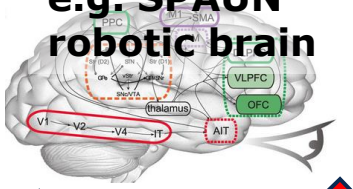
Brain-inspired **dynamic data sparsity**, i.e. ultra-efficient highly-parallel operation of AI algorithms on streaming data



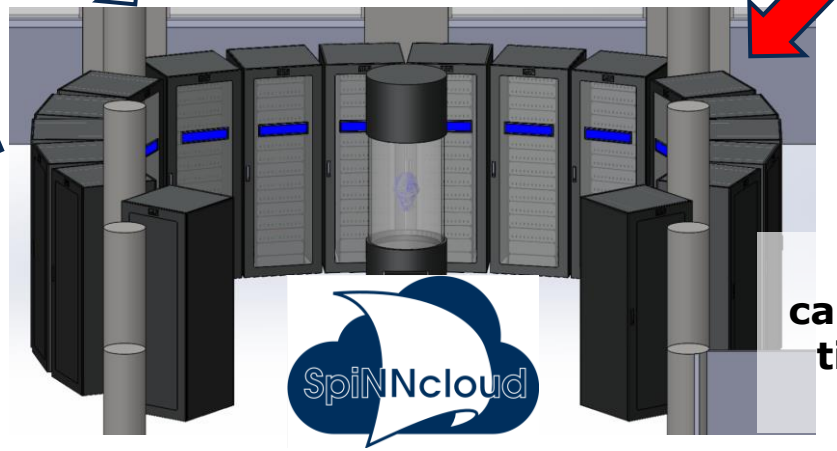
Largest real-time brain simulation platform worldwide,
3 PFLOPS CPU
0.4 ExaOPS in AI accelerator



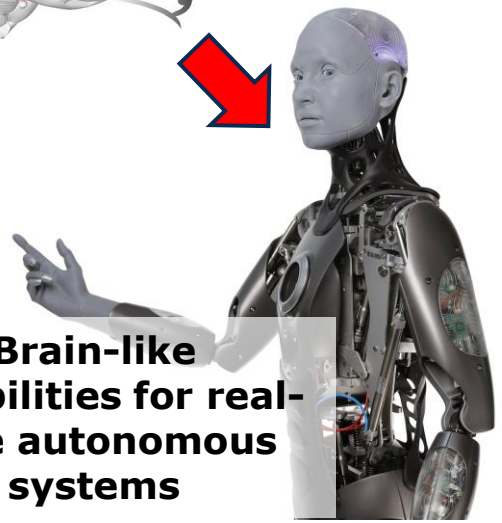
e.g. SPAUN robotic brain



Physical: 10^7 processors, 70.000 chips, 16 racks.

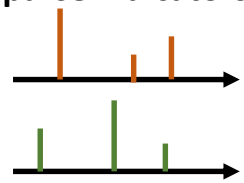


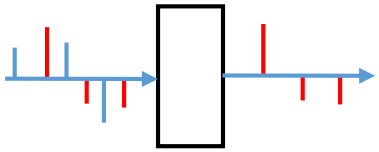
Brain-like capabilities for real-time autonomous systems



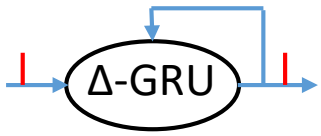
Processing

Robust, dynamically sparse indicators

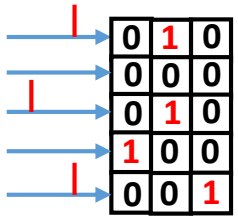




Memory-reduced streaming processing



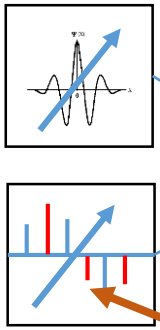
Simplified neurons for time encoding (e.g. LMU, ΔGRU)



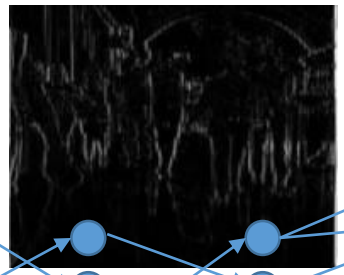
Accelerators (e.g. MAC-Array) used in sparse, event-based fashion

Architecture

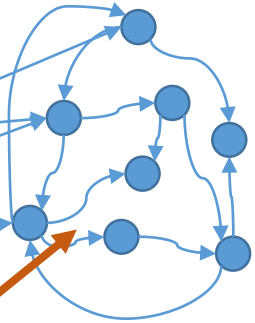
Trainable sparse extractors with event output



Feature combination and sparsification



Recurrent neural network



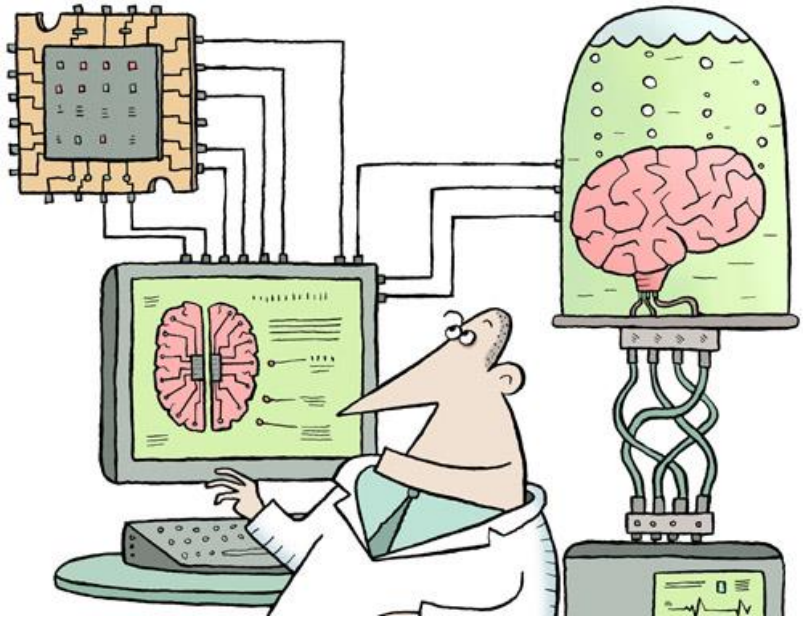
Target, e.g. object tracking, prediction, etc

Training with Sparsification

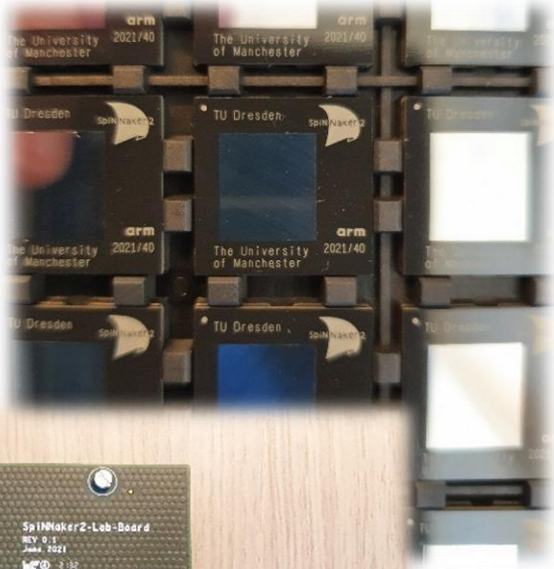
SpiNNedge ASICs:
Sparse Preprocessing and Neural Network Acceleration for Edge Applications (radar, video, audio, robotic, biomedical)

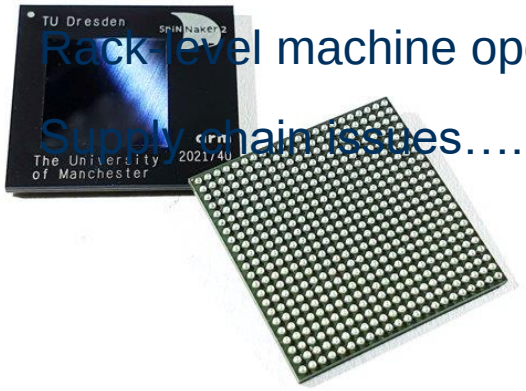
EVNN/EGRU:
Event-based RNN&DNN for distributed/ large-scale sparse AI applications

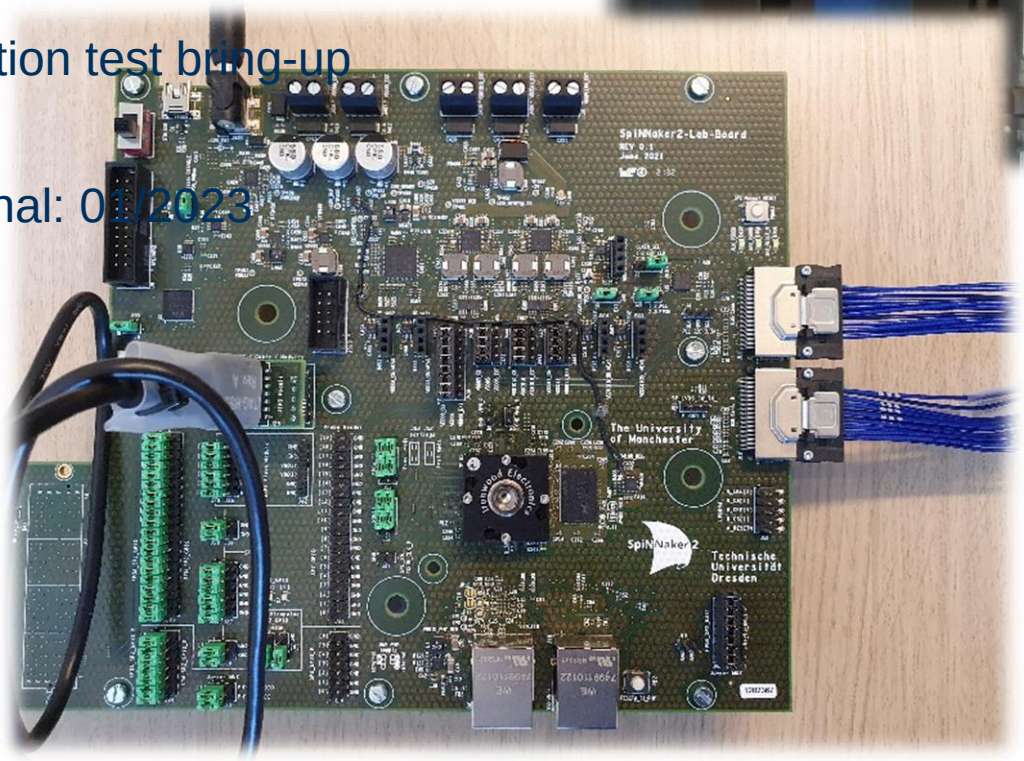
SpiNNaker2: Current State and Outlook



- Full-Mask Tapeout done in Globalfoundries 22FDX: 05/2021
- Construction for SpiNNaker2 housing started (water cooling, USV, floor strengthening, etc)
- 1st batch of ≈45k chips ordered for >5M core machine
- Lab tests successful, Production test bring-up ongoing
- Rack-level machine operational: 01/2023
- Supply chain issues….
- >100 persons involved: TUD,







Plenty of early results:

- Synaptic sampling as most complex plasticity rule ever on a neuromorphic chip
- DeepR and Eprop as competition to backprop
- Most efficient implementation of the Neural Engineering Framework
- Automotive Radar processing, robotics and tactile algorithms ...

	PL1 (0.50V, 150MHz)	PL2 (0.80V, 300MHz)	Comment
CoreMark Score (iterations/second)	40851	81711	152 PEs
CPU Gop/s	13.9	27.9	152 Pes, CoreMark Instruction Fetches
Energy efficiency [uW/MHz]	16.3	24.0	CoreMark benchmark, measured on toplevel with 152 PEs active, including toplevel and NoC overhead
MatMul performance (TOP/s)	2.24	4.50	152 PEs with HW acceleration
Tops/W	2.1	1.6	Matrix multiplication with HW accelerator, measured on toplevel with 152 PEs active, including Arm core, toplevel and NoC overhead
Chip-to-Chip	6x 12Gbit/s, bi-directional		

Update 3: SpiNNaker2 Software

- **SNN simulation using PyNN**
  - Will re-use large parts from SpiNNaker1 stack (pyNN.sPyNNaker)
  - Current work: Adaption of low-level software
  - Availability: 2023 for 48-node boards, earlier for single-chip system
  - **Lava integration** -> BMBF project with Intel
- **DNN processing using Apache TVM**
  - Use TVM compiler to map large DNNs on SpiNNaker2 systems
  - Utilize machine learning accelerator for Conv2D, Dense and ReLU; other layer types supported by code generation
  - Can load DNNs trained in any common framework (TensorFlow, Pytorch, ...)
  - Status: SW devolopment started, examples on single chip expected in next half year
- **Hybrid SNN/DNN**
  - Light-weight Python interface for SNNs or hybrid networks on single chip
  - Available: now, already in use by 3 external groups
  - Serves a prototype for scalable Hybrid NN framework (combination of PyNN and TVM)

```
from spinnaker2 import snn, hardware

neuron_params = {
        "threshold":1.,
        "alpha_decay":0.9,
        }

stim = snn.Population(
        size=10,
        neuron_model="spike_list",
        params={0:[1,2,3], 5:[20,30]},
        name="stim")

pop1 = snn.Population(
        size=20,
        neuron_model="lif",
        params=neuron_params,
        name="pop1")
```

TECHNISCHE UNIVERSITÄT DRESDEN

- **Outlook 1: SpiNNaker2 award**
  - 2*40.000€ award for one-year project showing off SpiNNaker2 capabilities
  - Low-level entry for early PhDs: Write a two-page outline, win award, have your own financing for a year
    - Offered on yearly basis
- Further outreach:
  - Ebrains integration
  - SpiNNaker2 workshops
  - Telluride this summer
  - etc

The SpiNNaker2 Award

The SpiNNaker2 Award offers a total of €80 000, split into a €40 000 prize each to the best two project proposals submitted that will demonstrate exciting and novel applications of the SpiNNaker2 neuromorphic hardware system [1, 2]. Proposals may be submitted by individuals or research groups worldwide, and will be judged on the ambition of the proposed application, the feasibility of the plan, and the track record of the proposer(s).

Following the design philosophy of the SpiNNaker2 systems, proposals should focus on real-time/closed loop interaction or should be motivated by a latency and energy-constrained setting. Usage of the hardware accelerator blocks [3,4] in SpiNNaker2 is recommended. The realization of proposed projects in a one-year time frame should be sketched. Award applications could include, but are not limited to:

- Multi-scale brain models
- Mobile/robotics applications
- Bio signal processing
- Novel combined artificial and spiking neural network processing paradigms

The submission deadline for the award is 00:00 UTC 30th November 2020. A project description in English of up to two A4 pages (10pt font, single spacing) is required and should include the following:

- Applicant(s) and affiliation(s) list
- Contact person information
- Brief description of own prior work and state of the art
- Description of your concept, its novelty and its potential impact
- Timeline and usage of funds
- Brief statement on dissemination

This award is sponsored by the "Dr. Stefan Weiße Stiftung" in cooperation with the Chair of Highly-Parallel VLSI Systems and Neuro-Microelectronics, TU Dresden, Germany (https://tu-dresden.de/ing/elektrotechnik/iee/hpsn/). The award selection is handled by a scientific jury headed by Christian Mayr (TU Dresden) and Steve Furber (University of Manchester). The selection process is by consensus vote. If no consensus is reached no award is made. The winner will be announced in January 2021. After a period of 12 months the winner is obliged to submit a final report and present the outcome at a SpiNNaker2 Workshop. During the project execution, SpiNNaker2 hardware and software support is provided by TU Dresden.

Send the document as a PDF file (max. size 2 MB) to spinnaker2-award@tu-dresden.de. Inquiries about the award should also be directed to the above email address.

SpiNNcloud

The platform for future Cognitive City applications

- Traffic Optimization
- Public Transport Management
- Autonomous Driving

- Smart infrastructure
- Environmental control

- Smart Security
- Emergency Services

- Deployment SpiNNcloud @ TUD with 8Mio€ EFRE/SMWK grant
- Integrated into Federal German AI supercomputing center (Scads.AI)
- Heavily used by Infineon AI center (Dresden)
- Separate machines to be integrated into Lausitz, Leipzig and Cottbus Supercomputing/Research Centers
- International user network with >40 Partnern, e.g. from the US: Johns Hopkins, UCI, Oculi Ltd, ABR Ltd)
- Usage: Massively parallel real time AI at high data rates: Autonomous Driving, Industry 4.0,...
- SpiNNaker2 and Loihi2 (Intel) are the two world-leading neuromorphic systems, representing DARPAs third wave of AI
- **Further Use Cases: Ultra-Fast drug screening, medical AI processing, Quantum computing emulation/Monte Carlo problems**
- **Roadmapplanning: SpiNNaker3/SpiNNaker2pro**

HP
FMI Friedrich Miescher Institute for Biomedical Research
Myelin S
Human Brain Project
BMW
KTH VETENSKAP OCH KONST
EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
POLITECNICO DI TORINO
UCL
UCI University of California, Irvine
ARM
BOSCH
JÜLICH Forschungszentrum
ScaDS.AI DRESDEN LEIPZIG
SPRIN-D BUNDESAGENTUR FÜR SPRUNGINNOVATIONEN
Europäische Union
Europa fördert Sachsen. EFRE Europäischer Fonds für regionale Entwicklung
GLOBALFOUNDRIES

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)*