SpiNNaker2 - An Energy efficient realtime neuromorphic compute system in 22FDX technology

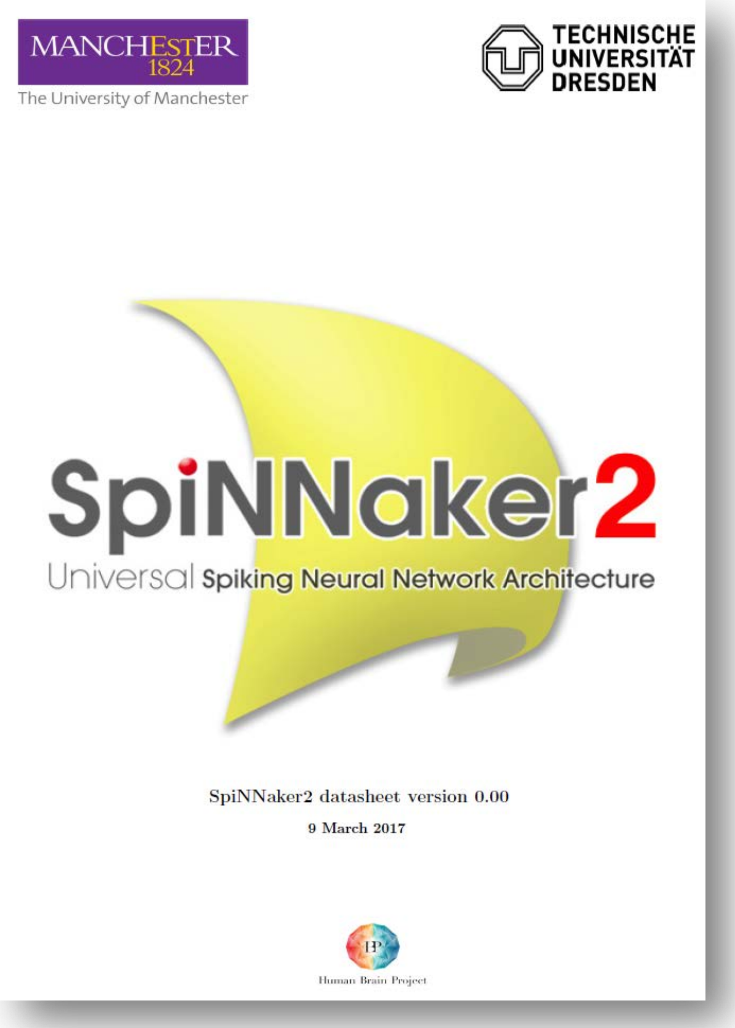
Sebastian Höppner, Johannes Partzsch, Christian Mayr, Steve Furber

Technische Universität Dresden, Germany
University of Manchester, UK

Arm Research Summit 2018

Outline

- SpiNNaker Overview
- SpiNNaker2 Hardware
- SpiNNaker2 Application Examples
- Conclusion



MANCHESTER 1824
The University of Manchester

TECHNISCHE UNIVERSITÄT DRESDEN

SpiNNaker2
Universal Spiking Neural Network Architecture

SpiNNaker2 datasheet version 0.00

9 March 2017

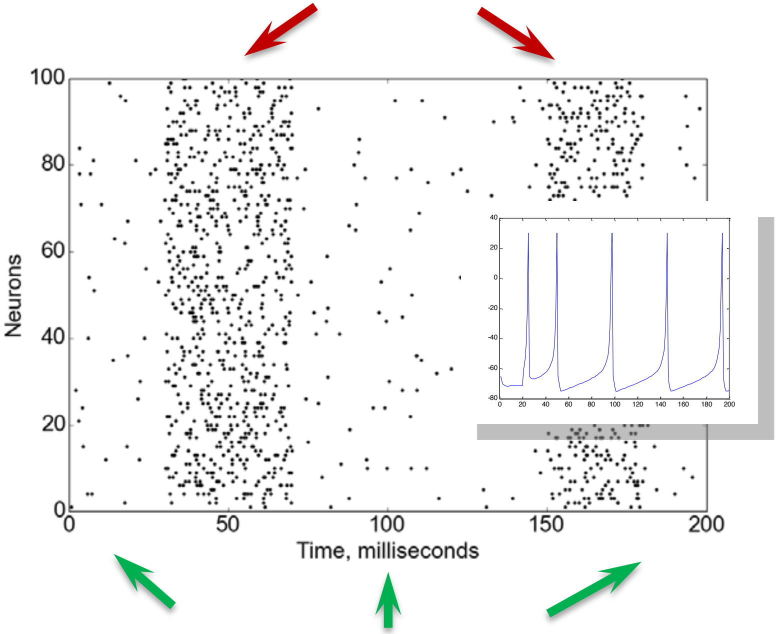
Human Brain Project

Neural Computation

To compute we need:

- Processing
  - Synaptic Updates
  - Neuron Computation
- Communication
  - Asynchronous Spike Communication
  - ‘Address Event’ Representation (AER)
- Storage
  - Synaptic weights
  - Axon ‘delay lines’
  - Neuron states

- High update rates
  - Peak processing load and spike communication bandwidth

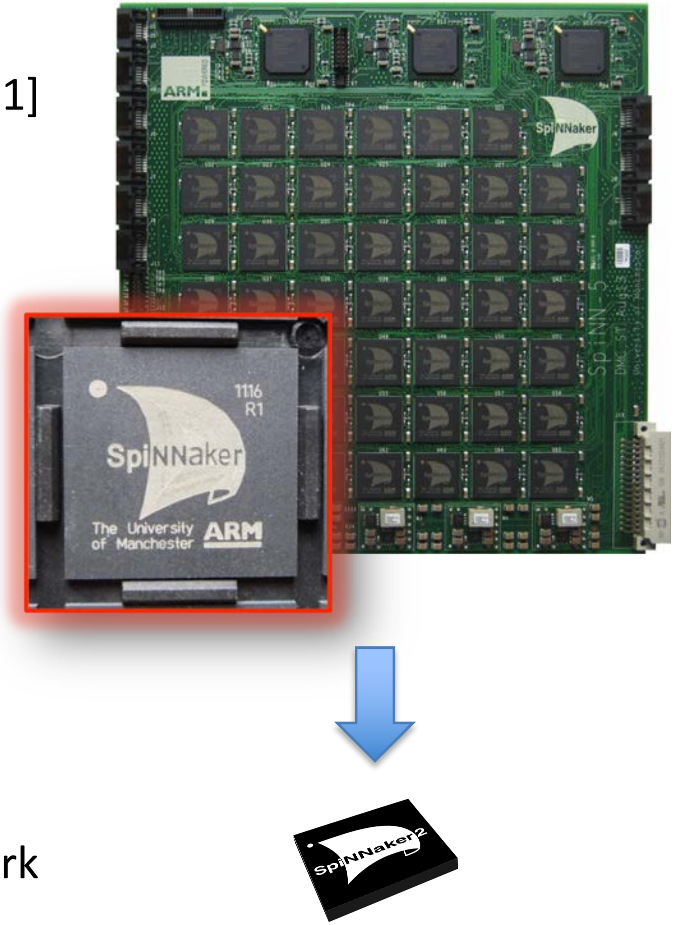


- Low update rates
  - Relaxed processing load and spike communication bandwidth

Latency requirements <u>≤1ms</u> for processing and communication

SpiNNaker

- Communication and memory centric architecture for efficient real-time simulation of spiking neural networks [1]
- Many-core (Arm based) architecture, 18 cores per chip
- **SpiNNaker** has a broad user base
  - ~40 systems in use around the world
  - Flexibility: adaptable network, neuron model & plasticity
  - Real-time: suits robotics & faster than HPC
  - System capacity of 10^9 neurons and 10^{12} synapses
  - Energy per synaptic event 10^{-8}J (HPC: 10^{-4}J)
- SpiNNaker uses 130nm CMOS technology
- Scope for improvement
  - on modern process (22FDX) [2]
  - Innovative circuit techniques to enhance throughput and energy efficiency for computation and communication
- SpiNNaker2 target: Enhance capacity for brain size network simulation in real time at >10x better efficiency



HBP *SpiNNaker2* Roadmap

2023

2022

2021

2020

2019

2018

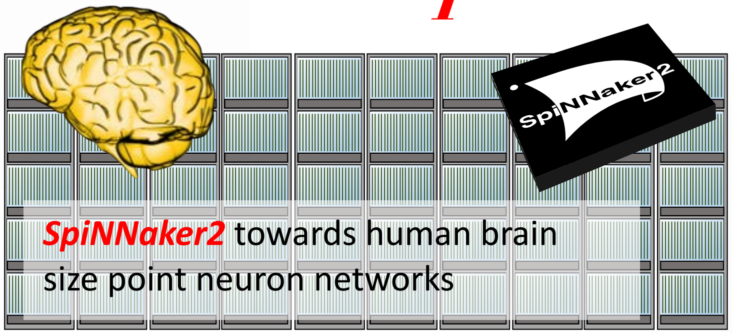
2017

2016

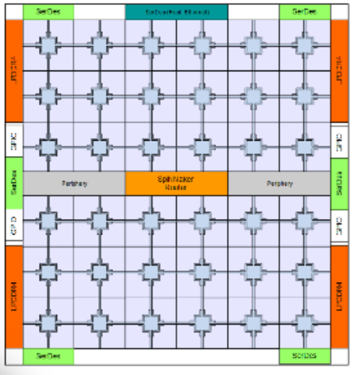
2015

2014

2013



SpiNNaker2 towards human brain size point neuron networks

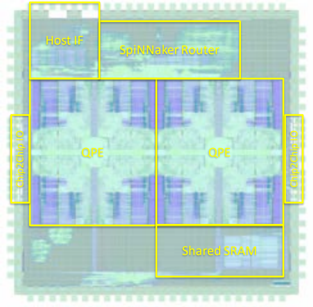


SpiNNaker2

- **144 Arm Cortex-M4F**
- power management
- SpiNNaker router
- low swing serial I/O
- 4x LPDDR4 memory IF
- 8GByte LPDDR4 PoP
- **22FDX** CMOS

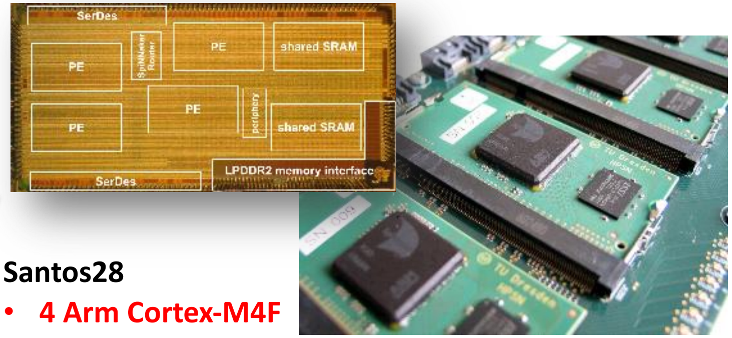
?

JIB2



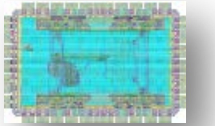
JIB1

- **8 Arm Cortex-M4F**
- SpiNNaker router,
- low swing serial I/O
- **22FDX** CMOS



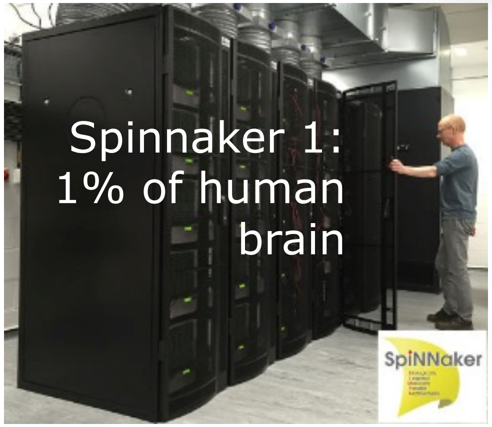
Santos28

- **4 Arm Cortex-M4F**
- Power management
- SpiNNaker router with SerDes
- LPDDR2 Memory Interface
- 28nm CMOS



NanoLink28

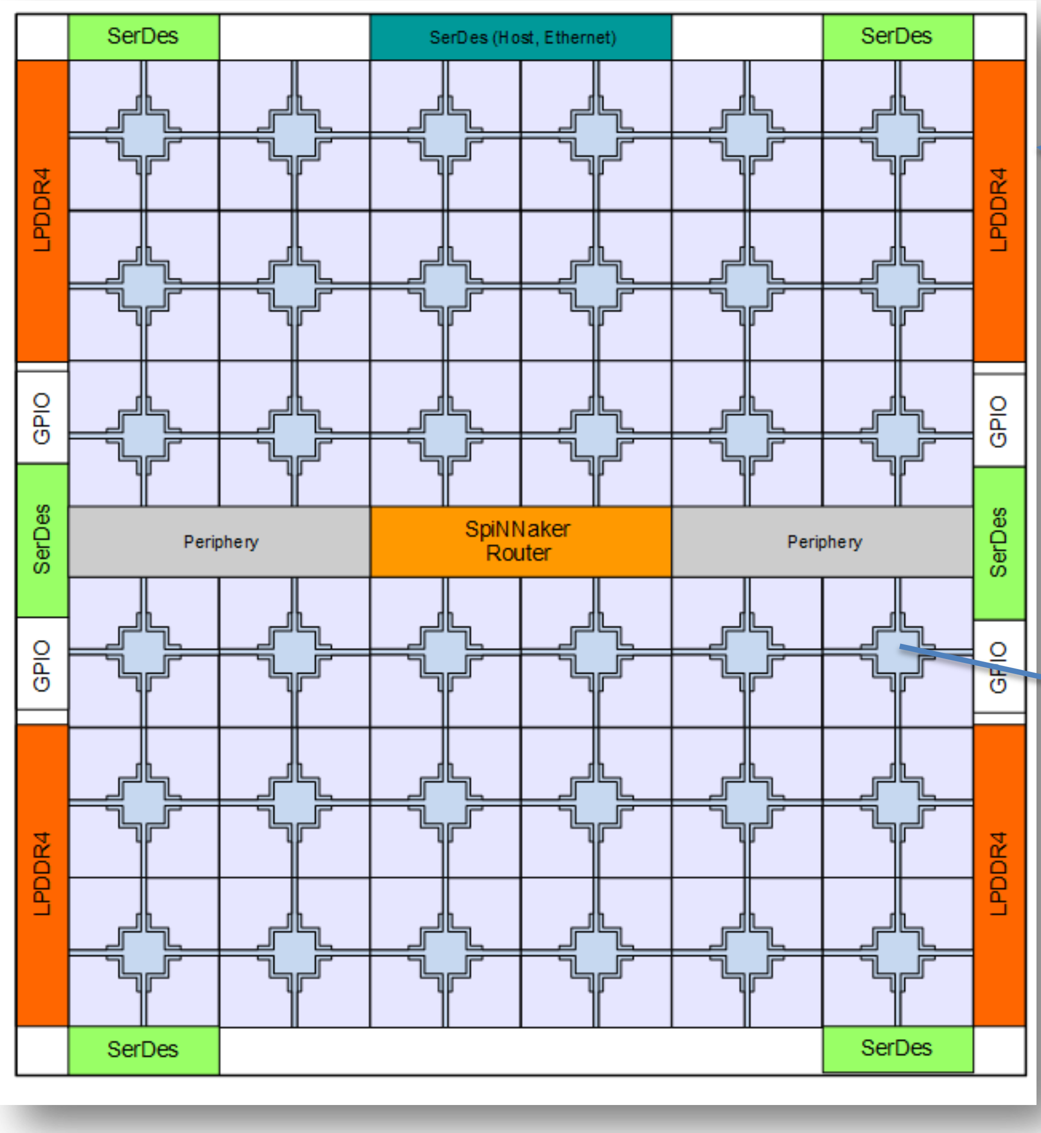
- SerDes Transceiver
- 28nm CMOS

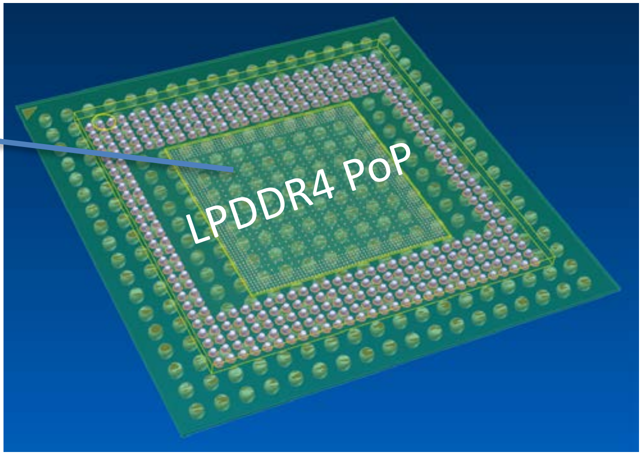


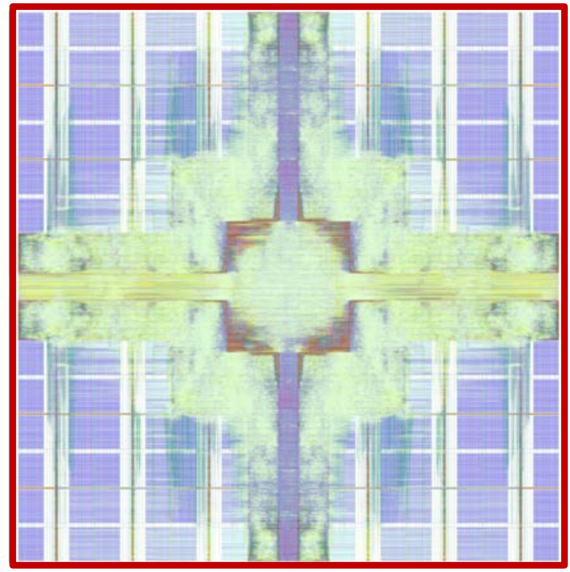
Spinnaker 1:
1% of human
brain

SpiNNaker2 Hardware

SpiNNaker2 Chip Overview







Processing Element

Dynamic Power Management

- DVFS and PSO [3]

Memory sharing

- Synchronous access to neighbor PEs

Multiply-Accumulate accelerator

- MAC array with DMA

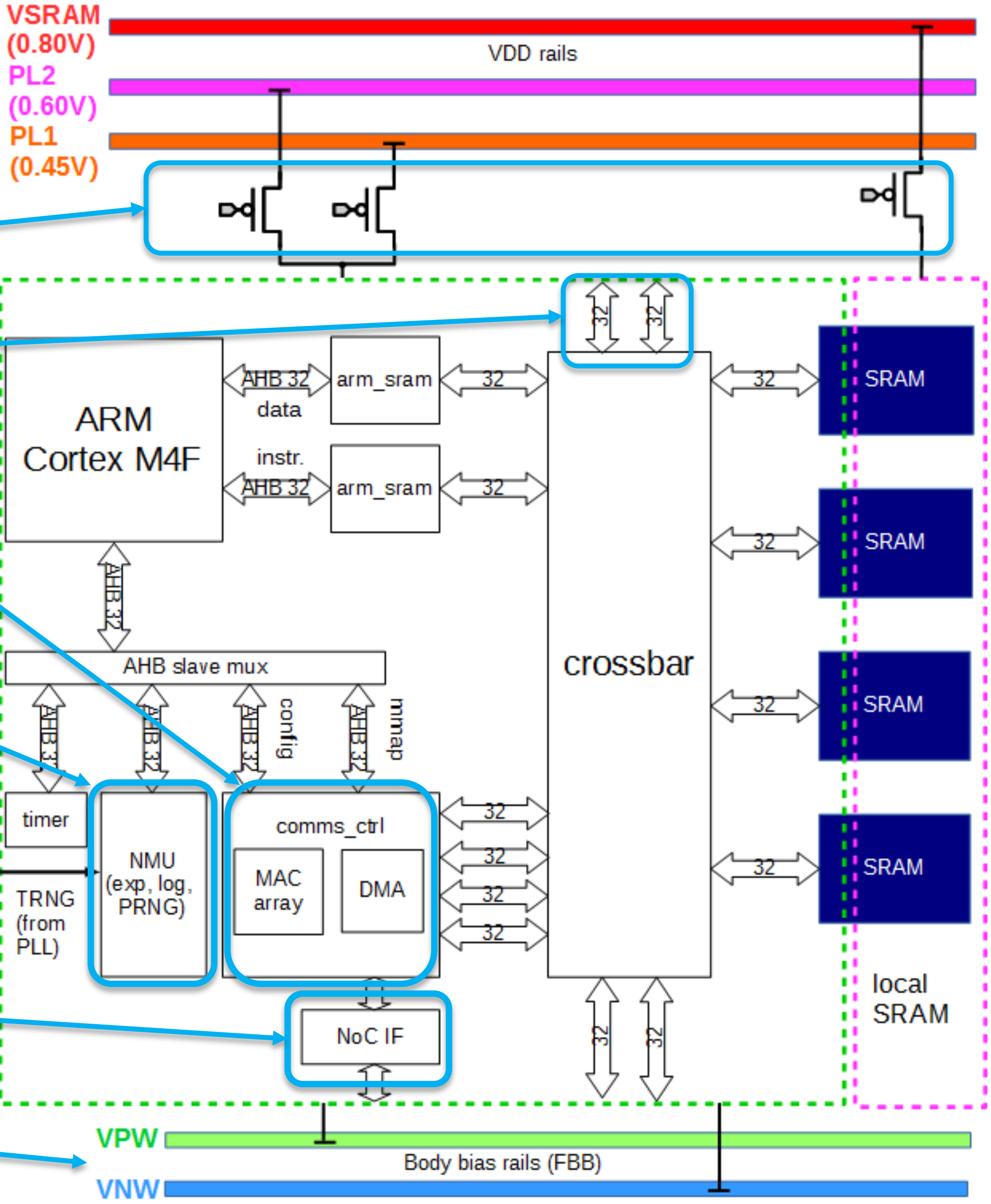
Neuromorphic accelerators

- Exp/log [4,7]
- Random numbers (PRNG, TRNG from ADPLL noise) [5]

Network-on-Chip

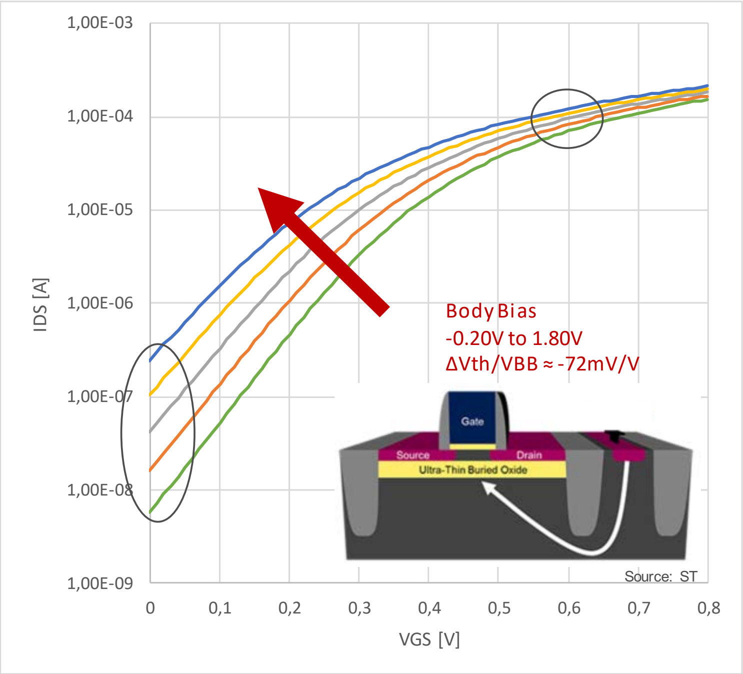
- On- and off-chip memory access
- SpiNNaker packet (spike) handling

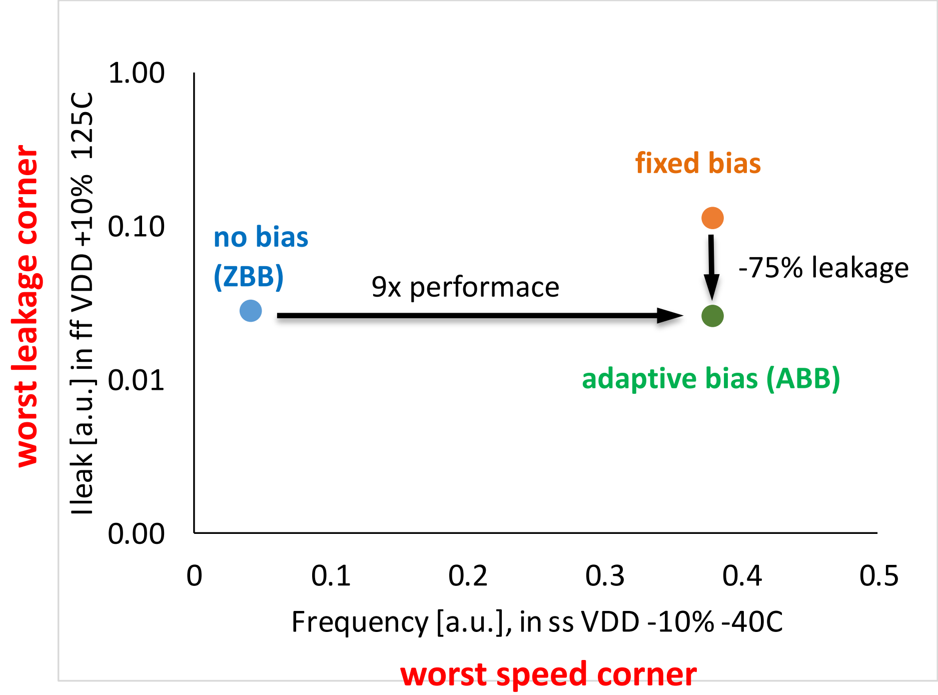
Adaptive Body Biasing



Ultra-Low-Voltage Design Enabled by ABB

- GLOBALFOUNDRIES 22FDX (FDSOI) technology [2]
- Adaptive body biasing (ABB) solution and foundation IP by Racyics [8]
- Enables ultra-low voltage operation down to 0.40V (0.36V worst-case) with guaranteed timing and power over PVT

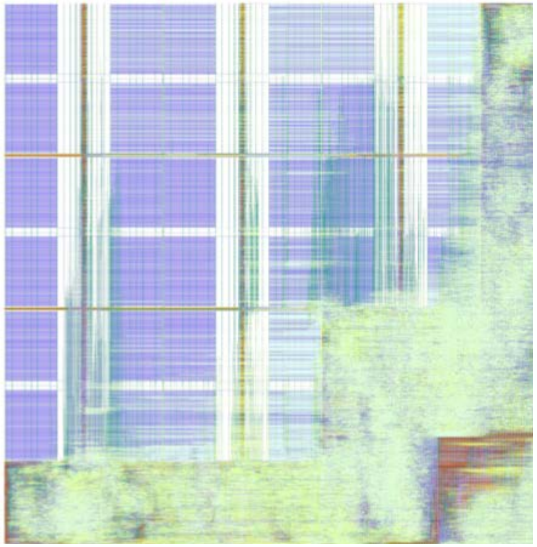


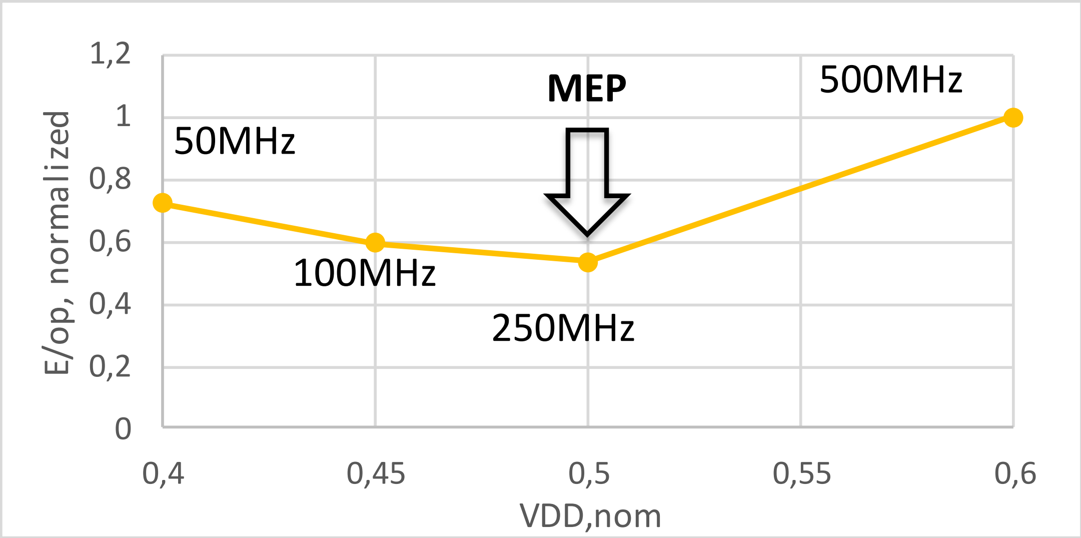


source: Racyics GmbH

PE Physical Implementation Strategy

- Synthesis and P&R studies and power-analysis for neuromorphic application scenario





Low-performance level (PL1)

- Operate at **Minimum Energy Point** (250MHz at 0.50V) or at ultra-low power mode (100MHz at 0.45V)

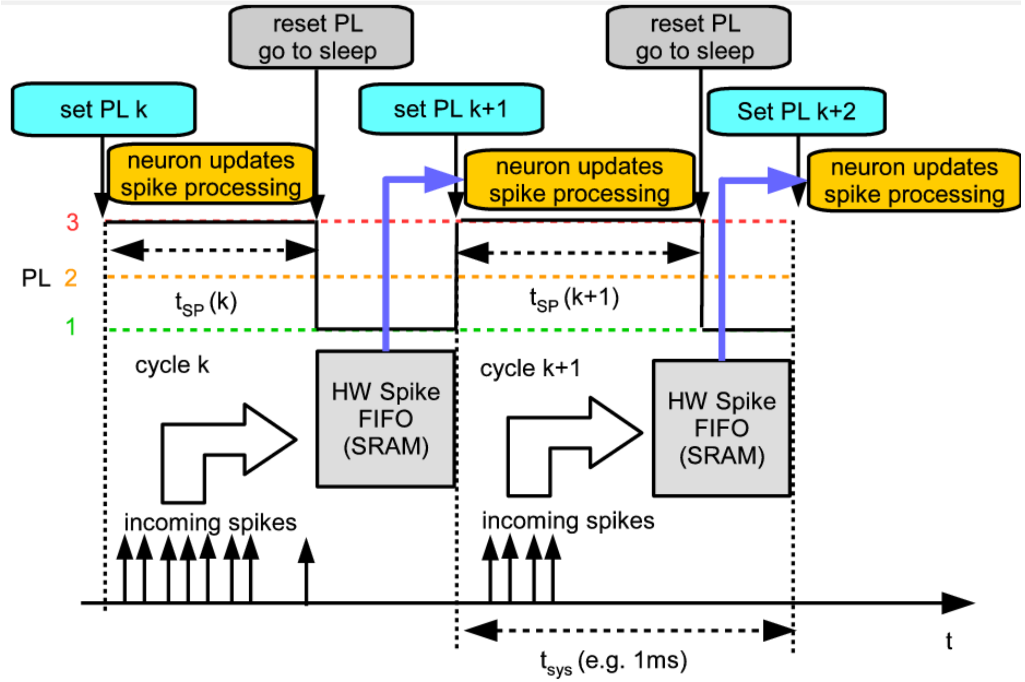
DVFS

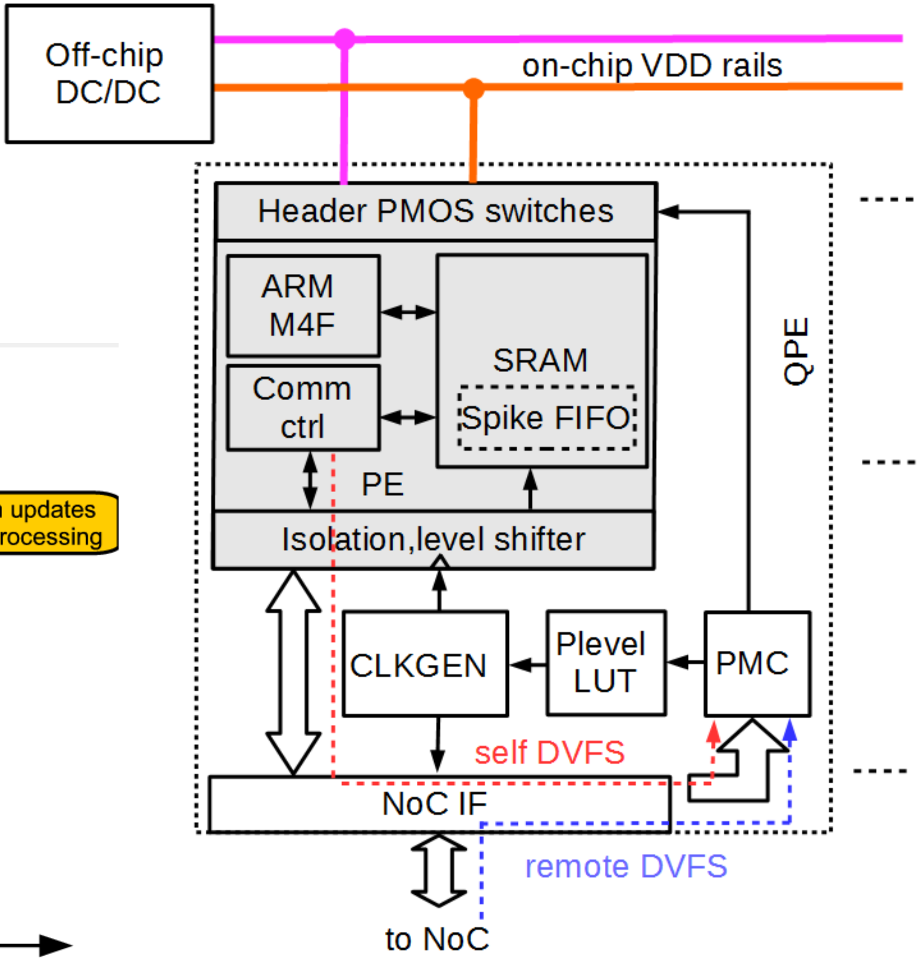
High-performance level (PL2)

- Operate at 500MHz at 0.60V for maximum peak performance for neuromorphic simulations

Neuromorphic Power Management

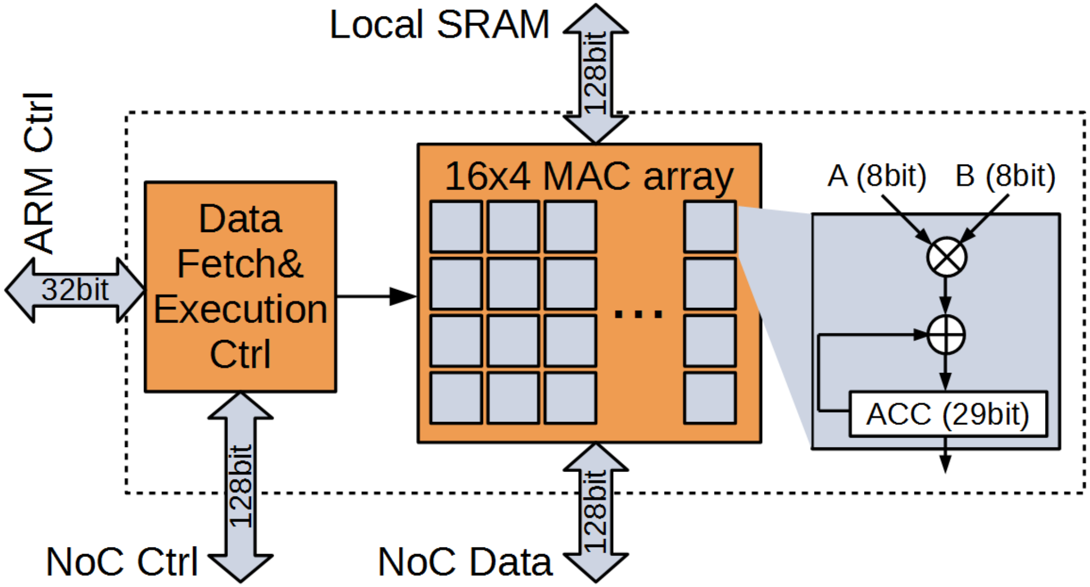
- **D**ynamic **V**oltage and **F**requency **S**caling
- **Fine-grained** (individually per PE)
- **Fast** DVFS (<100ns) PL change time [6]
- **Self**-DVFS PL change from software based on neuromorphic workload





Integrated MAC Accelerator

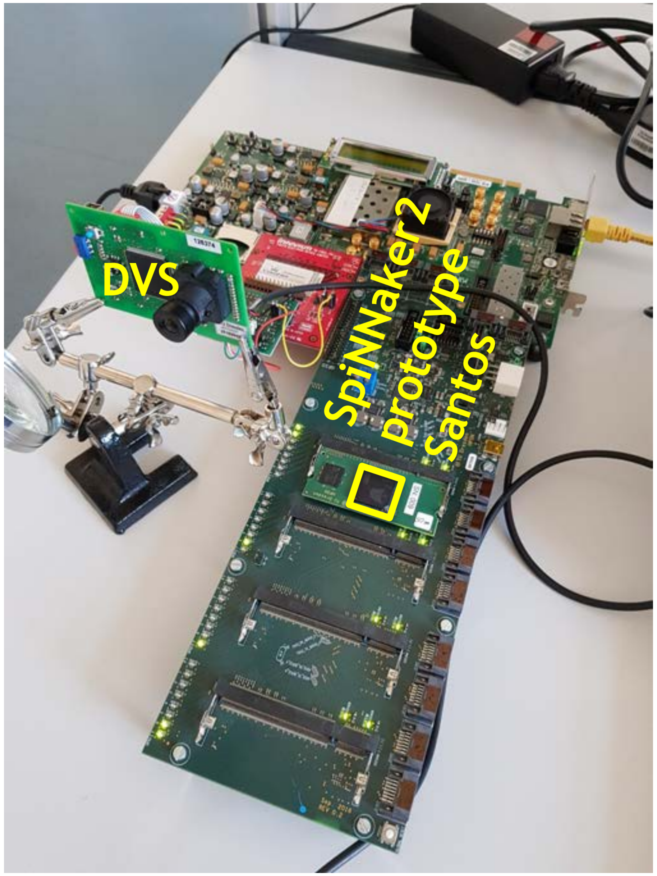
- 16x4 MAC array per PE
- Access local-SRAM and NoC
- Offloading matrix multiplication and convolution from the CPU
- Remote controlled operation possible



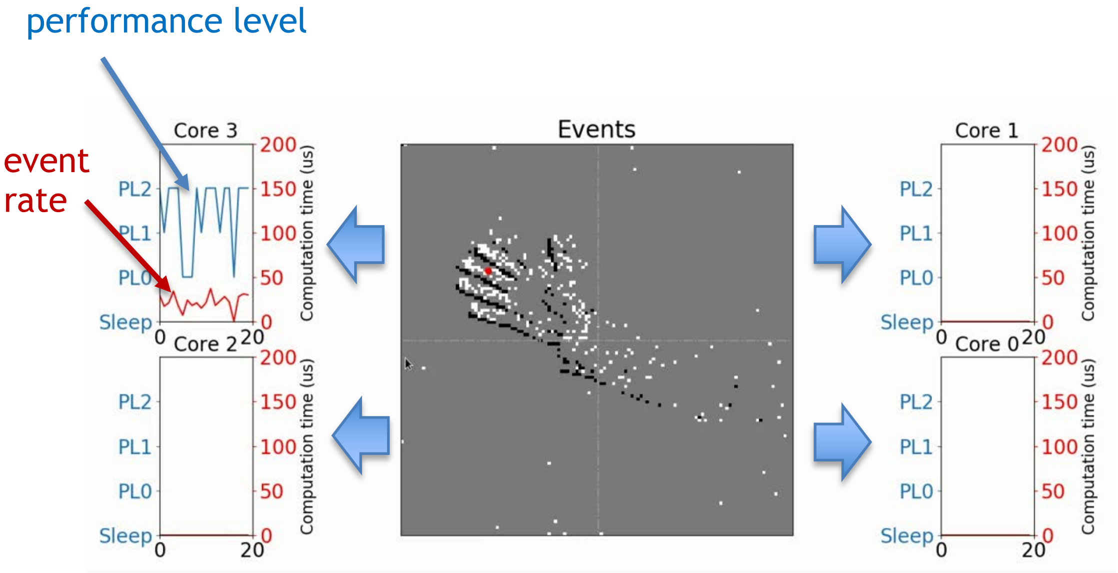
- Peak performance @250MHz:
  - 0.032 TOPS/PE → 4.6TOPS on ***SpiNNaker2*** at ≈ 0.72W PE power consumption → **6.4TOPS/W**

SpiNNaker2 Applications

Demo: DVS with DVFS



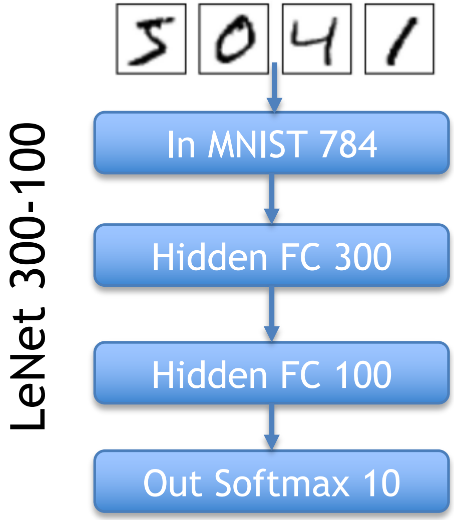
- Dynamic Vision Sensor (DVS) with Dynamic Voltage and Frequency Scaling (DVFS)
- Object tracking example mapped to 4 processors
- 0.2ms processing time steps with performance level adjustment per-step

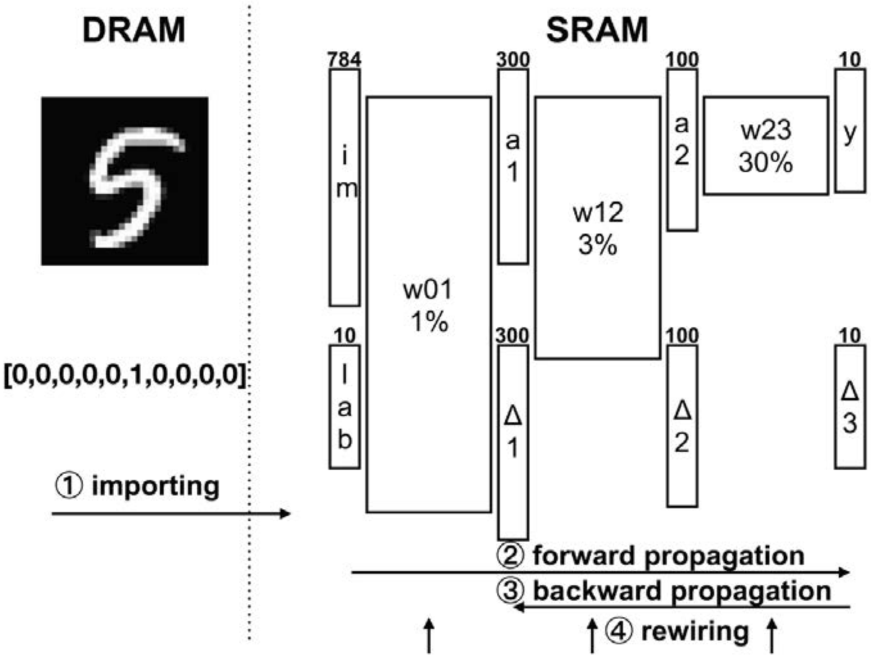


- DVFS results in > x4 power consumption reduction

Deep Rewiring

- Synaptic sampling as dynamic rewiring for rate-based neurons (deep networks)
- Ultra-low memory footprint even during learning
- Uses PRNG/TRNG, FPU, exp
  - → speed-up 1.5
- Example: **LeNet 300-100**
  - 1080 KB → 36 KB
  - training on local SRAM possible
  - ≈ 100x energy reduction for training on SpiNNaker2 prototype (28nm) compared to X86 CPU
  - → 96.2% MNIST accuracy for 0.6% connectivity

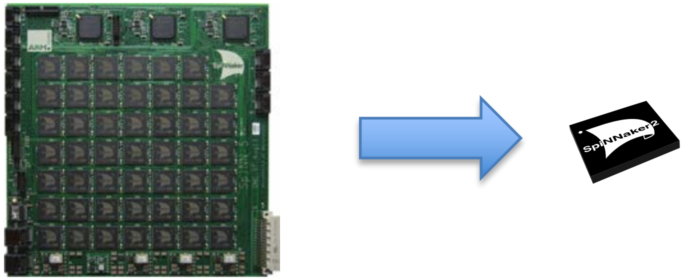




→ Details in [10,11]

Conclusion

- **Energy efficient digital many core approach for neuromorphics**
- **Motivated by advantages of a mix of current approaches:**
  - Processor based → flexibility
  - Fixed digital functionality as accelerators → performance
  - Low voltage (near threshold) operation enabled by 22FDX and ABB → energy efficiency
  - Event driven operation with fine-grained DVFS and energy proportional chip-2-chip links → workload adaptivity
- **Integrate a SpiNNaker 1 48 node board inside a single chip module**



Acknowledgment

The ***SpiNNaker2*** team

Technische Universität Dresden

Sebastian Höppner, Andreas Dixius, Stefan Scholze, Marco Stolba, Thomas Hocker, Stefan Hänzsche, Florian Kelber, Dennis Walter, Johannes Partzsch, Bernhard Vogginger, Johannes Uhlig, Georg Ellguth, Chen Liu, Ali Zeinolabedin, Yexin Yan, Stefan Schiefer, Stephan Hartmann, Love Cederstroem, Stephan Henker, Felix Neumärker, Sani Md Ismail, Christian Mayr

University of Manchester

Delong Shang, Gengting Liu, Dongwei Hu, Jim Garside, Mantas Mikaitis, Andrew Rowley, Luis Plana, Dave Lester, Simon Davidson, Steve Temple, Steve Furber

Thanks to ARM, Racyics and GLOBALFOUNDRIES

References

[1] S. B. Furber et al., "Overview of the SpiNNaker System Architecture," in IEEE Transactions on Computers, vol. 62, no. 12, pp. 2454-2467, Dec. 2013. doi: 10.1109/TC.2012.142

[2] R. Carter et al., "22nm FDSOI technology for emerging mobile, Internet-of-Things, and RF applications," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 2.2.1-2.2.4. doi: 10.1109/IEDM.2016.7838029

[3] S. Höppner et al., "Dynamic voltage and frequency scaling for neuromorphic many-core systems," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, 2017, pp. 1-4. doi: 10.1109/ISCAS.2017.8050656

[4] J. Partzsch et al., "A fixed point exponential function accelerator for a neuromorphic many-core system," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, 2017, pp. 1-4. doi: 10.1109/ISCAS.2017.8050528

[5] F. Neumarker, S. Höppner, A. Dixius and C. Mayr, "True random number generation from bang-bang ADPLL jitter," 2016 IEEE Nordic Circuits and Systems Conference (NORCAS), Copenhagen, 2016, pp. 1-5. doi: 10.1109/NORCHIP.2016.7792875

[6] S. Höppner, C. Shao, H. Eisenreich, G. Ellguth, M. Ander and R. Schüffny, "A power management architecture for fast per-core DVFS in heterogeneous MPSoCs," 2012 IEEE International Symposium on Circuits and Systems, Seoul, 2012, pp. 261-264. doi: 10.1109/ISCAS.2012.6271840

[7] Mantas Mikaitis, et al., Approximate Fixed-Point Elementary Function Accelerator for the SpiNNaker-2 Neuromorphic Chip, *submitted to ARITH25*

[8] www.makeChip.design

[9] D. Kappel et al., “A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning,” arXiv, 2017.

[10] G. Bellec et al., “Deep rewiring: Training very sparse deep networks”, arXiv, 2018

[11] Chen Liu et al., “Memory-efficient Deep Learning on a SpiNNaker 2 prototype”, *submitted*

Thanks for your attention