TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

Johannes Partzsch
Chair for Highly-Parallel VLSI Systems and Neuromicroelectronics

Neuromorphic Hardware - A System Perspective

NHR PerfLab Seminar Erlangen
15 April 2025

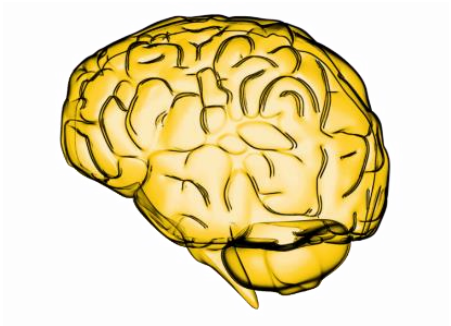
Outline

Core Ideas of Neuromorphic Hardware Revisited

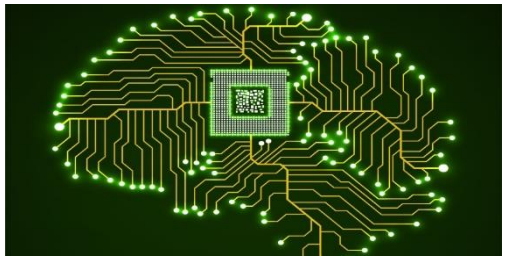
SpiNNcloud: The SpiNNaker2 System at TU Dresden

Outlook

Neuromorphic Hardware: Core Ideas

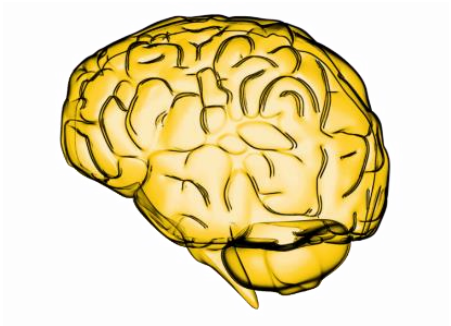


Rebuild, Learn and Apply

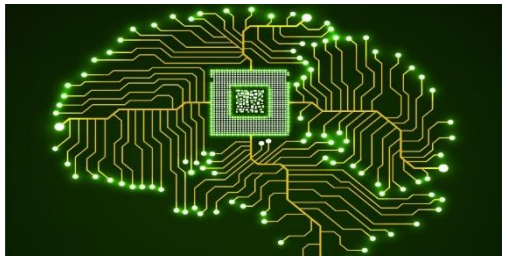


- Brain
- Brain Area
- Microcircuit
- Neuron
- Synapse
- Ion Channel

Neuromorphic Hardware: Core Ideas

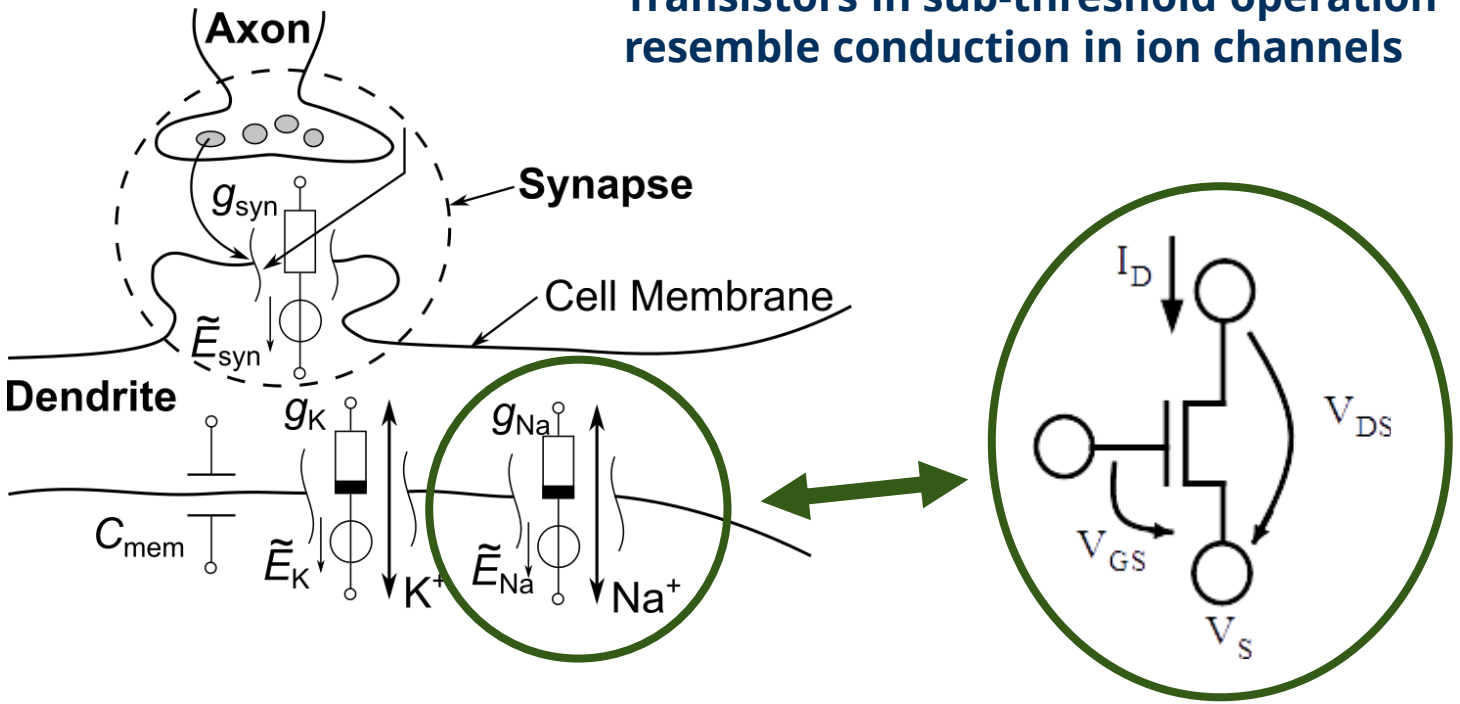


Rebuild, Learn and Apply

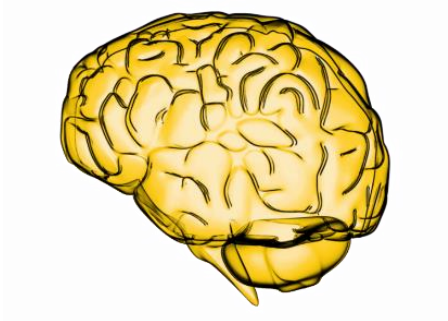


- Brain
- Brain Area
- Microcircuit
- Neuron
- Synapse
- Ion Channel

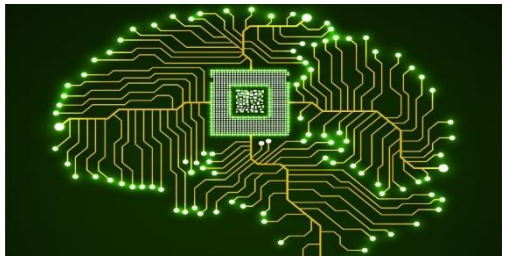
Transistors in sub-threshold operation resemble conduction in ion channels



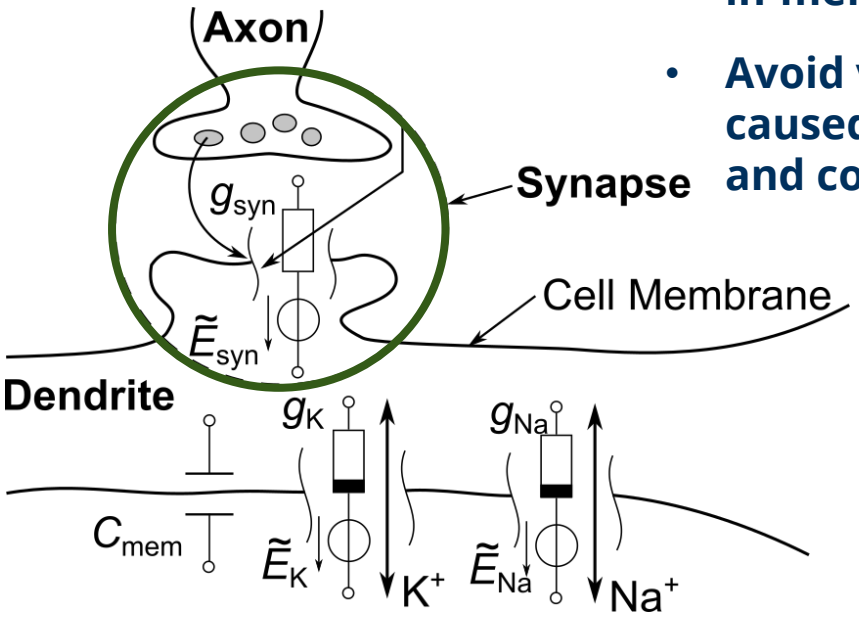
Neuromorphic Hardware: Core Ideas



Rebuild, Learn and Apply



- Brain
- Brain Area
- Microcircuit
- Neuron
- Synapse
- Ion Channel



- **Synapses combine weight storage and computation, i.e. perform in-memory computing**
- **Avoid von-Neumann bottleneck, caused by separation of memory and computation**

Neuromorphic Hardware: Large-Scale Systems vs. Core Ideas

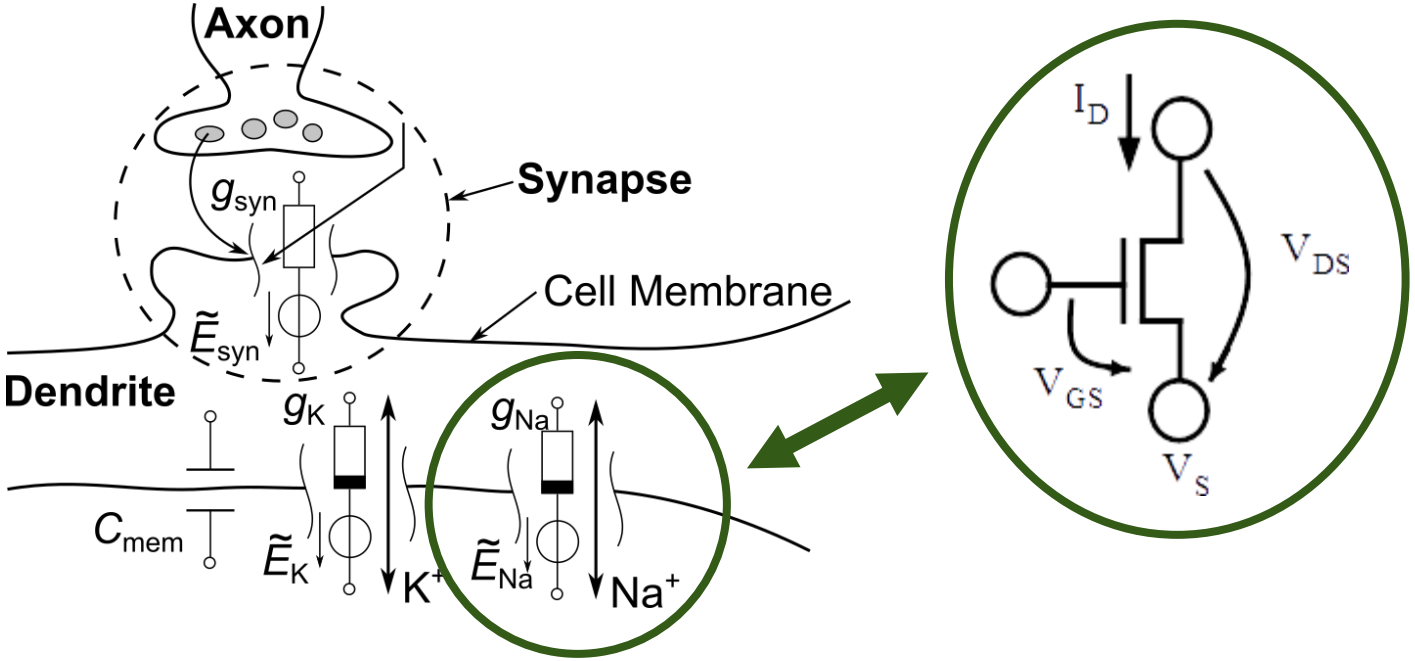
	Subthreshold	Analog computation	Merged weight storage& computation	Asynchronous Circuits
Neurogrid	✓	✓	(✓)	✓
BrainScaleS	(✓)	✓	✓	(✓)
IBM TrueNorth	✗	✗	✗	(✓)
Tianjic	✗	✗	✗	✗
Intel Loihi/Loihi2	✗	✗	✗	✓
SpiNNaker/SpiNNaker2	✗	✗	✗	(✓)

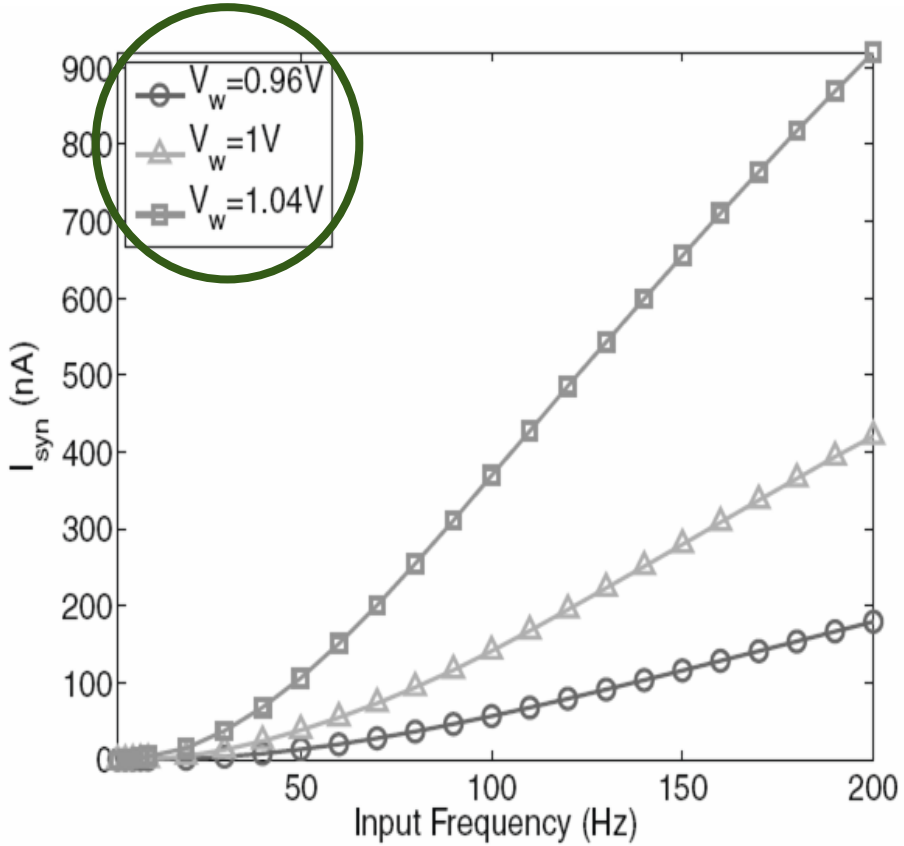
Core Ideas Revisited: Sub-Threshold Circuits

Super low voltages and currents = **potential for extraordinary energy efficiency in theory,**

BUT:

- Exponential voltage/current relation means **high susceptibility** to manufacturing tolerances and changes in bias voltages
- **High temperature dependence** (like in the brain…)





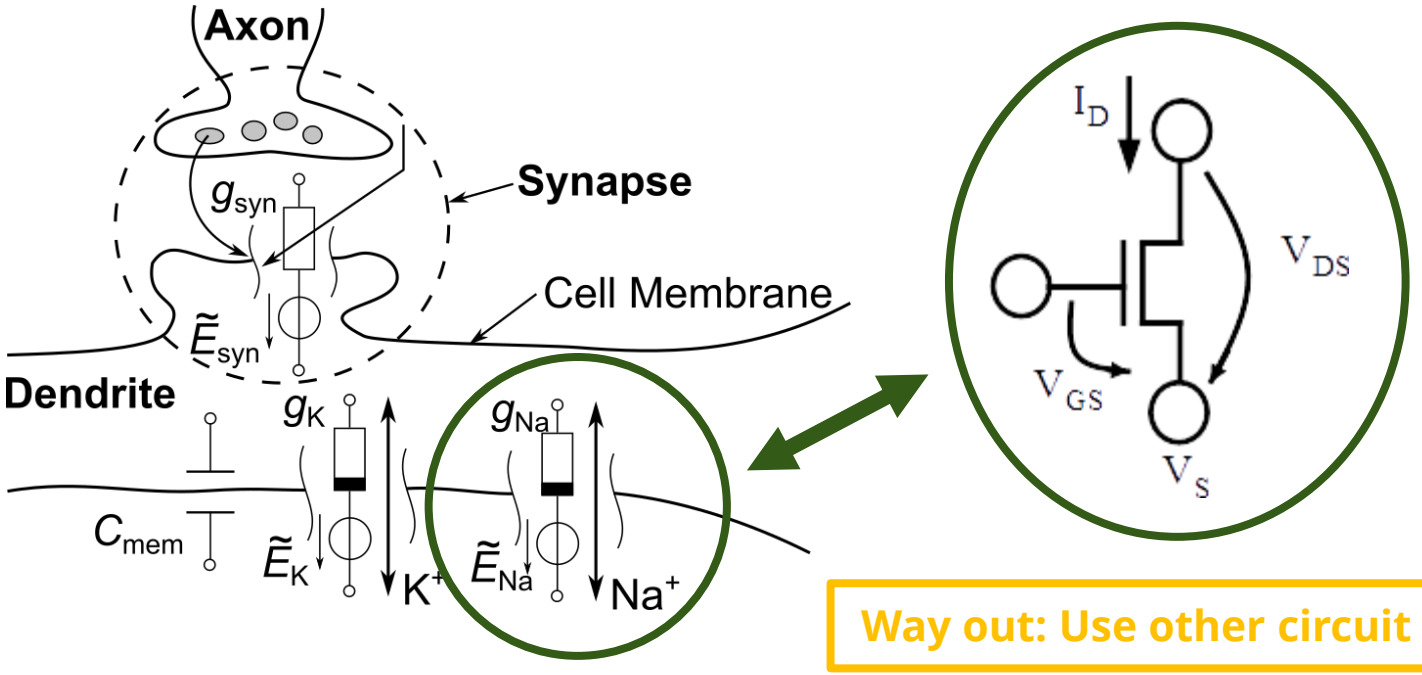
Synaptic Dynamics in Analog VLSI
Bartolozzi and Indiveri 2007

Core Ideas Revisited: Sub-Threshold Circuits

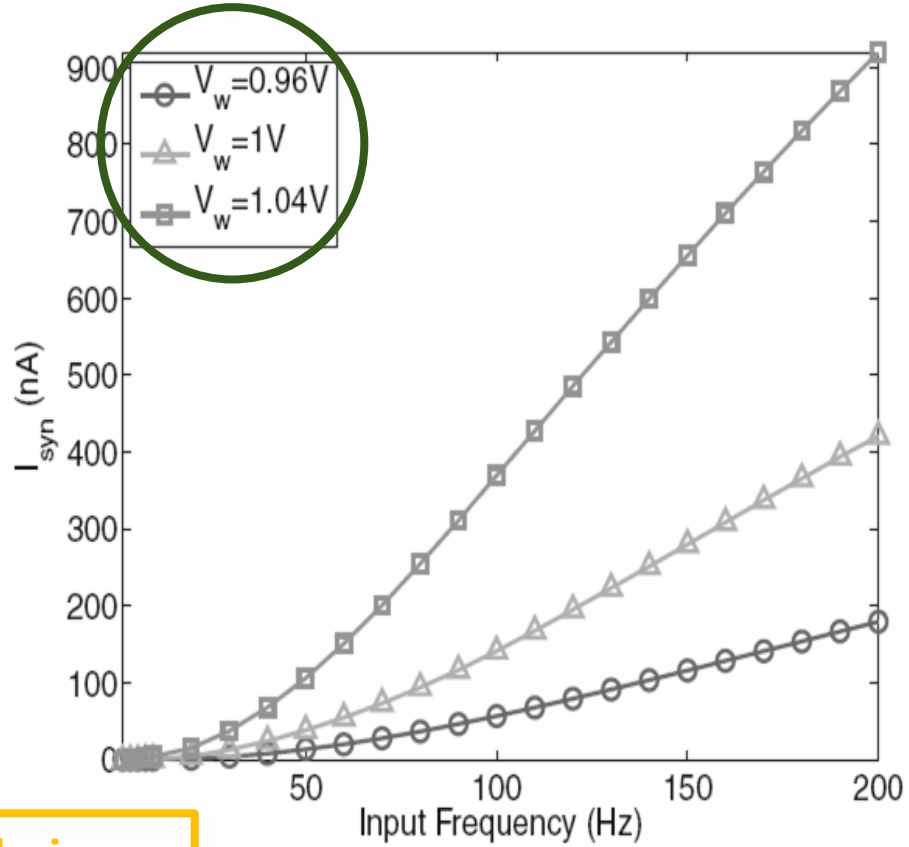
Super low voltages and currents = **potential for extraordinary energy efficiency in theory,**

BUT:

- Exponential voltage/current relation means **high susceptibility** to manufacturing tolerances and changes in bias voltages
- **High temperature dependence** (like in the brain…)



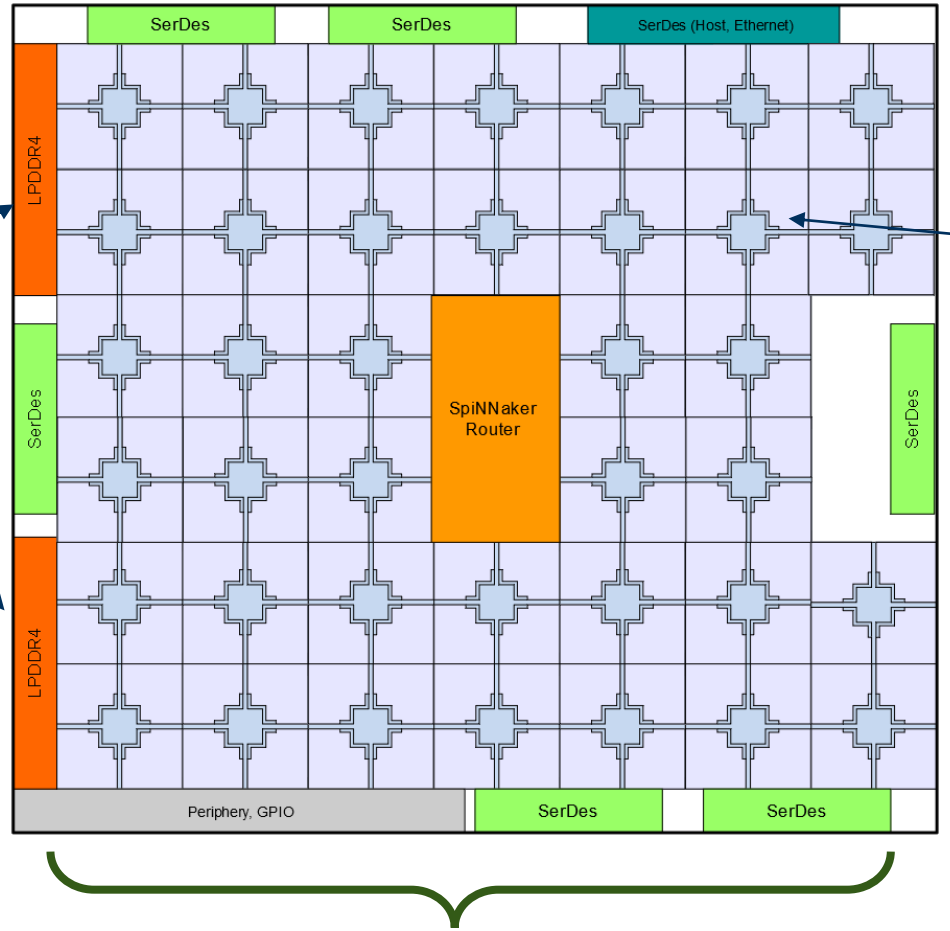
Way out: Use other circuit techniques



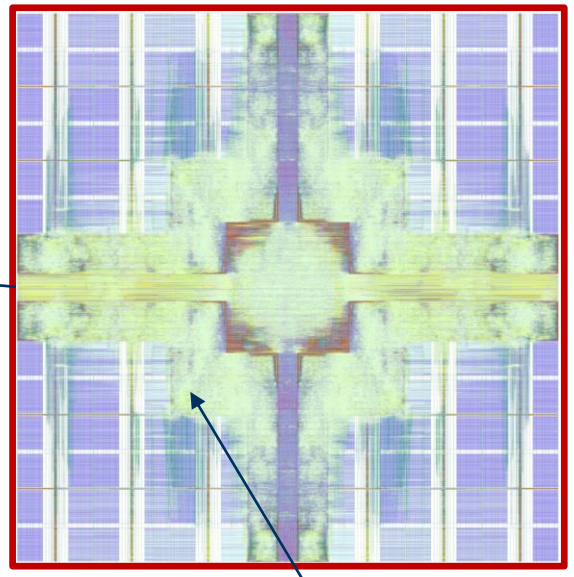
Synaptic Dynamics in Analog VLSI
Bartolozzi and Indiveri 2007

Core Ideas Revisited: von-Neumann Bottleneck vs. SpiNNaker2

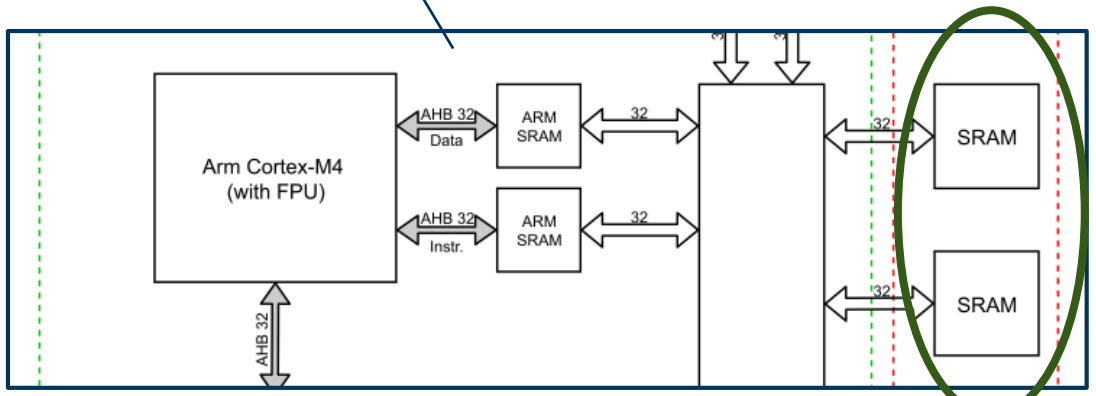
2 GByte



19 MByte



**Memory: 85%
of area per core**



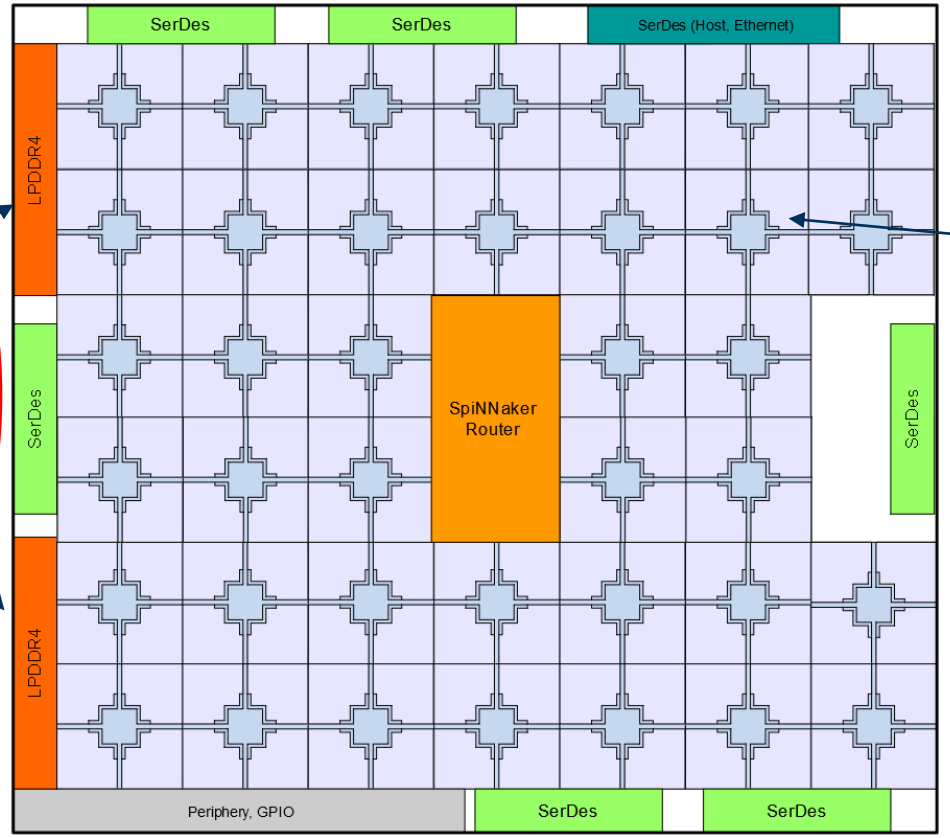
128 kByte

Core Ideas Revisited: von-Neumann Bottleneck vs. SpiNNaker2

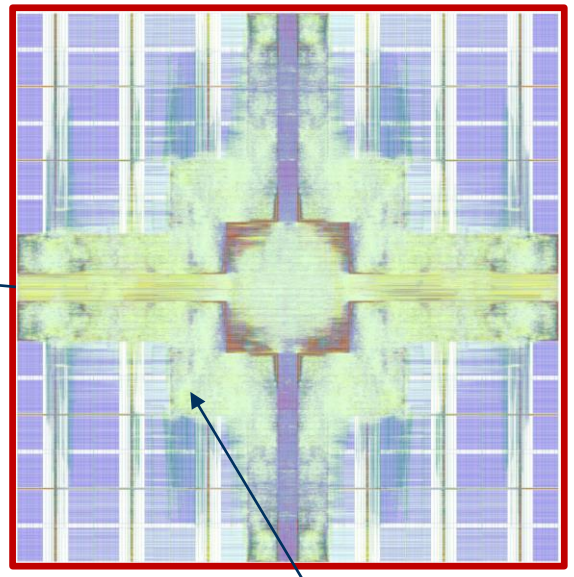
“big” von Neumann

2 GByte

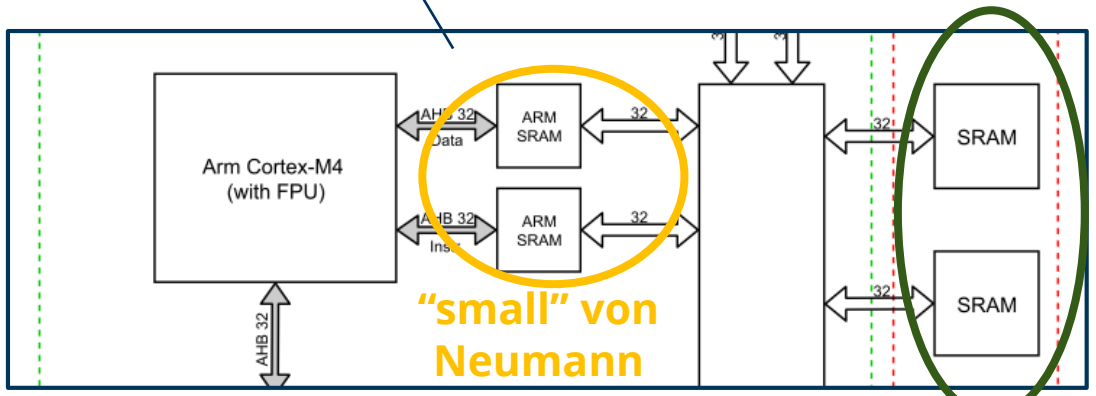
(>100 SpiNNaker2 Chips)



19 MByte



Memory: 85% of area per core



“small” von Neumann

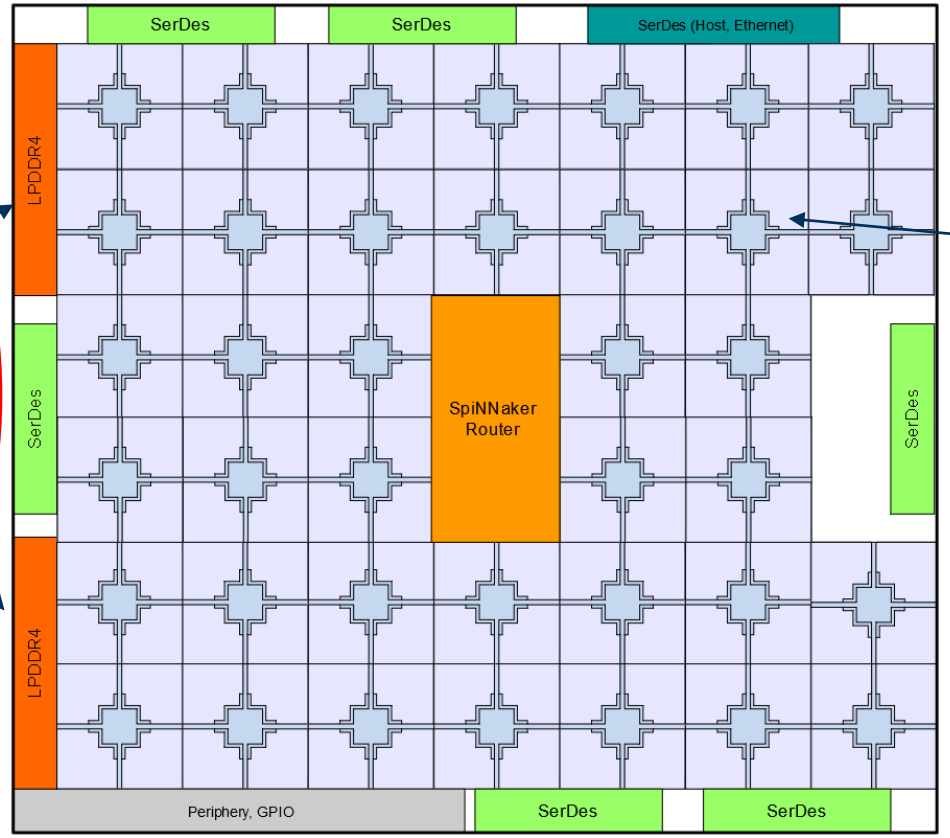
128 kByte

Core Ideas Revisited: von-Neumann Bottleneck vs. SpiNNaker2

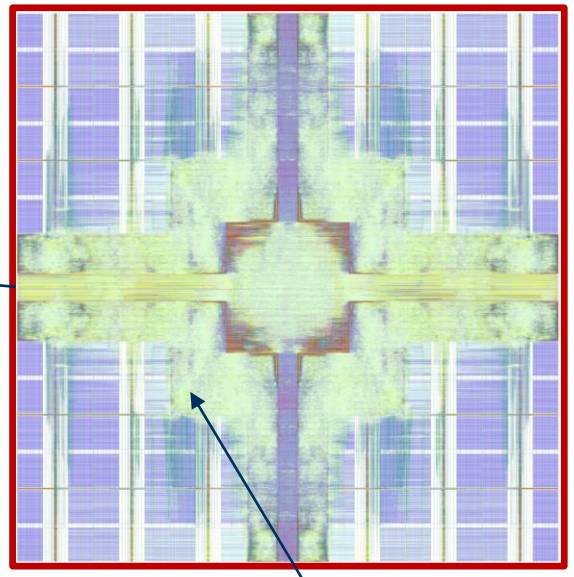
“big” von Neumann

2 GByte

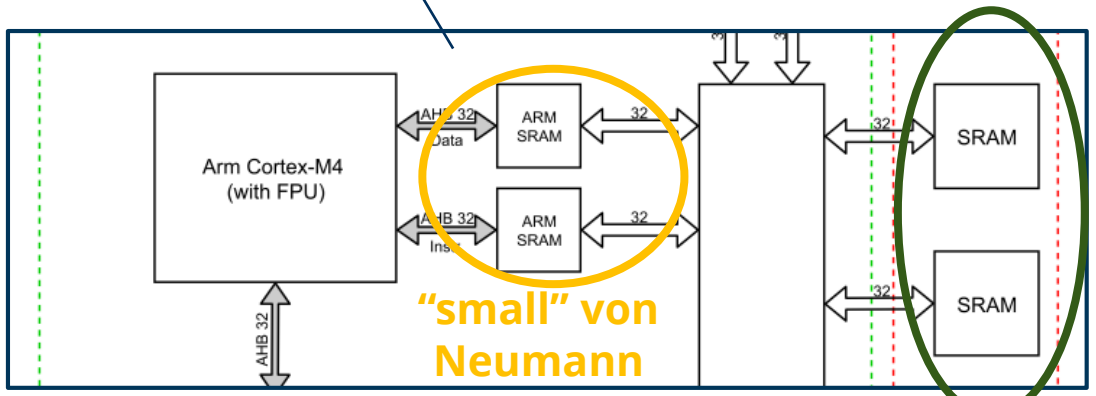
(>100 SpiNNaker2 Chips)



19 MByte



Memory: 85% of area per core



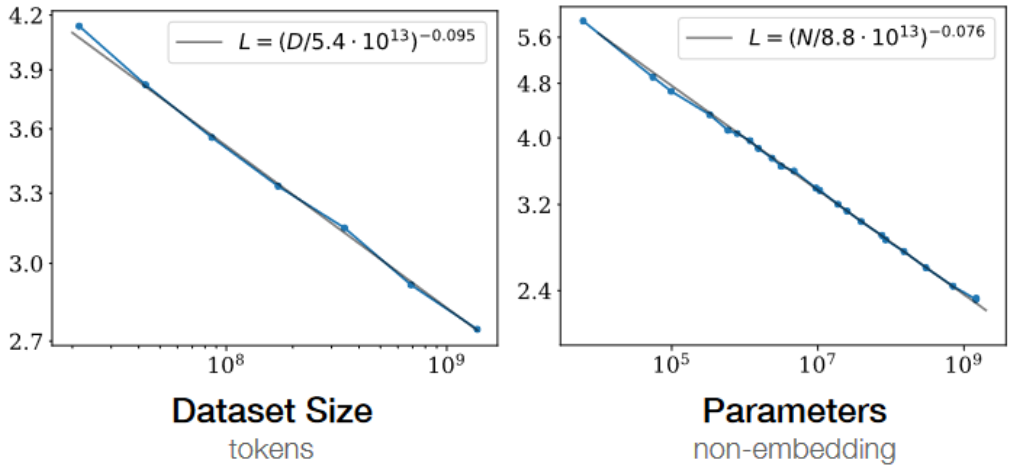
“small” von Neumann

128 kByte

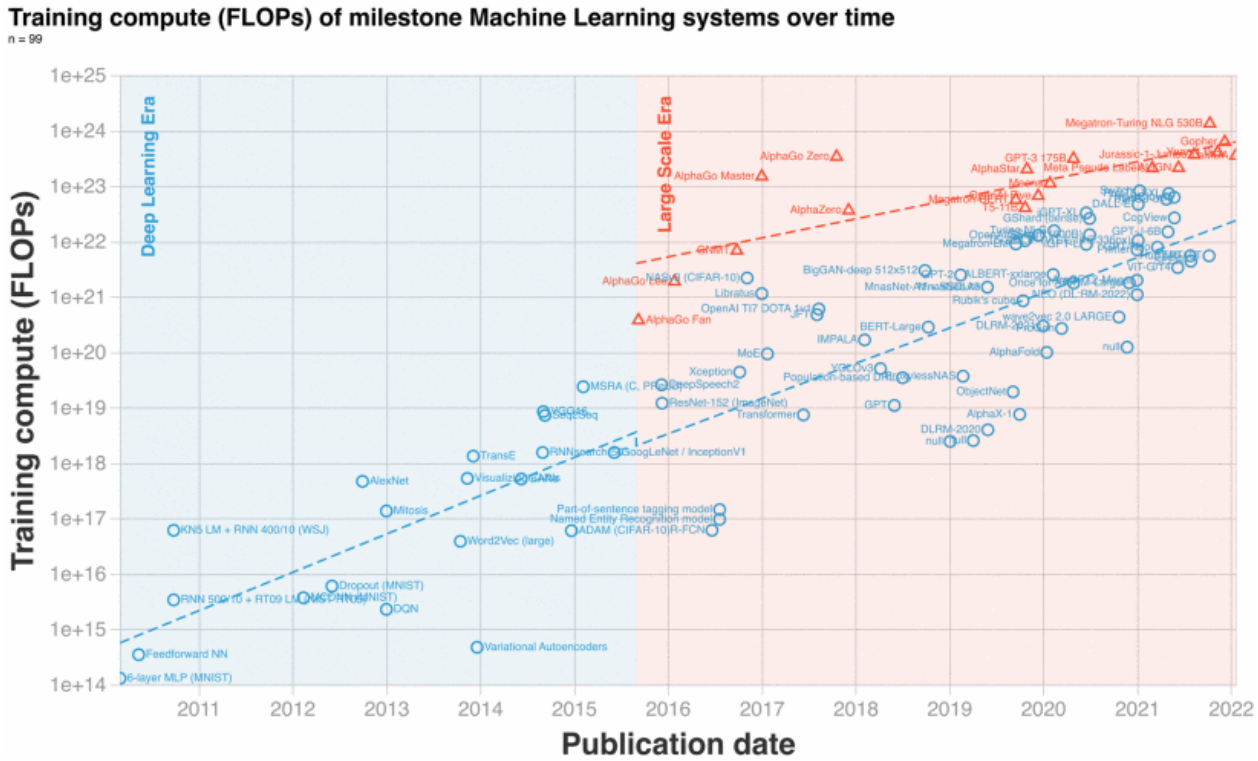
Ways out:
- **Compress models**
- **Emerging memories**
- **Stick to small size**

Indication from Deep Learning: Scale Matters…

Scaling Laws for Neural Language Models
Kaplan et al. 2020



Compute Trends Across Three Eras of Machine Learning
Sevilla et al. 2022

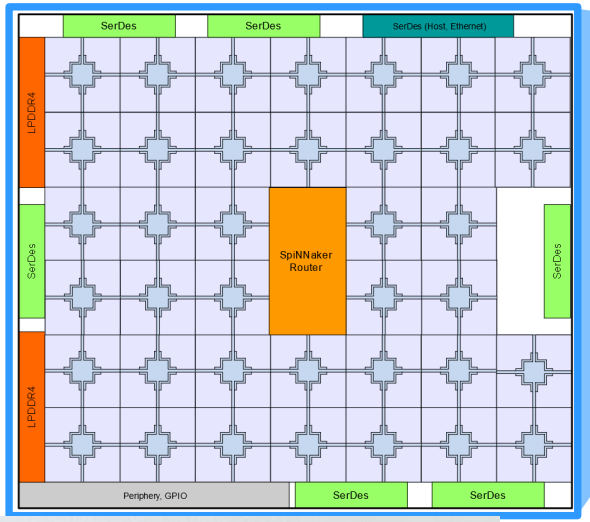


- Models become better with invested compute and memory
- Could the neuromorphic approach bring a significant gain in efficiency?

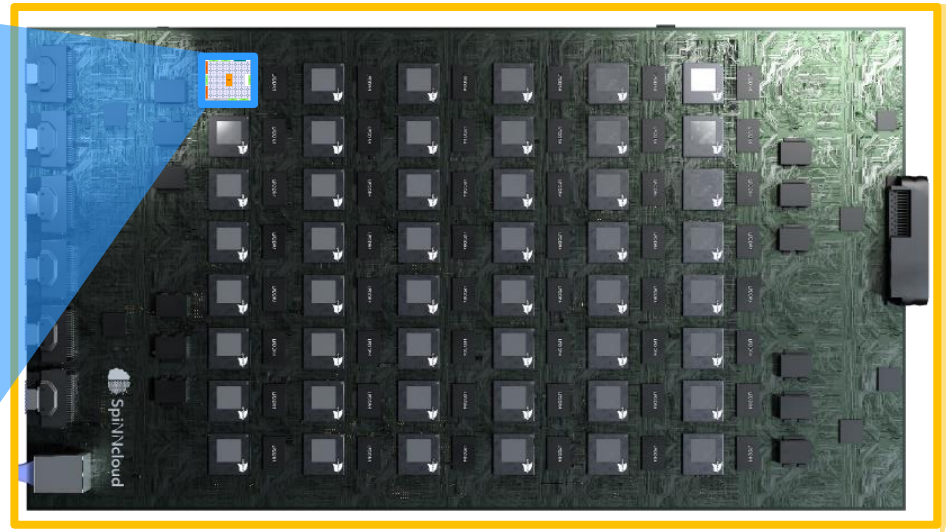
SpiNNcloud: Scaling of a Neuromorphic System

MANCHESTER 1824
The University of Manchester

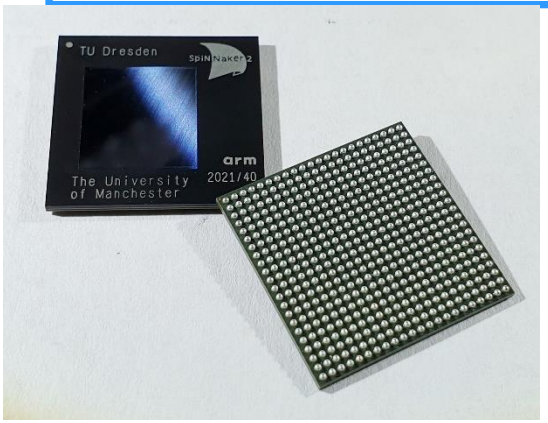
SpiNNaker2 Chip



SpiNNcloud Board



SpiNNcloud

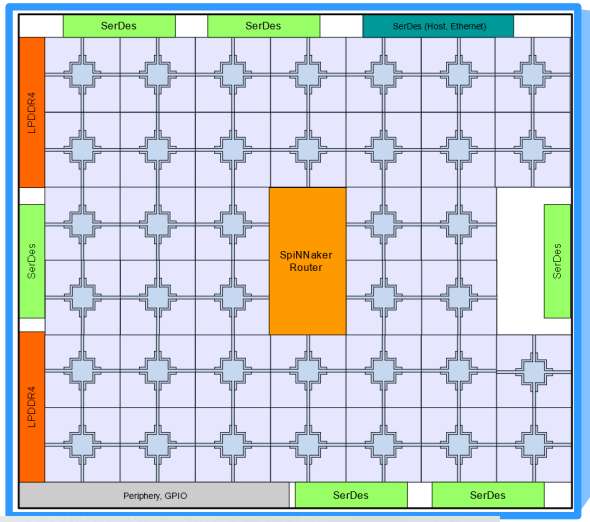


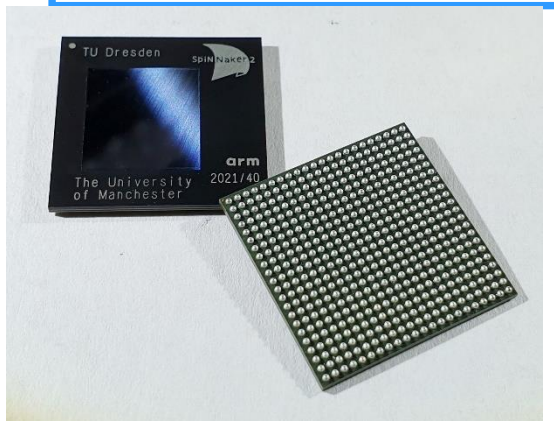
- 5 million cores (ARM processor+accelerators+memory) in 8 server racks (scaling target: 10 million cores in 16 server racks)
- approx. 35000 SpiNNaker2 chips
- Power budget at scaling target: 400kW
- Capacity for simulating >10 billion spiking neurons in real time

SpiNNcloud: Consequences of Scaling

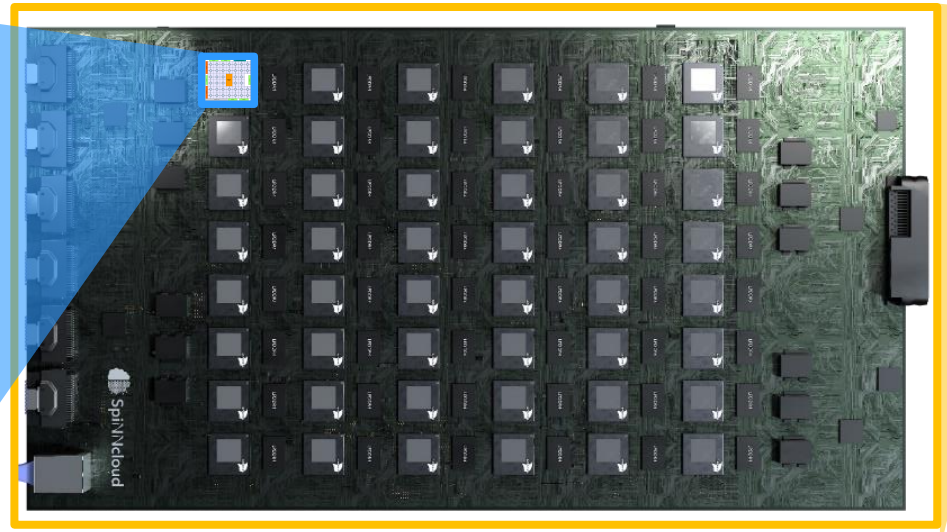
The University of Manchester

SpiNNaker2 Chip





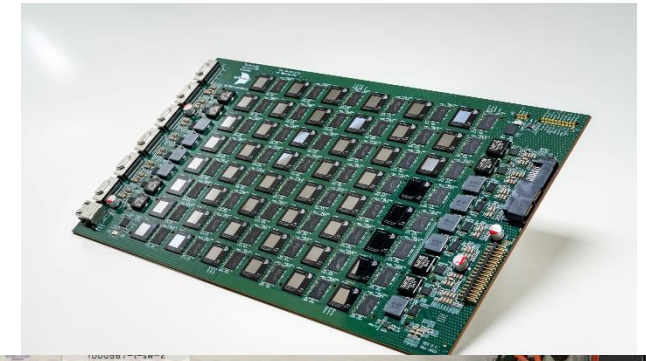
SpiNNcloud Board

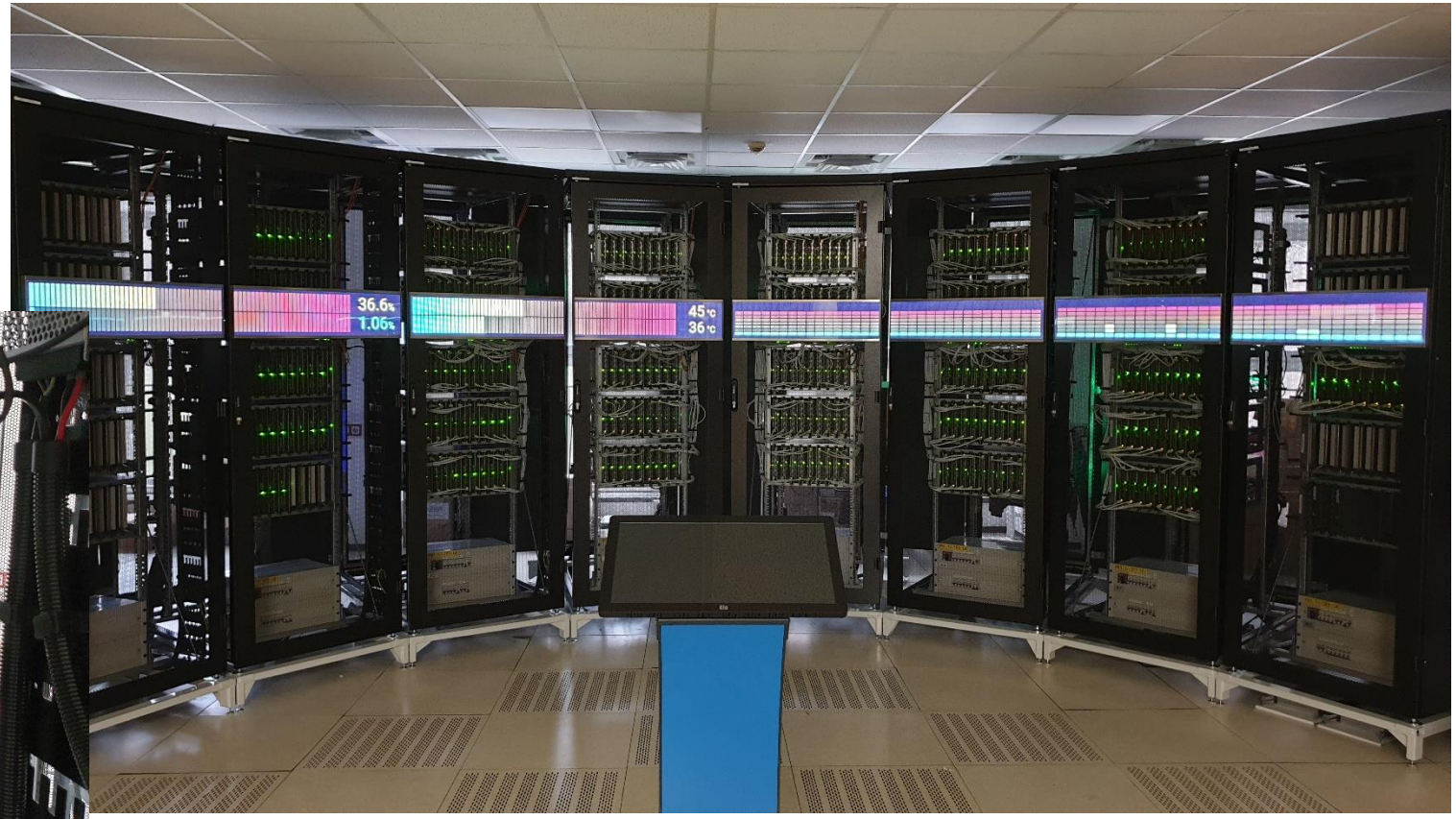


SpiNNcloud

- **Reliability**: single components need to work out of the box
- **Flexibility**: development is significant investment in time and resources – should be applicable to variety of use cases
- **Usability**: Making technology accessible to users is significant software effort
- **Neuromorphic challenge**: efficiently computing with events (spikes) means: minimize baseline power, cope with workload variations

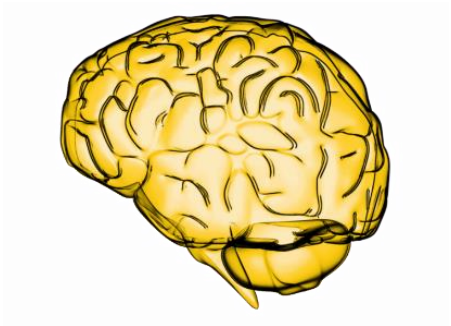
SpiNNcloud: State of Deployment



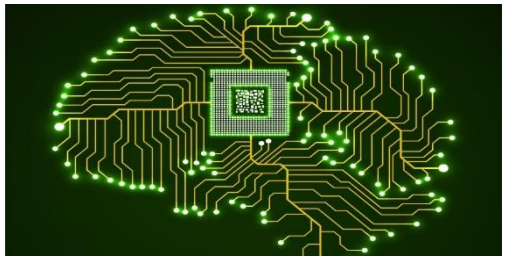


- (Almost) all boards 720 installed, cooling infrastructure up and running
- User access and operating software brought up currently

What makes SpiNNaker2 neuromorphic?



Rebuild, Learn and Apply



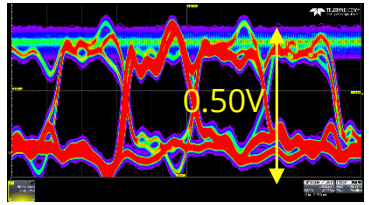
- Brain
- Brain Area
- Microcircuit
- Neuron
- Synapse
- Ion Channel

- Many tiny compute elements
- Workload adaptation
- Event communication

SpiNNaker2's main targets today:

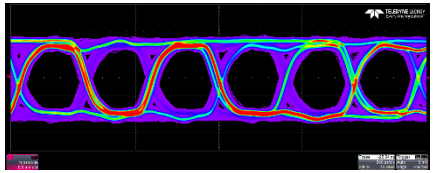
- Real-time spiking neural network simulation
- Neuromorphic algorithm development
- Experimental platform for computing with many tiny compute elements

SpiNNaker2 Chip Architecture



6x Serial Chip-2-Chip Links
- Only 0.50V swing
- 6 data lanes per link
- Up to 2.0Gbit/s per pin

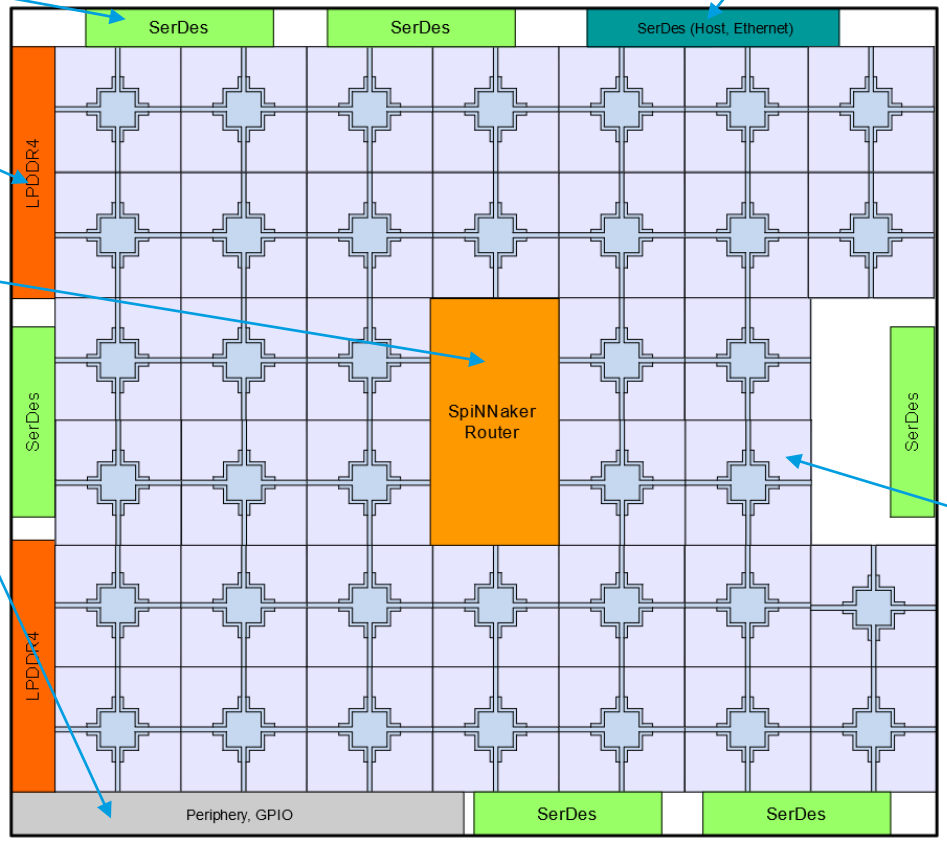
1x Host Interface
- LVDS SerDes
- 1Gbps Ethernet Connection



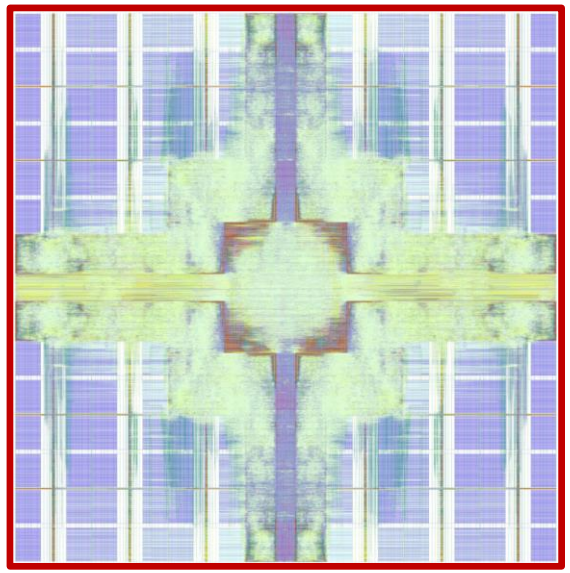
2x LPDDR4 Memory Interfaces
- 16-bit data width
- up to 2.4Gbit/s per pin

SpiNNaker Packet Router
- Lightweight event packet communication fabric

Periphery
- Flexible GPIO, SPI, I2C
- Arm M4F management processor
- Temperature sensor
- ABB generator



38x Quad-Core-Processing Elements
- in total 152 Arm M4F cores
- hardware accelerators
- dynamic power management
- adaptively body biased



- **All chip macros are implemented for "connect by place"**
- **No globally synchronous clocks**
- **Communication via 2 parallel NoCs (communication NoC, system NoC)**

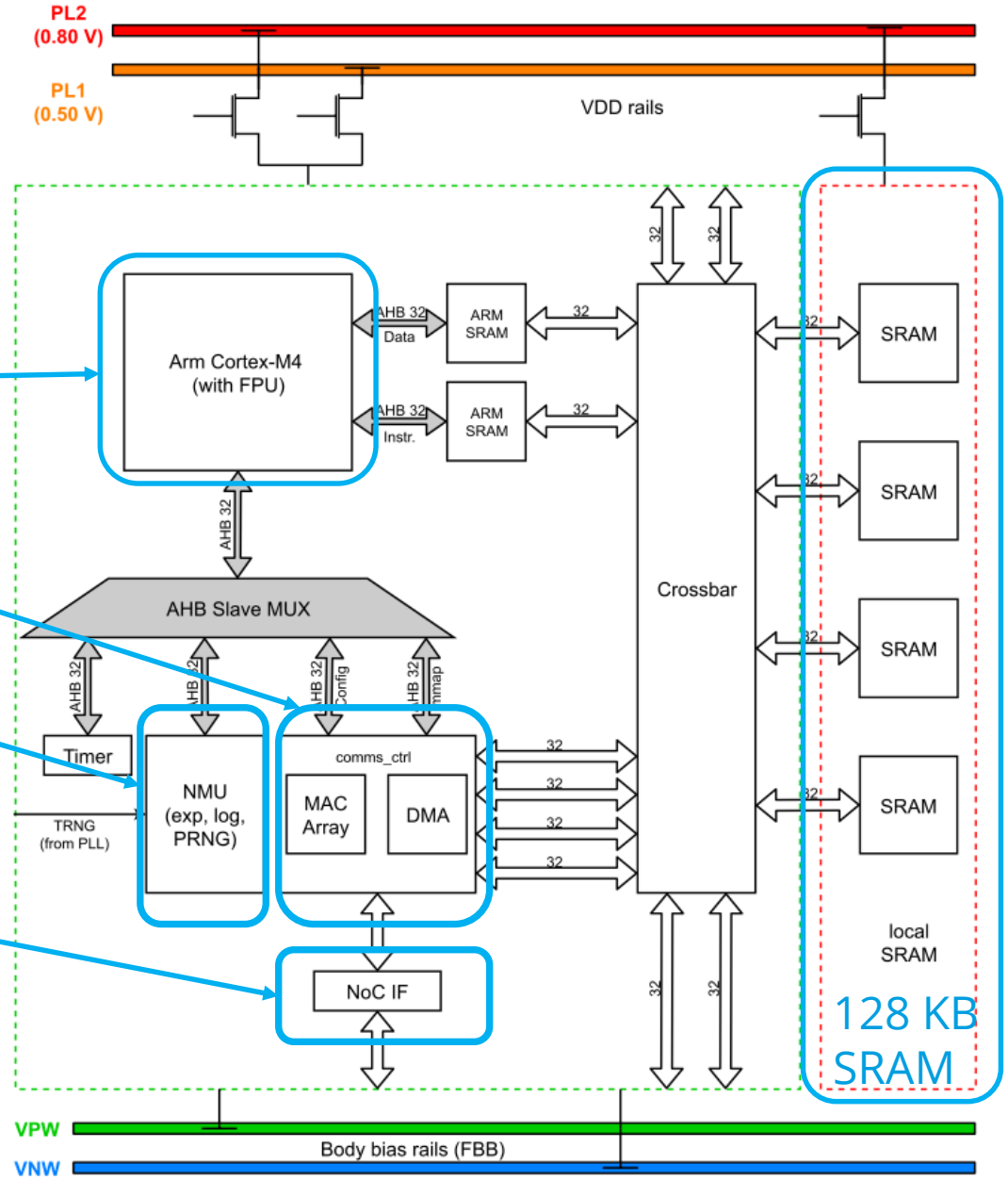
SpiNNaker2 Chip Architecture
Processing Element

ARM M4F processor
- Software realization of neurons and synapses
- Flexibility

Multiply-Accumulate accelerator
- MAC array with direct memory access (DMA)

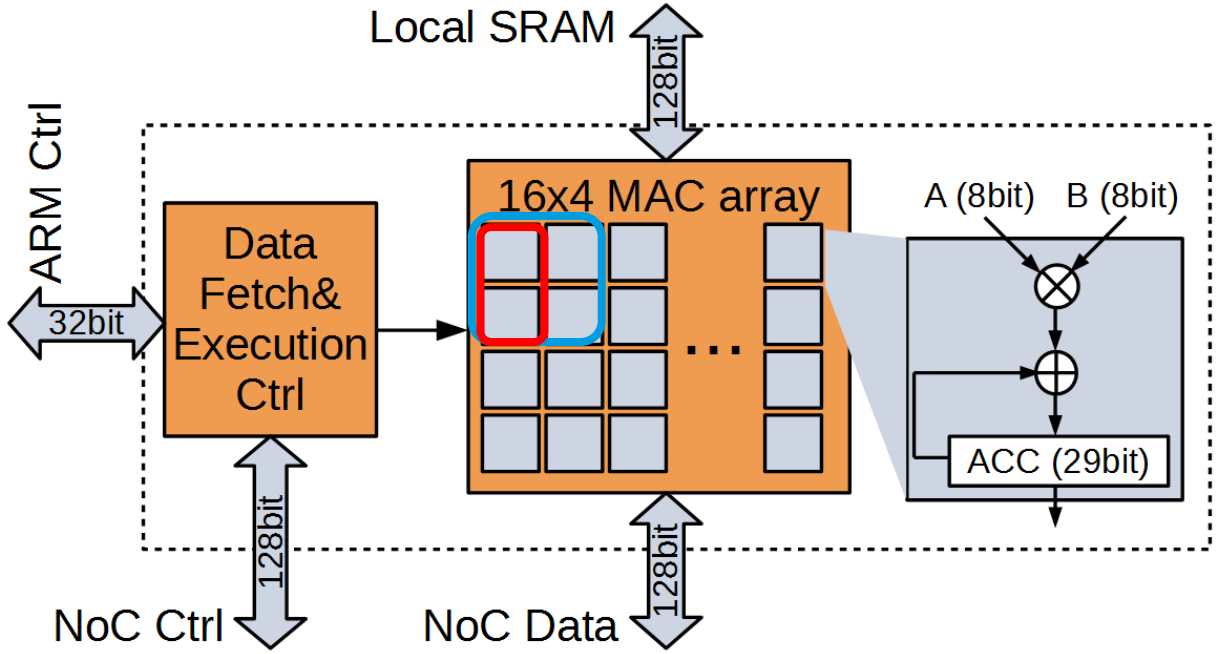
Neuromorphic accelerators
- Exp/log
- Random numbers (PRNG, TRNG)

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



SpiNNaker2 Chip Architecture
MAC Accelerators

- 16x4 MAC array per processing element
- operation modes:
  - matrix multiplication
  - 2D convolution
- ReLU and quantization to 8, 16 or 32 bit
- Combine 2 or 4 MAC units for 8x16 bit or 16bx16b integer operations:
  - Allows to trade-off resolution and throughput (performance)



- Only 7 % of silicon area of PE
- Makes SpiNNaker2 a competitive DNN inference chip:
  - 4.5 TOPS max throughput
  - 2.1 TOPS/W max energy efficiency

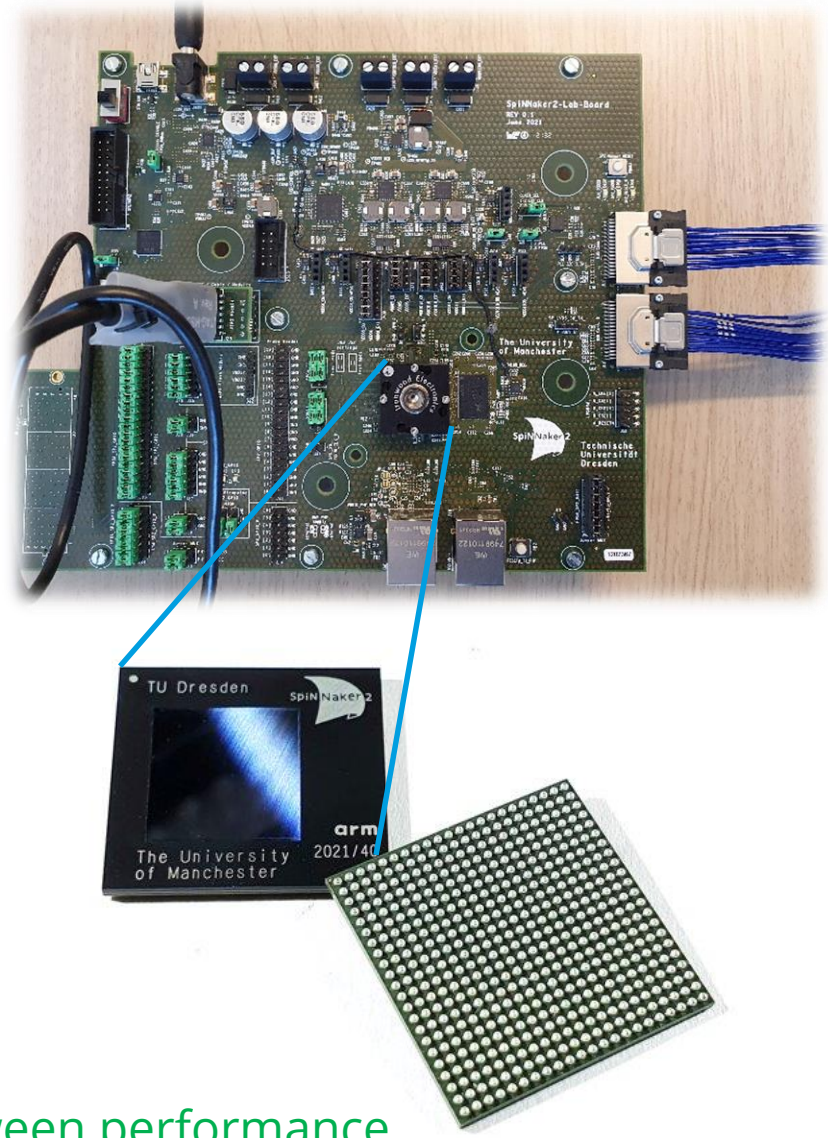
SpiNNaker2 Chip Architecture
Characteristics

- Optimized for minimum baseline power: ~250mW
- Enabled by Racyics ABB 0.5V IP
- Performance measurements (152 cores active):

	PL1 (0.50V, 150MHz)	PL2 (0.80V, 300MHz)
CoreMark Score (iterations/second)	40851	81711
CPU Gop/s	13.9	27.9
Energy efficiency [uW/MHz]	16.3	24.0
MatMul performance (TOP/s)	2.24	4.50
Tops/W	2.1	1.6

High CPU efficiency

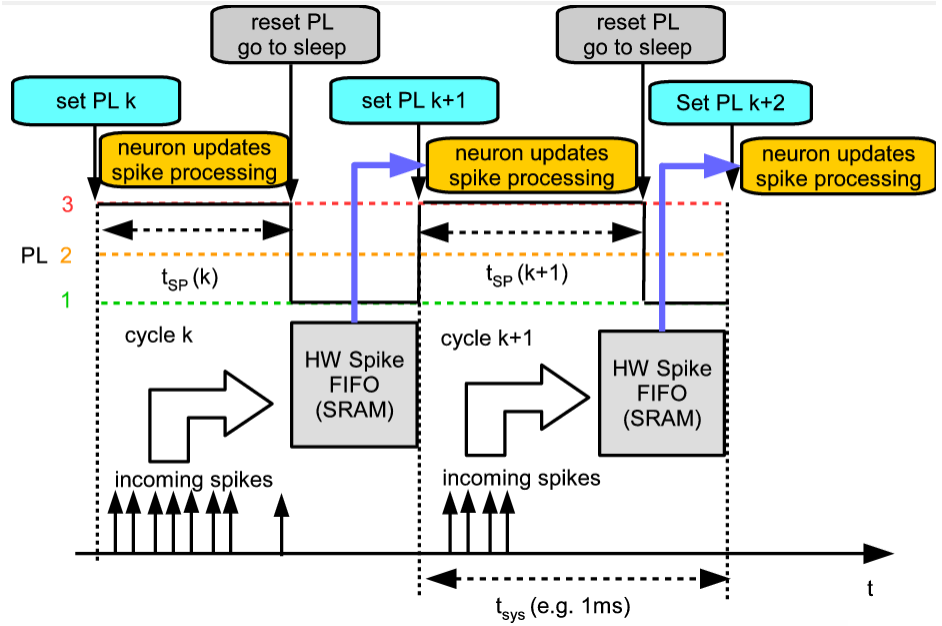
Trade-off between performance and efficiency for DNN inference

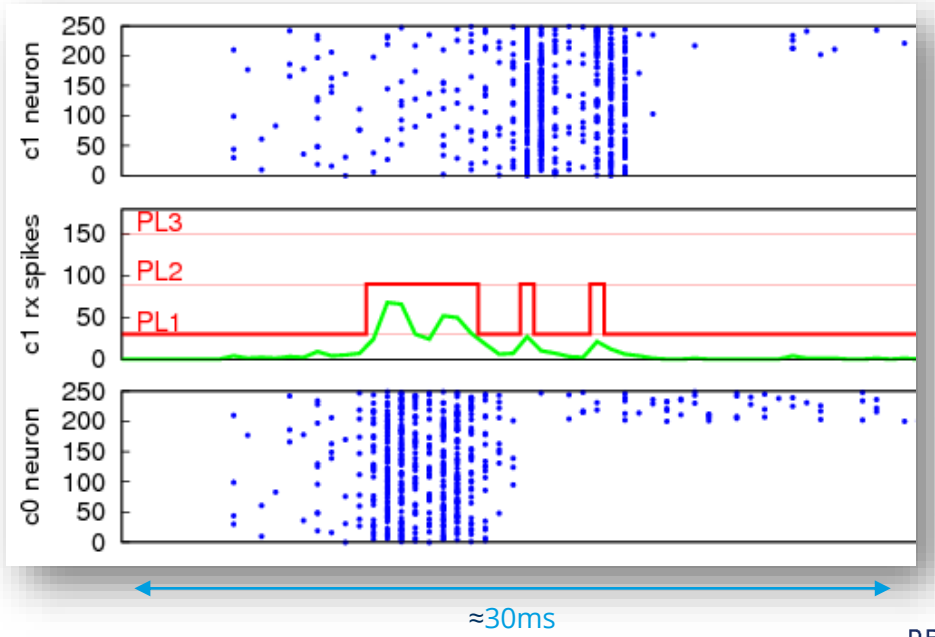


SpiNNaker2 Workload Adaptation

- **Fine-grained** (individually per PE) and **fast** DVFS (<100ns) performance level (PL) change
- **Self-DVFS:** performance level change from PE software, based on neuromorphic workload
- For spiking neuromorphic benchmarks:
  ≈ 90% of simulation cycles are processed at lowest PL
  - → maximum energy efficiency
- **System performance limit is reached at highest PL (only ≈ 2% of simulation cycles)**
  - → peak performance for real time achieved

Up to ≈50% PE power reduction, while still achieving peak performance for real time operation

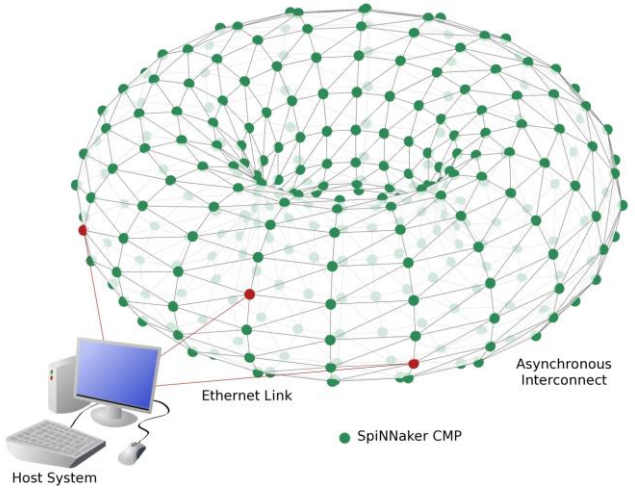


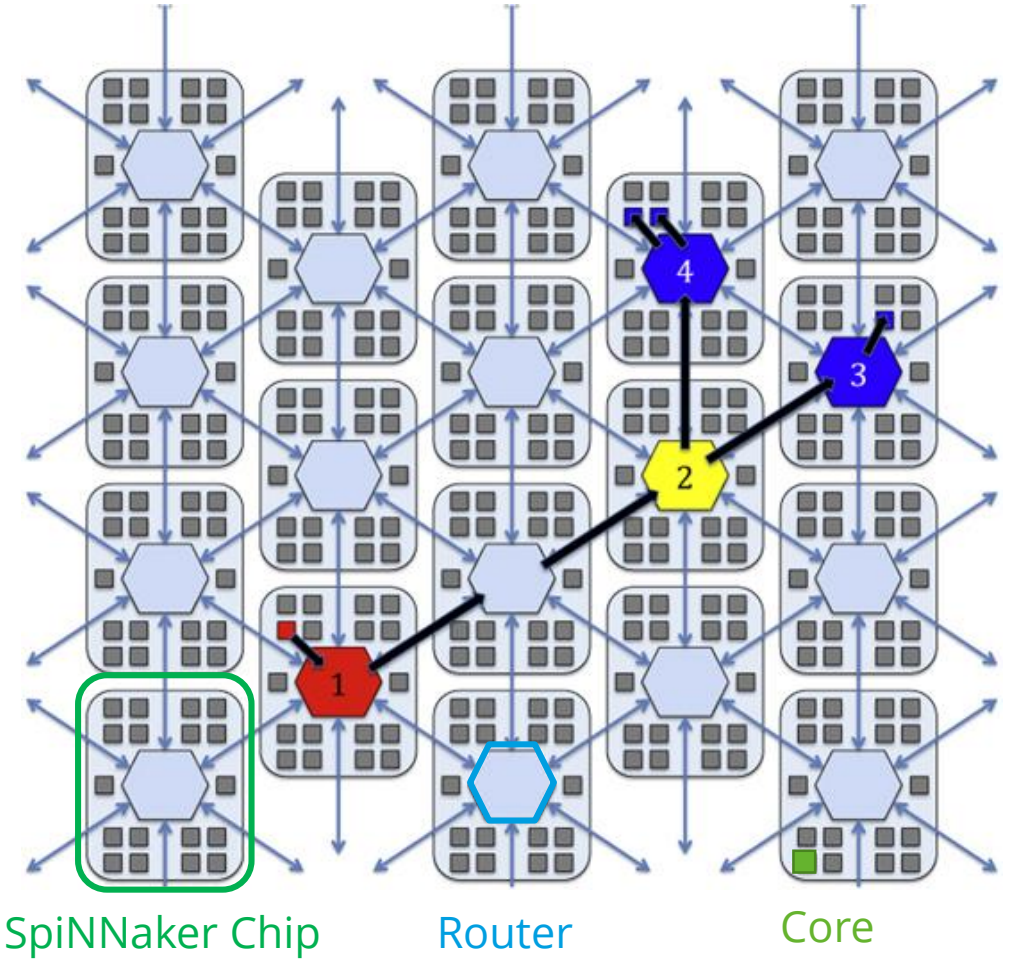


Synfire Chain benchmark, measured on SpiNNaker2 prototype

SpiNNaker Event Communication

- **Scalable communication adapted from SpiNNaker1**
  - Multi-cast routing of events by SpiNNaker Routers
  - SpiNNaker Multicast packets contain 32-bit key -> Address Event Reprensentation (AER) of each pre-synaptic neuron
  - Hexagonal mesh as torus
  - Asynchronous, real-time operation
  - Advancements in SpiNNaker2:
    - **Variable payload 0 – 128 bits per Multicast packet**
    - **Global Read/Write packets for efficient point-to-point data transfer**

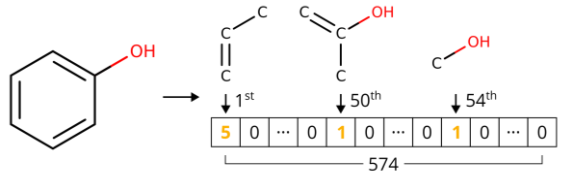


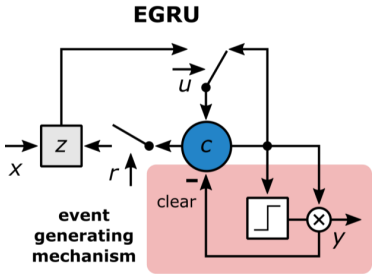


Large-scale neuromorphic computing systems
Steve Furber, 2016

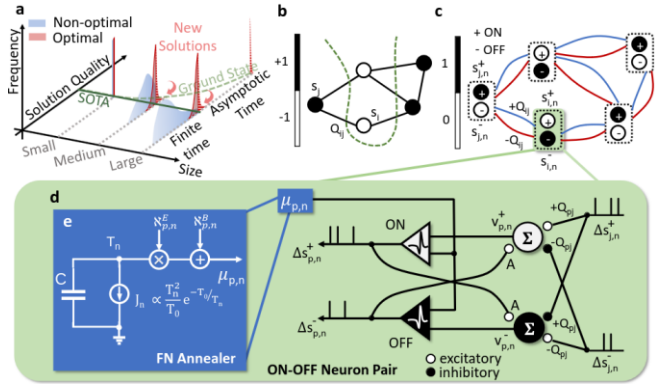
SpiNNcloud Applications (Selection)

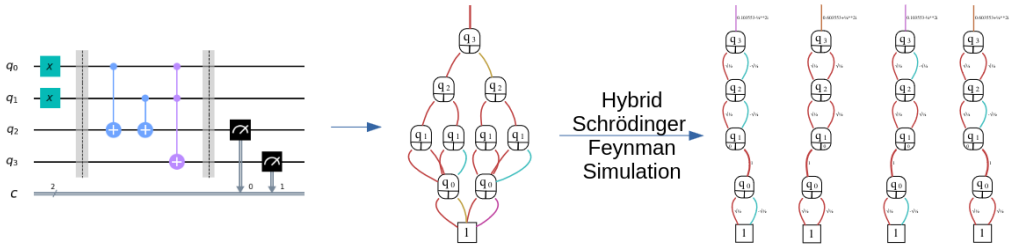
- **Drug discovery**: virtual high throughput screening of billions of molecules to find promising compounds
- **Efficient, event-based Deep Learning:** 18x less energy for language modelling with RNN than NVIDIA A100 (see later slides). Currently scaling up to Transformers and LLM
- **Optimization**: Efficient solving of **QUBO** (quadratic unconstrained binary optimization) problems with Spiking Neural Networks and random sources
- Scalable **Quantum Emulation** as bridge technology towards quantum computing





https://arxiv.org/abs/2406.05224

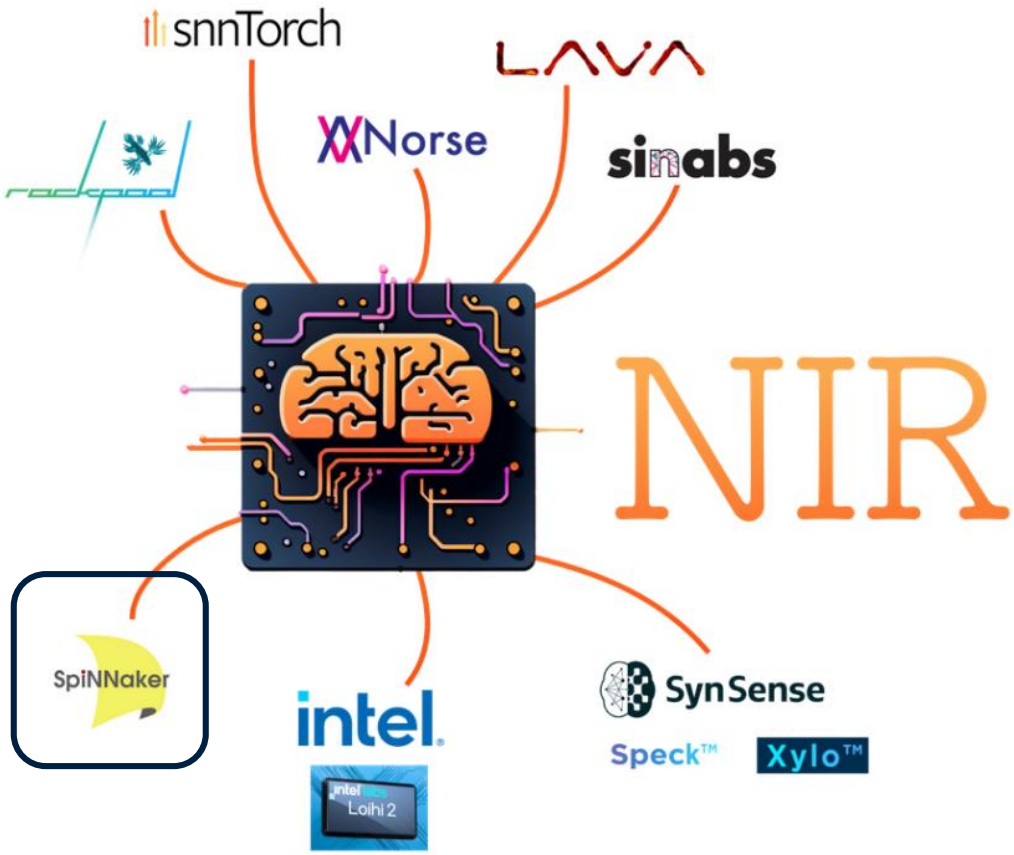




Quantum circuit | Decision diagram | Independent simple diagrams

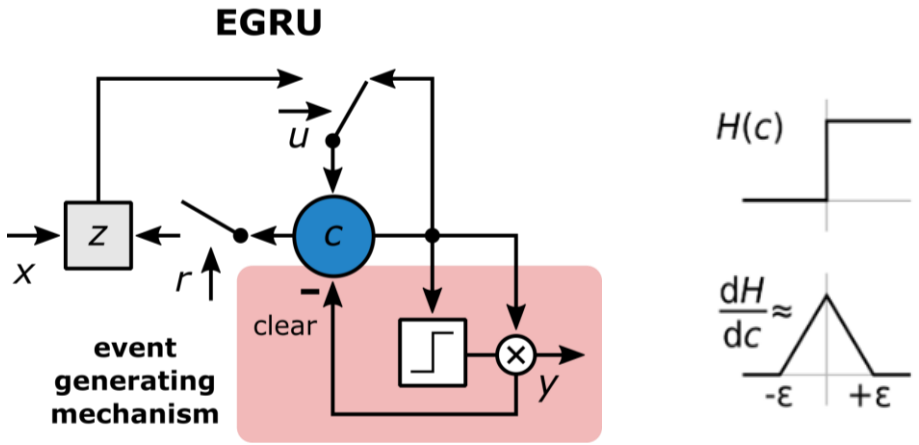
Neuromorphic Intermediate Representation (NIR)

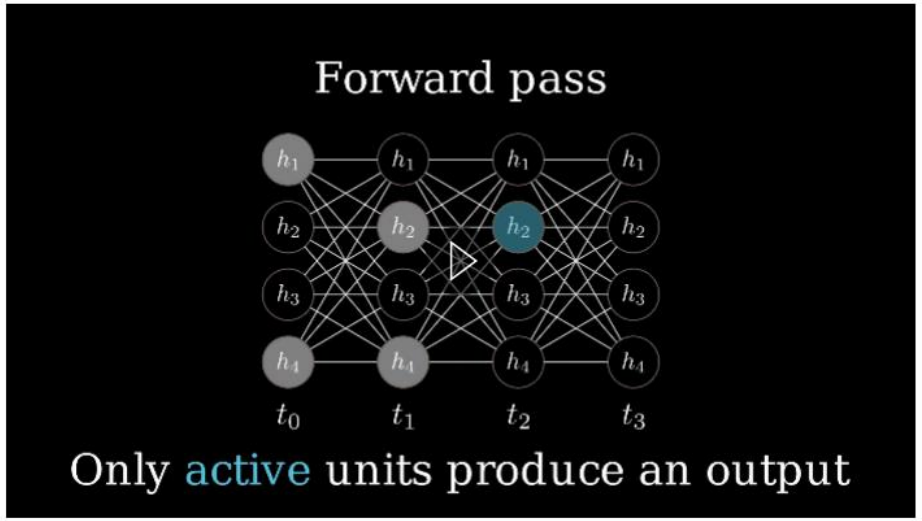
- Support to run deep SNN on SpiNNaker2:
  Train in any SNN training framework and convert via NIR (neuromorphic intermediate representation) to SpiNNaker2
- NIR was developed together with the neuromorphic community (started in Telluride 2023)
- Serves as an intermediate representation format similar to ONNX for DNN
- Goal: greater interoperability
- Supports 7 software frameworks (e.g. snnTorch, lava, Norse, NengoDL) and 4 hardware systems
- https://neuroir.org/



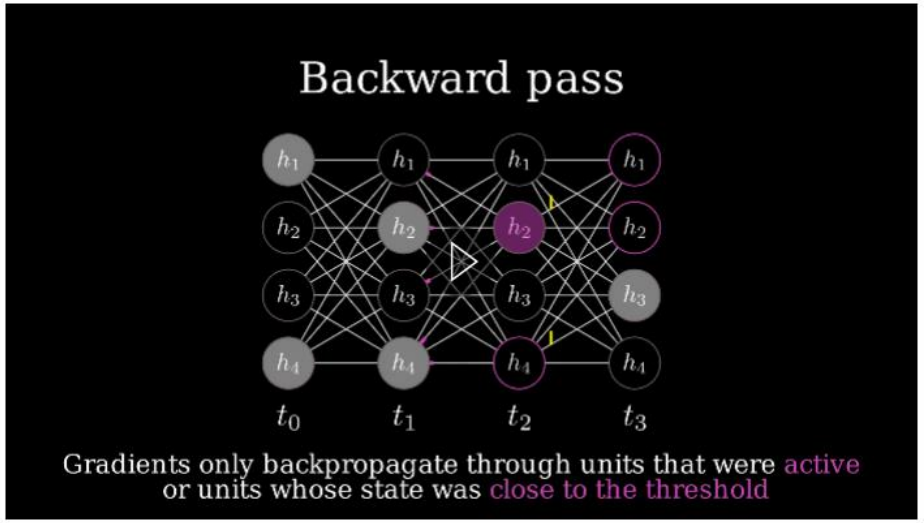
Outlook: The World Between Spiking and Deep Neural Networks

Event-Based Gated Recurrent Units (EGRU)





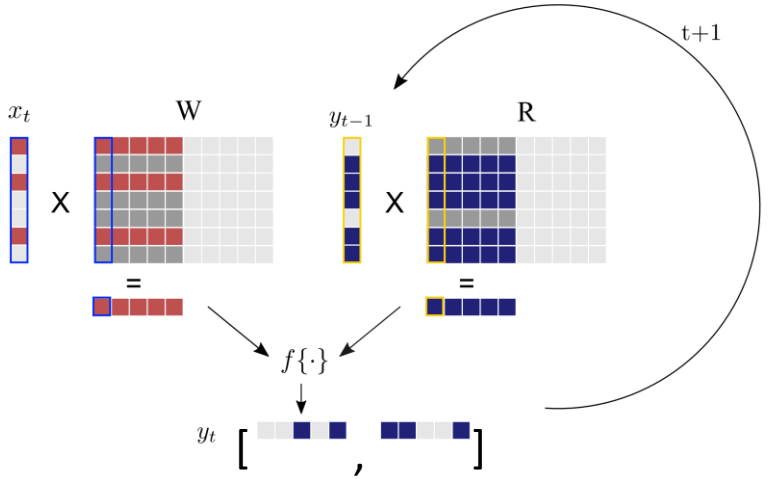
Efficient recurrent architectures through activity sparsity and sparse back-propagation through time
Subramoney et al., 2023



- Take well-functioning recurrent neural network, e.g. GRU
- Add an event-generating mechanism
- Train with surrogate gradients
- Achieves same performance as reference, but needs 80% less communication between layers
- No spikes, but “events” carrying a scalar value

Source of images: https://github.com/Efficient-Scalable-Machine-Learning/EvNN

Outlook: The World Between Spiking and Deep Neural Networks
Event-Based Gated Recurrent Units (EGRU)



Legend:

W	Input kernel
R	Recurrent kernel
x	Input
y	output
$f\{\ \}$	pointwise operation
[ ]	Broadcast and concatenate

Measurement	Nvidia A100	SpiNNaker2
Batch size	1	1
Power (W)	60	0.39
Time (mS)	19.9	170.25
Energy (J)	1.1935	0.0653
Test PPL	81.4	81.4

18x

EGRU LM on SpiNNaker2 comparison with GPU

value stored on PE

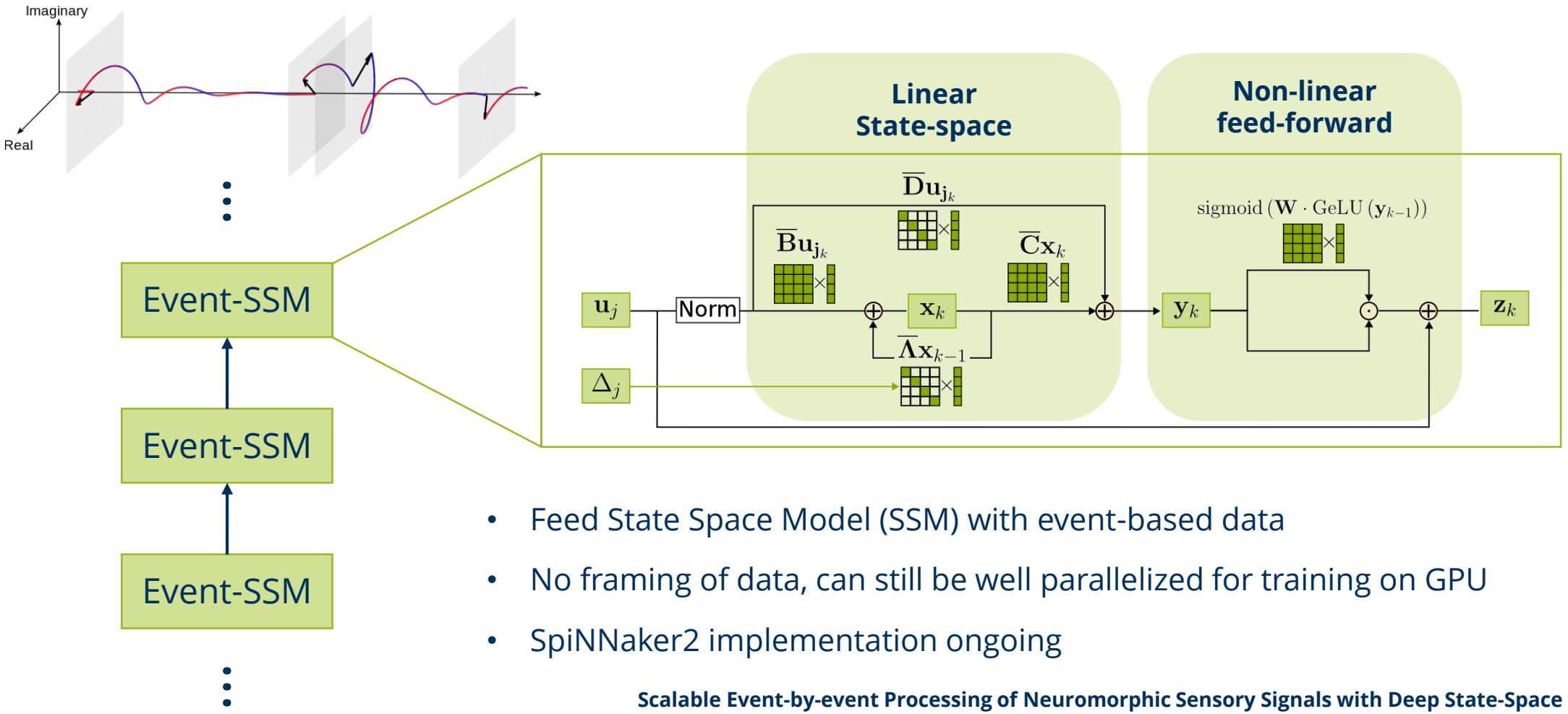
value stored but computation skipped

value not stored / zero

- Event communication in EGRU well matched to SpiNNaker2 communication fabric
- SpiNNaker2 outperforms A100 GPU for batch size 1
- GPU gets more efficient for higher batch size

Language Modeling on a SpiNNaker2 Neuromorphic Chip
Khan Nazeer et al., 2024

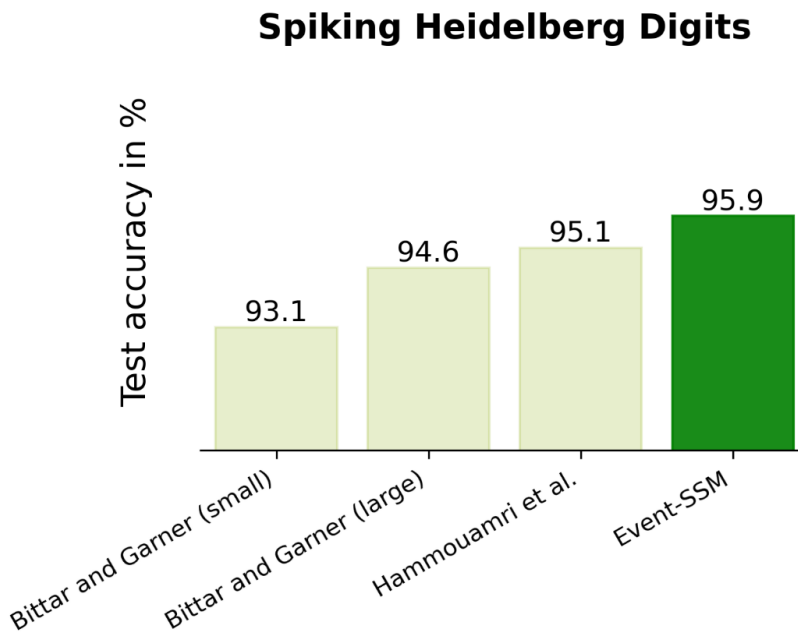
Outlook: The World Between Spiking and Deep Neural Networks
Event-Based State Space Models

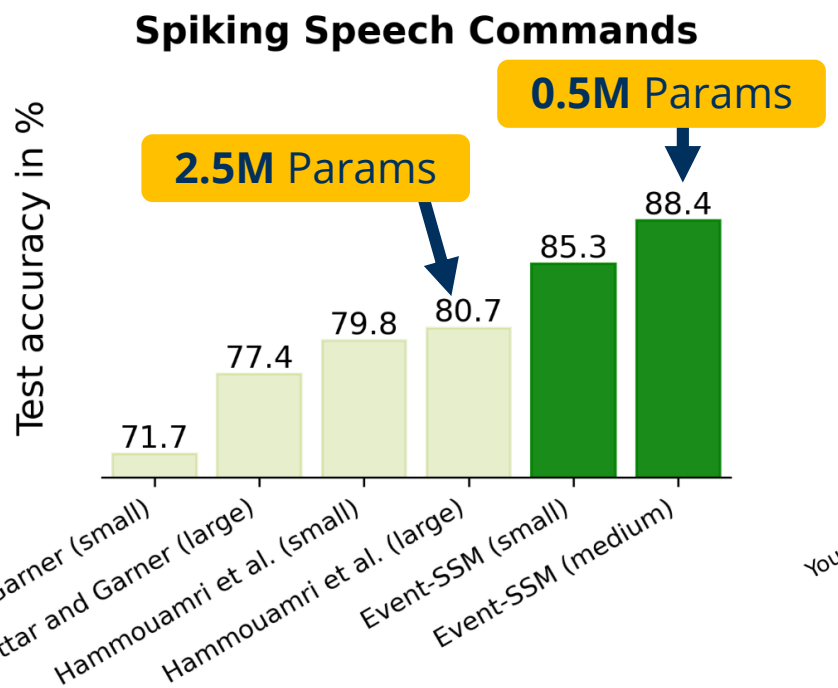


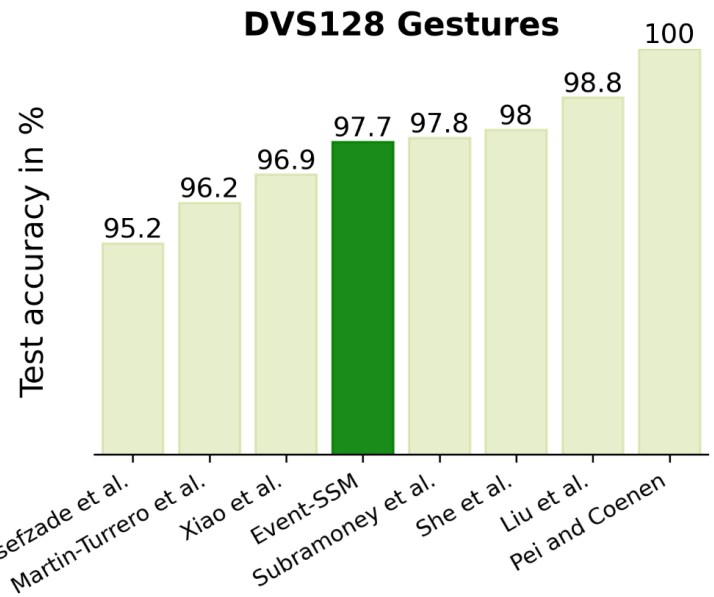
- Feed State Space Model (SSM) with event-based data
- No framing of data, can still be well parallelized for training on GPU
- SpiNNaker2 implementation ongoing

Scalable Event-by-event Processing of Neuromorphic Sensory Signals with Deep State-Space Models
Schöne et al. 2024

Outlook: The World Between Spiking and Deep Neural Networks
Event-Based State Space Models



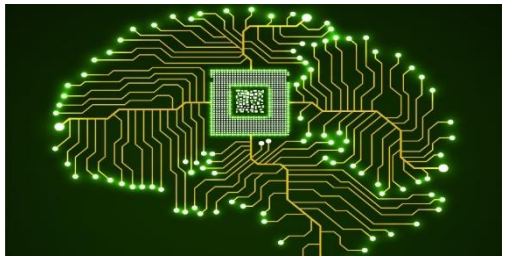




Scalable Event-by-event Processing of Neuromorphic Sensory Signals with Deep State-Space Models
Schöne et al. 2024

Conclusion

Rebuild, Learn and Apply



- Brain
- Brain Area
- Microcircuit
- Neuron
- Synapse
- Ion Channel

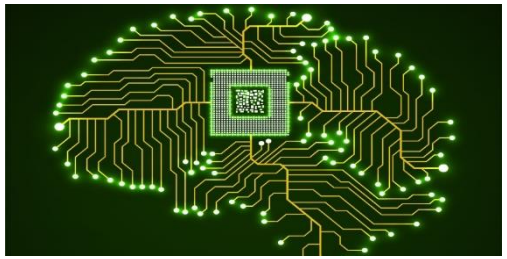
(Naive) human brain energy efficiency estimate: ~10-100 TOPS/W

- Assuming deep neural network abstraction: Synapse = multiply-accumulate operation
- 10^{14} synapses
- 1-10Hz firing rate
- 20W estimated power draw

Cutting-edge Deep Neural Network hardware is in the same order of magnitude

Conclusion

Rebuild, Learn and Apply



Inside neuromorphics…

Outside neuromorphics…

- Brain
- Brain Area
- Microcircuit
- Neuron
- Synapse
- Ion Channel

Event-based/spiking computing architectures

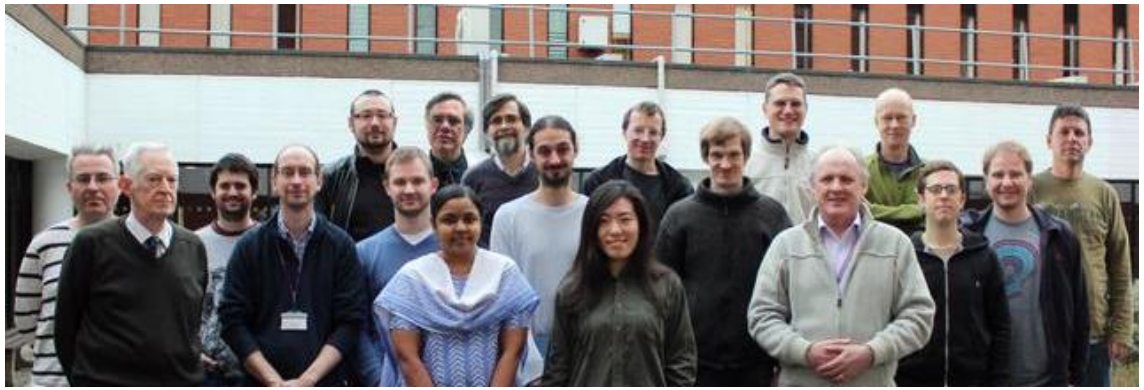
Memristive devices, in-memory computing

Sparsity in Deep Neural Networks

Industry Drive: embedded NVM, In-Memory Compute

The brains behind the brain

Team at Dresden led by Professor Dr. Habil. Christian Mayr



Team at Manchester led by Professor Steve Furber, the inventor of the ARM processor

Team at SpiNNcloud Systems GmbH is constantly growing ..