Event-Driven Neuromorphic Solutions for Efficient Dynamic Signal Processing

Dr. Federico Corradi

Assistant Professor, Neuromorphic Edge Computing Systems (NECS) Lab

Department of Electrical Engineering, Electronic Systems Group

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

October 17, 2024

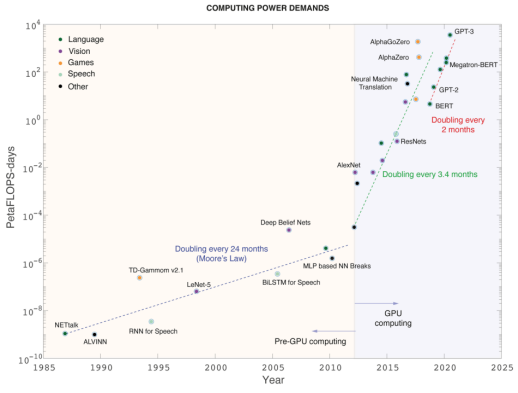
Outline

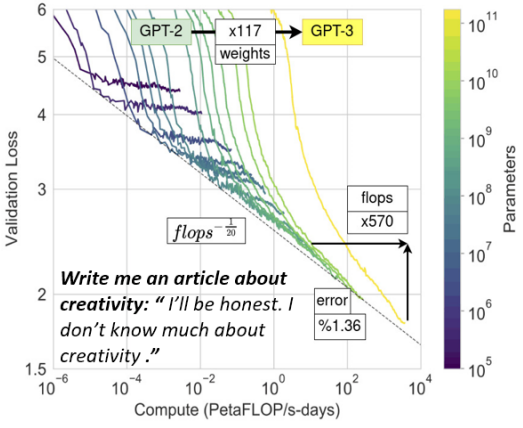
The Third Generation of Neural Nets
- Neurons Models
- Neural Coding Schemes

Training of Spiking Neural Networks
- Conversion: ANN to SNN Conversion
- Time-To-First-Spike Training of SNNs
- Direct Training: Back-propagation Through Time (BPTT)
- Direct Training: Forward Propagation Through Time (FPTT)

Neuromorphic Hardware
- Digital Neuromorphic Processors for The Edge
- Mixed-Signal Neuromorphic Processor (μBrain) for The Extreme Edge

Deep Learning Seems To Have No Limits





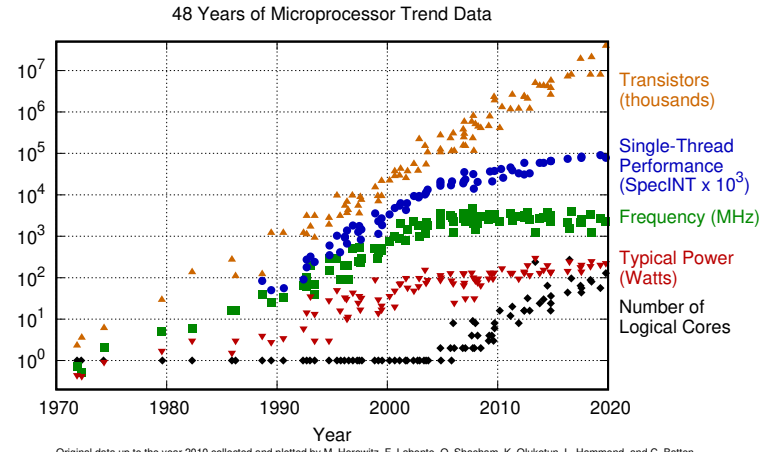
Energy Expensive (training):

- GPT-3: 1.29 GWh
- GPT4: 50 GWh $\rightarrow$ 40x

Diminishing returns in deep learning (i.e., scaling follows a power law)
[Boahen K., Nature, 2002]

TU/e

HW has its Limits: 50 Years of Microprocessors

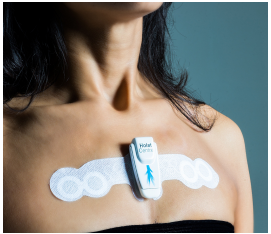


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2019 by K. Rupp

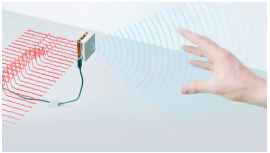
Physical limits of CMOS & Von Neumann Bottleneck

- Moore's Law
- Memory Wall
- Heat Wall

Today, Deep Learning is Not Embedded!

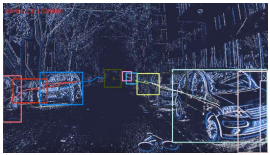


Biomedical



Mobile devices

Navigation



Fast Autonomy

Embedded Deep Learning Main Challenges

- Physical limits of CMOS & Von Neumann bottleneck
- Large volume of data, compute, and energy
- Brittle AI (fixed models)

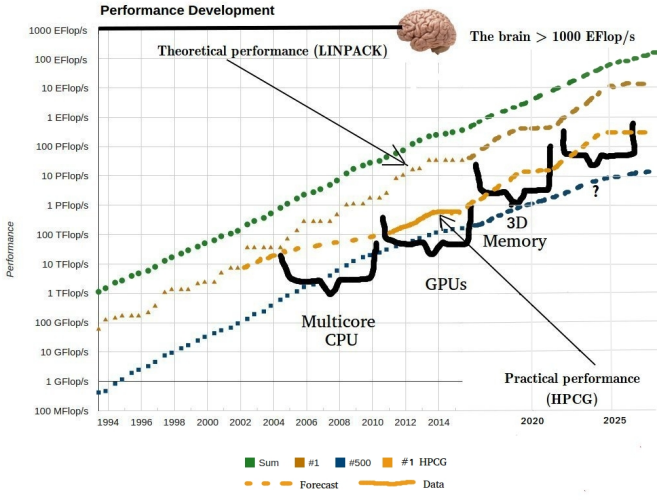
Brain, Any Limits?

Brain

- Speed = 1000 Exa Op/s*
- Power = 20 Watt
- Energy/operation = Power / Speed = $2x10^{-5}$ fJ/operation

Compare to Frontier

- Power = 22.7 MW
- Peak = 1.206 ExaFlop/s
- 8.7 Mcores, 680 m2
- Energy/op. = 19 pJ/op



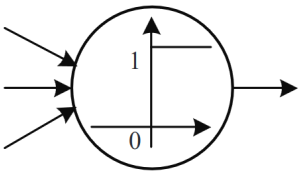
[Dettmers Tim, Blog]

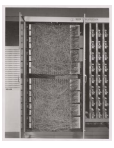
TU/e

Evolutionary Journey of Neural Networks Through Bioinspiration

First Generation

- No weights
- Threshold
- Exc/Inh inputs

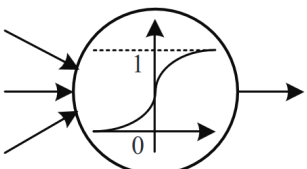




[McCulloch & Pitts, 1942]

Second Generation

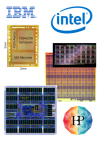
- Sigmoid / ReLU
- Real valued weights
- Differentiable



[Bridle J. 1989]

Third Generation

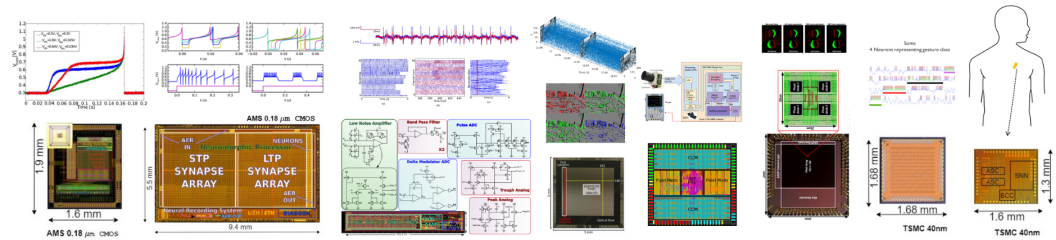
- Spiking neurons
- In/Out spike trains
- Time-based models



[Maas W., 1997] [Davies M., 2021]

TU/e

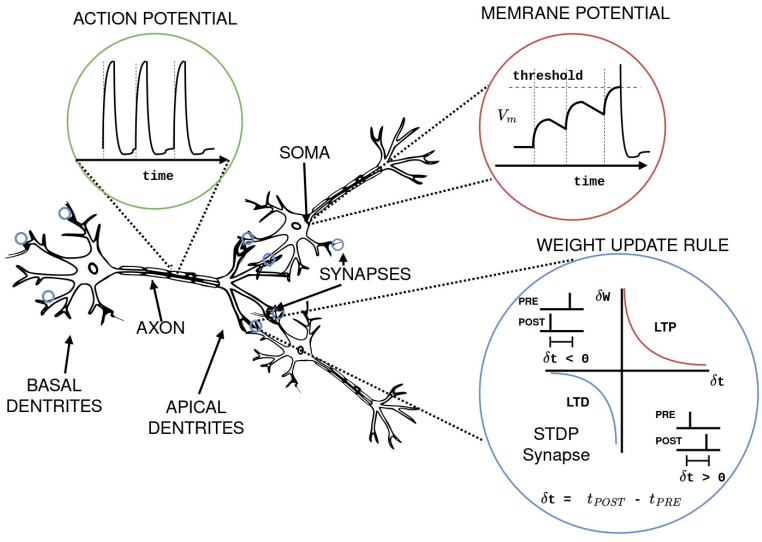
My Focus: Neuromorphic Sensing and Computing Systems



Co-design algorithms and spike-based neuromorphic sensing & computing architectures in CMOS and emerging technologies for edge applications.

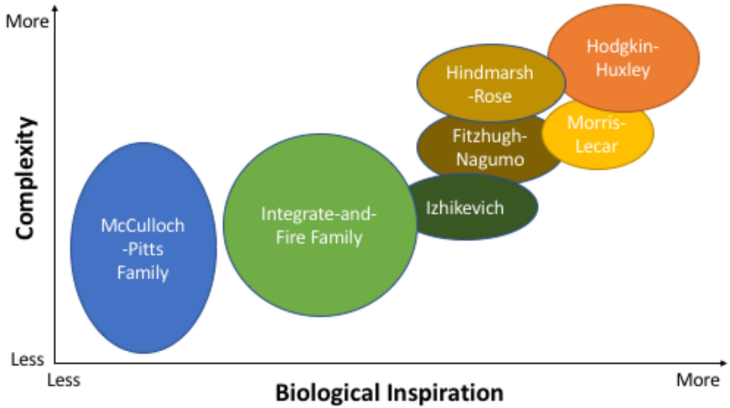
Federico Corradi | NECS Lab. | f.corradi@tue.nl

What is a Spiking Neural Network?



Federico Corradi | NECS Lab. | f.corradi@tue.nl

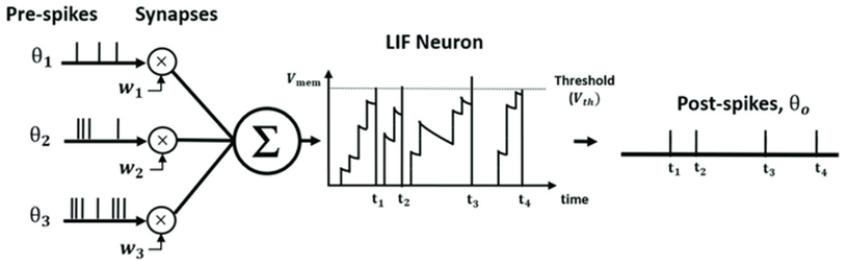
Neuron models: Complexity of Bio-realism



[Schuman et al, 2017]

TU/e

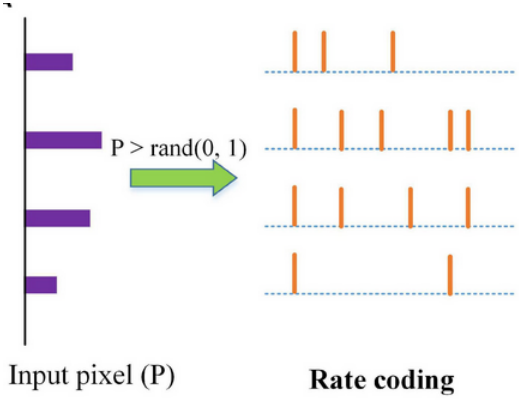
Leaky-Integrate-and-Fire (LIF) Neuron Model



LIF

- input spikes, output spikes (digital communication)
- membrane potential decay over time (stateful neuron)
- simple to implement (e.g., RC circuit)

Neural Coding Schemes

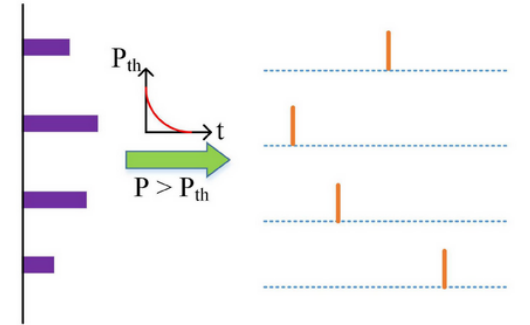


- Rate Coding, $\nu = \frac{N_{spikes}}{T}$

[Auge D. et al, 2021]

TU/e

Neural Coding Schemes

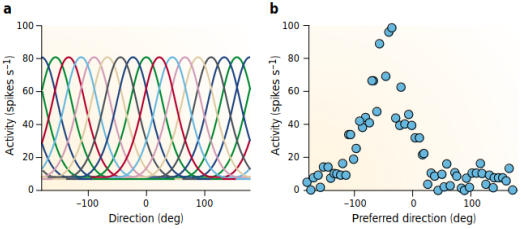


- Rate Coding, $\nu = \frac{N_{spikes}}{T}$
- Temporal Coding (TTFS, Inter-Spike-Interval)

[Auge D. et al, 2021] [Stanojevic A. et al, 2024]

TU/e

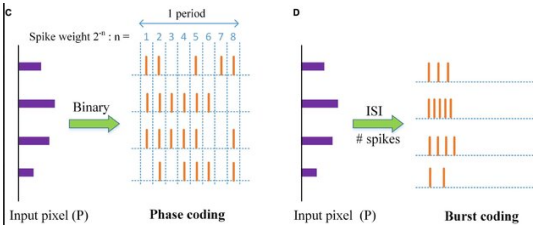
Neural Coding Schemes



- Rate Coding, $\nu = \frac{N_{spikes}}{T}$
- Temporal Coding (TTFS, Inter-Spike-Interval)
- Population Coding (neuron tuning curves)

[Auge D. et al, 2021] [Stanojevic A. et al, 2024] [Pouget et al, 2000]

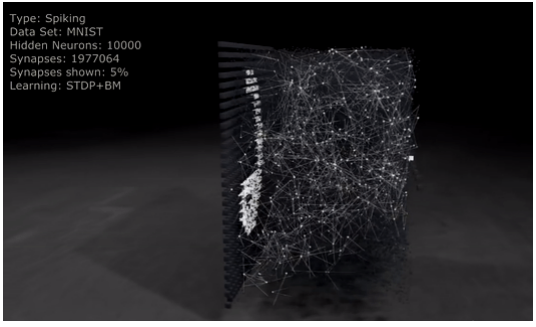
Neural Coding Schemes



- Rate Coding, $\nu = \frac{N_{spikes}}{T}$
- Temporal Coding (TTFS, Inter-Spike-Interval)
- Population Coding (neuron tuning curves)
- Sparse Coding, Phase Coding, etc...

[Auge D. et al, 2021] [Stanojevic A. et al, 2024] [Pouget et al, 2000]

Summary Part I: Spiking Neural Networks



SNN properties

- Fully parallel
- Spatiotemporal processing
- Asynchronous
- Sparse
- Additive weight operations
- Synaptic Plasticity

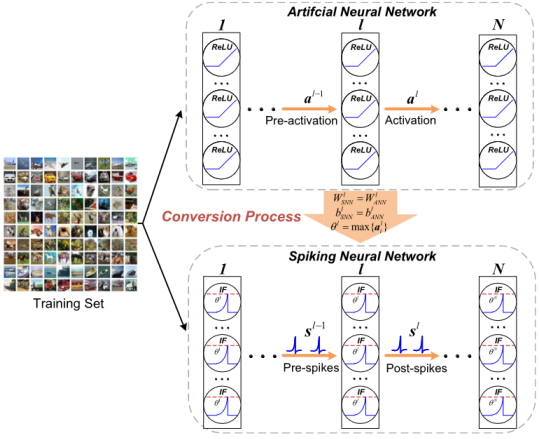
Outline

The Third Generation of Neural Nets
- Neurons Models
- Neural Coding Schemes

Training of Spiking Neural Networks
- Conversion: ANN to SNN Conversion
- Time-To-First-Spike Training of SNNs
- Direct Training: Back-propagation Through Time (BPTT)
- Direct Training: Forward Propagation Through Time (FPTT)

Neuromorphic Hardware
- Digital Neuromorphic Processors for The Edge
- Mixed-Signal Neuromorphic Processor (μBrain) for The Extreme Edge

Analog Neural Networks (ANN) to SNN Conversion



ANN to SNN

- ReLU (no bias)
- Mean-rate coding
- Train ANN using BP
- Copy weights into SNN
- Infer using SNN

[Diehl P. et al 2015] [Rueckauer B. et al. 2017] [Sengupta A. et al. 2019] [Wang Y. et al. 2024]

Analog Neural Networks (ANN) to SNN Conversion: Why Bother?

Energy efficiency comparison between ANN and SNN converted by our framework.

		$T = 2$	$T = 4$	$T = 8$	$T = 12$	$T = 16$
CIFAR10	OP_{ANN}	256.61M	256.61M	256.61M	256.61M	256.61M
	OP_{SNN}	89.07M$_{\pm 0.04M}$	177.5M$_{\pm 0.07M}$	351.89M$_{\pm 0.08M}$	526.09M$_{\pm 0.25M}$	700.41M$_{\pm 0.26M}$
	Avg Fire Rate	0.1834	0.1856	0.1857	0.186	0.1864
	Acc_{ANN}	95.67	95.67	95.67	95.67	95.67
	Acc_{SNN}	90.17$_{\pm 0.21}$	94.14$_{\pm 0.13}$	95.14$_{\pm 0.07}$	95.35$_{\pm 0.08}$	95.45$_{\pm 0.10}$
	Acc. Drop ×100%	−5.5%$_{\pm 0.21\%}$	−1.60%$_{\pm 0.13\%}$	−0.55%$_{\pm 0.07\%}$	−0.33%$_{\pm 0.08\%}$	−0.23%$_{\pm 0.10\%}$
	Energy Drop ×100%	−89.03%$_{\pm 0.01\%}$	−82.43%$_{\pm 0\%}$	−69.72%$_{\pm 0.01\%}$	−57.06%$_{\pm 0.02\%}$	−44.4%$_{\pm 0.02\%}$
CIFAR100	OP_{ANN}	256.98M	256.98M	256.98M	256.98M	256.98M
	OP_{SNN}	92.48M$_{\pm 0.02M}$	180.29M$_{\pm 0.03M}$	352.63M$_{\pm 0.09M}$	524.96M$_{\pm 0.23M}$	697.6M$_{\pm 0.12M}$
	Avg Fire Rate	0.1958	0.1908	0.1883	0.188	0.1882
	Acc_{ANN}	77.29	77.29	77.29	77.29	77.29
	Acc_{SNN}	64.89$_{\pm 0.29}$	73.58$_{\pm 0.17}$	76.35$_{\pm 0.22}$	77.13$_{\pm 0.08}$	77.22$_{\pm 0.17}$
	Acc. Drop ×100%	−12.4%$_{\pm 0.29\%}$	−3.71%$_{\pm 0.17\%}$	−1.22%$_{\pm 0.22\%}$	−0.21%$_{\pm 0.08\%}$	−0.09%$_{\pm 0.17\%}$
	Energy Drop ×100%	−87.95%$_{\pm 0\%}$	−81.21%$_{\pm 0.01\%}$	−68.76%$_{\pm 0.01\%}$	−56.41%$_{\pm 0.01\%}$	−44.06%$_{\pm 0.01\%}$

Energy Accuracy Trade-off

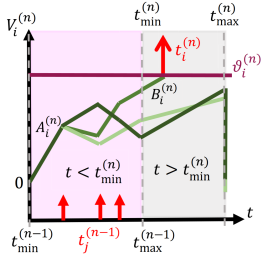
SNNs offers a good trade-off among energy and accuracy.

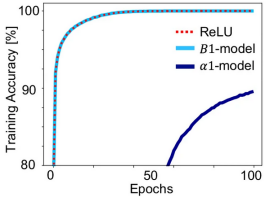
[Wang Y. et al. 2024]

TTFS Nets Have Exactly The Same Performance of ANN

Time-To-First-Spikes Nets

- SNN equivalent to feed-forward ReLu networks
- Same training behavior (equivalent gradient)
- Stateless network (frame-based)





nature communications

Article

https://doi.org/10.1038/s41467-024-51110-5

High-performance deep spiking neural networks with 0.3 spikes per neuron

Received: 20 November 2023

Accepted: 29 July 2024

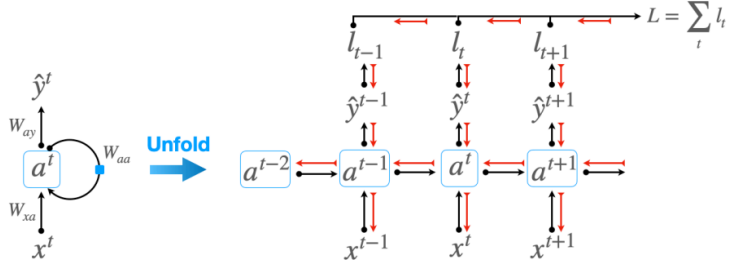
Published online: 09 August 2024

Ana Stanojevic[1,2], Stanisław Woźniak[1], Guillaume Bellec[2,3], Giovanni Cherubini[1], Angeliki Pantazi[1] & Wulfram Gerstner[2,3]

[Stanojevic et al, 2024]

TU/e

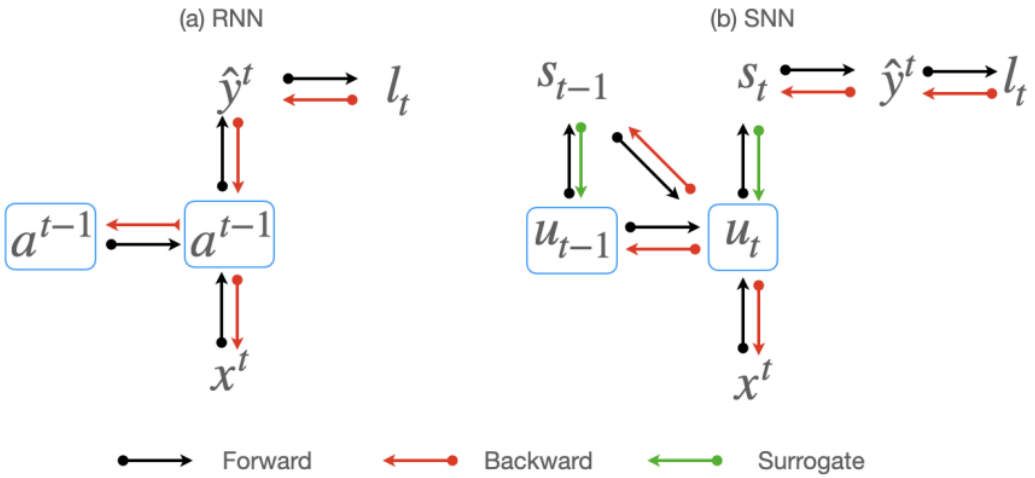
Spiking LIF Neurons: Backpropagation Through Time (BPTT)



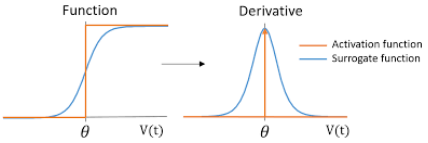
BPTT

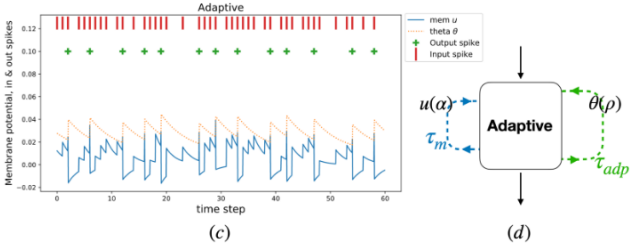
- Exploit Back-propagation
- Store the forward (black arrows)
- Unfold in time backward (red arrows)

Spiking LIF Neurons: Backpropagation Through Time (BPTT)



Spiking Recurrent Neural Networks





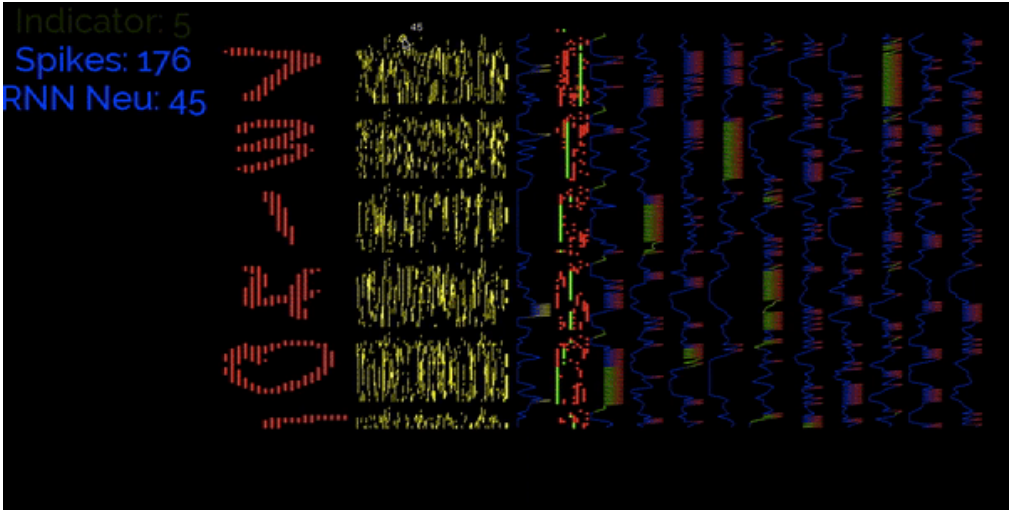
(c) (d)

Spiking Recurrent Neural Networks (SRNN)

SRNNs achieve competitive performance of classical recurrent neural networks (RNNs) while exhibiting sparse activity.

[Yin B., Corradi F., Bothe S., ICONS, 2021]

Spiking Neural Networks of LIF Neurons: SMNIST example



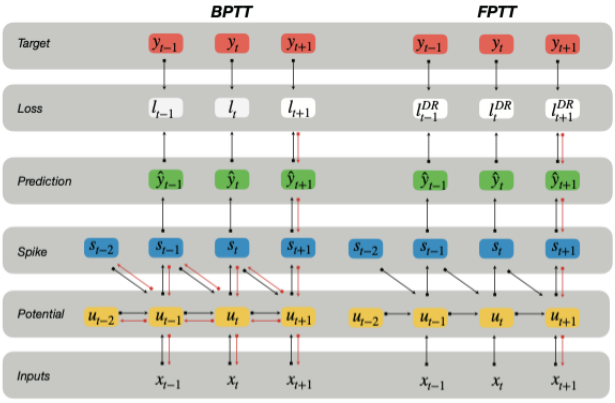
Federico Corradi | NECS Lab. | f.corradi@tue.nl
TU/e

Backpropagation Through Time (BPTT): Limitations

Algoritmic limitations of BPTT

- Cumbersome gradient computation path (via STE), and the approximation error surrogate gradient accumulates along time.
- Difficulty of back-propagating the gradient in deep spiking neural network models, vanishing gradient.
- Memory requirements for network training increases with the sequence lenght. It is challenging to train deep spiking neural network models on long sequences or when timestep become small.

Forwad Propagation Through Time (FPTT)



[Yin B., Corradi F., Bothe S., Nature Machine Intelligence, 2023]

Intuition

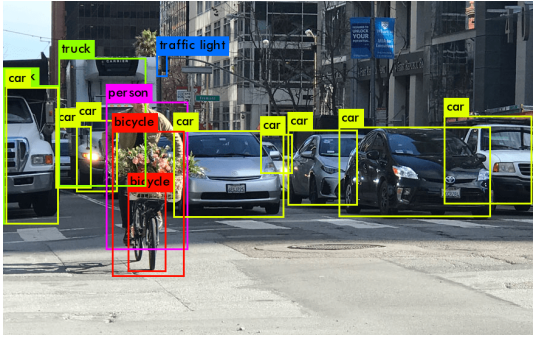
FPTT optimizes an instantaneous risk function, which includes a loss (as BPTT), and a dynamic regularization penalty (prev. observed losses).

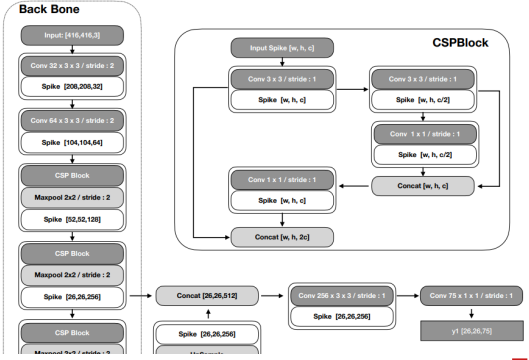
FPTT (online)

- Lessens surrogate gradient cumulative effect
- Lessens memory requirements
- Cost more computation

Deep Spiking Neural Networks Models (SpyYoloV4)

- 6.2 M spiking neurons
- Conv, fully connected, cross-stage partial subnets
- 14 M parameters
- Approaching DNNs



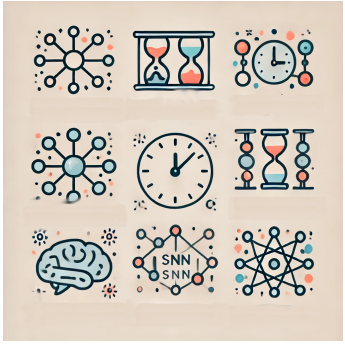


TU/e

Deep Spiking Neural Networks Models (SpyYoloV4)

[Yin B., Corradi F., Bothe S., Nature Machine Intelligence, 2023]

Summary part II



Training competitive and deep SNN today is possible!
Interesting properties for HW:

- Sparse communication
- Event-driven computation
- Temporal information processing
- Trade-off energy and accuracy
- Same accuracy as ANN!

Outline

The Third Generation of Neural Nets
- Neurons Models
- Neural Coding Schemes

Training of Spiking Neural Networks
- Conversion: ANN to SNN Conversion
- Time-To-First-Spike Training of SNNs
- Direct Training: Back-propagation Through Time (BPTT)
- Direct Training: Forward Propagation Through Time (FPTT)

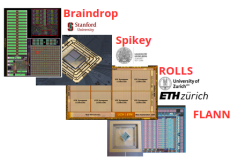
Neuromorphic Hardware
- Digital Neuromorphic Processors for The Edge
- Mixed-Signal Neuromorphic Processor (μBrain) for The Extreme Edge

Neuromorphic Hardware

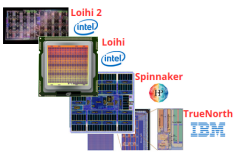
Neuromorphic Computing Systems

- Digital Neuromorphic Hardware
  - Software models.
  - Easy to design, scalable.
- Mixed-Signal Neuromorphic Hardware
  - Bio-realistic models.
  - Hard to design, noise, mismatch.
- Emerging Technologies
  - Non-volatile memories.
  - Spintronics.
  - New design styles (e.g., in-memory computing).

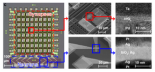
Analog neuromorphic architectures

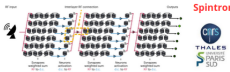


Digital neuromorphic architectures



Emerging neuromorphic architectures



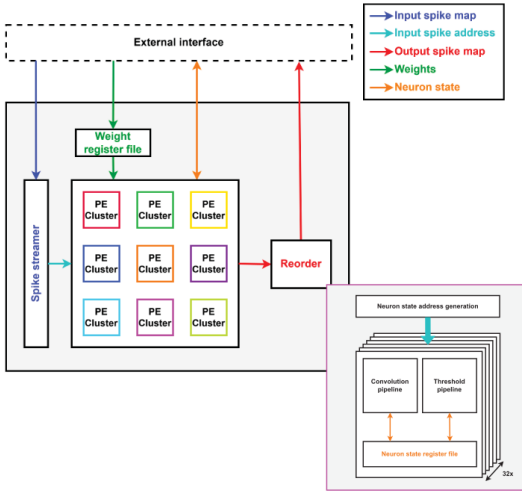


Mega: Digital SNN Accelerator for SpyYoloV4 (TU/e)

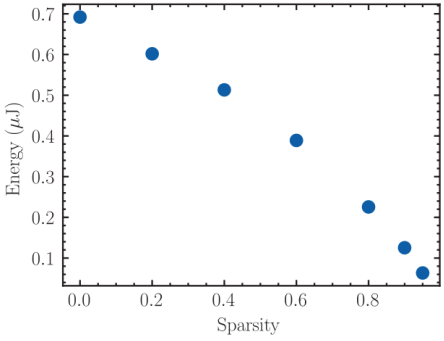
- 3x3 spiking convolution
  - Event-based: operations happen when there is a spike
  - Each PE cluster updates 32 neurons in parallel
- Memory hierarchy
  - Small, fast memories near the PEs
  - Larger memory for flexibility
- 0.501 pJ/SOP
  - Post-synthesis in 22 nm
  - 625 MHz @ 0.8V

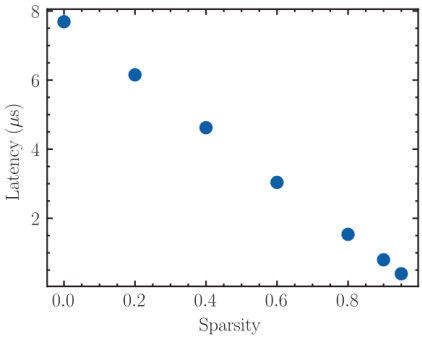
CONVOLVE

[Luiken R., unpublished]



Mega: What Happens if Spiking is Sparse?





- Energy and latency for a 96x48 spike map.
- SpyYoloV4 application.

[Luiken R., unpublished]

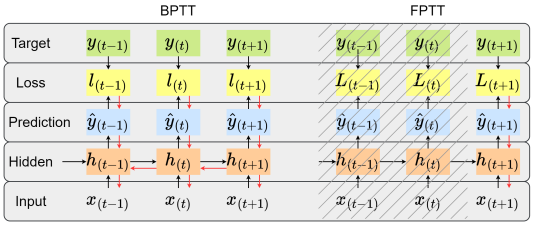
Recurrent Neural Network Training at the Edge

- traditionally in one go: update the weight **once** by all network states generated after processing the entire sequence

$$W_{new} = W_{old} - \eta \frac{\partial \sum_{t=1}^{T} l_{(t)}}{\partial W_{old}} \qquad (1)$$

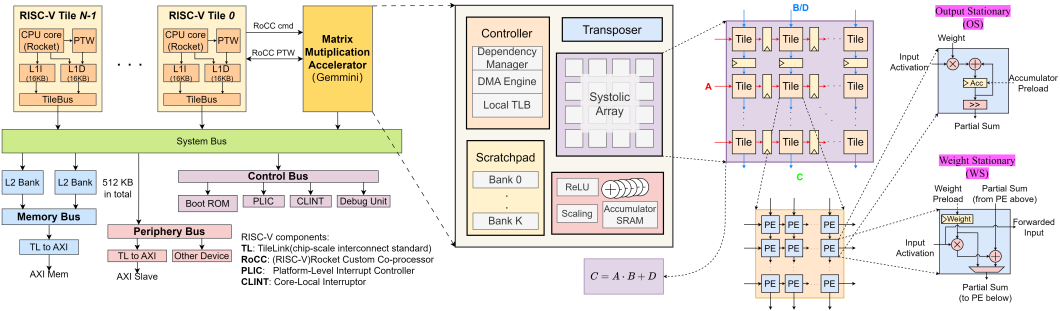
- Forward Propagation Through Time: update by the current network state **after every time step**; application of regularization term $R_{(t)}$ for stabilization

$$W_{(t+1)} = W_{(t)} - \eta \frac{\partial (l_{(t+1)} + R_{(t+1)})}{\partial W_{(t)}} \qquad (2)$$



- How to efficiently accelerate FPTT computations by designing an edge hardware architecture?
- Explore several level of partitioning for the sequence.
- Sequential MNIST Divide it into K parts: 1,2,7,14,28,56

Customized Embedded Platform for FPTT Training



Architecture

- RISC-V & Systolic Arrays
- Brain Float 16 (BF16), low memory req.
- Weight Stationary only

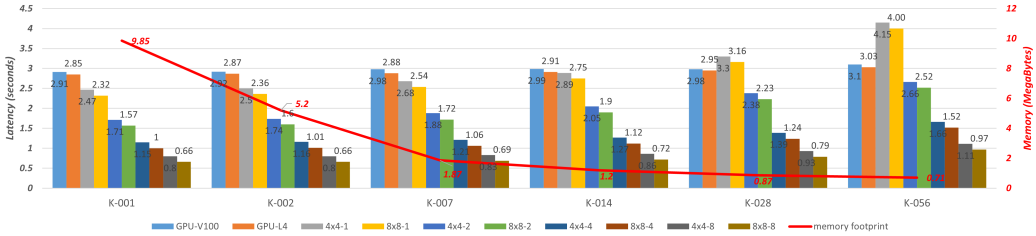
Explore System Scaling

- 2 mesh sizes of the systolic array in Gemmini: 4x4, 8x8
- 4 types of CPU cores: 1,2,4,8

[Zhang Y. Gomony D.M., Corporral H., Corradi F., VLSISOC, 2024]

TU/e

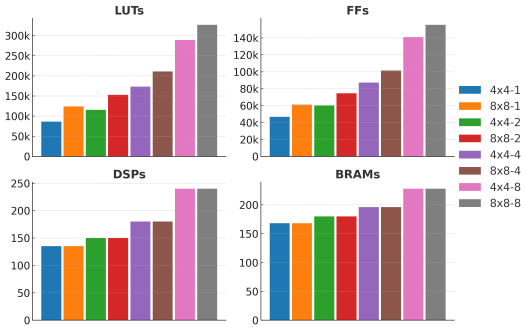
Results: Training at the Edge (FPTT) Architecture

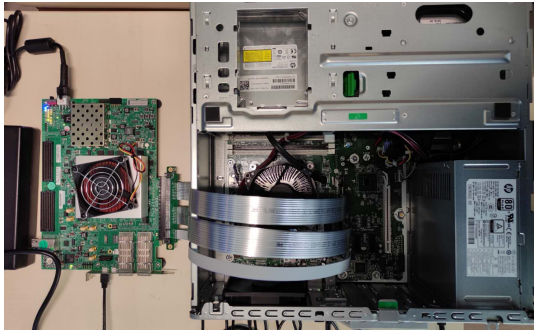


- For most clusters of K, customized architectures outperform two GPU platforms in latency.
- Trade-off between latency and resource utilization, given any cluster of K.
- Degree of partition: for the cluster, a benchmark of larger K, i.e., finer partition over a sequence leads to a slight latency increase but significant memory saving.

[Zhang Y. Gomony D.M., Corporral H., Corradi F., VLSISOC, 2024]

Results: Training at the Edge (FPTT) Architecture





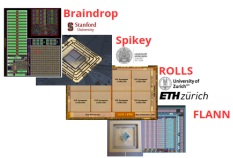
Cycle-accurate hardware simulation in FPGA.
Latency and Memory Savings when exploiting FPTT with a Compressed BF16 architecture.
[Zhang Y. Gomony D.M., Corporral H., Corradi F., VLSISOC, 2024]

Neuromorphic Hardware

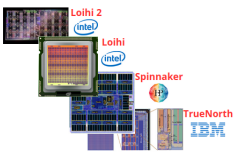
Neuromorphic Computing Systems

- Digital Neuromorphic Hardware
  - Software models.
  - Easy to design, scalable.
- Mixed-Signal Neuromorphic Hardware
  - Bio-realistic models.
  - Hard to design, noise, mismatch.
- Emerging Technologies
  - Non-volatile memories.
  - Spintronics.
  - New design styles (e.g., in-memory computing).

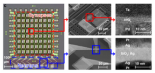
Analog neuromorphic architectures

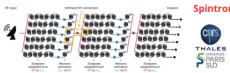


Digital neuromorphic architectures



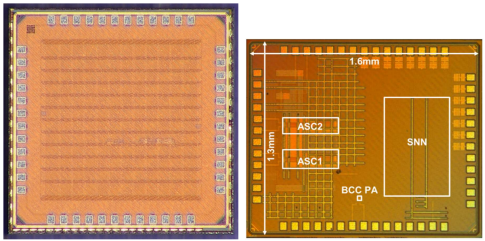
Emerging neuromorphic architectures





Neuromorphic Computing Hardware for the Extreme Edge

Synthesizable mixed-signal

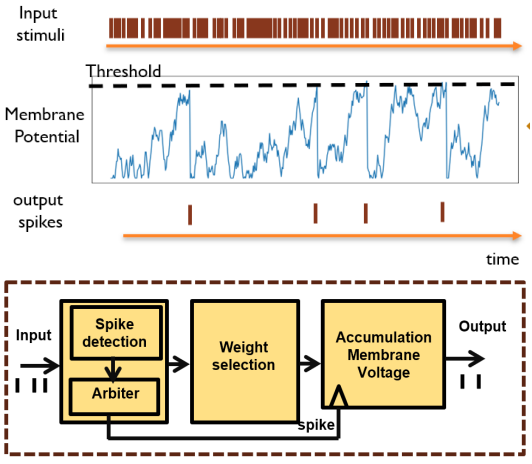


[Stuijt J. ... Corradi F., Front. Neurosc., 2021] [He Y., JSSC, 2022]

Synthesizable mixed-signal:

- Benefits:
  - Fully asynchronous
  - Fast development-to-deployment cycle
  - Ultra-low-power
- Downsides:
  - Larger area
  - Difficult to train

μBrain: Fully Synthesizable Mixed-signal Neuromorphic Processor

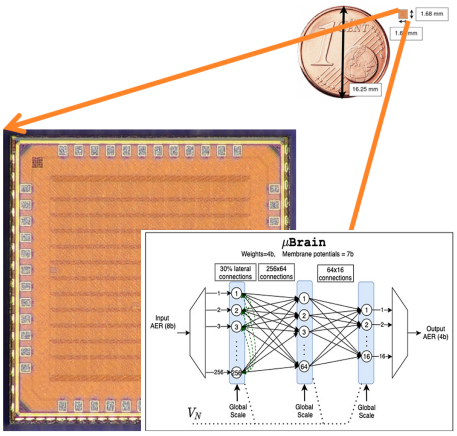


LIF

- Self-timed (asynchronous)
- Programmable synapses
- Scalable
- Assumption: arbitration delays are in the order of *ns*, while incoming spikes are spaced in μs or even *ms*

[Stuijt J., ... Corradi F., Front. Neurosc., 2021]

μBrain: Fully Synthesizable Mixed-signal Neuromorphic Processor

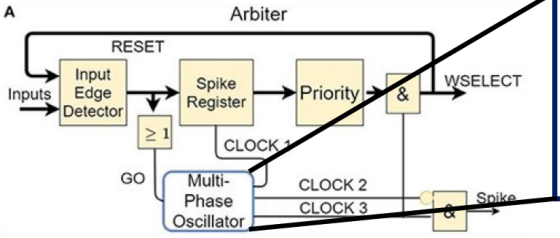


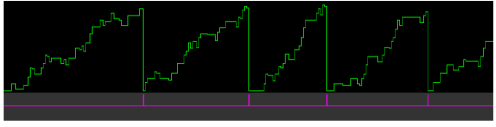
μBrain

- $1.68mm^2$ (60% syn, 4% neu, 36% arb)
- 300 neu, 200000 syn
- $\leq 100\ \mu W$
- Programmable, Fully parallel, Asynchronous
- High-fidelity (SW vs. HW)

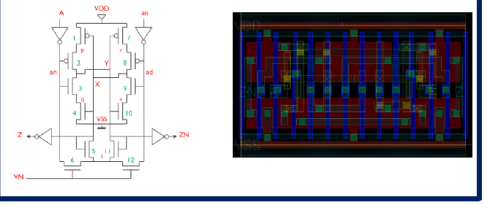
TU/e

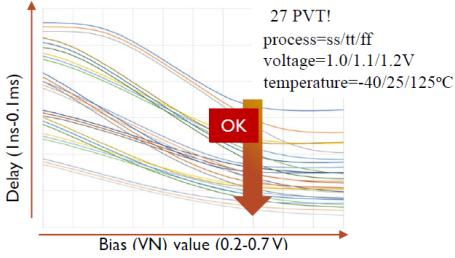
μBrain: Self-timed Neurons





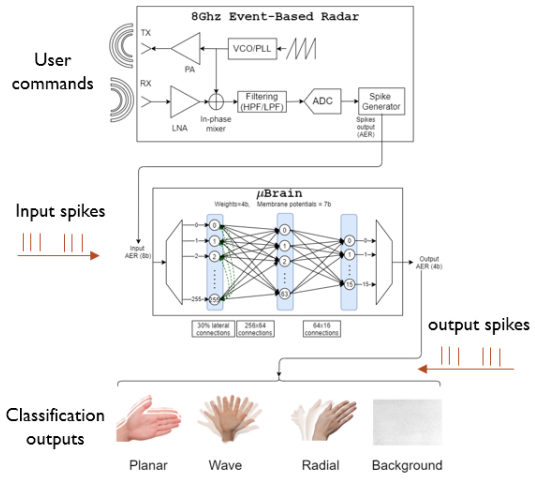
Neurons delay cell (mixed signal design)

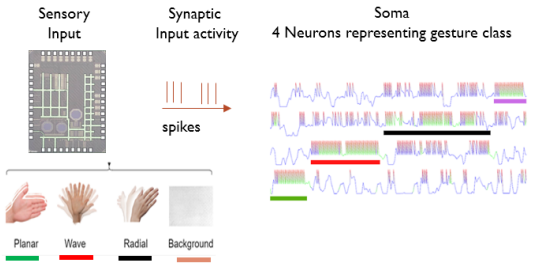




[Stuijt J., ..., Corradi F., Front. Neurosc., 2021]

μBrain: Radar Signal Processing





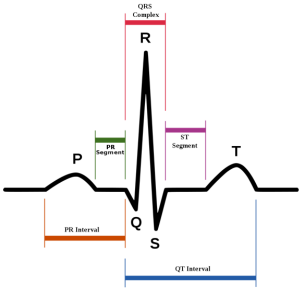
Performance

- Always-on ($< 100\mu$W compute)
- $> 93\%$ accuracy
- High-fidelity SW vs HW
- Near- and in-sensor computing

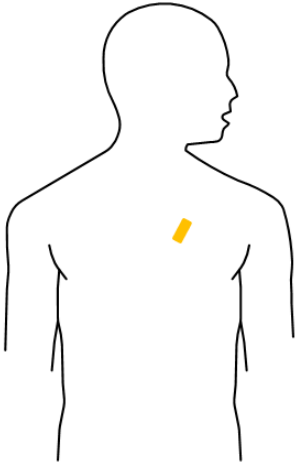
[Stuijt J., ... , Corradi F., Front. Neurosc., 2021]

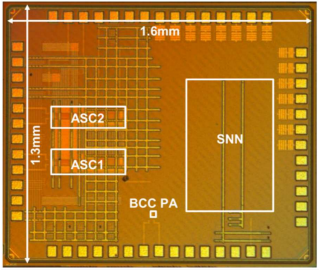
TU/e

μBrain: Implantable Biomedical Devices



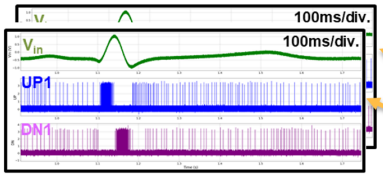
Always-on monitoring
Temporal feature extraction
for anomaly detection

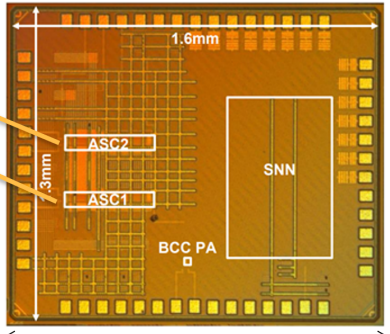




Federico Corradi | NECS Lab. | f.corradi@tue.nl

μBrain: Implantable Biomedical Devices



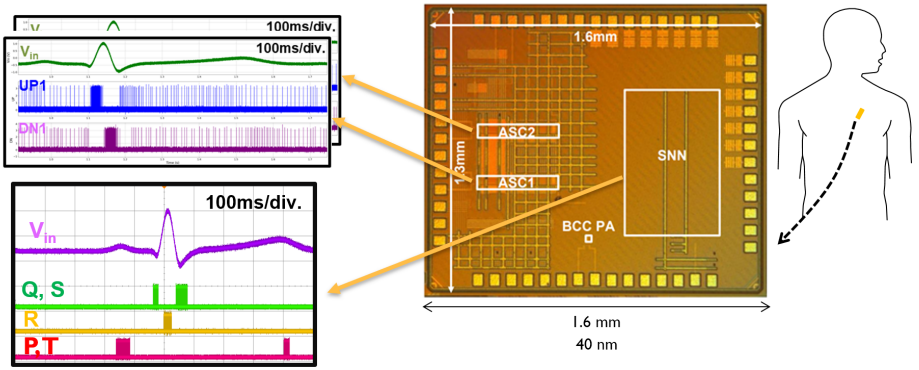


- Analog to Spike Converter: bio-inspired!
  - Lower data rate (compression)
    - Active on demand

Re-synthesized μBrain [He Y. et al, JSSC, 2022]

- 100 neurons, 7000 syn, $\leq$ 50 μW
- Analog-to-spike converters (asynchronous)

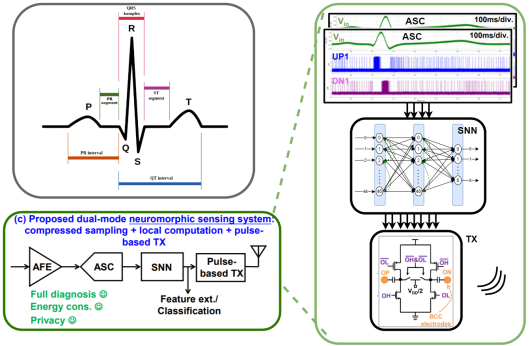
μBrain: Implantable Biomedical Devices



Re-synthesized μBrain [He Y. et al, JSSC, 2022]

- Temporal feature extraction
- With an accuracy of less than 1 millisecond!

μBrain: Implantable Biomedical Devices



	This work		[2] Luo, TBIOCAS'19	[3] Kim, TBIOCAS'14		Reveal LINQ [10]
Tech. (nm)	40		130	180		N.A.
Supply (V)	0.9/1/1.1		0.9	1.2		N.A.
System	ADC+SNN+TX		AFE+ADC+ DSP+TX	AFE+ADC+DSP+TX		N.A.
Applications	**Insertable**		Wearable	Wearable		Insertable
	Daily monit.	Diagnostic		Daily monit.	Diagnostic	
ADC architecture	LC		SAR	SAR		N.A.
ADC sample rate × nr. of bits (bps)	~ 140 (event-driven) × 2b =280		256 × 12b =3072	512/64 (Adaptive) × 12b =877 (7× compression)		256× 16b =4096
ADC ENOB (bits)	5	9.6	10.32	N.A.		N.A.
TX technology	Pulse based BCC		QPSK/BFSK	BLE		Bluetooth
Power con. (µW)						
TX	**1.2**	**1.5**	69.4~317.4[B]	1000	13302	-
DSP/NN/Memory	13	15	-	42	0	-
ADC	14	34	4.6	18	38	-
AFE	-	-	4.6	21	21	-
Total (excl. AFE)	**28.2**	**50.5**	74~322	1060	13340	N.A.[C]
Nr. of ECG chan.	2		2	3		N.A.
Core area (mm^2)	**0.32**[A]		6.25	23.5		N.A.
System area (mm^2)	**72**		N.A.	~500		322[D]
On-chip labeling	**PQRST**		No	R		R
Label precision (ms)	**<1**		>4[E]	>2[E]		>4[E]

[A] No AFE. [B] 0.38% TX duty cycling. [C] 3-yr battery life. [D] Include ECG electrodes. [E] Limited by sampling rate.

Accurate and efficient spike-based electrocardiogram monitoring.
[He Y. et al, JSCC 2022]

Neuromorphic Edge Computing Systems: Summary

Summary

- Diverse solutions for different requirements (Digital, Mixed-Signals, In-Memory Computing)
  - Trade-off: Flexibility, Energy, Memory
- Asynchronous tiny SNNs chips for IoT applications achieve competitive performance!
  - Always-on (on-demand), ULP ($\leq$ 100 μW)
  - Consume data locally, ASIC for specific tasks

Neuromorphic Edge Computing Systems: Summary

Summary

- Diverse solutions for different requirements (Digital, Mixed-Signals, In-Memory Computing)
  - Trade-off: Flexibility, Energy, Memory
- Asynchronous tiny SNNs chips for IoT applications achieve competitive performance!
  - Always-on (on-demand), ULP ($\leq$ 100 μW)
  - Consume data locally, ASIC for specific tasks

What's next?

- Scaling to large SNN models.
- Heterogenous computing architectures.
- In-memory computing with emerging technologies.

Neuromorphic Edge Computing Systems: Summary

Summary

- Diverse solutions for different requirements (Digital, Mixed-Signals, In-Memory Computing)
  - Trade-off: Flexibility, Energy, Memory
- Asynchronous tiny SNNs chips for IoT applications achieve competitive performance!
  - Always-on (on-demand), ULP ($\leq 100\ \mu W$)
  - Consume data locally, ASIC for specific tasks

What's next?

- Scaling to large SNN models.
- Heterogenous computing architectures.
- In-memory computing with emerging technologies.

Collaborators

- Sander Bothe (CWI)
- Yao-Hong Liu (IMEC)
- Manil Dev Gomony (TU/e)

Students

- Yuming He - PhD (IMEC)
- Rick Luiken - PhD (TU/e)
- Bojian Yin - PostDoc (TU/e)

CWI NWO imec CONVOLVE TU/e

Interested? Reach out!

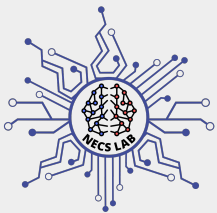
Neuromorphic Edge Computing Systems Lab

Federico Corradi

Eindhoven University of Technology

f.corradi@tue.nl, +31(0)402 472 556

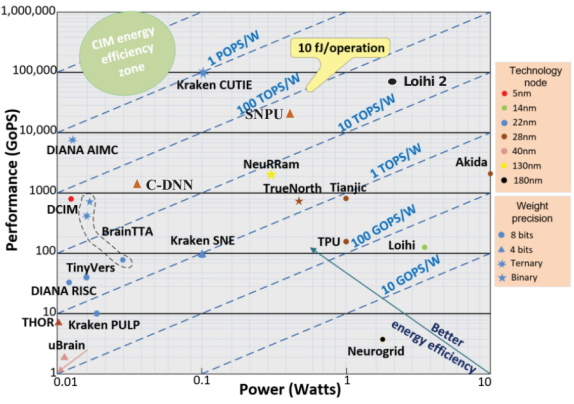
Neuromorphic Edge Computing Systems Lab



SOTA in Edge-AI Processor/Accelerator HW (ANNs and SNNs)

SOTA

- Mostly Artificial Neural Nets (ANN)
- Some Spiking Neural Nets (SNN)
- Weight Precision (8/4/Ternary/1)
- Technology Nodes (5-180nm)

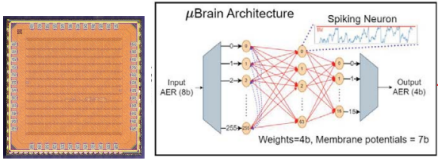


[Gomony D. M. et al, 2023]

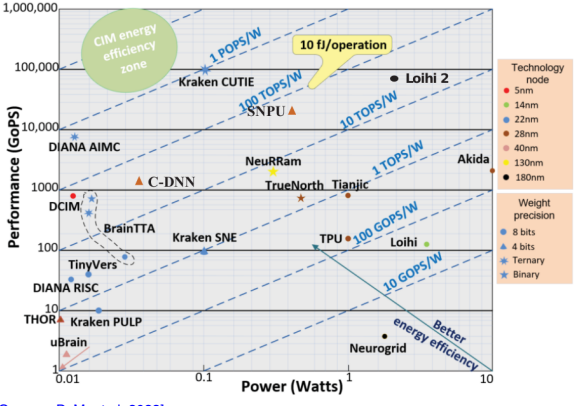
μBrain: SNN

μBrain (IMEC)

- 40 nm, 1.4mm^2
- Event based architecture
- Asynchronous design (no clock)
- Without schedules, clocks, state machines
- Extreme low power, not en-eff.



[Stuijt J. et al, Front. Neurosc., 2021]



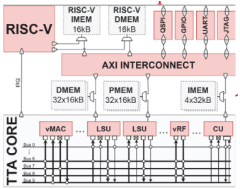
[Gomony D. M. et al, 2023]

TU/e

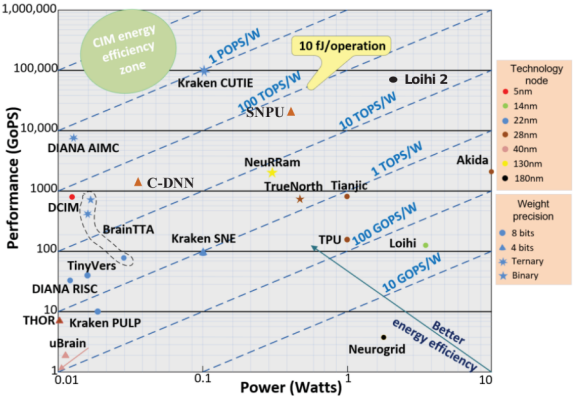
BrainTTA - Flexible & Mixed-precision (TU/e)

BrainTTA (TU/e)

- TTA-based accelerator, 22nm
- Fully-programmable (C-compiler)
- SNPU
- Flexible precision:
  - INT8, ternary, binary
  - 405 / 67 / 35 fJ/op



[Molendijk M. et al, 2023]



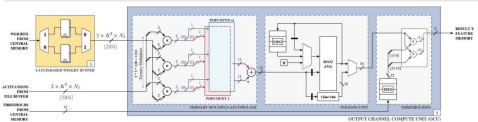
[Gomony D. M. et al, 2023]

TU/e

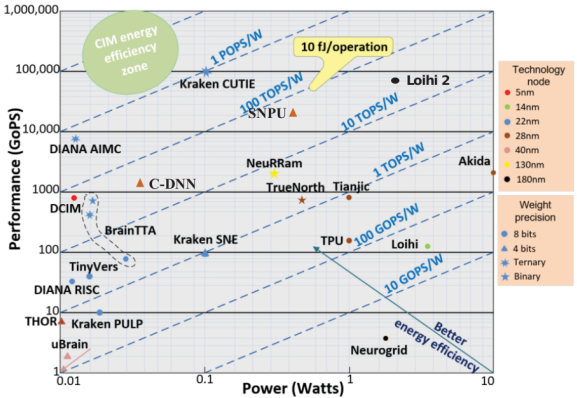
CUTIE: Ternary DNN (ETH)

CUTIE (ETH)

- ANN: Ternary (-1,01) Inf.
  - exploit zero surpression
  - ternary compression
- 1.6 bit / symbol
- complete unrolled loops in HW
- 3x3*128 = 1152 conv. size
- 22 nm, 2.5-4 fJ/op (%Sparsity dependent)



[Scherer M, et al., 2021



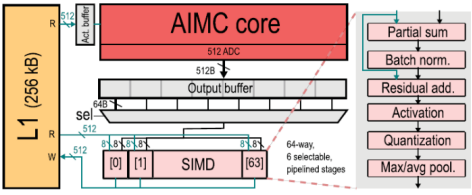
[Gomony D. M. et al, 2023]

TU/e

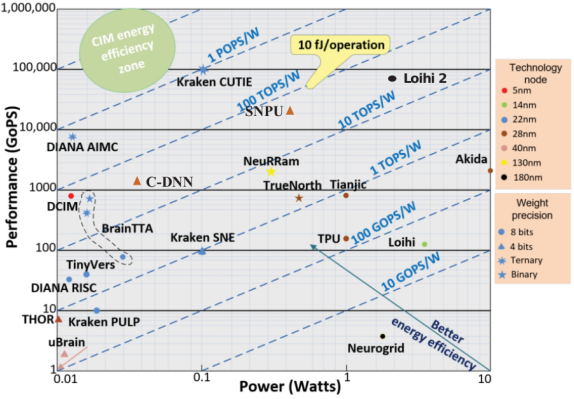
DIANA: Mixed-Signal In-memory Compute (IMEC)

DIANA (IMEC)

- Precision-scalable digital core
- Analog In-Memory Core (AIMC)
  - INT 8/4/2
  - Ternary weights, 7-b activation
  - Programmable SIMD
  - 22nm 1.7 fJ/Op (I/W/O= 7/1.5/6-bit)



[Houshmand P. et al. , 2022]



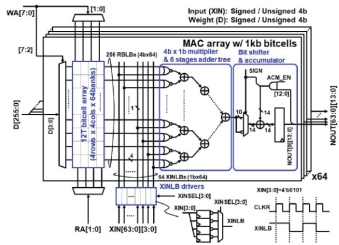
[Gomony D. M. et al, 2023]

TU/e

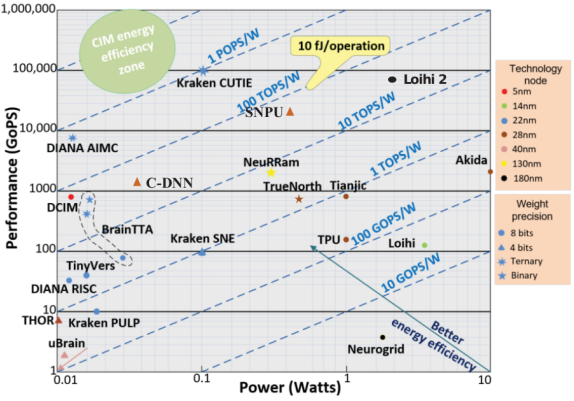
Digital CIM - SRAM-based

Digital (CIM)

- 12T bitcell based architecture
- Flexible precision
- INT 8/4, 5nm, 15.8 / 3.9 fJ/Op
- MAC using 1-b at a time



[Fujiwara H., et al, 2022]



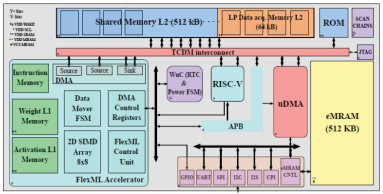
[Gomony D. M. et al, 2023]

TU/e

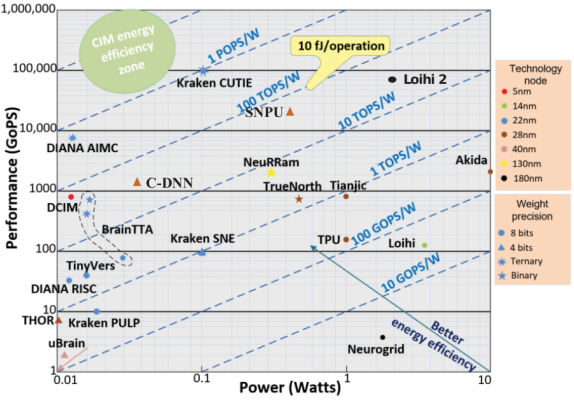
TinyVers - Embedding MRAM

TinyVers

- Supports various DNN layers, to traditional ML models like SVM
- RISC-V + ML Acc.
- Flexible-precision scalable digital accelerator, max 17 TOPS/W ∼59 fJ/Op (Int2)



[Vikram J et al, 2022]



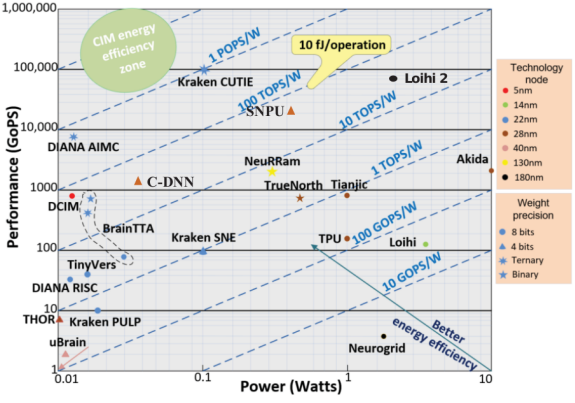
[Gomony D. M. et al, 2023]

TU/e

SOTA in Edge-AI Processor/Accelerator HW: Summary

SOTA

- ANN chips approaching 1 fJ/Op
  - However:
    - Requires complete unrolling (CUTIE) and/or AIMC (DIANA)
    - System overhead often neglected
- Flexibility has its price:
  - E.g. BrainTTA suffers at least one-order in energy efficiency
  - DRAM overhead not included
- SNNs have high potential
  - However:
    - at best in the 35 fJ/Op range
    - need to catch-up



[Gomony D. M. et al, 2023]

TU/e

Interested? Reach out!

Collaborators

- Henk Corporaal (TU/e)
- Manil Dev Gomony (TU/e)

CONVOLVE NWO imec

Neuromorphic Edge Computing Systems Lab

dr. Federico Corradi

Eindhoven University of Technology

f.corradi@tue.nl, +31(0)402 472 556

Neuromorphic Edge Computing Systems Lab