Spiking Neural Networks *on FPGA* for Edge AI Applications

Dr. Federico Corradi

Assistant Professor, Neuromorphic Edge Computing Systems (NECS) Lab

Department of Electrical Engineering, Electronic Systems Group

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

December 12, 2023

Outline

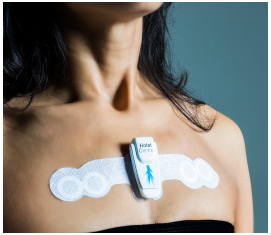
Challenges in the Future of Computing and Edge AI

My Research Focus: Neuromorphic Computing & Engineering

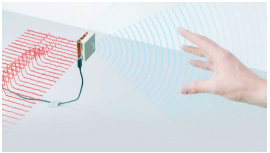
Spiking Neural Networks as *models* of computation

Spiking neural networks in FPGA

Edge artificial intelligence applications

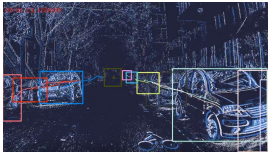


Biomedical



Mobile devices

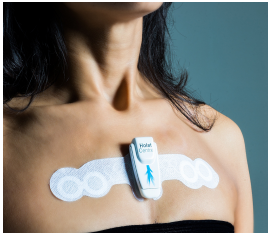
Navigation



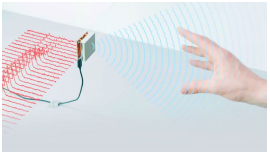
Fast Autonomy

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

Edge artificial intelligence applications

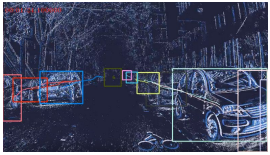


Biomedical



Mobile devices

Navigation

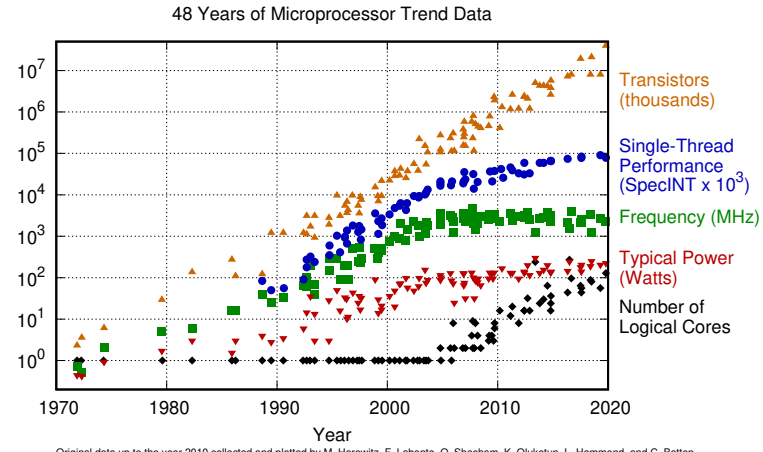


Fast Autonomy

Today, Deep Learning is Not Embedded!

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

Challenges in the future of computing and edge AI



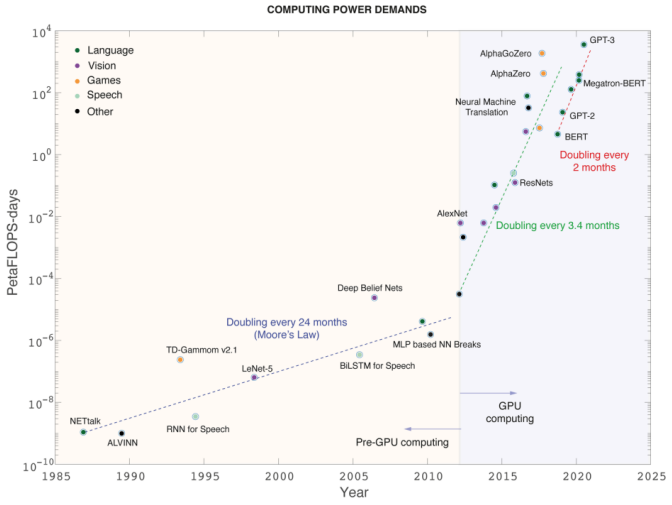
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2019 by K. Rupp

Physical limits of CMOS & Von Neumann Bottleneck

- Moore's Law
- Memory Wall
- Heat Wall

TU/e

Challenges in the future of computing and edge AI



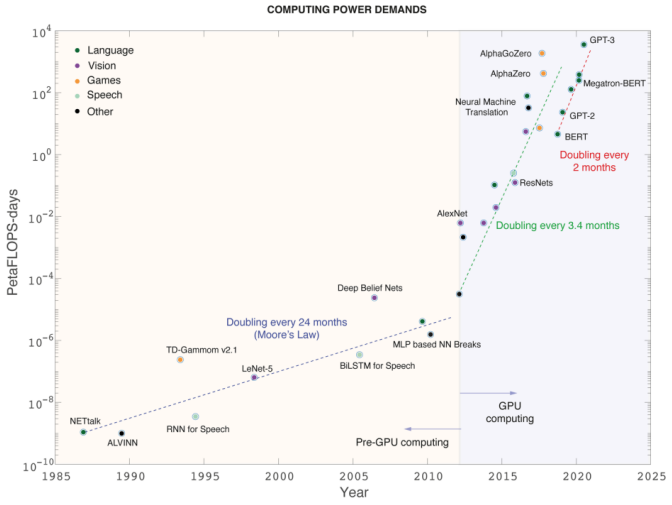
[Mehonic and Kenyon 2022)], [AKCP], [Boahen K. 2022]

Large Volume of Data, Compute, and Energy

- Training DNNs has extremely high demands in terms of power consumption.

TU/e

Challenges in the future of computing and edge AI



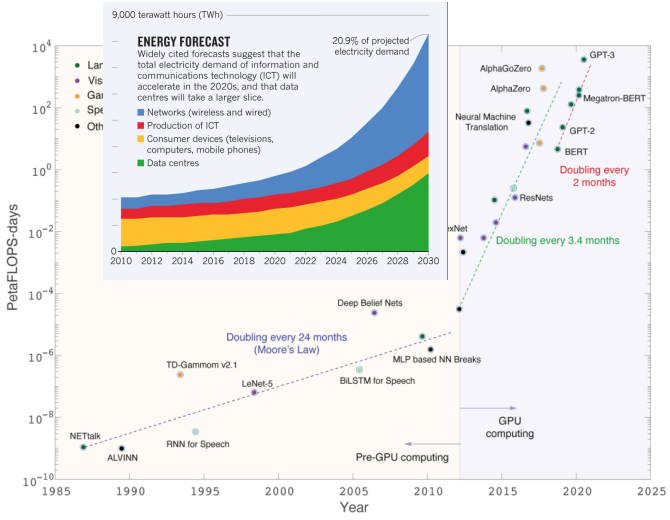
[Mehonic and Kenyon 2022)], [AKCP], [Boahen K. 2022]

Large Volume of Data, Compute, and Energy

- Training DNNs has extremely high demands in terms of power consumption.
- According to conservative estimates, training chat GPT-4 over $63 million.

TU/e

Challenges in the future of computing and edge AI



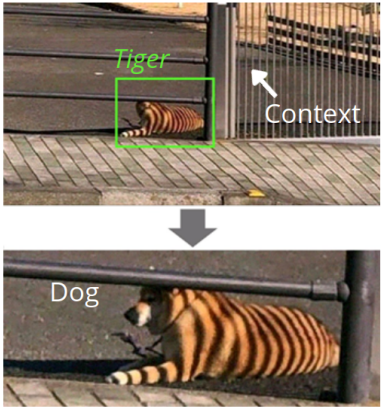
[Mehonic and Kenyon 2022)], [AKCP], [Boahen K. 2022]

Large Volume of Data, Compute, and Energy

- Training DNNs has extremely high demands in terms of power consumption.
- According to conservative estimates, training chat GPT-4 over $63 million.
- Cloud energy consumption has more than quadrupled from the advent of GPU use for DNN training.

TU/e

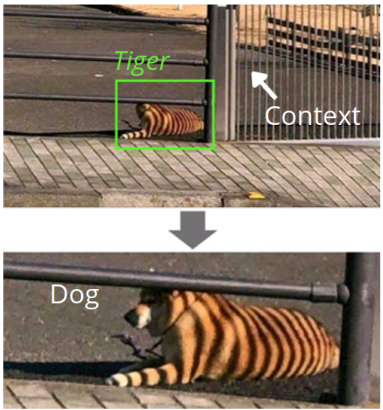
Challenges in the future of computing and edge AI

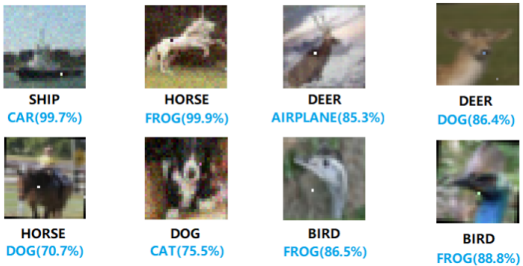


Brittle AI

- Narrow and brittle A.I. (performs well only in predefined situations).

Challenges in the future of computing and edge AI





Brittle AI

- Narrow and brittle A.I. (performs well only in predefined situations).
- AI models can be easily fooled.

Outline

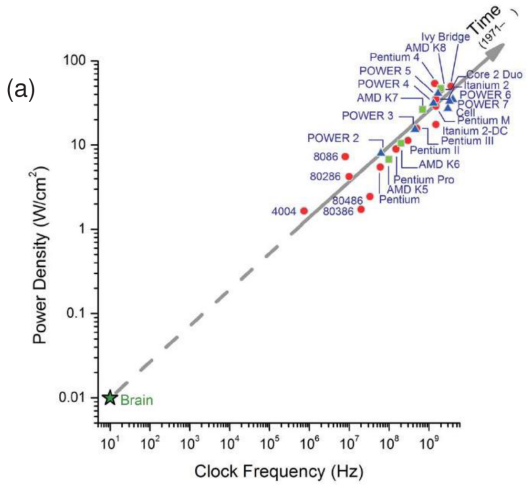
Challenges in the Future of Computing and Edge AI

My Research Focus: Neuromorphic Computing & Engineering

Spiking Neural Networks as *models* of computation

Spiking neural networks in FPGA

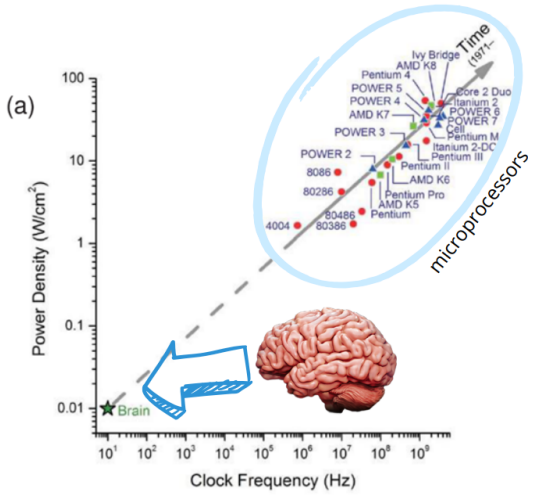
Paths beyond the current limits in computing and AI



Merolla et al. Science, 2014

TU/e

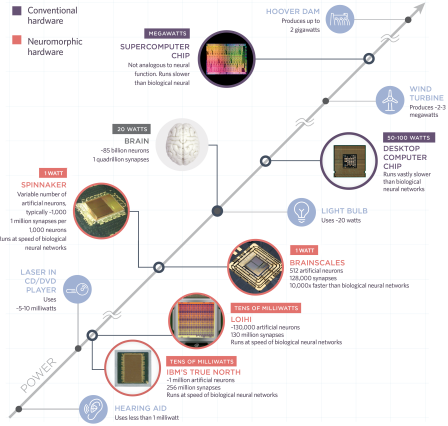
Paths beyond the current limits in computing and AI



Merolla et al. Science, 2014

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

Paths beyond the current limits in computing and AI

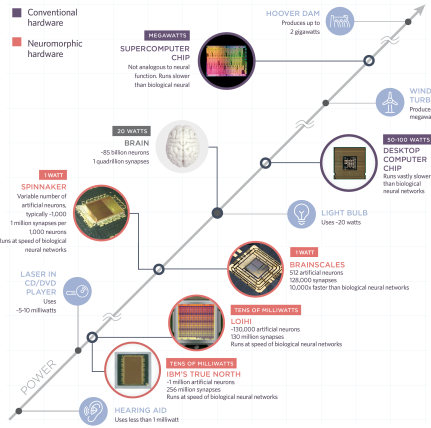


[The Scientist, 2019]

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

TU/e

Paths beyond the current limits in computing and AI



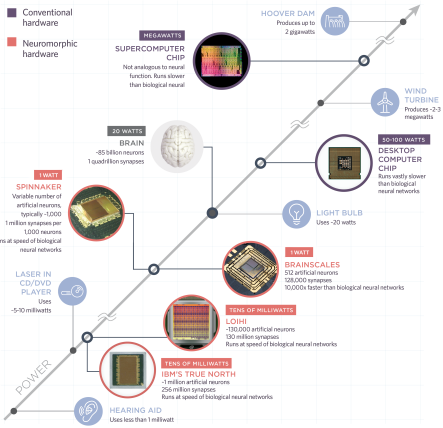
Neuromorphic Computing

Efficient computer architectures for brain-inspired models of computation (e.g., spiking neural networks)

[The Scientist, 2019]

TU/e

Paths beyond the current limits in computing and AI



Neuromorphic Computing

Efficient computer architectures for brain-inspired models of computation (e.g., spiking neural networks)

Today's digital neuromorphic hardware

- Parallel processing
- Distributed memory
- Sparse encoding
- Data-flow communication

[The Scientist, 2019]

TU/e

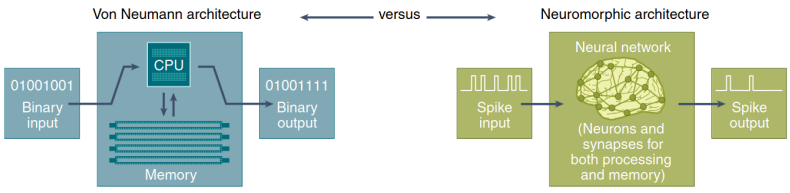
Neuromorphic computing systems

Today's AI and compute architectures

- Physical limits of CMOS & Von Neumann bottleneck.
- Large volume of data, compute, and energy.
- Brittle AI.

Neuromorphic computing

- Exploit the physics of CMOS and emerging technologies.
- Event-based, in-memory computing, low-power.
- Brain-inspired computing.



[Schuman C. D. et al, Nature 2022]

Outline

Challenges in the Future of Computing and Edge AI

My Research Focus: Neuromorphic Computing & Engineering

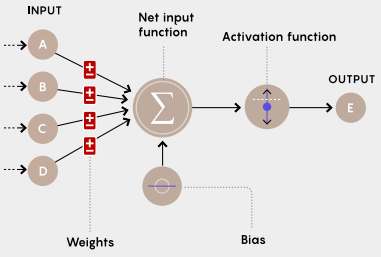
Spiking Neural Networks as *models* of computation

Spiking neural networks in FPGA

Artificial vs natural intelligence

Artificial neural networks

simulate abstract brain-inspired computing algorithms on digital time-multiplexed computing substrates.



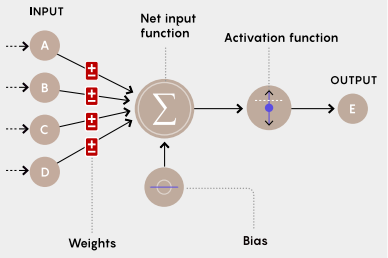
[Quanta Magazine]

TU/e

Artificial vs natural intelligence

Artificial neural networks

simulate abstract brain-inspired computing algorithms on digital time-multiplexed computing substrates.

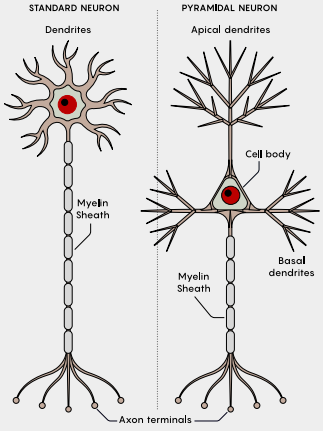


[Quanta Magazine]

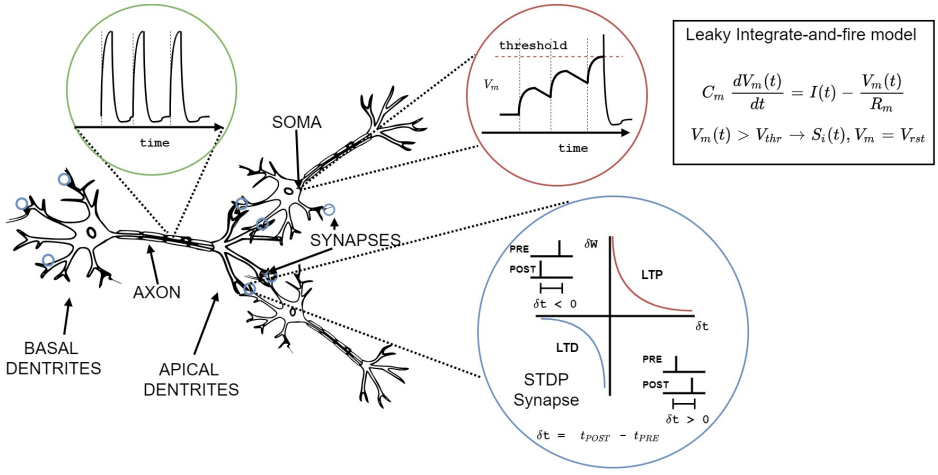
Biological neural networks

They use physical (real) time and circuit-dynamics to compute through their time evolution. The morphology and structure of the network determines its functionality.

The physical hardware substrate is the algorithm.



Spiking neural networks as *models* of computation



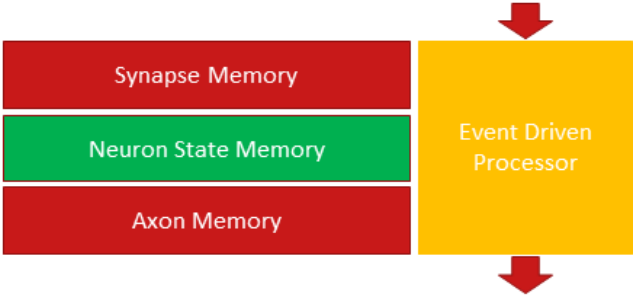
Leaky Integrate-and-fire model

$$C_m \frac{dV_m(t)}{dt} = I(t) - \frac{V_m(t)}{R_m}$$

$$V_m(t) > V_{thr} \rightarrow S_i(t), V_m = V_{rst}$$

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

Neuromorphic Computing *digital hardware template*

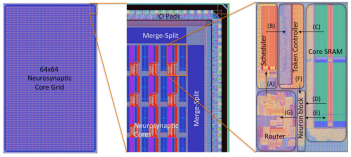


Typical digital architecture

- One core updates the state of many neurons (von Neumann)
- Synapse memory (weights)
- State memory (membrane potential)
- Axon memory (destination)

Fully Digital Neuromorphic Architectures *ASICs*

TrueNorth, IBM 2014

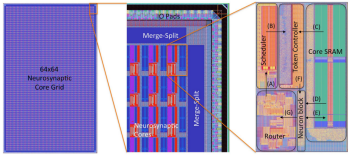


[Akopyan et al, 2015]

- $430mm^2$, 28nm CMOS
- 1 M neurons (4,096) cores, 2
- 256 M synapses
- Real-time (no learning)

Fully Digital Neuromorphic Architectures *ASICs*

TrueNorth, IBM 2014



[Akopyan et al, 2015]

- 430mm^2, 28nm CMOS
- 1 M neurons (4,096) cores, 2
- 256 M synapses
- Real-time (no learning)

Loihi, Intel 2018

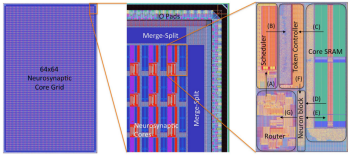
[Davies et al, 2018]

- 60mm^2, 14nm FinFet
- 130000 neurons (128) cores + 3 x86 cores
- 128 M synapses
- Real-time (online learning)

TU/e

Fully Digital Neuromorphic Architectures *ASICs*

TrueNorth, IBM 2014



[Akopyan et al, 2015]

- 430mm^2, 28nm CMOS
- 1 M neurons (4,096) cores, 2
- 256 M synapses
- Real-time (no learning)

Loihi, Intel 2018

[Davies et al, 2018]

- 60mm^2, 14nm FinFet
- 130000 neurons (128) cores + 3 x86 cores
- 128 M synapses
- Real-time (online learning)

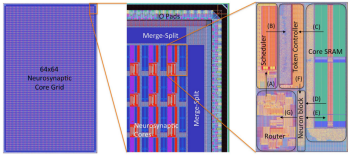
Loihi-2, Intel 2022

[Tomshardware]

- 31mm^2, 4nm FinFet
- 1M neurons (128) cores + 3 x86 cores
- 120 M synapses
- Real-time (online learning)

Digital Neuromorphic Architectures

TrueNorth, IBM 2014



[Akopyan et al, 2015]

Loihi, Intel 2018

[Davies et al, 2018]

Loihi-2, Intel 2022

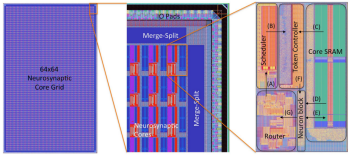
[Tomshardware]

Digital Neuromorphic Hardware

- Global Asynchronous Local Synchronous

Digital Neuromorphic Architectures

TrueNorth, IBM 2014



[Akopyan et al, 2015]

Loihi, Intel 2018

[Davies et al, 2018]

Loihi-2, Intel 2022

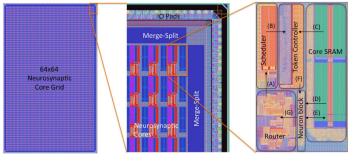
[Tomshardware]

Digital Neuromorphic Hardware

- Global Asynchronous Local Synchronous
- Event-driven Processing

Digital Neuromorphic Architectures

TrueNorth, IBM 2014



[Akopyan et al, 2015]

Loihi, Intel 2018

[Davies et al, 2018]

Loihi-2, Intel 2022

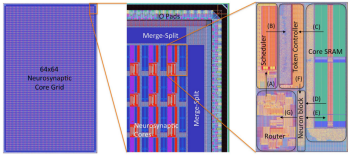
[Tomshardware]

Digital Neuromorphic Hardware

- Global Asynchronous Local Synchronous
- Event-driven Processing
- Near Memory Computing

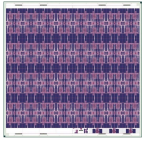
Digital Neuromorphic Architectures

TrueNorth, IBM 2014



[Akopyan et al, 2015]

Loihi, Intel 2018



[Davies et al, 2018]

Loihi-2, Intel 2022

[Tomshardware]

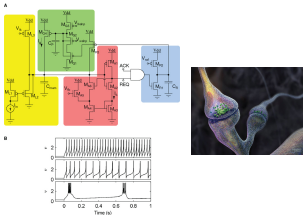
Digital Neuromorphic Hardware

- Global Asynchronous Local Synchronous
- Event-driven Processing
- Near Memory Computing
- Scalable Connectivity Interfaces

Neuromorphic Edge Computing Systems Lab.'s *research lines*

Neuromorphic Devices

- Subthreshold analog
- In-memory computing
- Emerging materials integration (ReRAM)

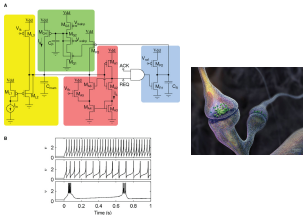


TU/e

Neuromorphic Edge Computing Systems Lab.'s *research lines*

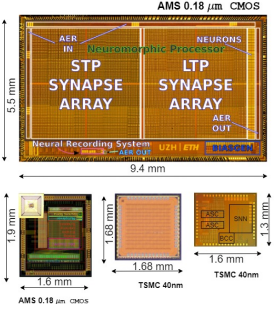
Neuromorphic Devices

- Subthreshold analog
- In-memory computing
- Emerging materials integration (ReRAM)



Neuromorphic Computing & Engineering

- Application driven
- Custom ASIC (SNNs)
- **SNNs in FPGAs**

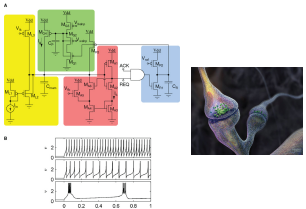


TU/e

Neuromorphic Edge Computing Systems Lab.'s *research lines*

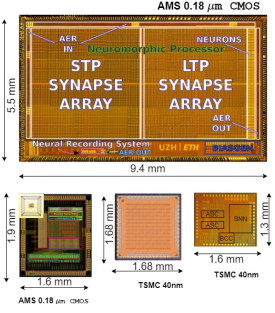
Neuromorphic Devices

- Subthreshold analog
- In-memory computing
- Emerging materials integration (ReRAM)



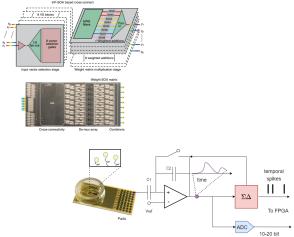
Neuromorphic Computing & Engineering

- Application driven
- Custom ASIC (SNNs)
- **SNNs in FPGAs**



Neuromorphic Beyond CMOS

- Ultra-fast electro-photonics SNNs
- Synthetic biological sensors



TU/e

Outline

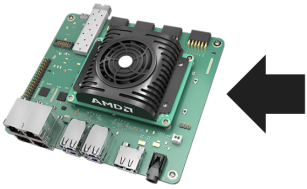
Challenges in the Future of Computing and Edge AI

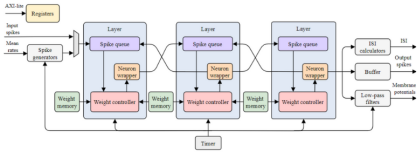
My Research Focus: Neuromorphic Computing & Engineering

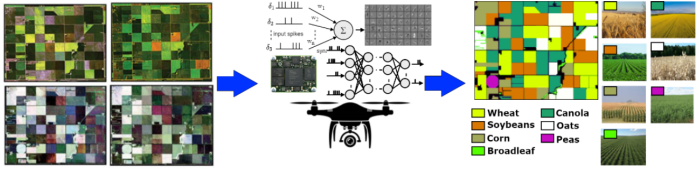
Spiking Neural Networks as *models* of computation

Spiking neural networks in FPGA

Spiking neural networks in digital reprogrammable *hardware*







Digital *simulations*

Neurons and synapses are physically implemented in cheap and massively parallel Field-Programmable Gate Array (FPGA).

[Corradi F. et al. 2021]
[Irmak H. et al. 2021]
[Sankaran et al. 2022]

Spiking neural networks in digital reprogrammable *hardware*

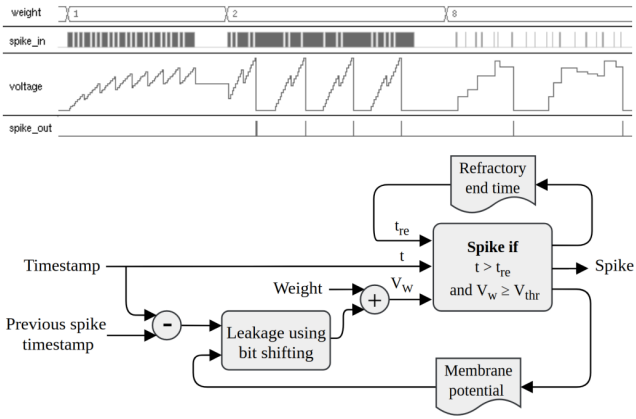
Name	Neuron Model	Driven	Network Topology	Algorithm
Bluehive [22]	Izhikevich	Time-driven	Feed-forward and Recurrent	Mean-rate
FDF [23]	Conductance	Time-driven	Feed-forward	Mean-rate
n-Minitaur [24]	LIF	Event-driven	Feed-forward	Mean-rate
Pani [9]	Izhikevich	Time-driven	Recurrent	Mean-rate
Tsinghua [12]	LIF	Hybrid (time and event)	Feed-forward	Mean-rate
Gyro [13]	LIF	Event-driven	Feed-forward and Recurrent	Mean-rate
Irmak et al. [25]	LIF and ReLU	Hybrid (time and event)	CNN, MLP and SNN	ANN and Mean-rate
This work	Simplified LIF	Event-Driven	Feed Forward and Recurrent	Single spike control (BPTT)

Table 1: Spiking neural networks on FPGA.

FPGA *simulations* of spiking neural networks

In literature, we can find many neuron models, execution modes, network topologies and algorithms: **FPGAs** are perfect for experimenting with SNNs computing systems!

Spiking neural networks in digital reprogrammable *hardware*

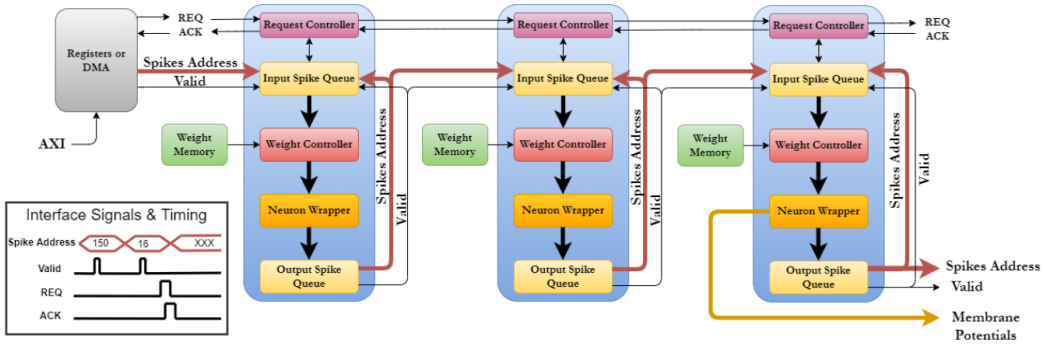


Leaky-Integrate-and-Fire Neuron

- Event-driven
- Programmable synapses
- Scalable (Fully parallel execution)
- Assumption: arbitration delays are in the order of *ns*, while incoming spikes are spaced in μs or even *ms*

[Corradi F. et al. 2021]

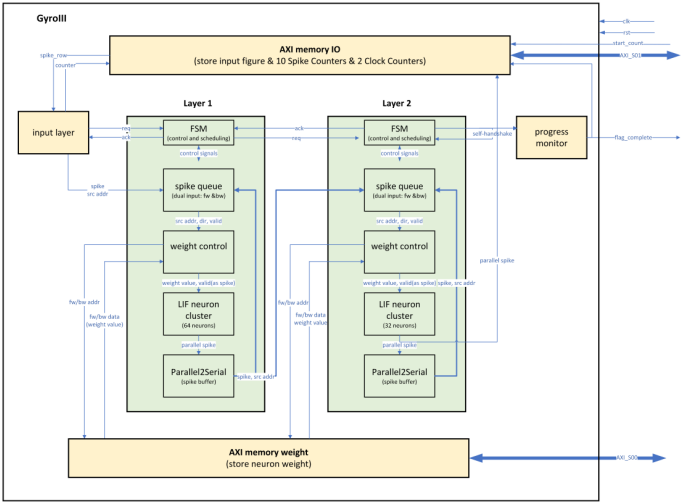
Spiking neural networks in digital reprogrammable *hardware*



A fully-spatial, fully-parallel, and event-driven spiking neural networks in FPGA.

[Corradi et al. 2021]
[Sankaran et al. 2022]

Spiking neural networks in digital reprogrammable *hardware*

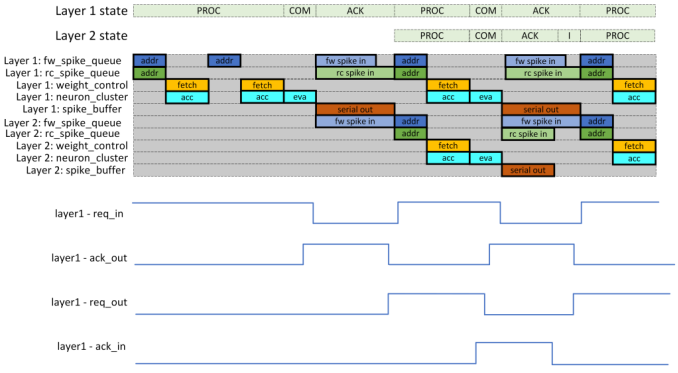


- Batch processing (I/O mem)

[Corradi et al. 2021]
[Sankaran et al. 2022]

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

Spiking neural networks in digital reprogrammable *hardware*

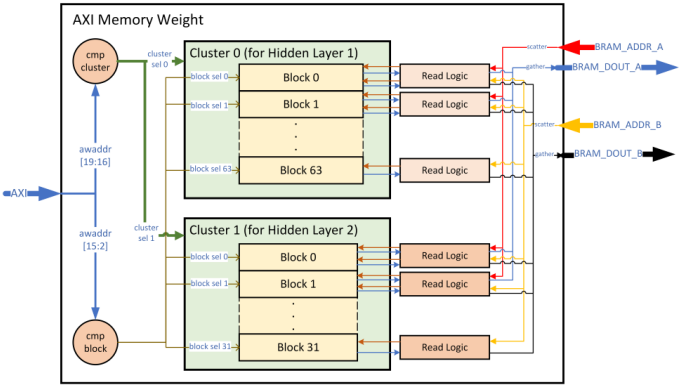


- Batch processing (I/O mem)
- Execution pipeline (layer-wise)

[Corradi et al. 2021]
[Sankaran et al. 2022]

TU/e

Spiking neural networks in digital reprogrammable *hardware*

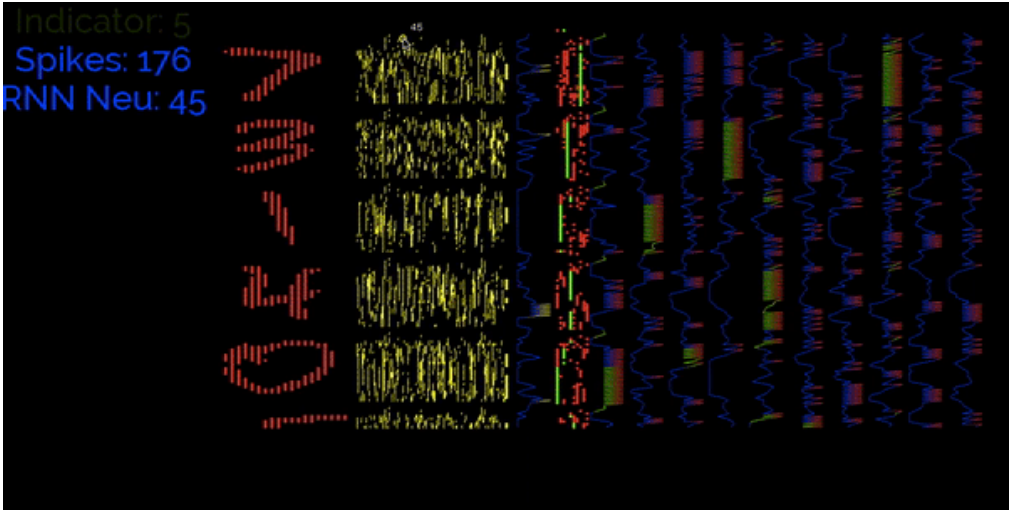


- Batch processing (I/O mem)
- Execution pipeline (layer-wise)
- On-chip memory (Bram/URAM)

[Corradi et al. 2021]
[Sankaran et al. 2022]

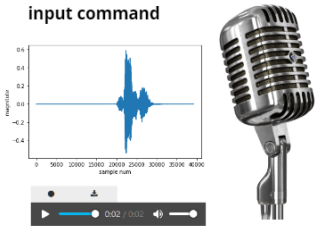
TU/e

Spiking neural network *workloads*



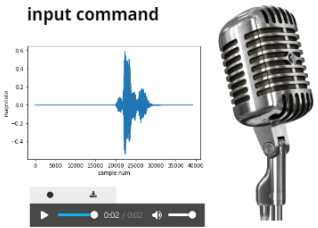
Federico Corradi | NECS Lab. | f.corradi@tue.nl

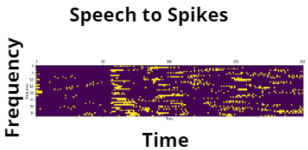
Spiking neural network for *keyword spotting* (GSC)



The pipeline includes pre-processing the audio signal and converting it into spikes, followed by processing it with a spiking neural network (250x250x250x36).

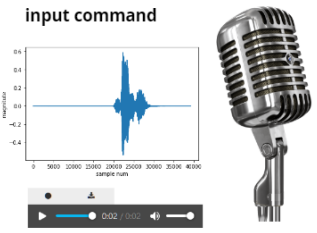
Spiking neural network for *keyword spotting* (GSC)

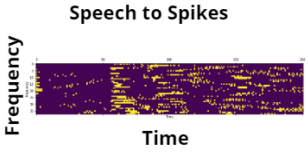


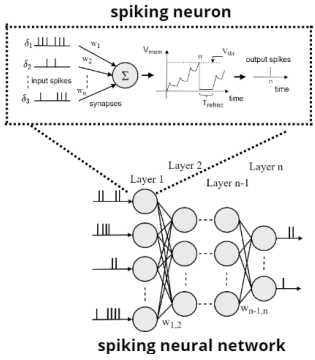


The pipeline includes pre-processing the audio signal and converting it into spikes, followed by processing it with a spiking neural network (250x250x250x36).

Spiking neural network for *keyword spotting* (GSC)

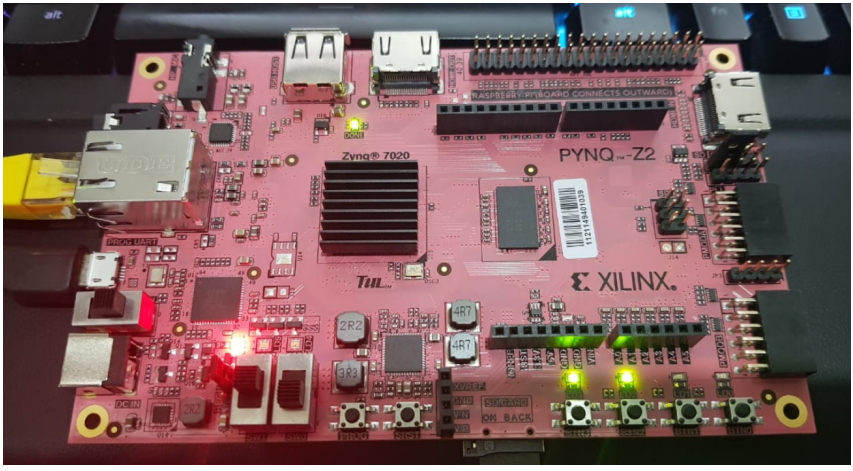






The pipeline includes pre-processing the audio signal and converting it into spikes, followed by processing it with a spiking neural network (250x250x250x36).

Spiking neural network for keyword spotting (GSC) *deployment*



Exploit ARM processor for data I/O and FPGA acceleration for SNN processing.

Spiking neural network for keyword spotting (GSC) *deployment*

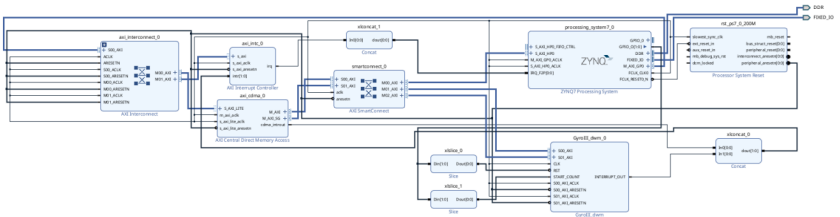


Figure 4: PYNQ system block diagram

Resource	Utilitization	Available	Utilization %
LUT	43918	53200	82.55
LUTRAM	11260	17400	64.71
FF	34082	106400	32.03
BRAM	140	140	100.00
BUFG	1	32	3.13

Table 1: Hardware resource utilization of design on PYNQ

Spiking neural network keyword spotting (GSC) *results*

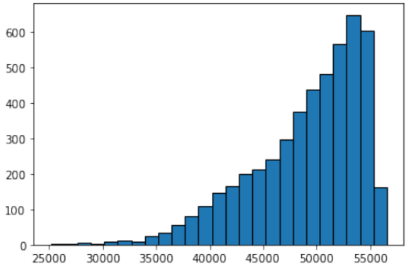


Figure 5: Histogram of processing time of inputs from the test set on the pynq FPGA

Performances

- $\sim$ 1.08 milliseconds @ 50 Mhz

Spiking neural network keyword spotting (GSC) *results*

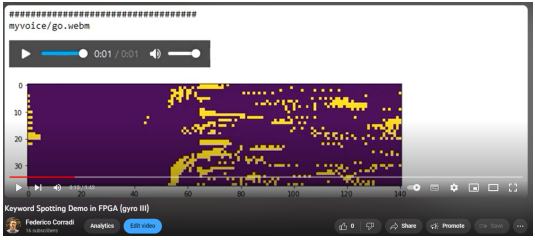
Experiment	Train accuracy	QAT accuracy	Quantization accuracy
Adam 0.004	0.6783	0.7144	0.7075
Adam 0.002	0.7207	0.7167	0.7167
Adam 0.001	0.7159	0.7274	0.7260
Adam 0.0005	0.7388	0.7491	0.7492
Adam 0.00025	0.7216	0.7547	0.7563
SGD 0.004/0.9	0.6948	0.7287	0.7272
SGD 0.002/0.9	0.7153	0.7509	0.7350
SGD 0.001/0.9	0.7261	0.7312	0.7289
SGD 0.0005/0.9	0.7310	0.7594	0.7578
SGD 0.00025/0.9	0.7155	0.7393	0.7292
SGD 0.004/0.95	0.7281	0.7330	0.7319
SGD 0.002/0.95	0.7070	0.7441	0.7351
SGD 0.001/0.95	0.7330	0.7547	0.7517
SGD 0.0005/0.95	0.6951	0.7455	0.7507
SGD 0.00025/0.95	0.7404	0.7289	0.7323
SGD 0.004/0.99	0.7263	0.7426	0.7404
SGD 0.002/0.99	0.7108	0.7459	0.7425
SGD 0.001/0.99	0.7162	0.7313	0.7329
SGD 0.0005/0.99	0.7088	0.7468	0.7445
SGD 0.00025/0.99	0.7233	0.7442	0.7433

Performances

- $\sim$ 1.08 milliseconds @ 50 Mhz
- Accuracy vs quantiation trade-offs [Eissa S. et al. 2023]

TU/e

Spiking neural network keyword spotting (GSC) *results*

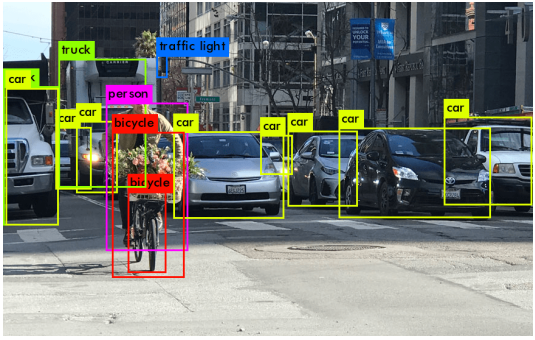


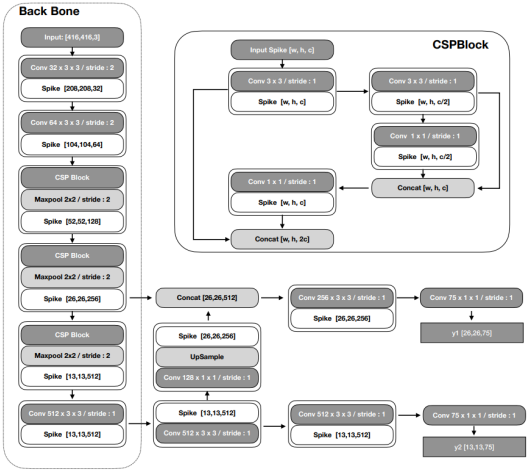
Performances

- $\sim$ 1.08 milliseconds @ 50 Mhz
- Accuracy vs quantiation trade-offs [Eissa S. et al. 2023]
- It works ok! [YouTube Demo]

Scaling up: deep Spiking Neural Networks Models

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



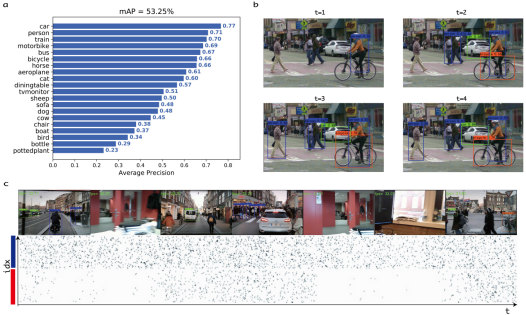


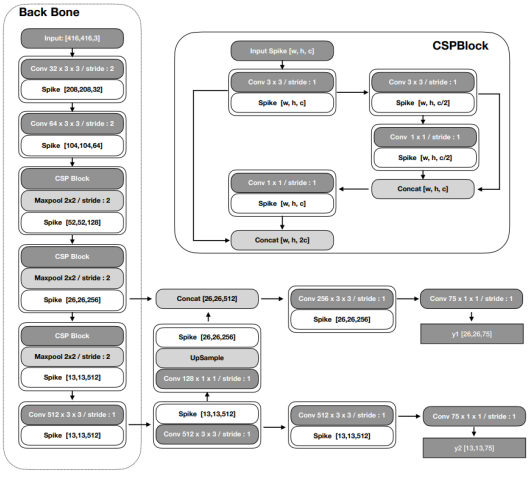
[Yin et al, 2023 NMI]

TU/e

Scaling up: deep Spiking Neural Networks Models

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

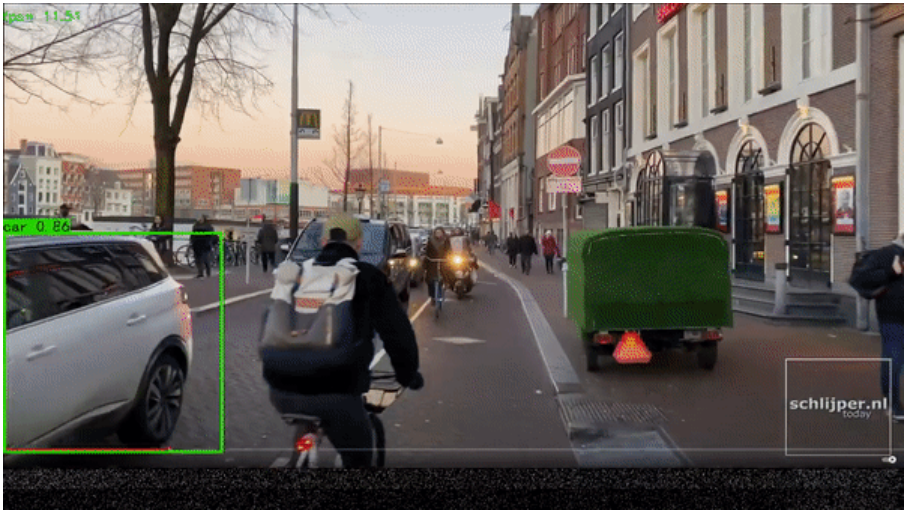




[Yin et al, 2023 NMI]

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

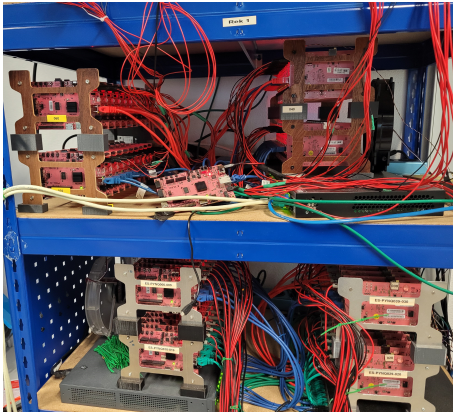
Deep Spiking Neural Networks Models



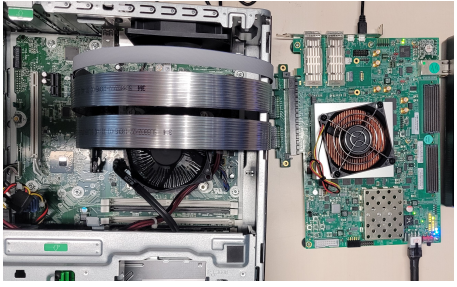
[Yin et al, 2023 NMI]

TU/e

Afraid of not fitting?



Pynq Z2 cluster



The Virtex UltraScale+ FPGA VCU118

Scaling up...

At TU/e, we have enough resources!

TU/e

Thank you!

Interested? Reach out!

to be continued..

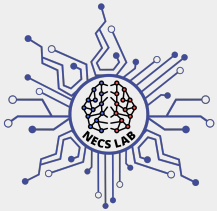
Neuromorphic Edge Computing Systems Lab

Federico Corradi

Eindhoven University of Technology

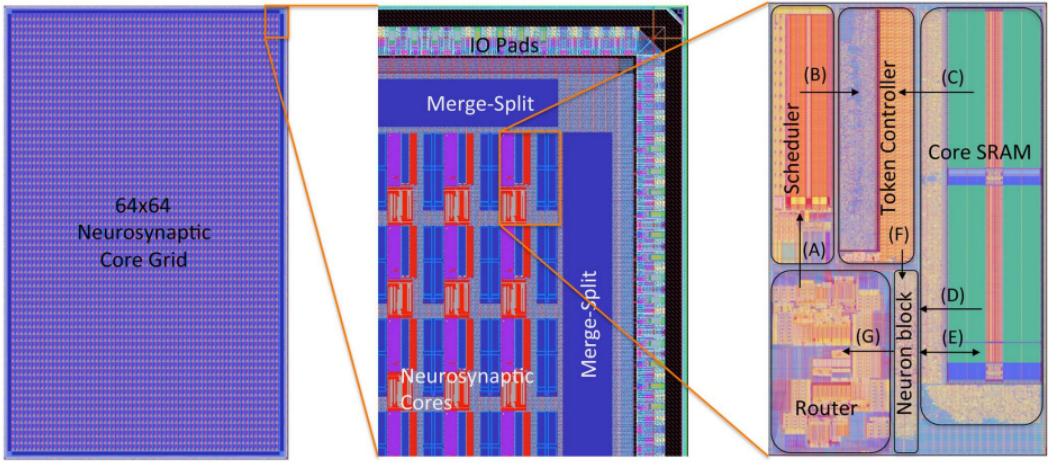
f.corradi@tue.nl, +31(0)402 472 556

Neuromorphic Edge Computing Systems Lab



Digital Neuromorphic Hardware

TrueNorth, IBM 2014

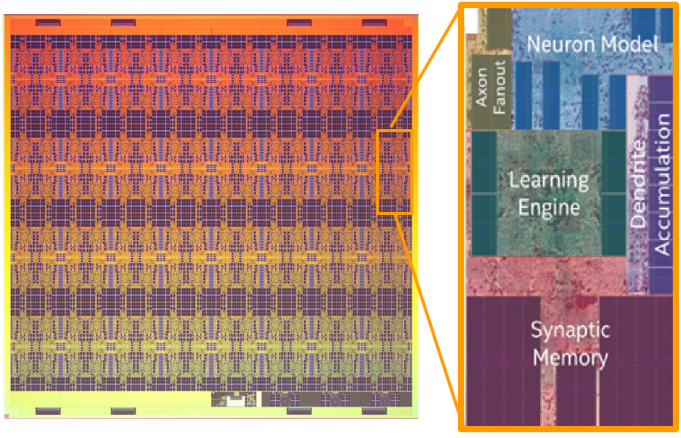


[Akopyan et al, 2015]

TU/e

Digital Neuromorphic Hardware

Loihi, Intel 2018

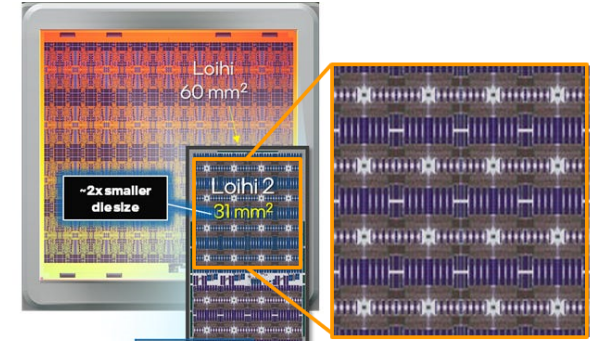


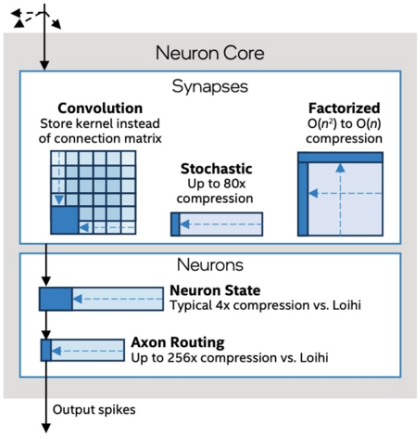
[Davies et al, 2018]

TU/e

Digital Neuromorphic Hardware

Loihi-2, Intel 2022



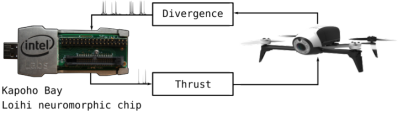


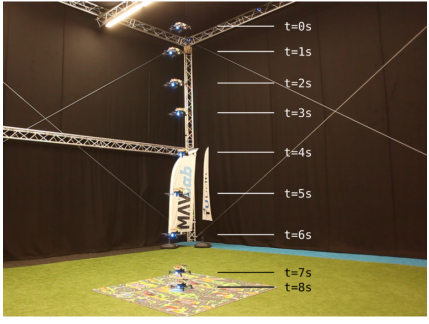
TU/e

Digital Neuromorphic Hardware: Applications

Real-Time bio-inspired navigation

[Schoepe et al, 2021]





[Dupeyroux et al, 2020] [Youtube Video]

TU/e