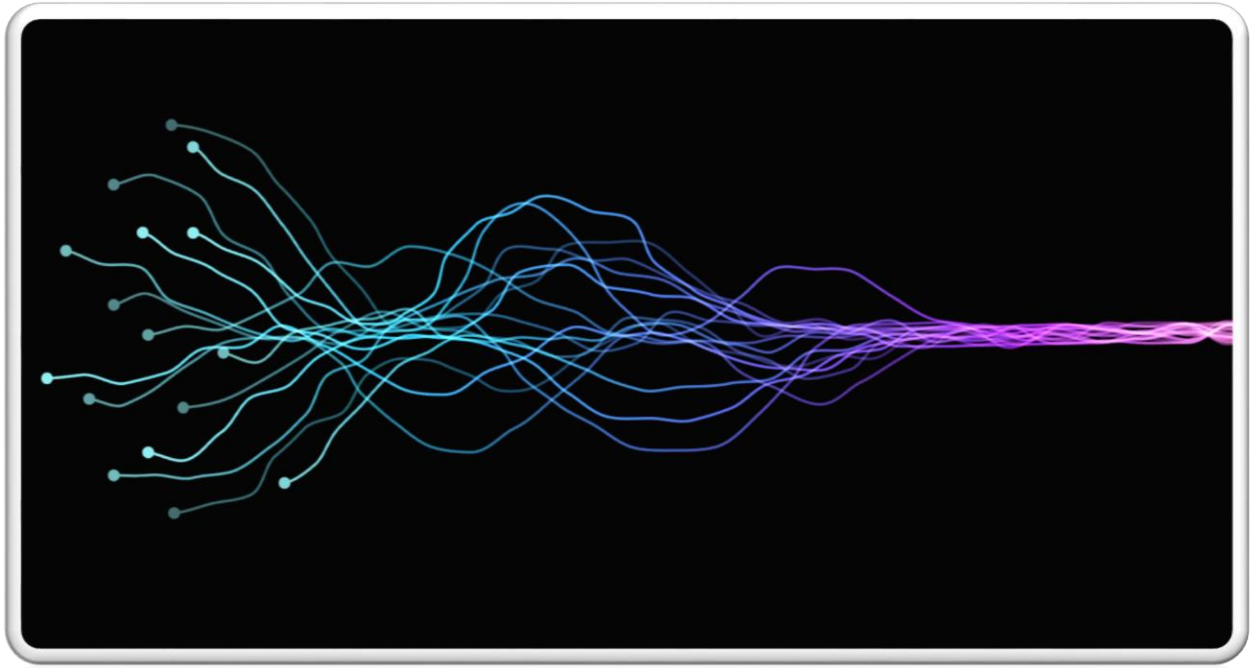
Edge AI
How to bring AI at the edge of the physical world

Benoît Miramond / LEAT



UNIVERSITÉ CÔTE D'AZUR

Outline

- LEAT research lab
- Edge AI
- Different Edge lines
  - Edge lines and properties
  - Some examples studied at LEAT
- Smart sensors: When AI touches the physical world
  - A matter of energy
- The bio-inspired approach
- Conclusion

LABORATOIRE D'ELECTRONIQUE
ANTENNES ET TELECOMMUNICATIONS

Laboratoire d'Electronique, Antennes et Télécommunications

Unité Mixte de Recherche UMR7248
Université Côte d'Azur et CNRS

cnrs advancing the frontiers

cremant

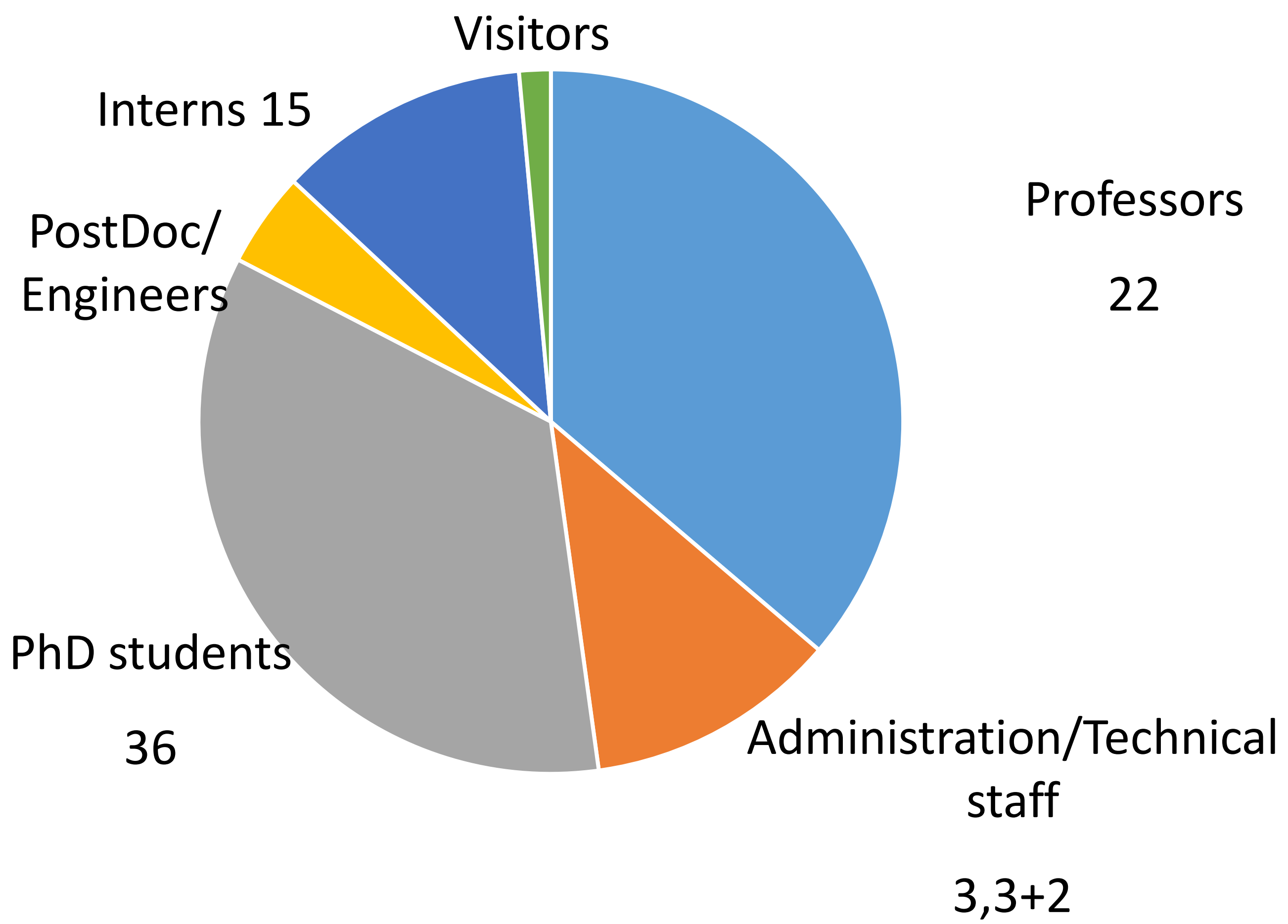
UNIVERSITÉ CÔTE D'AZUR

Location

Campus SophiaTech

Composition of the laboratory

Members : ~80 (June 2022)



Activities

Publications / Conferences

Patents / R&D

3 teams

Electronics/ Computer Sciences

Lab. Com. UCA/CNRS-Orange Labs

cremant

- Co-directors : Ph. Ratajczak (Orange Labs)
  F. Ferrero (UCA-CNRS)

- Joint research center
  - Orange Labs :
    Unité de recherche ANT: Antennes (Orange Labs Sophia)
    Unité de recherche WAVE: Interactions Ondes-corps humain(Orange Labs Paris)

- 2012-2022 Subjects of research
  - Integrated Antennas
  - Communications from 60 to 120 GHz
  - Sensors and sensor networks
  - New materials, electromagnetic modeling and applications

UNIVERSITÉ CÔTE D'AZUR

orange™

cnrs

Academic Collaborations

Industrial Collaborations

orange™

DGA

NAVAL GROUP

SORIN GROUP
AT THE HEART OF MEDICAL TECHNOLOGY

BOWEN

AIRBUS

technicolor

TCL

eDF

ST
life.augmented

NXP

GROUPE RENAULT

synopsys®

THALES

menta

Atos Bull

WIKA

Valeo

intel®

PROPHESEE
METAVISION FOR MACHINES

Research Teams

- **ISA**: **I**maging and **A**ssociated **A**ntennas **S**ystems
  Imagerie microondes et **S**ystèmes d'**A**ntennes

- **CMA**: **A**ntenna **D**esign and **M**odeling
  Conception et **M**odélisation d'**A**ntennes

- **EDGE**: **E**dge computing & **DiG**ital syst**E**ms
  Systèmes **N**umériques et **C**alcul embarqué

EDGE Team

Edge computing & **D**i**G**ital **E**lectronics

EDGE research axis

1. **eBrain** - **e**mbedded **B**io-inspi**R**ed **A**rtificial **I**ntelligence and **N**euromorphic Architectures

2. **eWISE** - **e**nergy-aware **WI**reless **S**ensor n**E**tworks

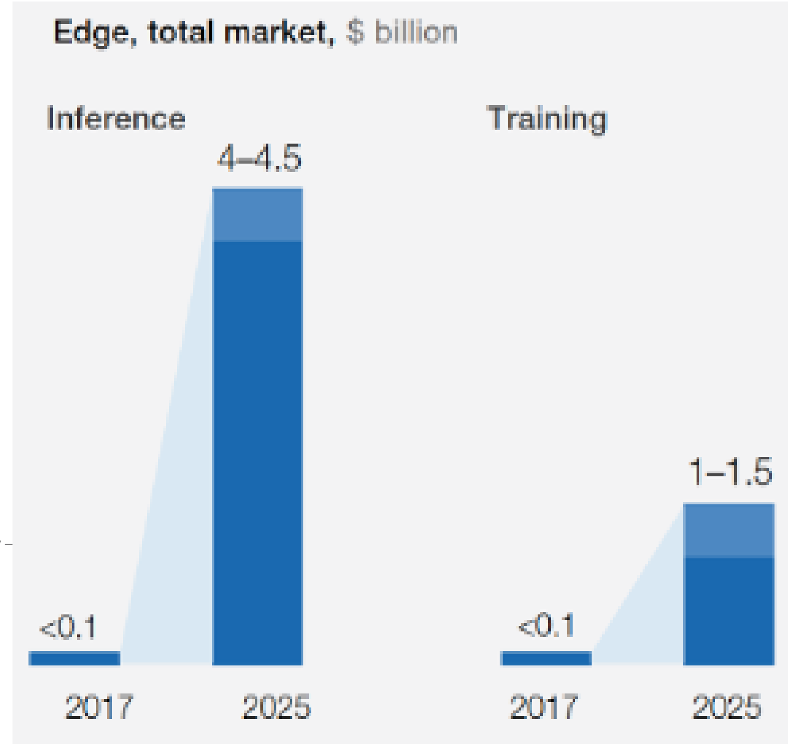
3. **eSoC** - **e**nergy efficiency of **SoC**

E-Health, Smart City

IoT, wearables

Autonomous cars

From embedded systems to Edge Intelligence



(1) McKinsey –

Data volume explodes with AI, 5G, IoT

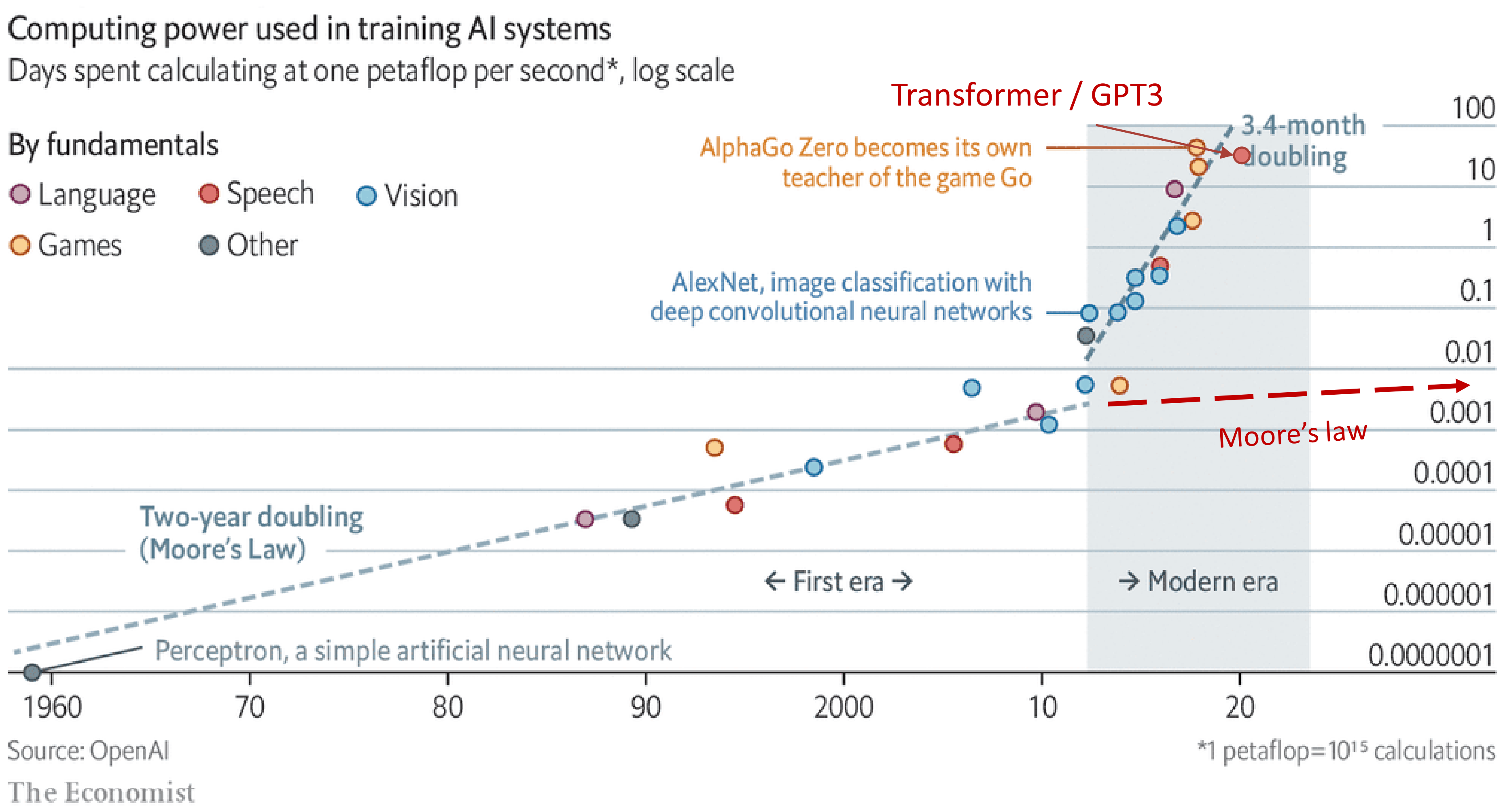
- ONLY 25% of usable data reach a datacenter
- 75% of data must be analyzed on site immediately

The impact in France and Europe will be immense in Aerospace, Automotive, Defense, Telecom,...

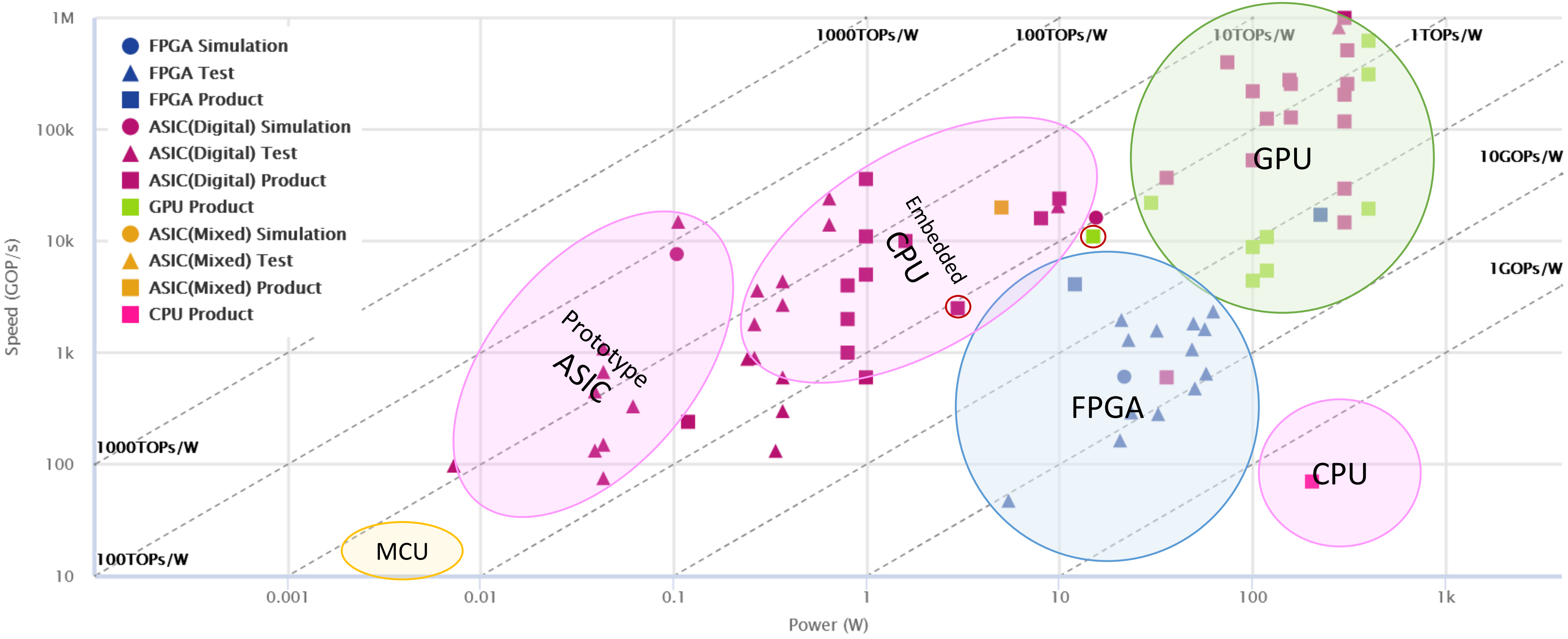
AI / Edge processors market has important growth
GPUs and FPGAs should not dominate this market.

1 Application-specific integrated circuit.
2 Central processing unit.
3 Field programmable gate array.
4 Graphics-processing unit.

A contrasted picture on Cloud Artificial Intelligence

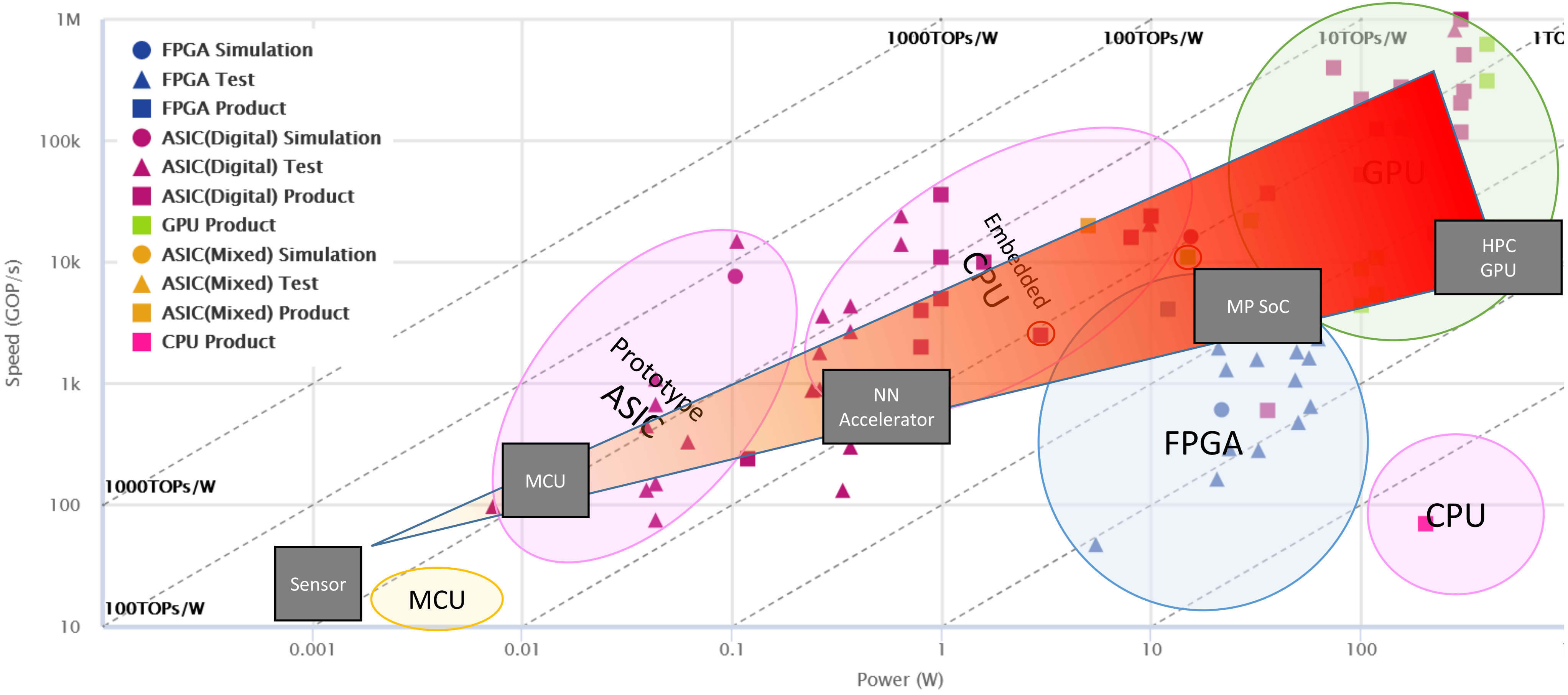


Digital Neural Network Accelerators



https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

Digital Neural Network Accelerators

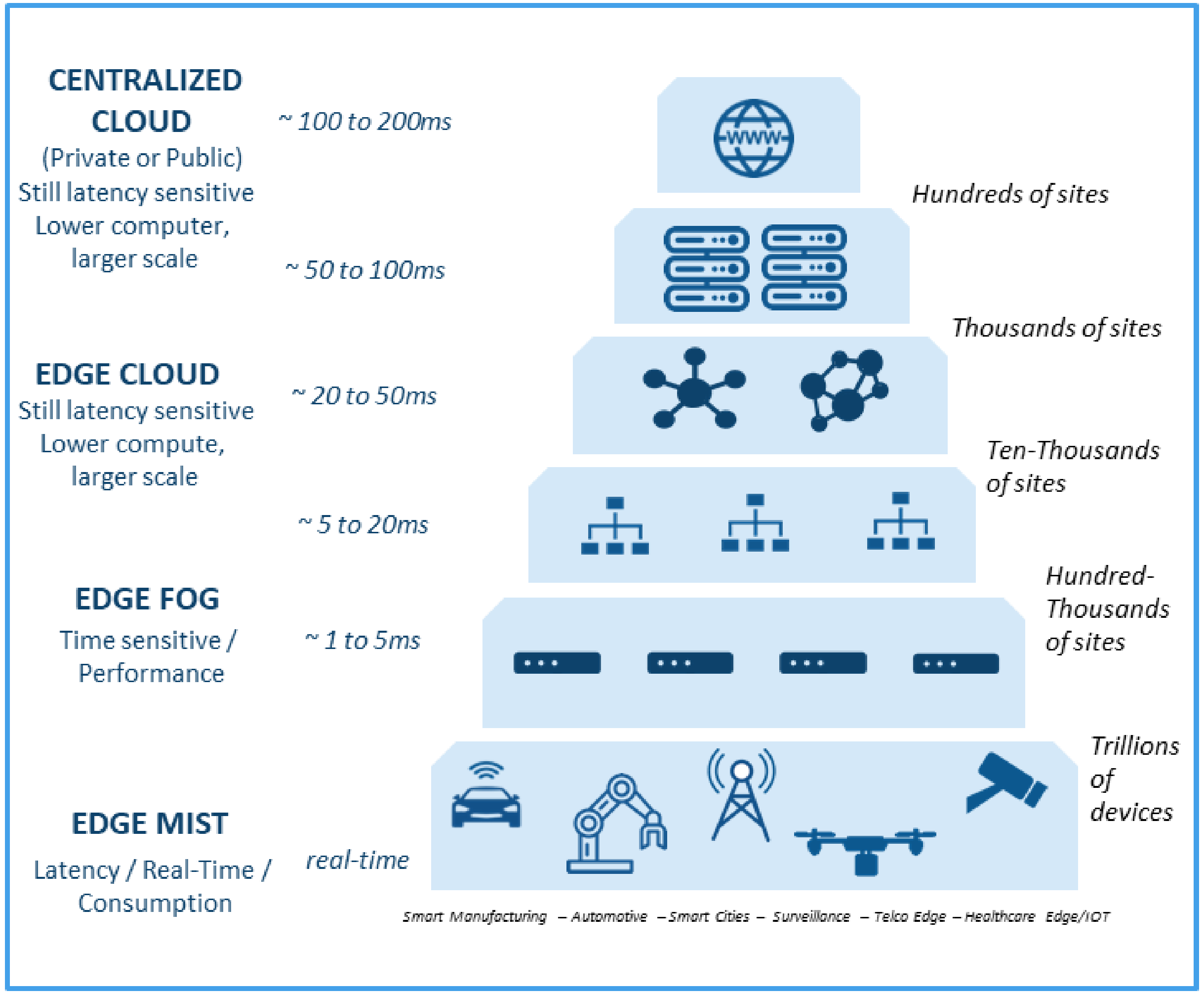


https://nicsefc.ee.tsinghua.edu.cn/projects/neural-network-accelerator/

- Specialized chips for AI calculation in the cloud
  - Nvidia GPU, US
  - Google TPU, US
  - Baidu Kunlun, CH
  - GraphCore, EN
  - Intel Movidius, US
  - Cerebras, US => 300.000 cores per wafer, 15kW
- At the Edge
  - NVIDIA Jetson can provide 11 T FLOPs, dissipating up to 15 W
  - Myriad X 4TOPS dissipating up to 1,5 W
  - Google Coral = 4 TOPS for 2W
  - …

Edge Lines

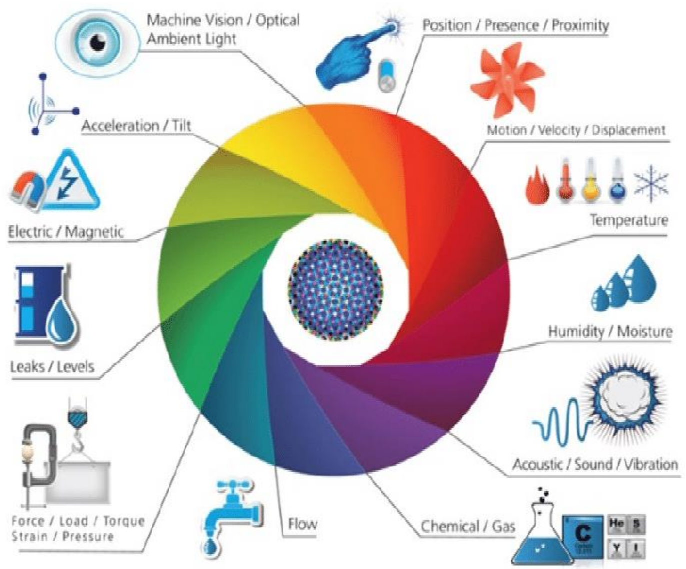
Edge Lines and their specific constraints



	Memory	Computation	Power	Efficiency
Edge Servers	GB	1 Tops	100 W	10 Gops/W
Gateway	MB	100 Gops	1 W	100 Gops/W
IoT Nodes	Hundreds of kB	1 Gops	1 mW	1 000 Gops/W

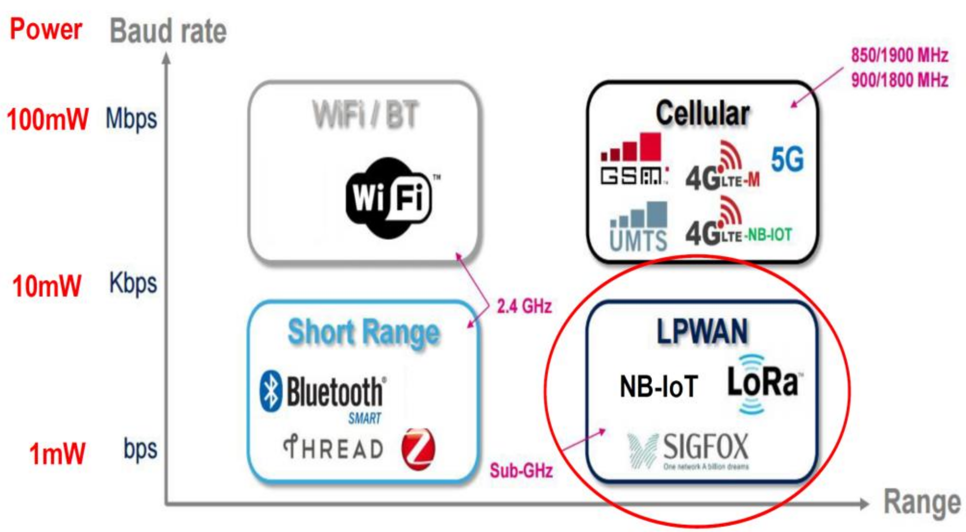
Key elements of IoT sensors

Sensors



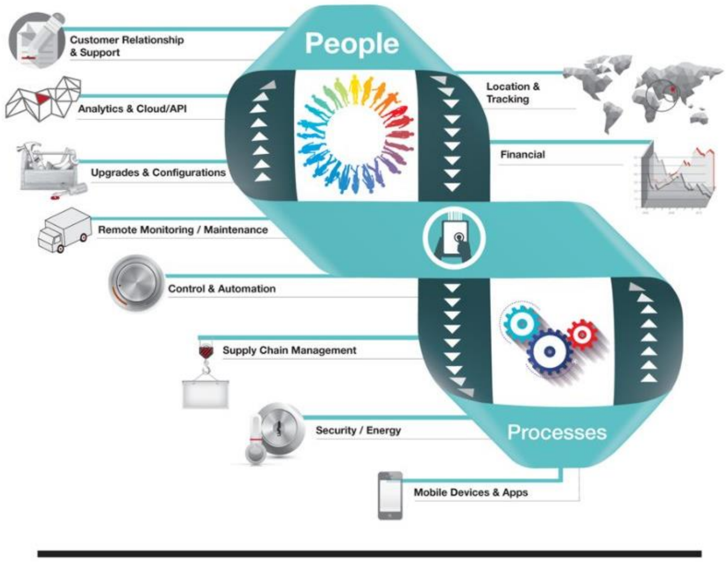
Captures a discrete representation of the dynamics of the physical world

Connectivity



Transmits the sensors data through wireless communication

Persons & process



Provides the information to people or process the raw data into more abstract information

When EdgeAI enables smart sensing

Fusion of AI, embedded sensors and connectivity

Smart sensors

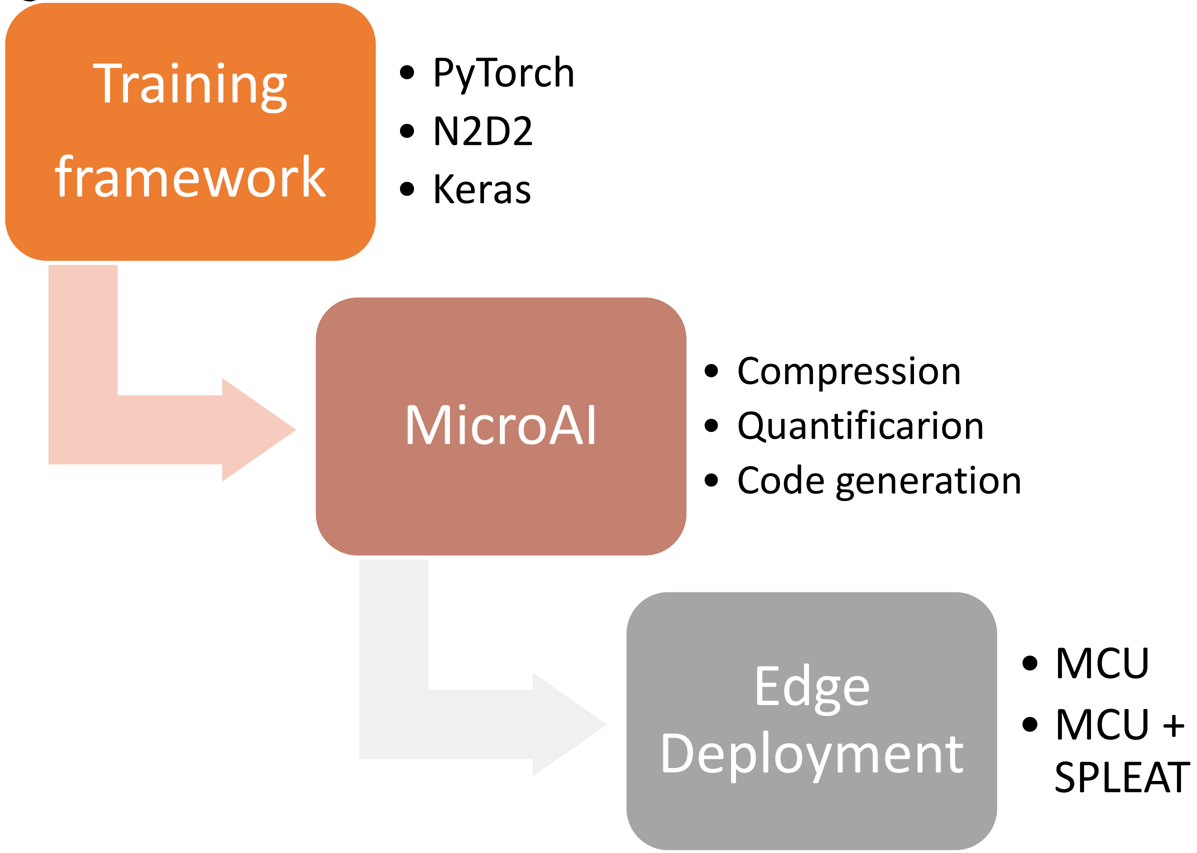
- Edge AI also offers the possibility to embed near-sensor processing

- **By bringing AI closer to the sensor, the goal is**
  - **To reduce the amount of data to communicate**
  - **To lower the global energy consumption of the digital infrastructure**
  - **To reduce latency for decising making (close or open loop)**

- Integrating AI into (near to) the sensor needs to specifically work at different scales
  - Algorithm/training: explore neural architecture that reduce parameters/computation
  - Embedded preparation: compression, quantization of the network
  - Electronic hardware: design and optimize the electronic architecture to support the neural network => Hw/Sw Codesign

The LEAT codesign flow for Edge AI

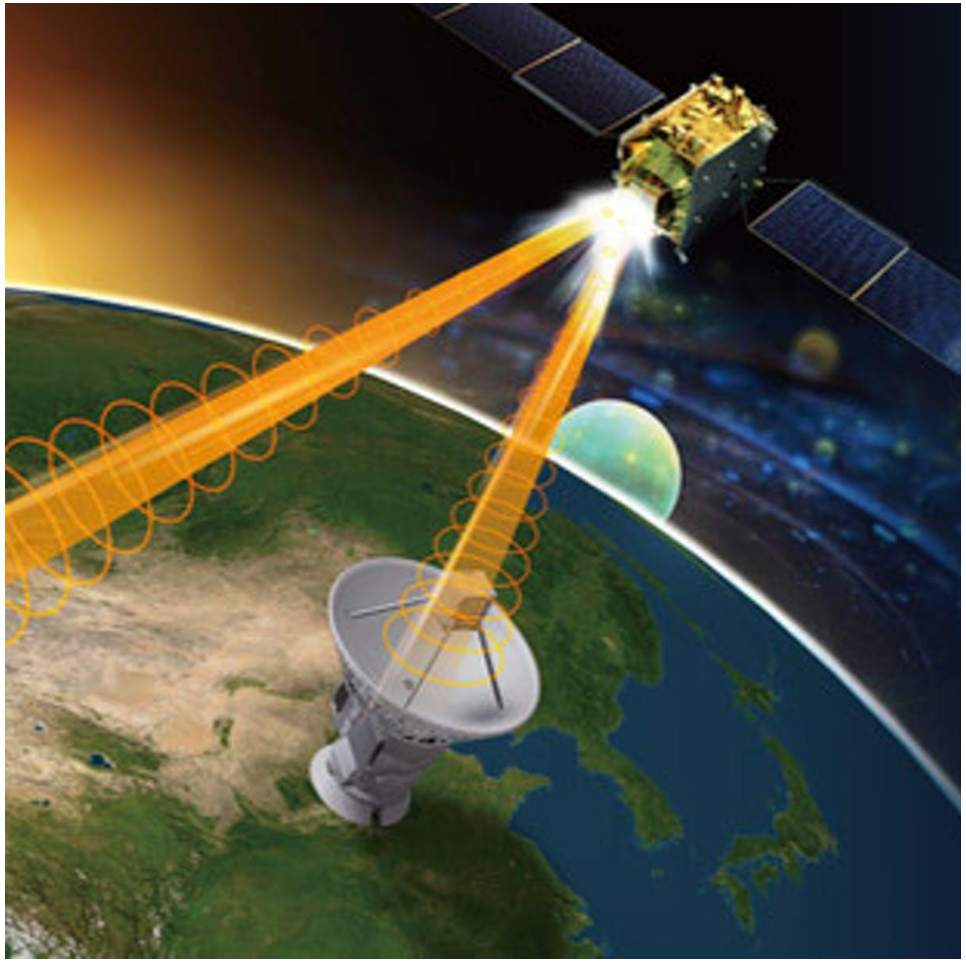
- Complete Solution*: from Training to Edge*
- Training of networks (frameworks PyTorch, Keras, N2D2)
- Embedded preparation of ANN with MicroAI
  - Quantification des SNN
  - Automatic code generation
  - Open-source: https://bitbucket.org/edge-team-leat/microai_public
- Hardware accelerator: next gen AI
  - Convolutionnal networks
  - Reprogrammable Architecture
  - Signal and Image processing applications



Quantization and deployment of deep neural networks on microcontrollers, PE Novac, GB Hacene, A Pegatoquet, B Miramond, V Gripon, Sensors 21 (9), 2984, 2021
SPLEAT: SPiking Low-power Event-based ArchiTecture for in-orbit processing of satellite imagery,, N. Abderrahmane B. Miramond, IJCNN 2022

Example of near-sensor classification

CIAR



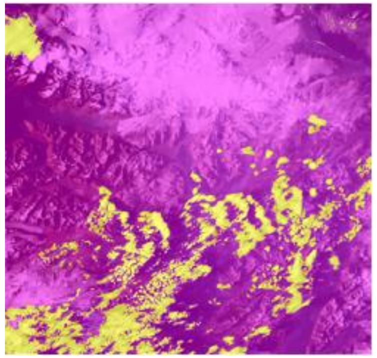
Send the entire image

1

VS.

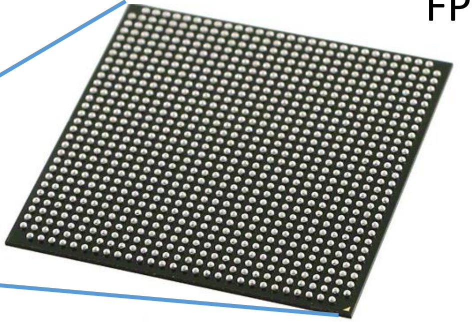
Send only the images without clouds, fire, …

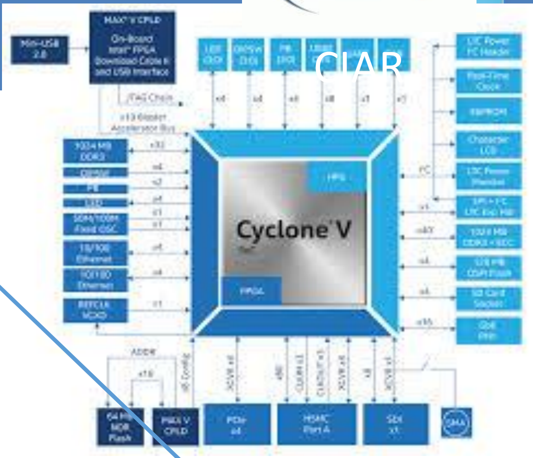
2



What is the on-board scientific experience ?

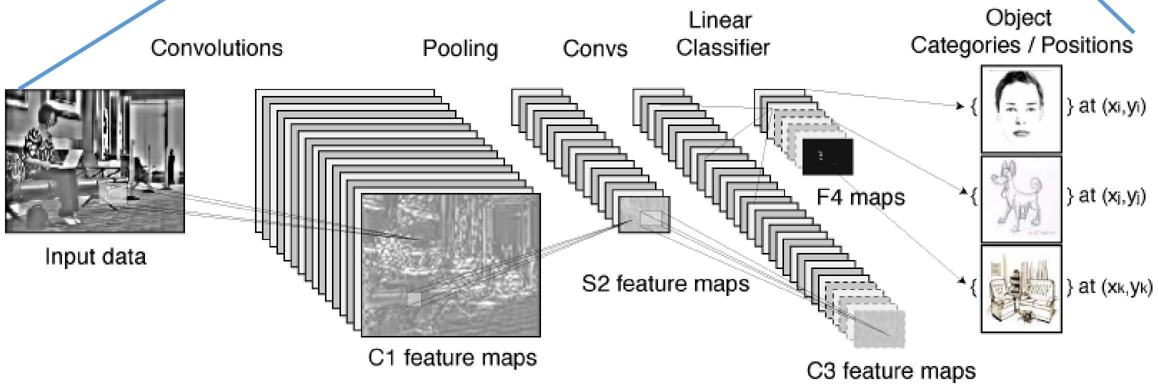
FPGA Electronic device



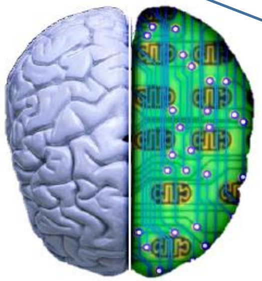


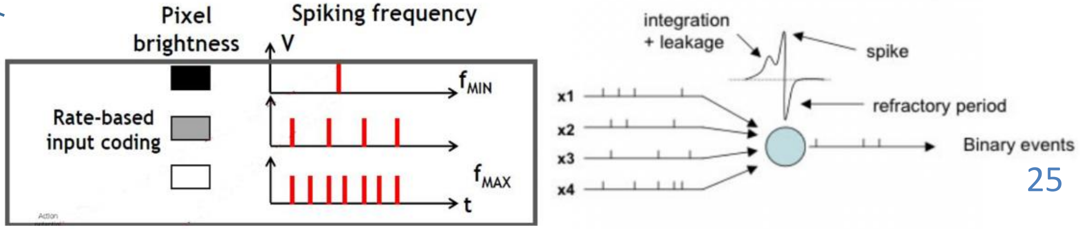
More details in the publication:
"An Hybrid Neural Network on FPGA for Embedded Satellite Image Classification", Edgar Lemaire et al., *IEEE International Symposium on Circuits and Systems (**ISCAS**), 2020*

1. Artificial Neural Network



2. Bio-inspired **Spiking Neural Network**



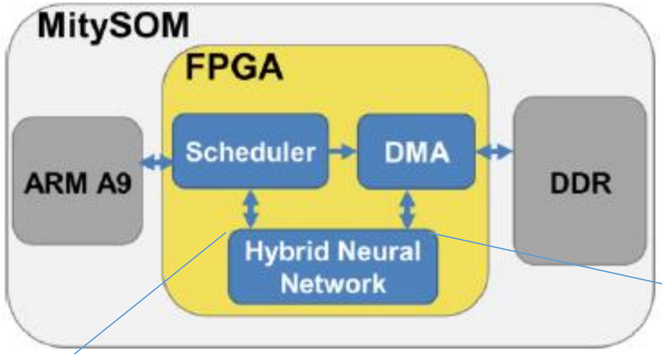


From Spiking Neural Networks to bio-inspired machine-learning

THALES

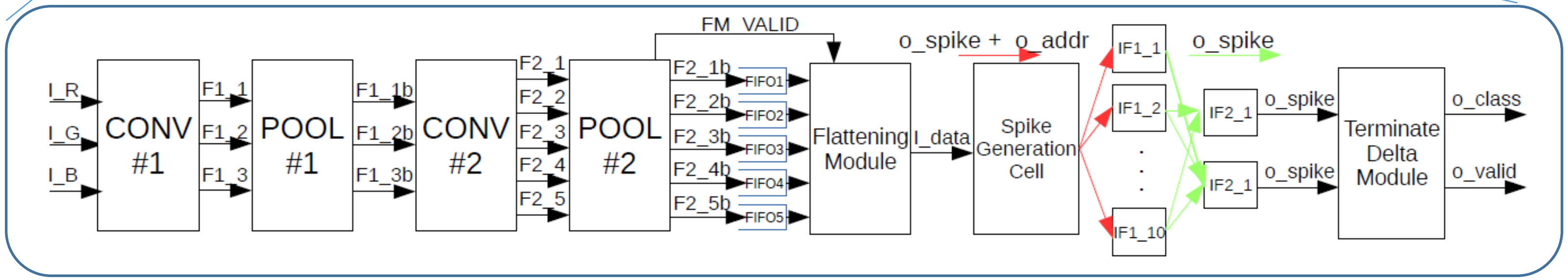
CIAR

Rate coding + CNN to SNN Conversion



MitySOM FPGA board

OPS-SAT: an ESA CubeSat



	Formal CNN	Hybrid CNN
Logic Cells occupation (%)	71%	59%
Working clock Frequency (MHz)	100	100
Recognition rate (%)	88	87
Average latency per image (μs)	25	43
LUTs in classif. stage (#)	9292	1572
Registers in classif. stage (#)	4320	1134
Block Memory Bits in classification stage (#)	0	3120
Power dissipation (mW)	1248.44	1192.66

Next step: full spike architecture with SPLEAT (SPiking Low-energy Event-based ArchiTecture) 1k -> 500k synapses

E Lemaire, M Moretti, L Daniel, B Miramond, P Millet, An FPGA-based Hybrid Neural Network accelerator for embedded satellite image classification, IEEE International Symposium on Circuits and Systems 2020

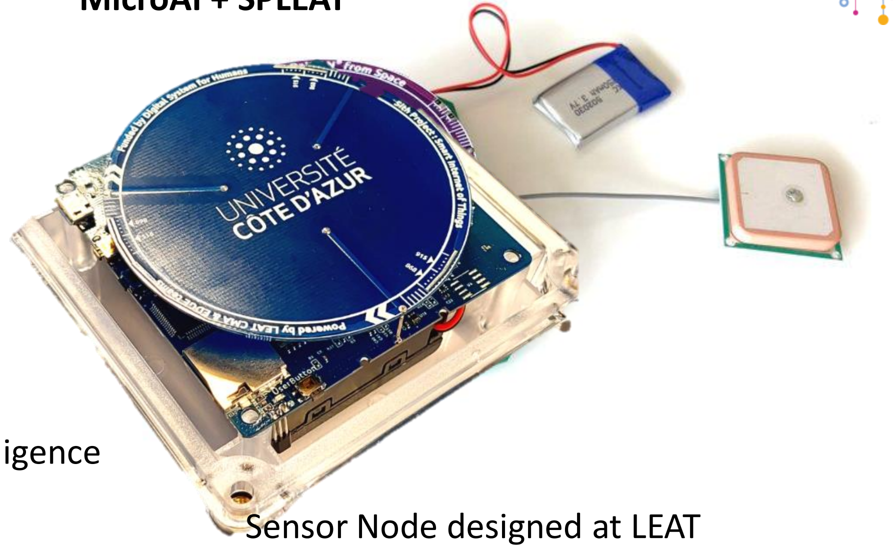
[L. Khacef, N. Abderrahmane and B. Miramond. Confronting machine-learning with neuroscience for neuromorphic architectures design. In International Joint Conference on Neural Networks (IJCNN). 2018]

Example of distributed AI with Satellite IoT

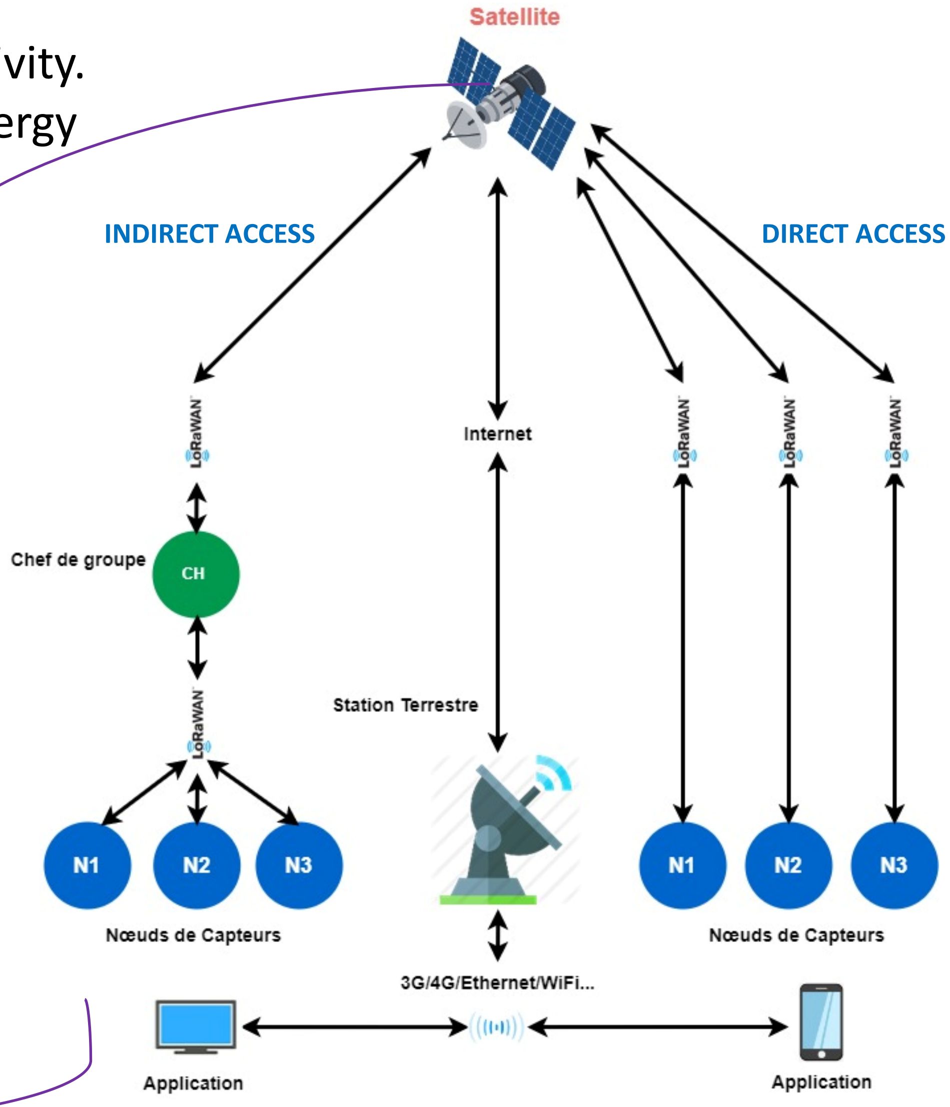
Ground sensors connected to a satellite.
End nodes embed sensors, MCU, battery and LoRA connectivity.
Each node embeds EdgeAI and has to be autonomous in energy

MicroAI + SPLEAT

Satellite LoRa Gateway
Embedded AI



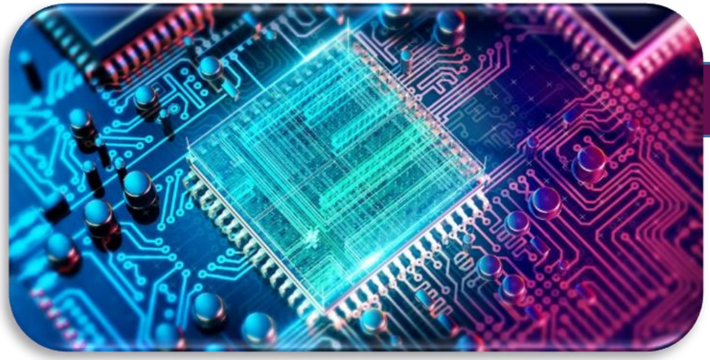
Sensor Node designed at LEAT



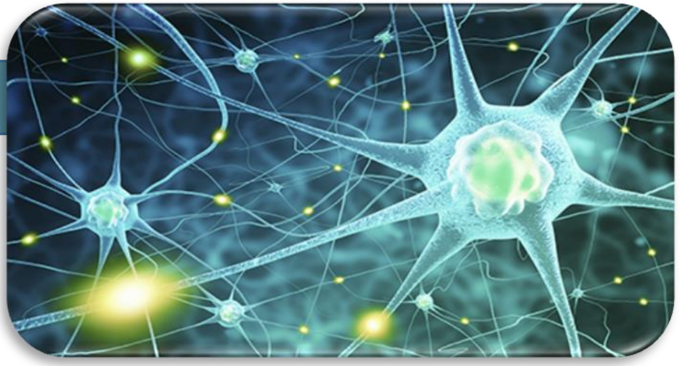
Ns: Nodes
CH: Cluster Head
AI: Artificial Intelligence

I. Abdoulaye, L. Rodriguez, C. Beleudy, B. Miramond, Embedded Artificial Neural Network for Data Prediction in Efficient Wireless Sensors Networks, ASPAI 2022

The bio-inspired approach at LEAT



Electronics



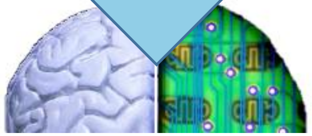
Cognitive Neurosciences

LEAT
LABORATOIRE D'ELECTRONIQUE
ANTENNES ET TELECOMMUNICATIONS

- Neuromorphic
- Embedded AI
- Smart IoT

Bio-inspired Artificial Intelligence

- Spiking Networks
- Brain plasticity
- Self-Organization



ebrAIn

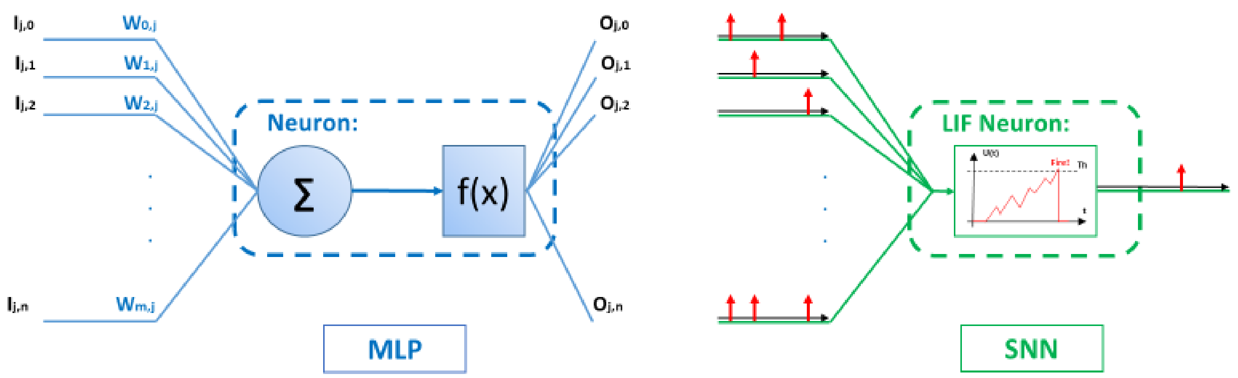
Embedded Bio-inspiRed Artificial Intelligence and Neuromorphic systems

UNIVERSITÉ CÔTE D'AZUR

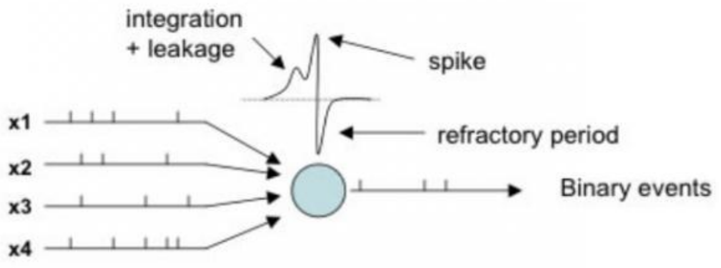
Bio-inspired computing with Spiking Neural Networks

- Spiking neural networks are the main subject of exploration in the domain of bio-inspired computing.
- **Main technical reasons:**
  - Impulsion coding
  - Temporal integration operations
  - Asynchronous behaviour
  - Decentralized learning rules
  - Bio-mimetic approach

CNN vs SNN with Leaky Integrate and Fire neurons

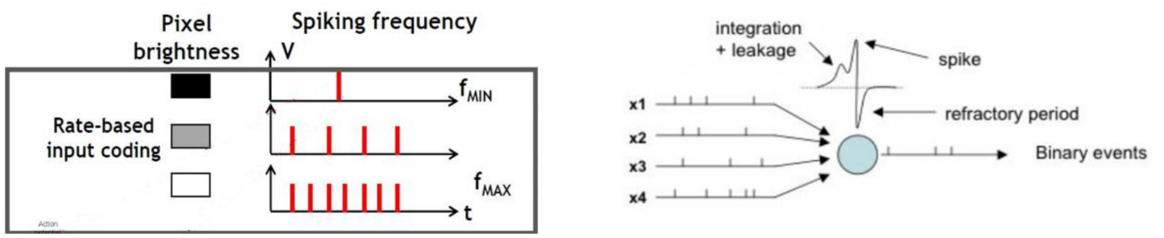


- **Main scientific questions**:
  1. How to code efficiently information in spikes ?
  2. Define new neural models: How to train those networks ?
  3. How to capture event-based data ?

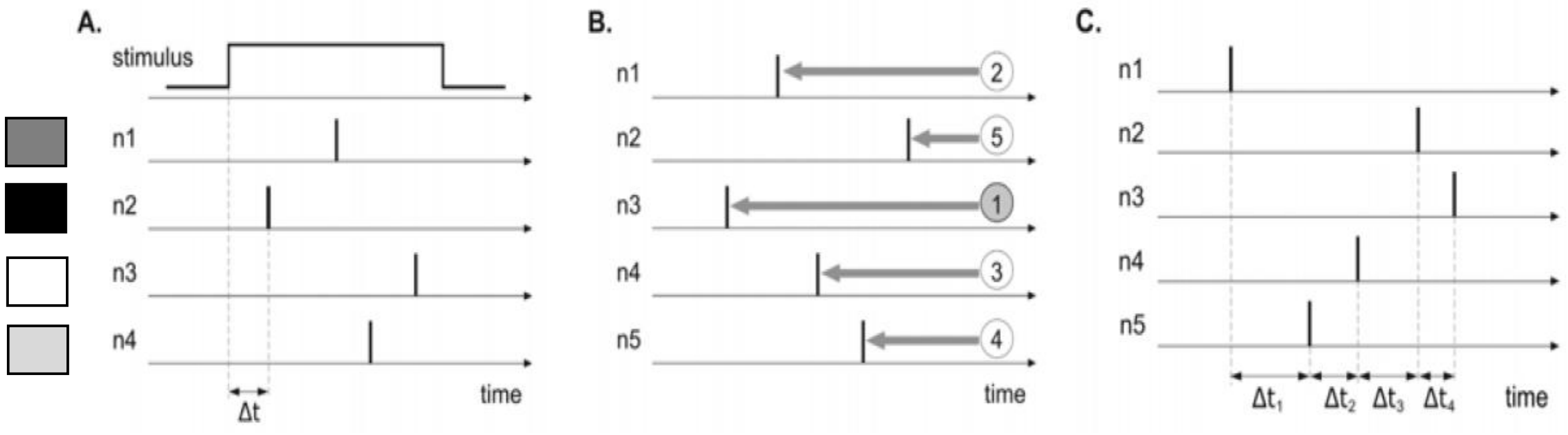


Question 1: Spike coding

- Rate coding
  - Rate coding find the average spiking frequency of a neuron over a certain timeframe



- Time coding
  - the neuron output is encoded in the temporal information of individual spikes.
    - time to first spike – TTFS (A),
    - rank order coding – ROC (B),
    - latency coding (C)



Ponulak, Filip & Kasiński, Andrzej. (2011). Introduction to spiking neural networks: Information processing, learning and applications. Acta neurobiologiae experimentalis. 71. 409-33.

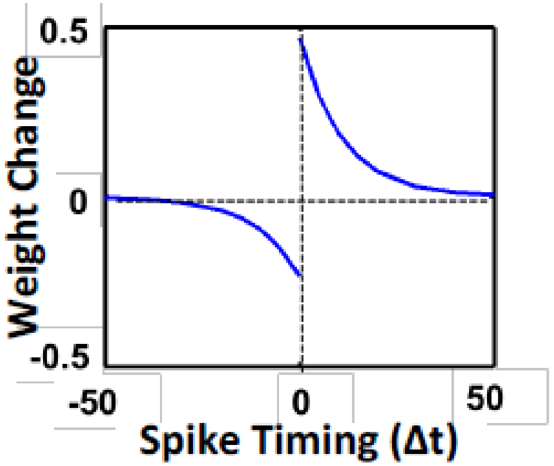
Question 2: Training spiking neural networks

1
STDP

2
Conversion

3
Spike
Backprop

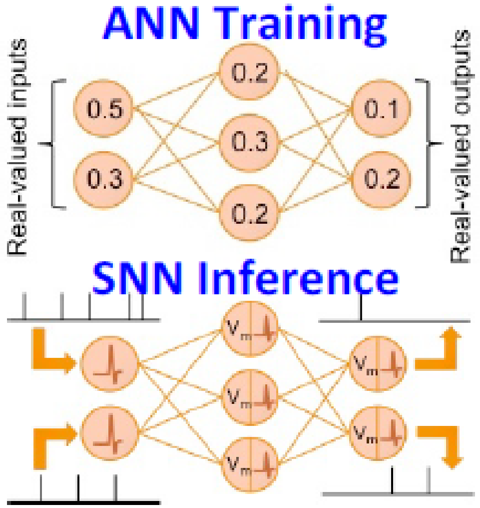
STDP Learning



Pros: Unsupervised local learning

Cons: Limited accuracy and shallow networks

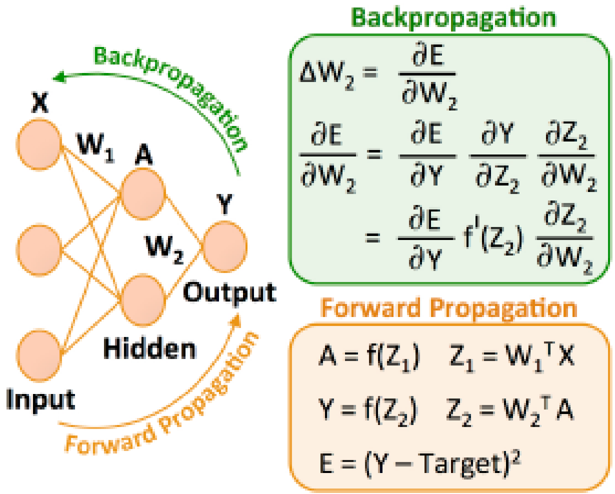
ANN-SNN Conversion



Pros: Takes advantage of standard ANN training

Cons: Incurs higher inference latency

Backpropagation



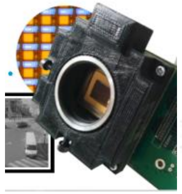
Pros: Competitive accuracy and lower inference latency

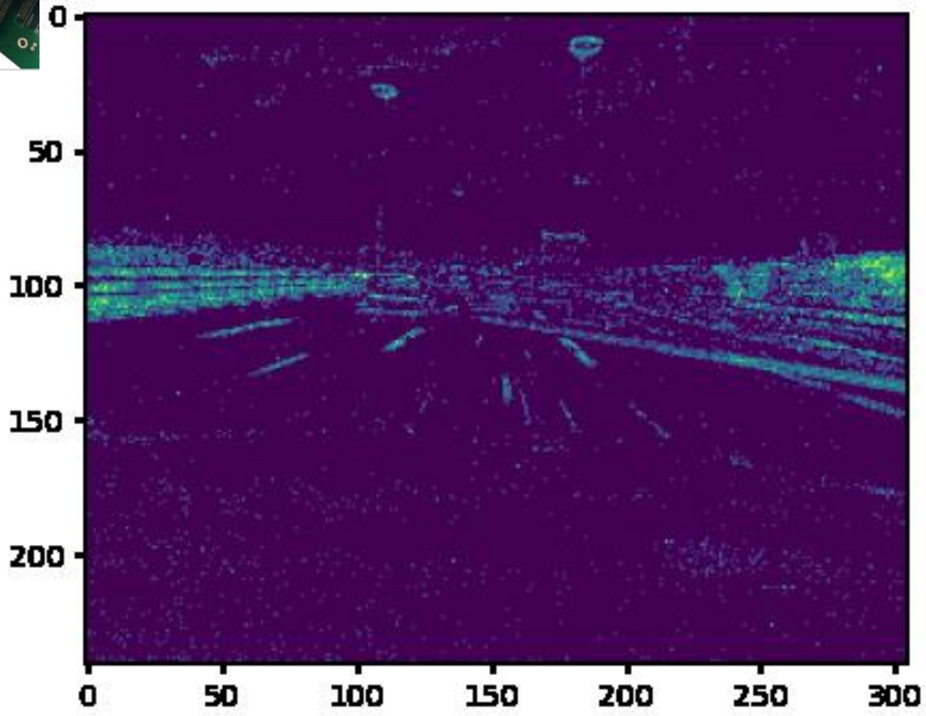
Cons: Higher training effort

G. Srinivasa, et. Al, TRAINING DEEP SPIKING NEURAL NETWORKS FOR ENERGY-EFFICIENT NEUROMORPHIC COMPUTING, ICASSP, 2020

Question 3. How to capture event-based data

DeepSee: industrial ANR Project





Event-based cameras (EBC)

Main technical reasons

- Event-based representation
- Sparse inputs
- High temporal sensibility (µs)
- High Dynamic Range (HDR)

Main scientific questions

- How to train SNN from event-based data ?
- How to take advantage of input sparsity ?

Application in image processing

Classification / Object Detection / Optical flow …

Main scientific results

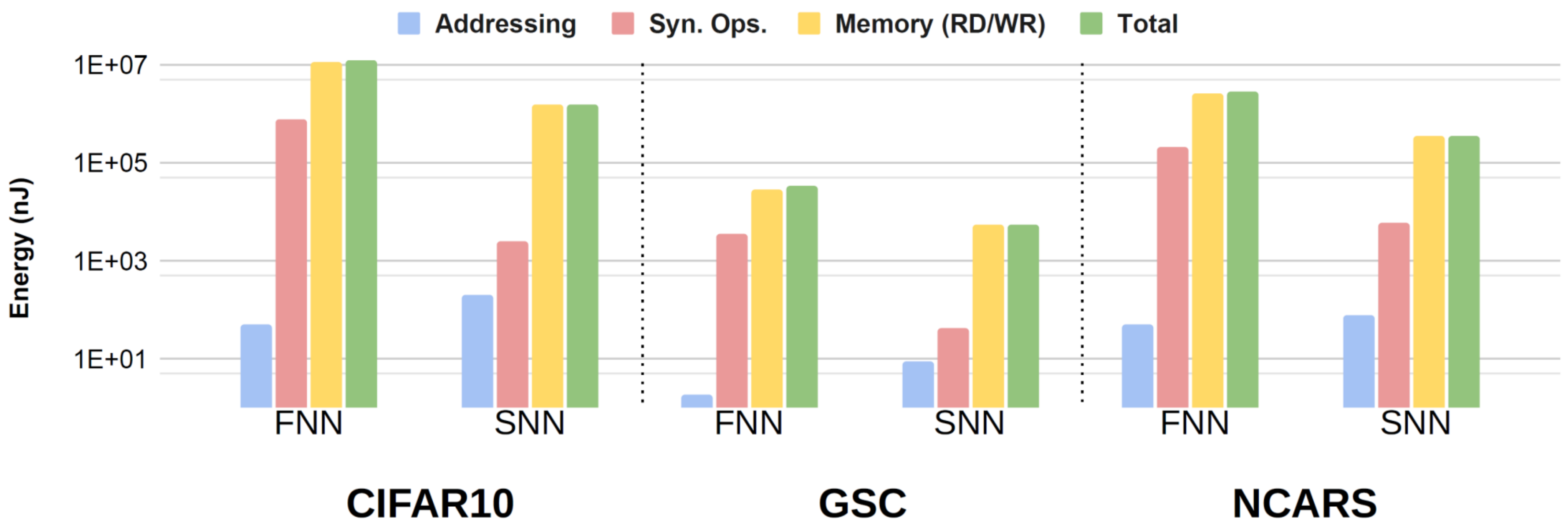
SNN with sparse convolutions [2]

First Spiking network for Object Detection on EBC [3]

[2] Learning from event cameras with sparse spiking convolutional networks, Loïc CORDONE, Sonia FERRANTE, Benoît Miramond; IJCNN 2021

[3] Object Detection with Spiking Neural Networks on Automotive Event Data, Loïc CORDONE, Benoît Miramond; IJCNN 2022

Comparison between CNN and SNN



	CIFAR 10	GSC	NCARS
Energy consumption reduction (ASIC 45 nm)	7.85 x	6.25 x	8.02 x
Spike Rate (vs. CNN)	0.1	0.14	0.08

An Analytical Estimation of Spiking Neural Networks Energy Efficiency , Edgar Lemaire, Loïc Cordone, Andrea Castagnetti, Pierre-Emmanuel Novac, Jonathan Courtois and Benoît Miramond, 29th International Conference on Neural Information Processing (ICONIP 2022)

Conclusion

Conclusion

- The combination of Edge AI and sensors
  - makes AI to the contact of the physics of the real world
  - Addresses the question of the energy consuption reduction of AI

- **By bringing AI closer to the sensor, the goal is**
  - **To reduce the amount of data to communicate**
  - **To lower the global energy consumption of the digital infrastructure**
  - **To reduce latency for decising making (close or open loop)**

- Original approach and promising results on bio-inspired AI thanks to
  - Greater sparsity
  - Event-based processing (specific neuromorphic hardware)
  - Reduced power consumption
  - And a large amount of unexplored features in the brain

- Remaining challenges for
  - EdgeAI training
  - Neuromorphic architectures
  - Realistic application demonstrations

EdgeAI, let’s play !

The field of possibilities is only limited by your imagination

Plug

Train

Embed

Play

Repeat …

IDEX <u>**Sith**</u> project, F. Ferrero, L. Rodriguez, B. Miramond

« l'organisation, la chose organisée, l'action d'organiser, et le résultat sont inséparables ».

Paul Valéry

Questions ?

[LEAT Lab, eBrain group: https://leat.univ-cotedazur.fr/ebrain/](https://leat.univ-cotedazur.fr/ebrain/)