Session Traitements

NEUROMORPHIC CIRCUITS

Lorena ANGHEL

TIMA

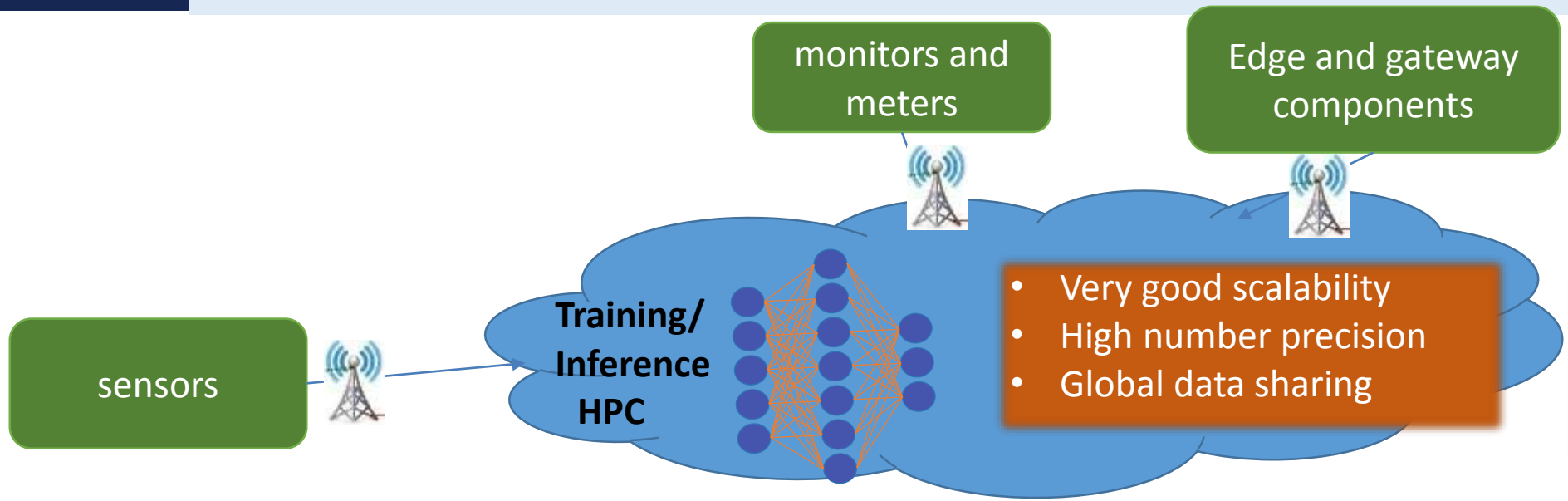
WORKSHOP
L'Intelligence Naturelle au coeur
des enjeux de l'Intelligence Artificielle

NeuroCoG
Univ. Grenoble Alpes

Pôle GRENOBLE Cognition

CNRS - Grenoble INP - UJF

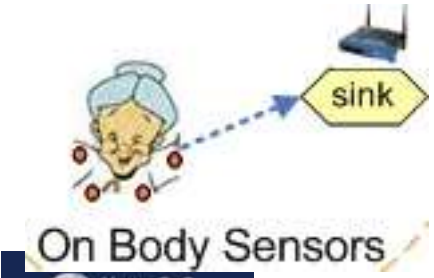
From Cloud DNN to Embedded Hardware DNN



- Huge memory capacity
- High maintenance and cooling cost
- High Communication latency

Deep Intelligent Solutions

data collection, predictions, diagnosis, classification, anomaly detection, correlations, pattern detection, decision making



e-health

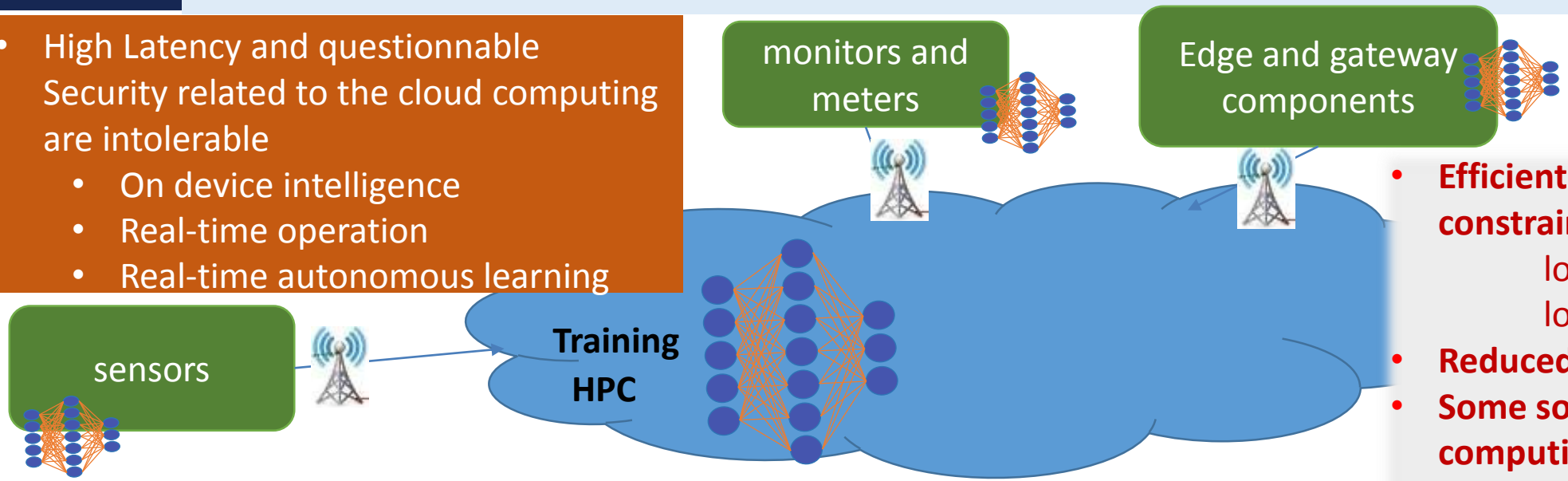
smart cities

Manufacturing

autonomous vehicles

From Cloud DNN to Embedded Hardware DNN

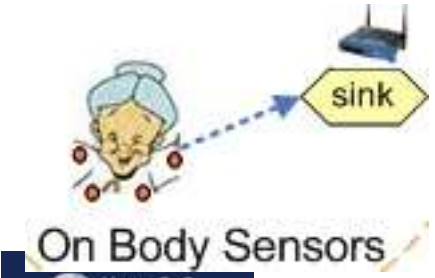
- High Latency and questionnable Security related to the cloud computing are intolerable
  - On device intelligence
  - Real-time operation
  - Real-time autonomous learning



- **Efficient deployment on HW constrained ressources**
  - low power operation
  - low memory capacity
- **Reduced number precision**
- **Some solutions need to change computing paradigm**
  - non Von-Neuman computation

Deep Intelligent Solutions

data collection, predictions, diagnosis, classification, anomaly detection, correlations, pattern detection, decision making



e-health

smart cities

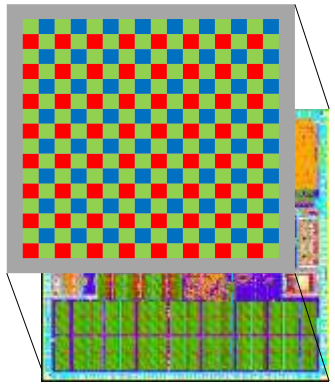
Manufacturing

autonomous vehicles

HW Solutions for Embedded Vision AI

How to design « smart » imagers, mixing acquisition and processing in a single device ?

Imager



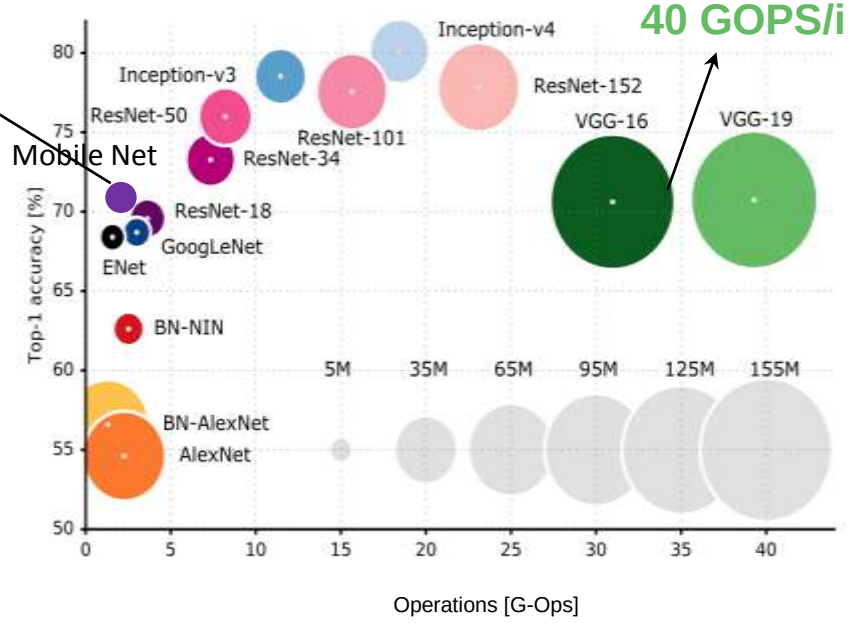
Processing

The Imager and Processing dies are 'stacked' to form a single device. The computing resources are limited :

- 1.5 MB to 12 MB memory
- Power << 100 mW

4.2 M parameters (32-bit)
569 MMAC/image

140M parameters (32-bit)
40 GOPS/image

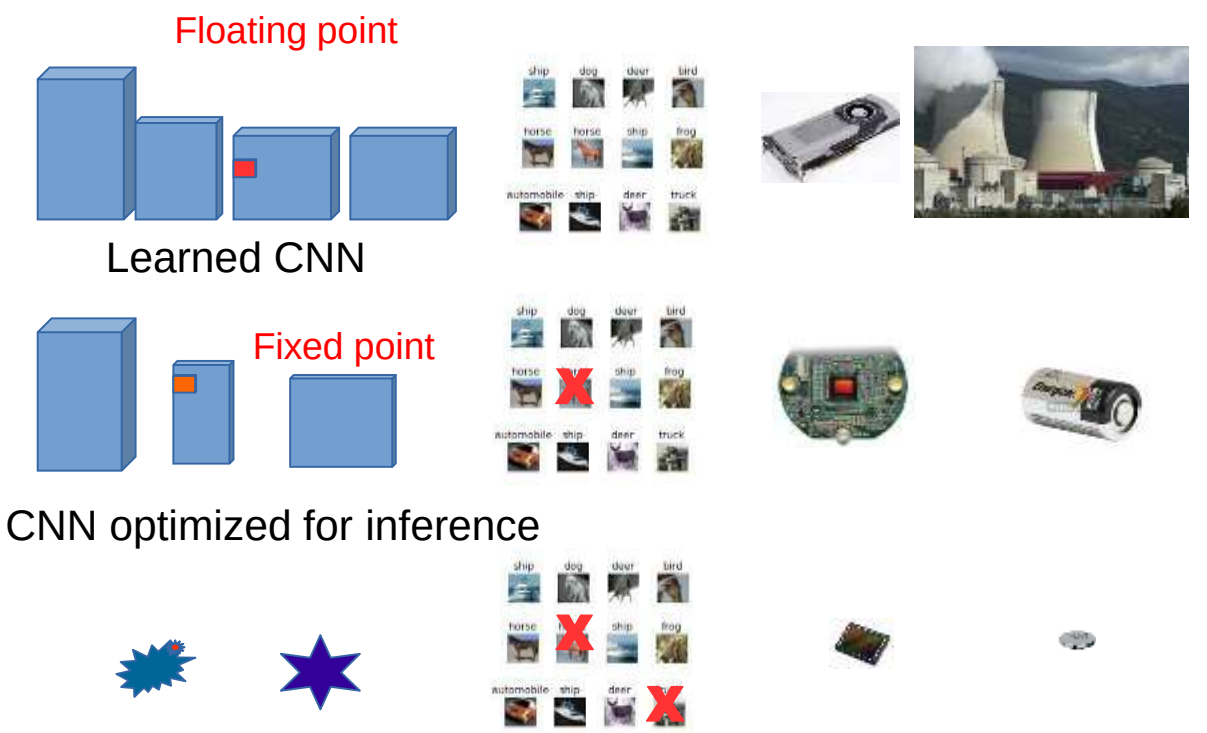


CNNs requires huge computing power and memory footprint (coefficients and temporary data)

HW Solutions for Embedded Vision AI

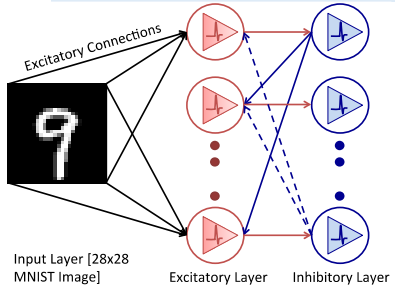
How to design « smart » imagers, mixing acquisition and processing in a single device ?

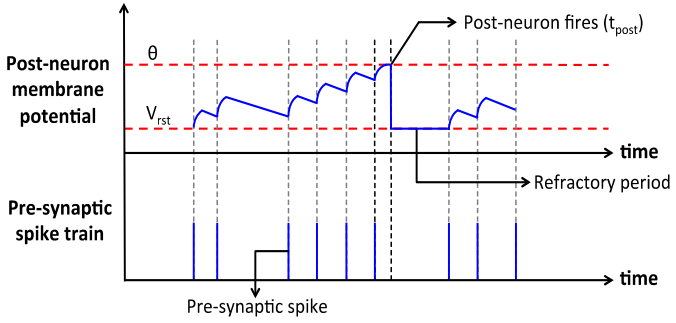
- Optimize for non-functional metrics : memory, computing power, power consumption
  - ➔ **impact on CNN architecture and learning**
- **CNN compression**/decompression
  - ➔ Save computing power
  - ➔ Save memory throughput and footprint
- Efficient **hardware** (specialized operators), and embedded **software** (customized processors) for CNN inference
  - ➔ Exploit all levels of **parallelism** and reduce overheads
  - ➔ Dedicated internal hardware support (decompression)

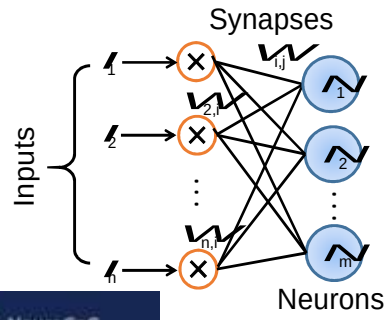


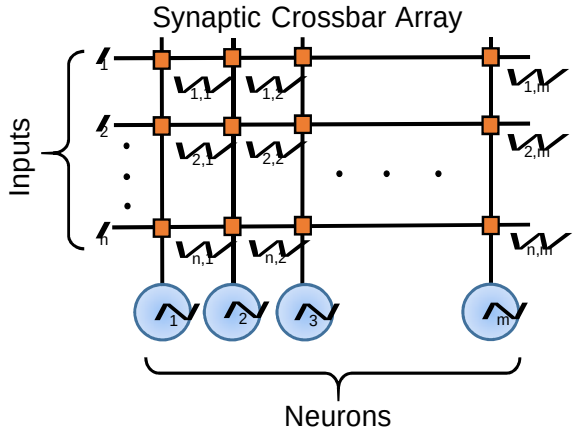
Compressed CNN **Data and coefficients transforms**

Brain-Inspired HW Neuromorphic Solutions









Spiking Neural Networks

- Mimik brain-like computation – use data encoded as spikes
- Bio-inspired learning rule – un-suppervised Spike Timing Dependent Plasticity
- Higher energy efficiency – event driven computation (20x less - fJ)

Applications

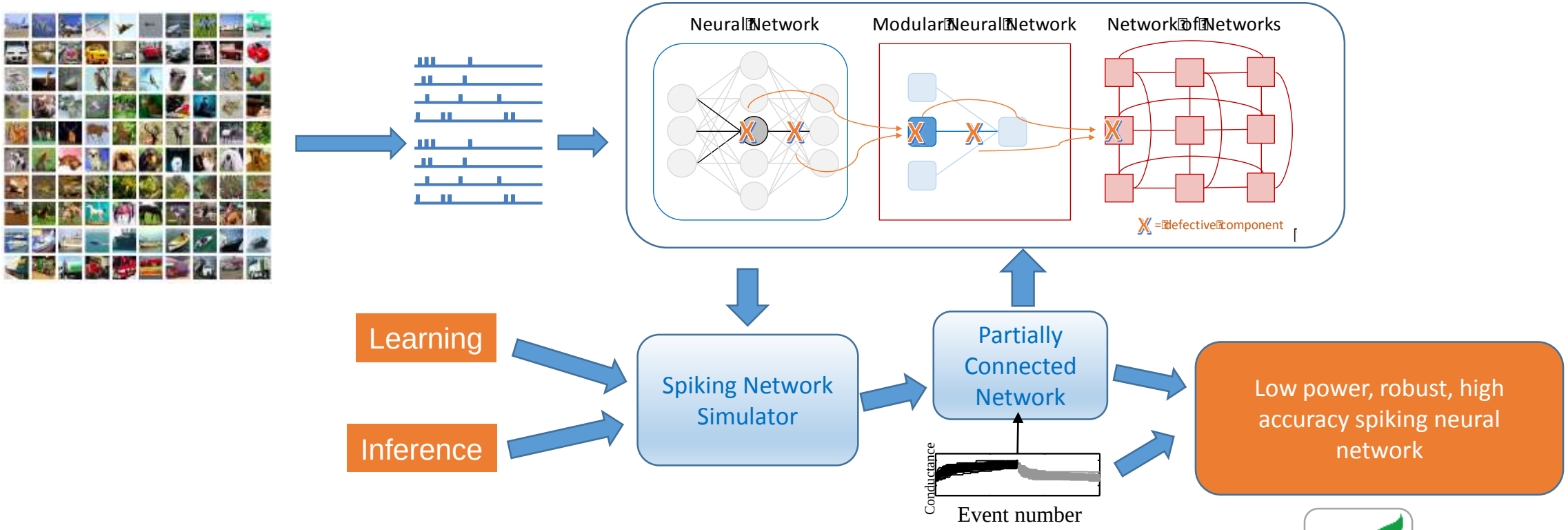
- Sparse data and multi modal information processing, belief, probabilistic and prediction engines, pattern recongnition
- IOT, mission critical systems, high security data, personalized medecine

Approaches

- Kernel Based computation – one kernel per NN layer
- Use hybrid CMOS-emerging devices (Magnetic/Resistive RAM)
  - getting closer to the stochastic neuron and inference and decision making resembling bayesian inference
- Use Computing-in-Memory Paradigm

Brain-Inspired HW Neuromorphic Solutions

Design of Ensemble Architectures based on massive network prunning defined by learning, defect, reliability and security attack strategies



Able to support unsupervised and local/distributed learning

Variability, defects, design constraints, test and security threats

Interactions and Collaborations

- **Applications**
  - Computer vision
  - Speech and language/text recognition
  - Information management from sensors (wireless sensing in e-health) - prediction
  - Biomimetic methods and applications in medical imaging and computer aided surgery
  - Autonomous vehicles (cars, robotics, etc)

ST Microelectronics
TIMC
LPNC

- **Technology and Architectures**
  - CMOS technology last generations
  - Emerging technologies: magnetic, resistive, nanowire
  - Non von-Neuman computing paradigms

ST Microelectronics,
SPINTEC Lab
CEA

- **New Algoritmic Approaches**
  - Statistical Modelling and Big Data approaches for better energy efficiency

Contacts

Lorena Anghel, lorena.anghel@univ.grenoble-alpes.fr
Stephane Mancini, stephane.mancini@univ.grenoble-alpes.fr
Ioana Vatajelu, ioana.vatajelu@univ.grenoble-alpes.fr