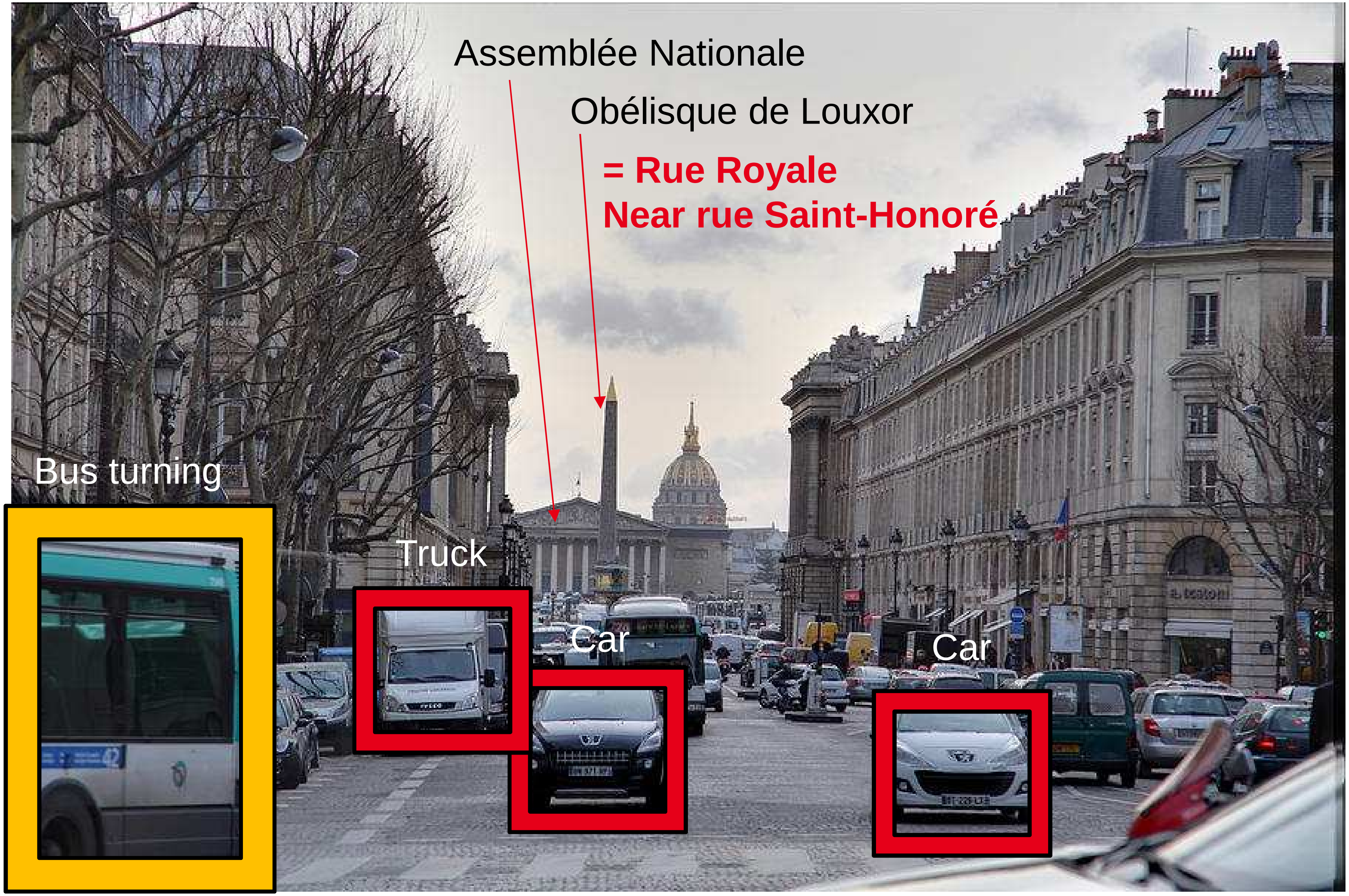
BRAIN-INSPIRED COMPUTING FOR ADVANCED IMAGE AND PATTERN RECOGNITION

Leti Devices Workshop | Marc DURANTON | December 4, 2016

IMAGE RECOGNITION: KEY FOR FUTURE APPLICATIONS



ImageNet: Classification

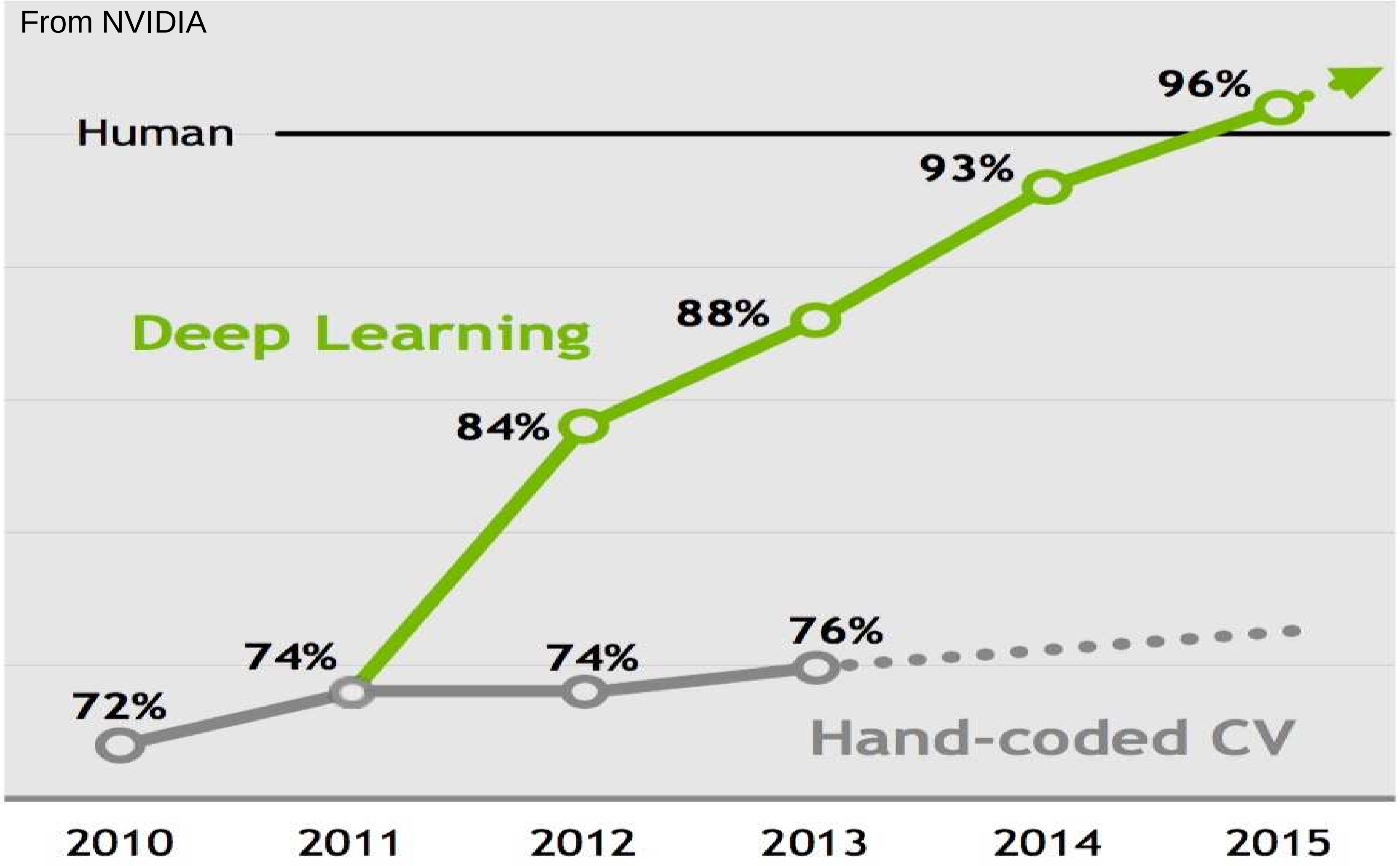
- Give the name of the dominant object in the image
- Top-5 error rates: if correct class is not in top 5, count as error
  - Black:ConvNet, Purple: no ConvNet

2012 Teams	%error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

2013 Teams	%error
Clarifai (NYU spinoff)	11.7
NUS (singapore)	12.9
Zeiler-Fergus (NYU)	13.5
A. Howard	13.5
OverFeat (NYU)	14.1
UvA (Amsterdam)	14.2
Adobe	15.2
VGG (Oxford)	15.2
VGG (Oxford)	23.0

2014 Teams	%error
GoogLeNet	6.6
VGG (Oxford)	7.3
MSRA	8.0
A. Howard	8.1
DeeperVision	9.5
NUS-BST	9.7
TTIC-ECP	10.2
XYZ	11.2
UvA	12.1

COMPETITION ON IMAGENET: SINCE 2012, CONVOLUTIONAL NEURAL NETWORKS (CNN) ARE LEADING!

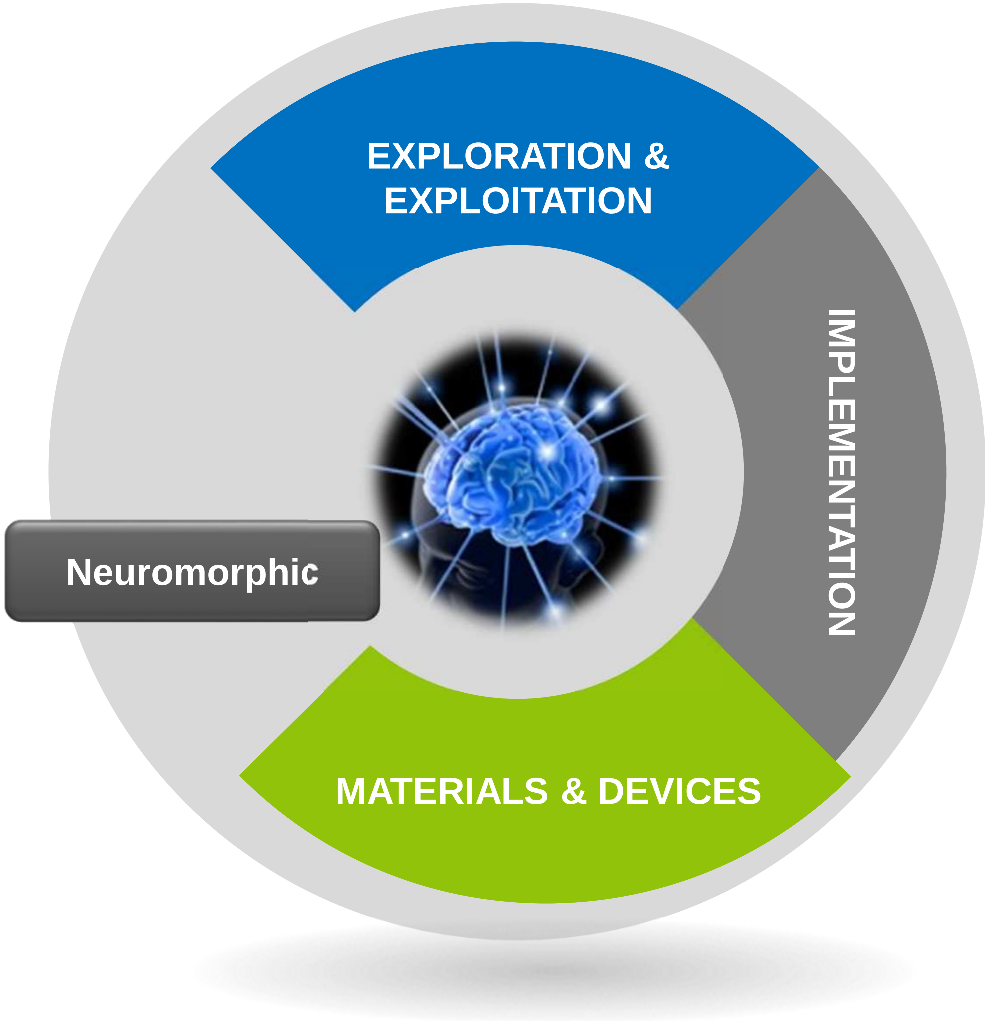


Deep Learning is Everywhere (ConvNets are Everywhere)

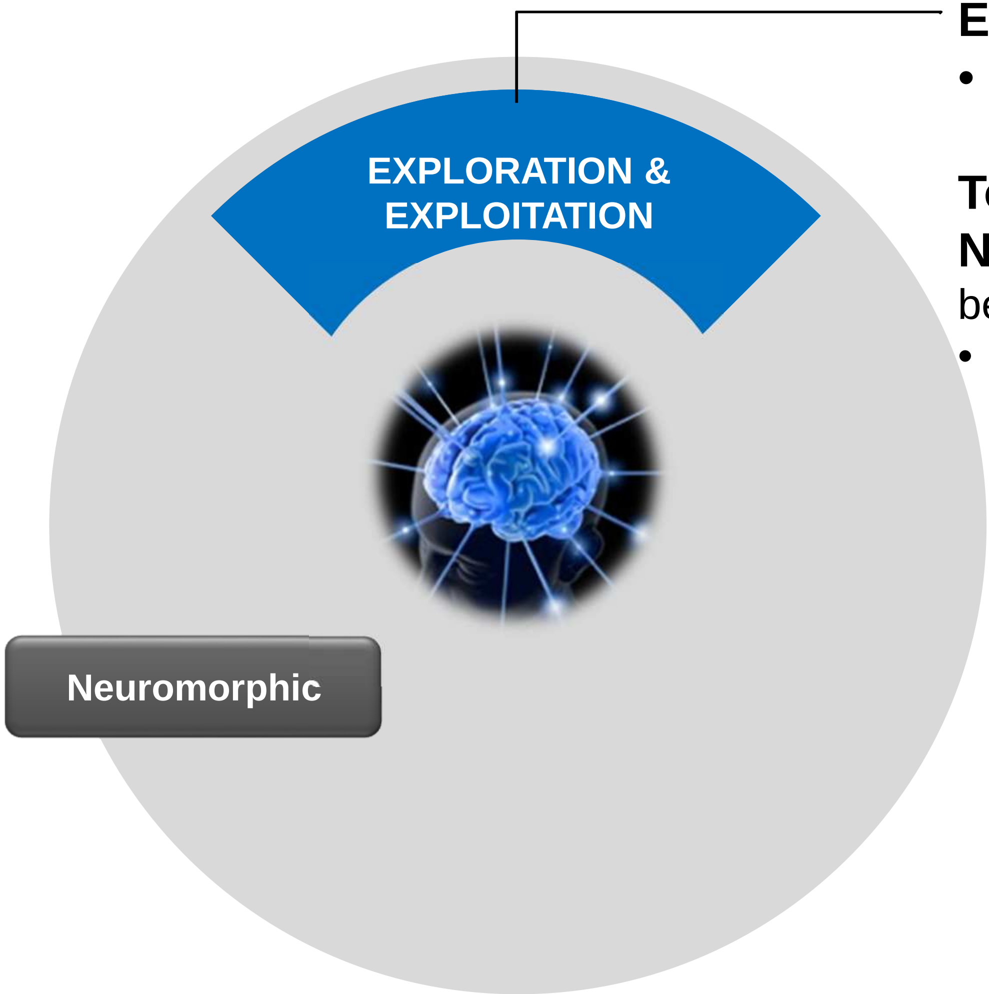
Y LeCun

- **Lots of applications at Facebook, Google, Microsoft, Baidu, Twitter, IBM…**
  - Image recognition for photo collection search
  - Image/Video Content filtering: spam, nudity, violence.
  - Search, Newsfeed ranking
- **People upload 800 million photos on Facebook every day**
  - (2 billion photos per day if we count Instagram, Messenger and Whatsapp)
- **Each photo on Facebook goes through two ConvNets within 2 seconds**
  - One for image recognition/tagging
  - One for face recognition (not activated in Europe).
- **Soon ConvNets will really be everywhere:**
  - self-driving cars, medical imaging, augemnted reality, mobile devices, smart cameras, robots, toys…..

DEEP LEARNING AND NEUROMORPHIC SYSTEMS AT LETI AND LIST



DEEP LEARNING AND NEUROMORPHIC SYSTEMS AT LETI AND LIST



Exploitation of Deep Neural Networks

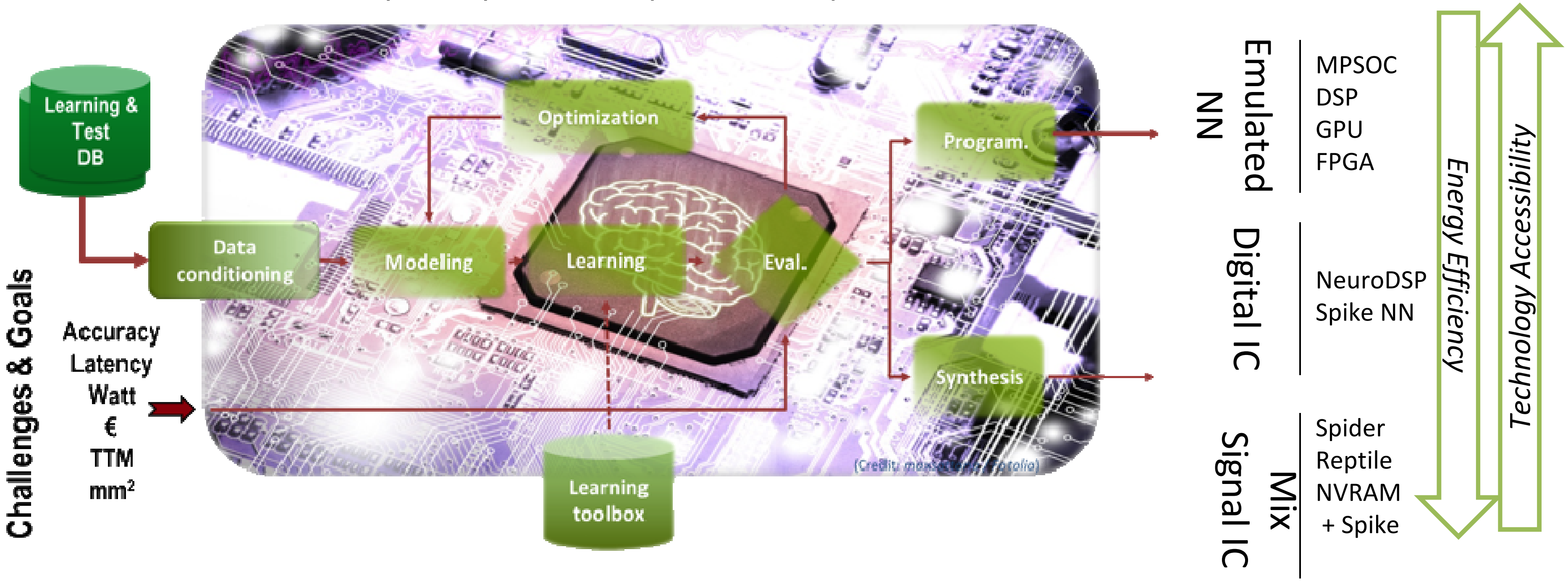
- Image recognition, annotation and indexing

Tools for fast and accurate Neural Network (NN) exploration & Architecture benchmarking: ***N2D2***

- Neural Network exploration (including with spike coding and new materials)

N2D2: PLATFORM FOR DEVELOPING DEEP NEURAL NETWORK APPLICATIONS

- **N2D2** is a **platform to design and generate deep neural network** (DNN) and to select the computing platform which fit best application needs
- Fast benchmarking of Components Off the Shelf and exports to dedicated ASIC:
  - Parallel processors (OpenCL, OpenMP)
  - GPU (OpenCL, Cuda, CuDNN)
  - FPGA (RTL, HLS)
  - Leti & List specific processors (like **P-Neuro**)

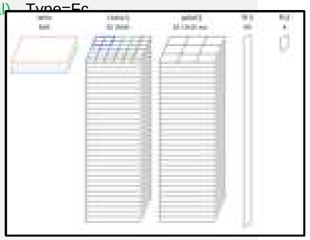


FAST AND ACCURATE NN EXPLORATION

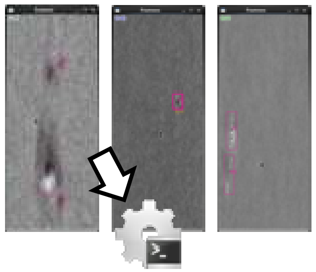
Automated architecture mapping and benchmarking tool flow

1) Deep network builder

```
; Environment
[env]
SizeX=8
SizeY=8
ConfigSection=env.config

[env.config]
ImageScale=0

; First layer (convolutional)
[conv1]
Input=env
Type=Conv
KernelWidth=3
KernelHeight=3
NbChannels=32
Stride=1

; Second layer (pooling)
[pool1]
Input=conv1
Type=Pool
PoolWidth=2
PoolHeight=2
NbChannels=32
Stride=2

; Third layer (fully connected)
[fc1]
Input=conv2
Type=Fc
```

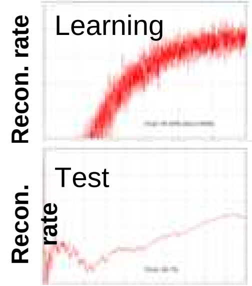


2) Learning a database

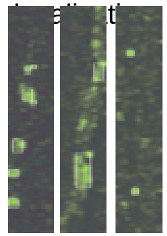


N2D2 software framework

3) Analysis of network performances

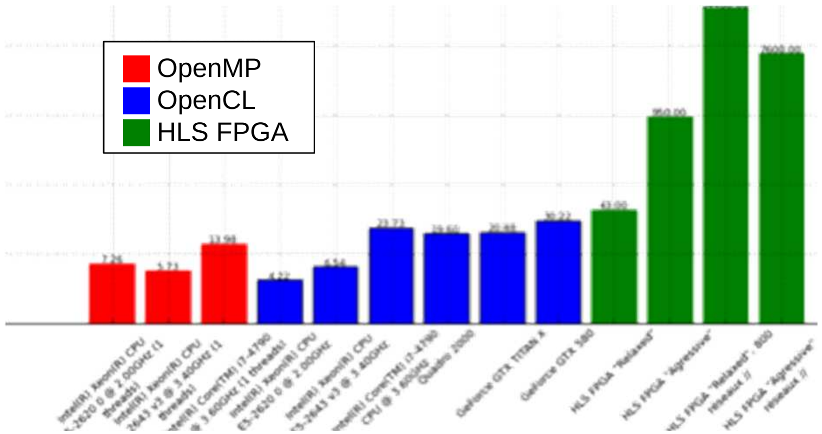


Output categories and localization



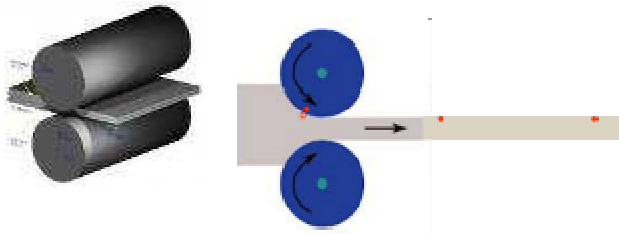
Inference phase

4) CPU, GPU and FPGA-based real-time implementation



➔ Wide targets range, perfs and power metrics

EXAMPLE OF INDUSTRIAL APPLICATION of N2D2: ROLLING MILL



CONSTRAINTS

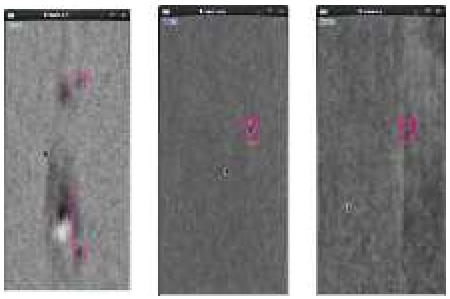
- Real time with very high throughput (20m/s)
- Tiny defect (~mm) with low contrast
- Complex environment (oil vapor, few space for inspection..)

SOLUTION

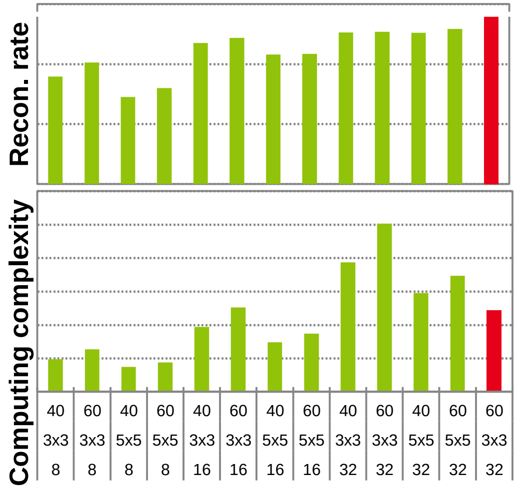
- Database labelling and Processing
- Fast NN topology Exploration
- Performance vs complexity analysis

➔ Real time performance achievable on FPGA (direct code generation)

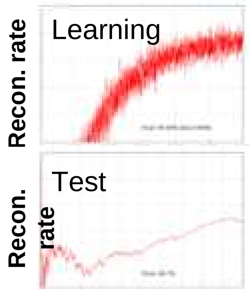
1) Defects labeling and visualization

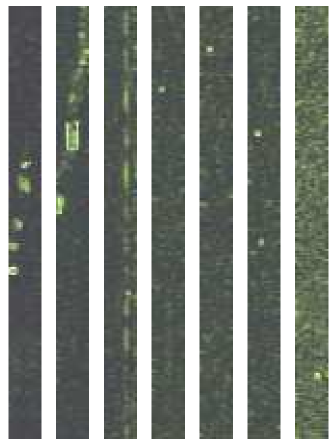


2) NN Exploration and benchmarking

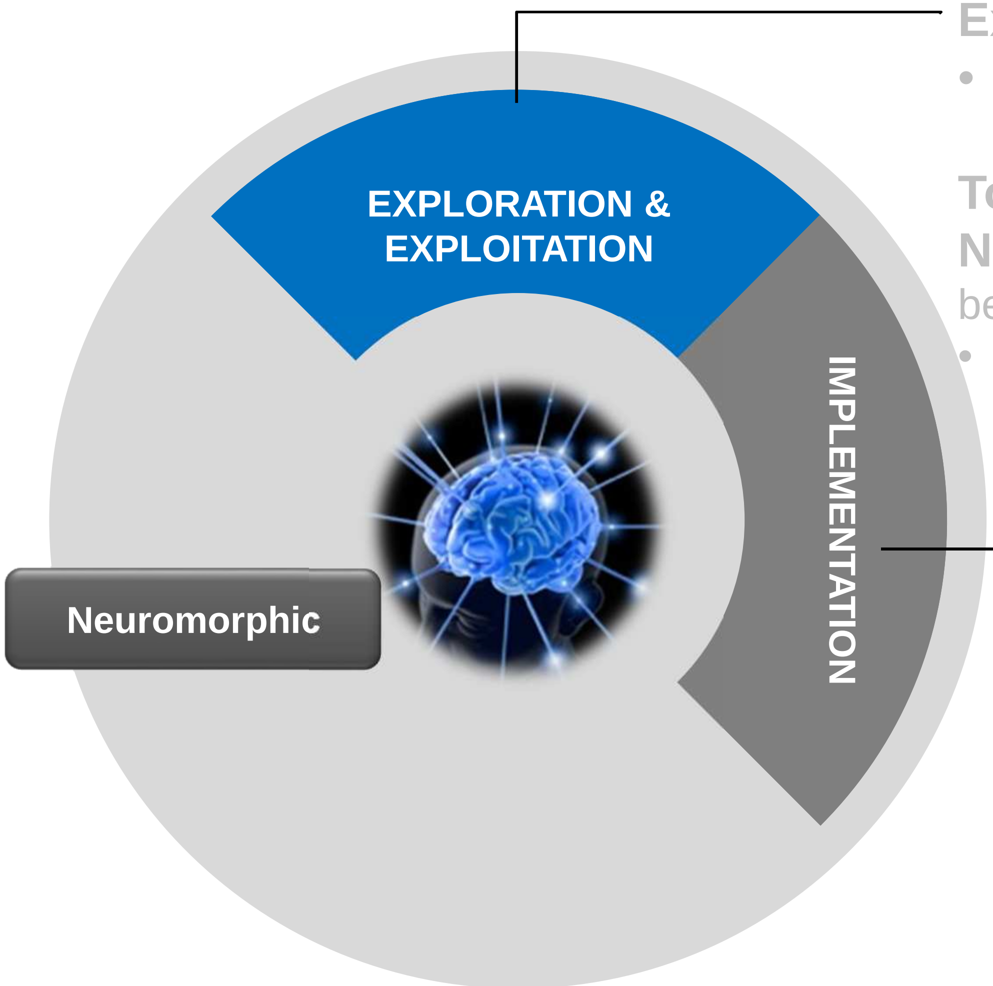


3) Defects identifications after NN learning





DEEP LEARNING AND NEUROMORPHIC SYSTEMS AT LETI AND LIST



Exploitation of Deep neural Networks

- Image recognition, annotation and indexing

Tools for fast and accurate Neural Network (NN) exploration & Architecture benchmarking: ***N2D2***

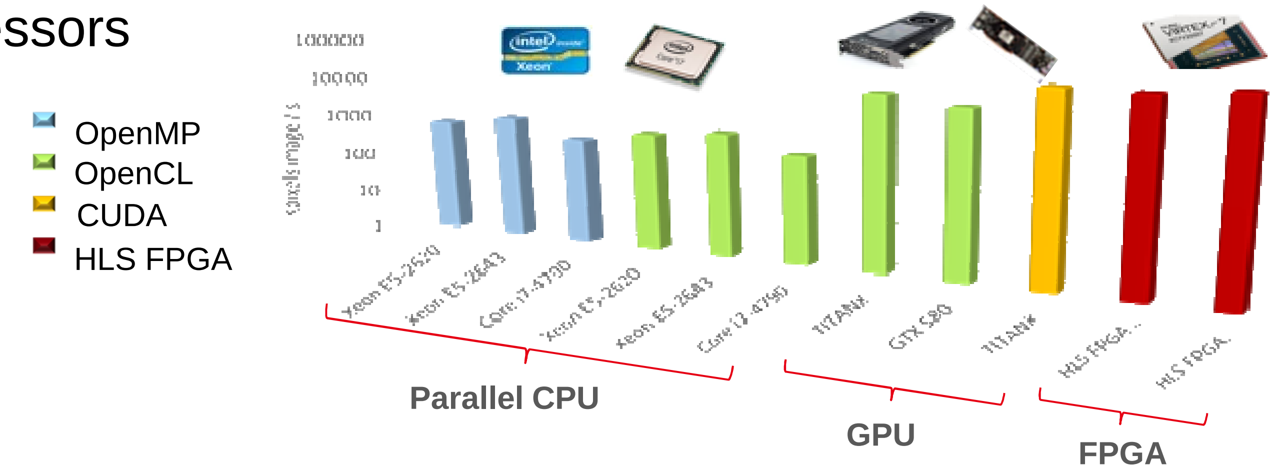
- Neural Network exploration (including with spike coding and new materials)

Diversity of implementations:

- Software solution / GPU
- Reconfigurable devices / FPGA
- Dedicated implementations
  - Full CMOS and binary coding: ***P-NEURO***
  - Full CMOS and “spike coding”
  - Using new materials

N2D2 and P-Neuro: complete solution for Deep Learning in smart nodes

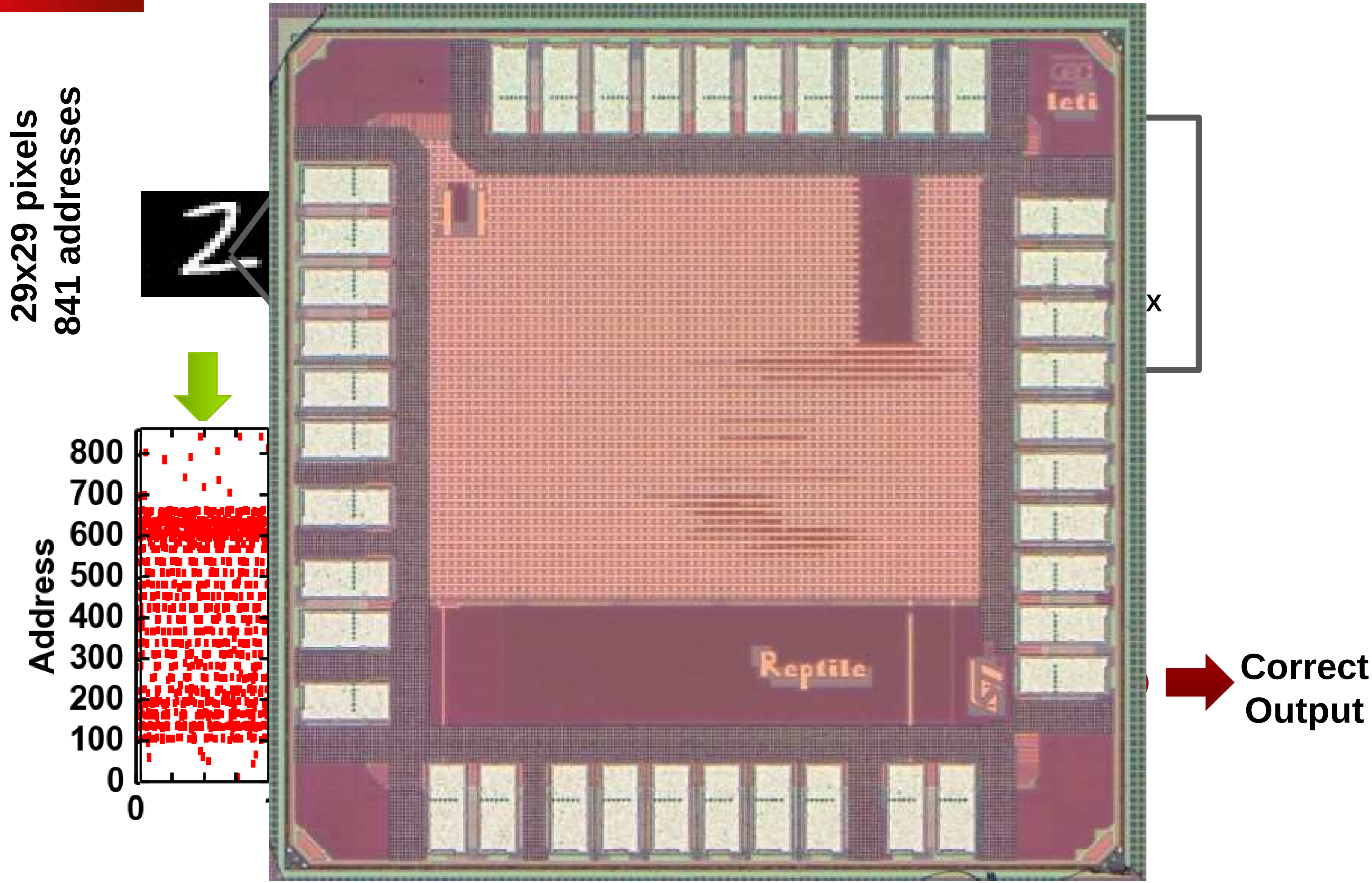
- Fast benchmarking of Components Off The Shelf:
  - Parallel processors
  - GPU
  - FPGA (HLS)



- Performance of ***P-Neuro*** neural network processing unit
  - Example on Faces extraction,
    - Database of 18000 images
  - Comparison of 5 different architectures
  - Focus on energy efficiency
  - Expected performance of ***P-Neuro***:
    - FDSOI 28nm, 1GHz
    - 1.8 TOPs/W, <0.5 mm^2 (4 cores)
    - Fully scalable from 1 to 1024 cores
    - Ready for integration in **smart nodes**

Target	Frequency	Energy efficiency
Quad ARM A7	900 MHz	380 images/W
Quad ARM A15	2000 MHz	350 images/W
Tegra K1	850 MHz	600 images/W
Intel I7	3400 MHz	160 images/W
P-Neuro (FPGA)	**100 MHz**	**2 000 images/W**
P-Neuro (ASIC)	**500 MHz**	**125 000 images/W**

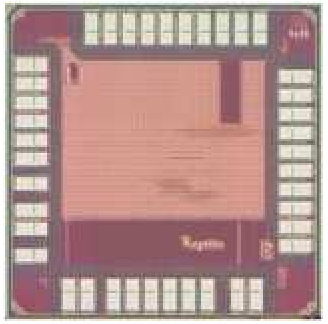
SPIKE-BASED CODING

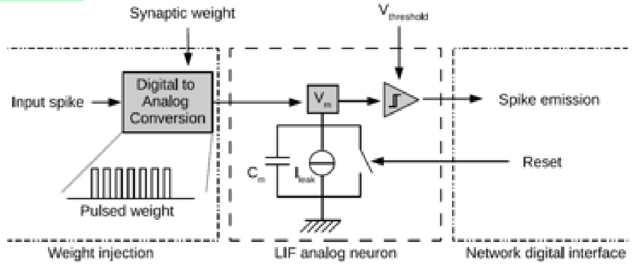


THE PROMISES OF SPIKE-CODING NN

- Reduced computing complexity and natural temporal and spatial parallelism
- Simple and efficient performance tunability capabilities
- Spiking NN best exploit NVMs such as RRAM, for massively parallel synaptic memory

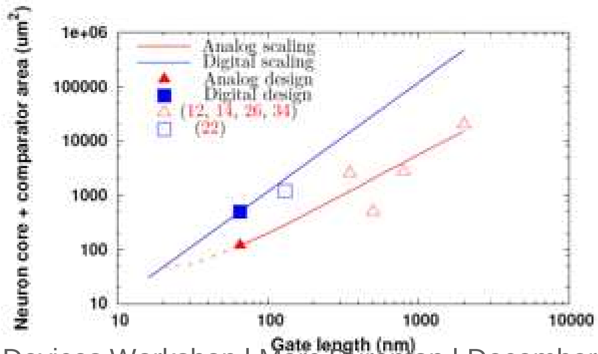
	Formal neurons	Spiking neurons
Base operation	- Multiply-Accumulate (MAC)	+ Accumulate only
Activation function	- Non-linear function	+ Simple threshold
Parallelism	- Spatial multiplexing	+ Spatial and temporal multiplexing



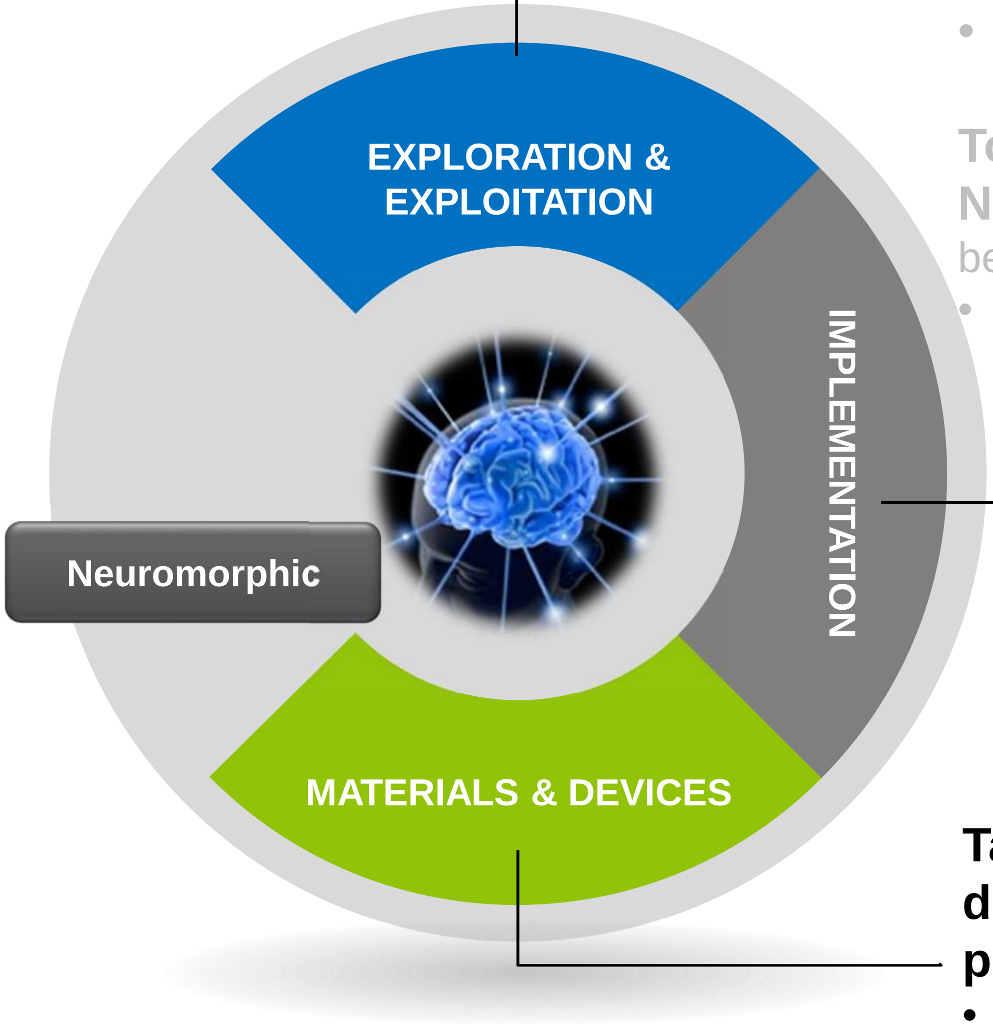


- **Two test chips implemented in 65nm**
  - Reptile: 3 tiles of 12 neurons
  - Spider: 25 tiles of 12 neurons

- **Advanced technology nodes**
  - Comparison of Analog and Digital neurons
  - Gain of Analog neuron (less area) reduces
    → Curves cross at 22nm node



DEEP LEARNING AND NEUROMORPHIC SYSTEMS AT LETI AND LIST



Exploitation of Deep neural Networks

- Image recognition, annotation and indexing

Tools for fast and accurate Neural Network (NN) exploration & Architecture benchmarking: ***N2D2***

- Neural Network exploration (including with spike coding and new materials)

Diversity of implementations:

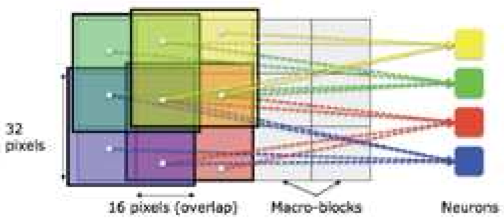
- Software solution / GPU
- Reconfigurable devices / FPGA
- Dedicated implementations
  - Full CMOS and binary coding: ***P-NEURO***
  - Full CMOS and “spike coding”
  - Using new materials

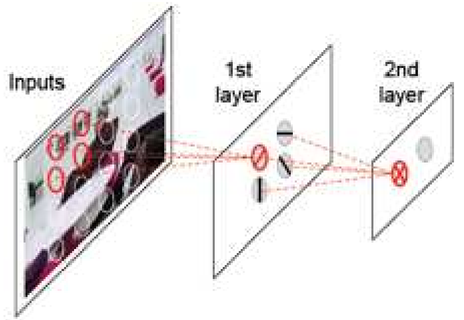
Take full advantage of advanced devices to break the density and power issues:

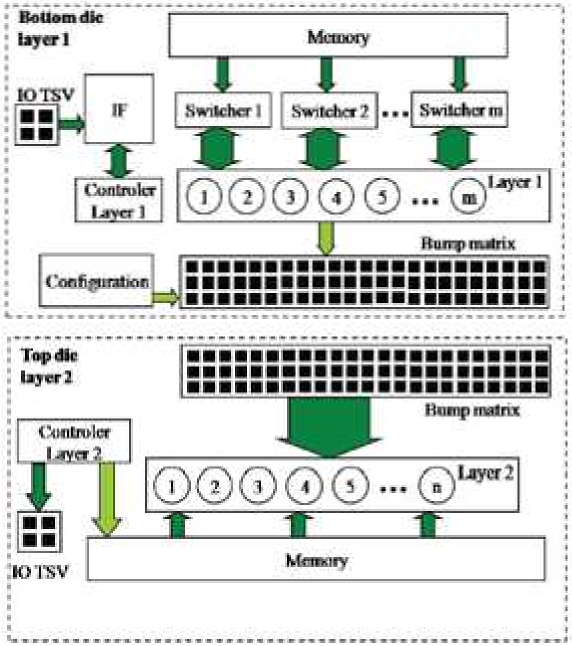
- 3D integration, CoolCube^{TM}.
- RRAM, PCM and new devices,

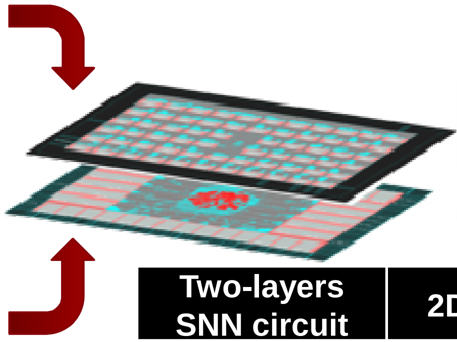
3D SPIKING NEURAL NETWORK

- **Neural Networks**
  - Naturally 3D for 2D inputs, layers optimally distributed in stacked dies
  - Vertical connections between layers: minimizes interconnect length, avoid routing congestion
- **NEMESIS 3D two-layers SNN test chip**
  - 1st layer: 48 macro-block neurons, 1024 synapses per neuron (49 152 total)
  - 2nd layer: 50 fully connected neurons, 2 400 synapses

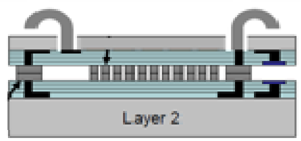








Nemesis Test Chip
ALTIS 130nm
CuCu bonding

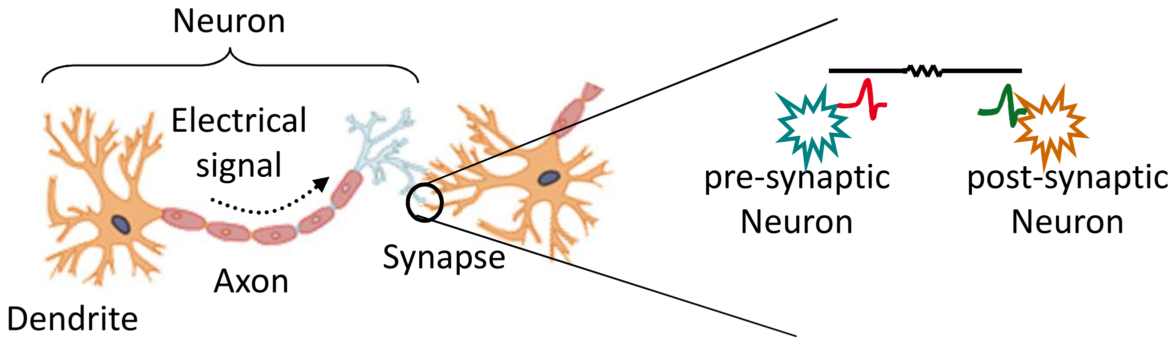


Two-layers SNN circuit	2D	3D
Total area (mm²)	7,97	3,63 (-54%)
Power (mW)	428	354 (-17%)
Critical path (ns)	9,00	6,63 (-26%)

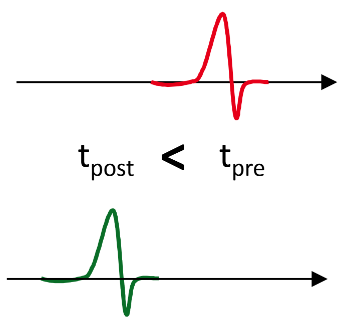
[B. Belhadj, R. Heliot, P. Vivet, CASSES'2014]

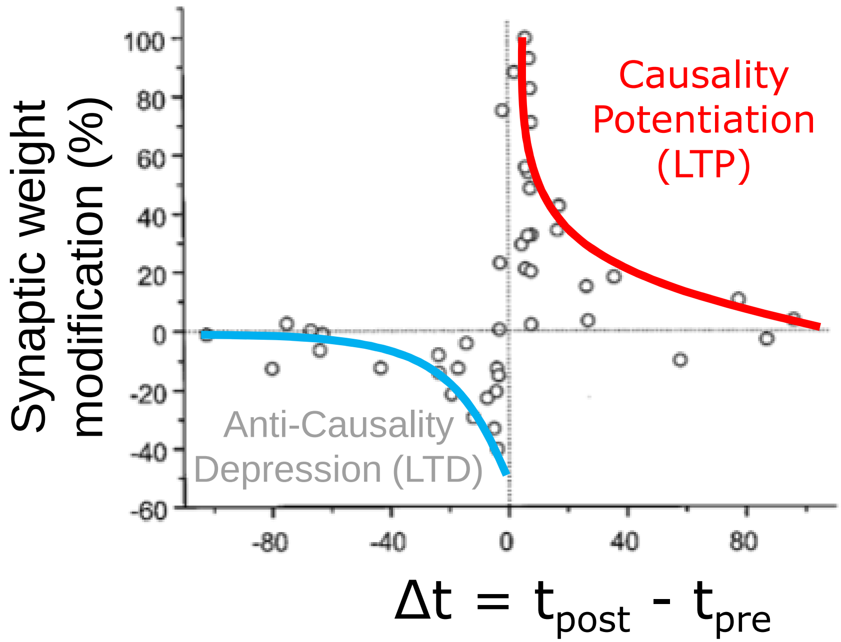
➔ 3D offers 2x better total area and 25% better power efficiency vs 2D

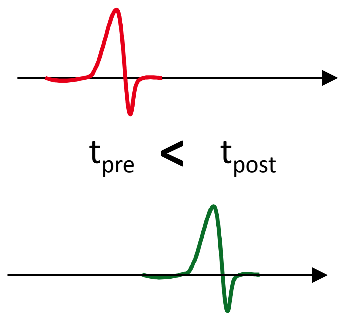
LEARNING FROM NEUROSCIENCE: A STDP (SPIKE TIMING DEPENDENT PLASTICITY) PRIMER



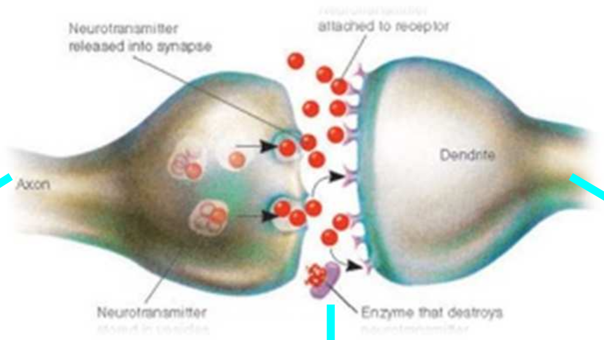
STDP = correlation detector
➔ Possible learning model of the brain?







NEW ELEMENT: RRAM AS SYNAPSES



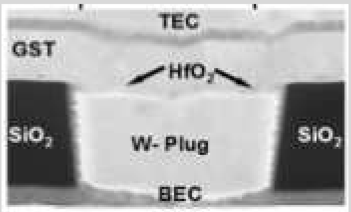
Thermal effect

Electrochemical effect

Electronic effect
oxygen vacancies

PCM

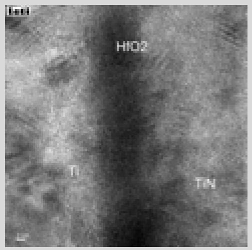
GST
GeTe
GST + HfO_2



M.Suri, et. al, IEDM 2011
M.Suri, et. al, IMW 2012 , JAP 2012
O.Bichler et al. IEEE TED 2012
M.Suri et al., EPCOS 2013
D.Garbin et al., IEEE Nano 2013

OXRAM

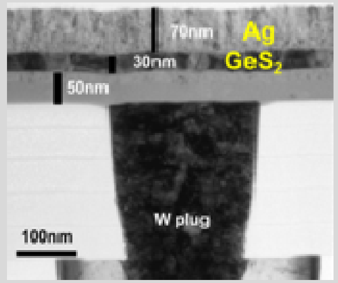
TiN/HfO_2/Ti/TiN



D.Garbin et al. IEDM 2014
D.Garbin et al., IEEE TED 2015

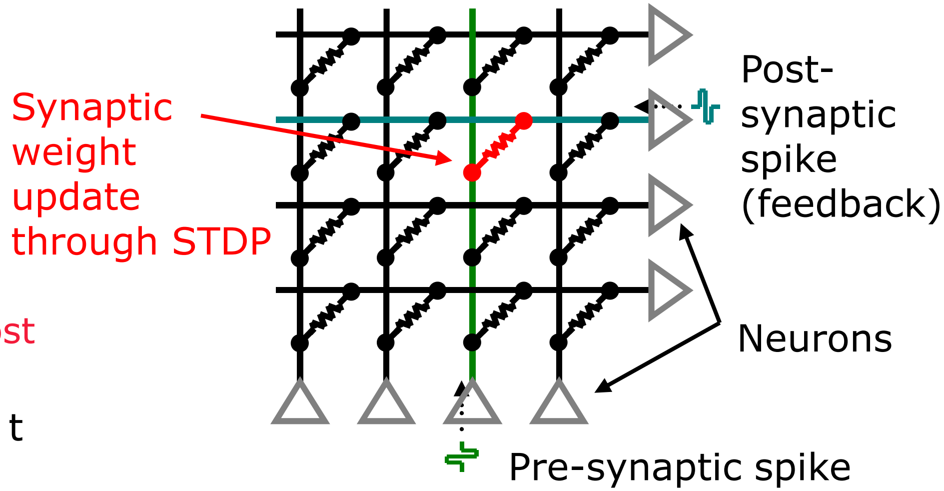
CBRAM

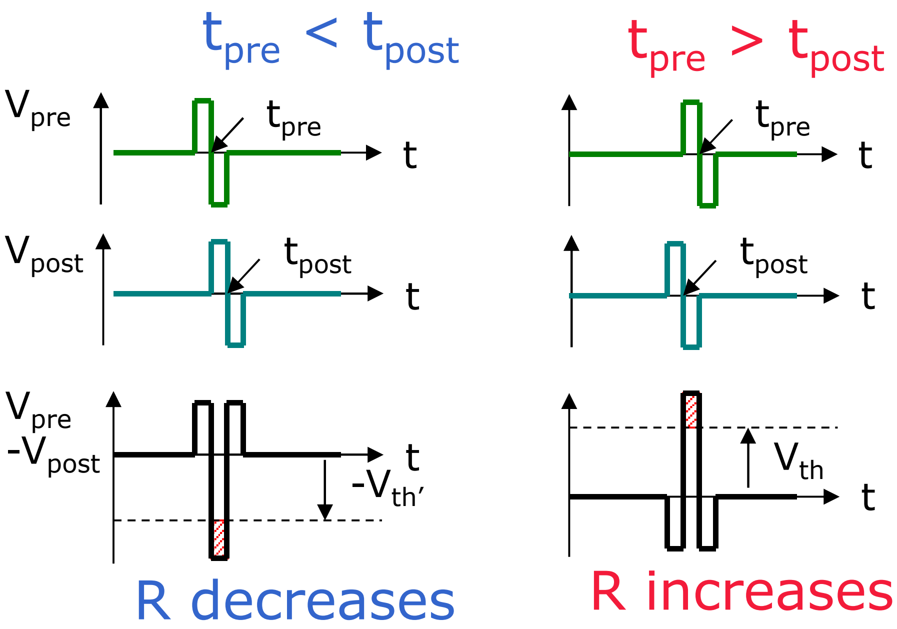
Ag / GeS_2

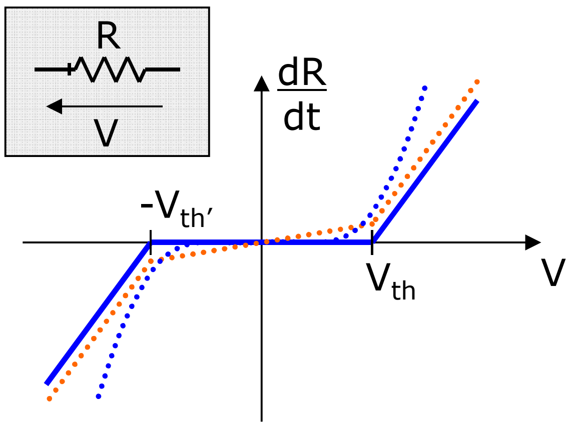


PRINCIPLE CROSSBARS OF MEMRISTORS

First Proposed by Snider(1)

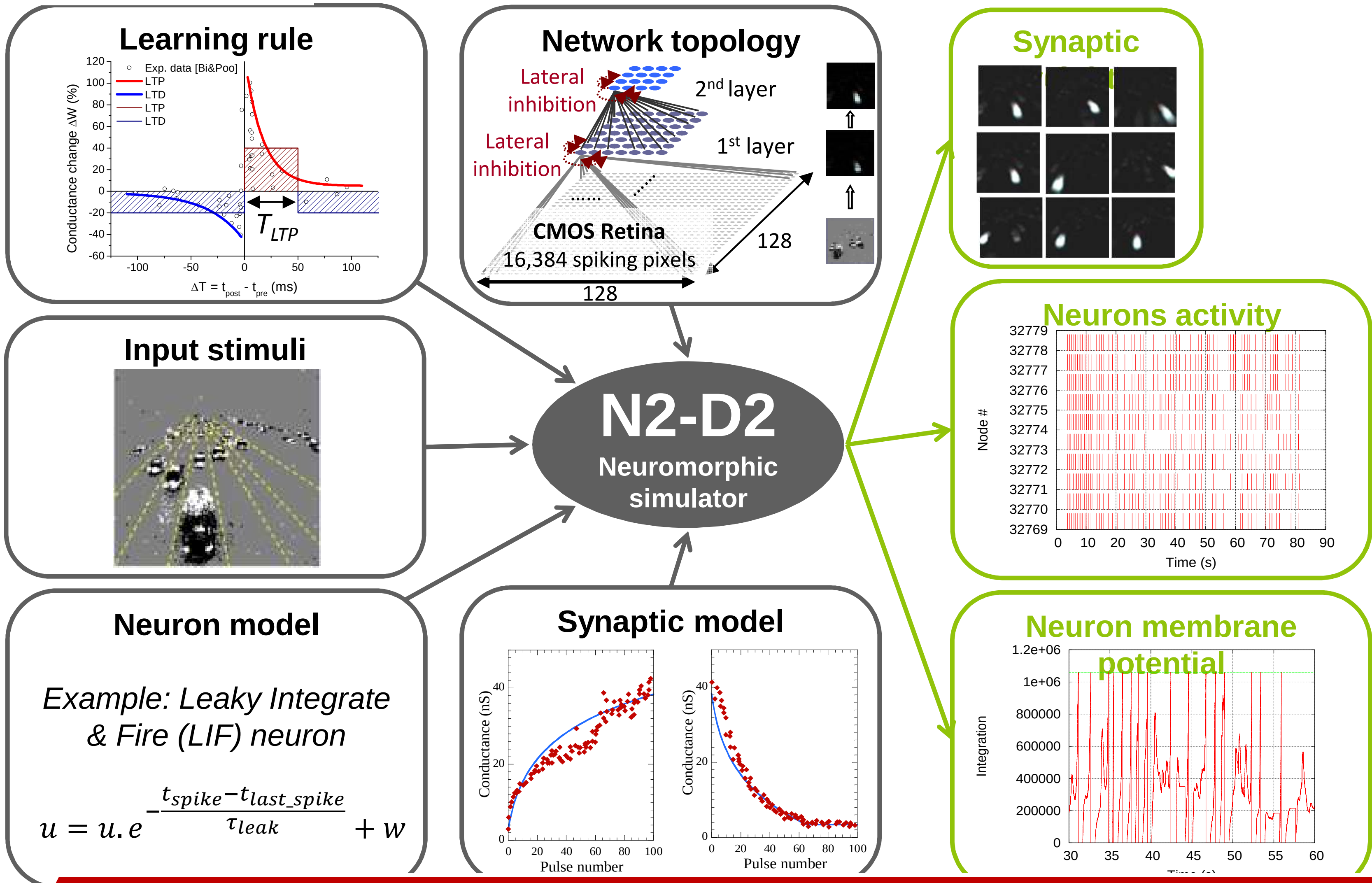






1. G. Snider, *Nanoscale Architectures*, 2008
2. B. Linares-Barranco et al, *Nature Precedings*, 2009

BIO-INSPIRED MODELS EXPLORATION



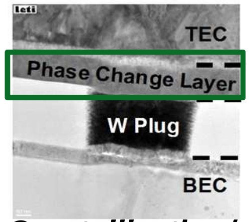
➔ Complete tool flow for bio-inspired synapses, neurons and learning rules network simulations

[O. Bichler et al., NanoArch'2014]

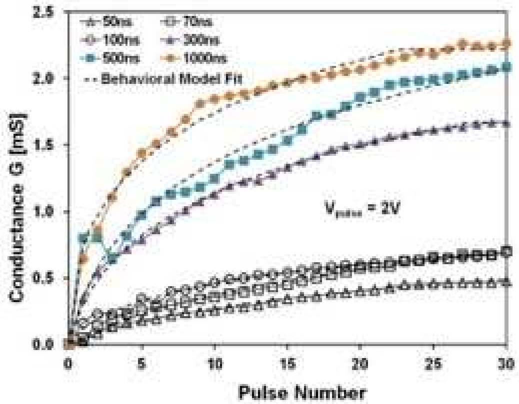
NVM SYNAPSES IMPLEMENTATIONS

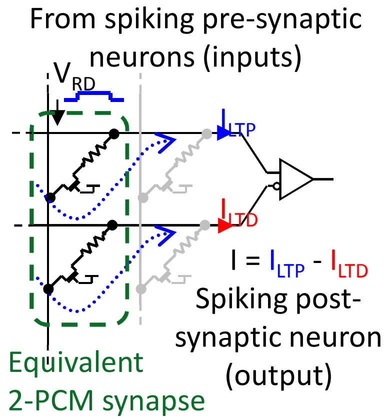
2-PCM synapses for unsupervised cars trajectories extraction

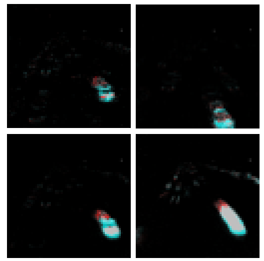
PCM



Crystallization/ Amorphization



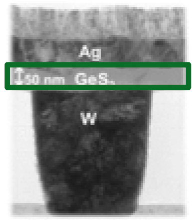




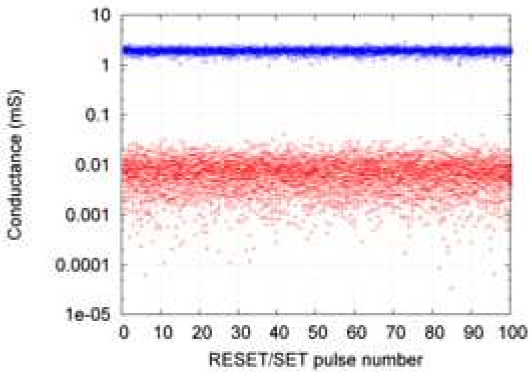
[O. Bichler et al., Electron Devices, IEEE Transactions on, 2012]

CBRAM binary synapses for unsupervised MNIST handwritten digits classification with stochastic learning

CBRAM



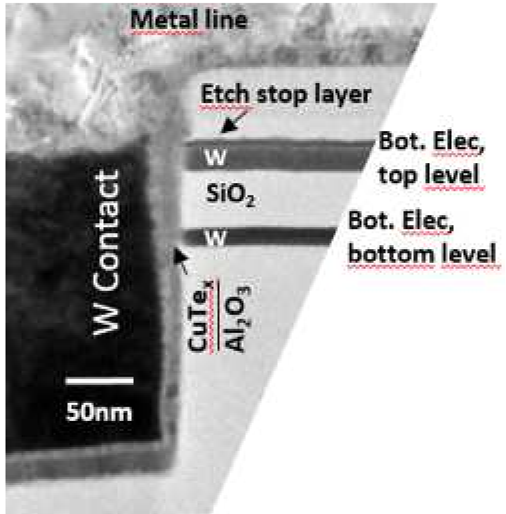
Forming/Dissolution of conductive filament



[M. Suri et al., IEDM, 2012]

EXAMPLE OF ON-GOING INVESTIGATIONS: VRRAM FOR NEUROMORPHIC APPLICATIONS

- **Investigation of VRRAM based on CBRAM stack**
  - 2 levels (proof of concept)
  - 16 levels (goal)
  - 1 select transistor per level (proof of concept)
  - Integrated selector (goal)
  - CBRAM most suitable R for neuromorphic
  - OxRAM also analysed



- **Design: support development for VRRAM**
  - **<u>High Density</u>**: Estimate the maximum size of a VRRAM-based array supposing to have an integrated selector [E. Cha, ISCAS 2014]
  - **<u>Neuromorphic</u>**: propose a circuit dimensioning for the neuromorphic approach presented at IEDM 2015 (1TnR pillar ~ Synapse, NO Selector)

AN EU COLLABORATIVE PROJECT: NEURAM3

Objective:

- Fabricate a chip implementing a neuromorphic architecture that supports state-of-the-art machine learning algorithms and spike-based learning mechanisms.

Features:

- 28nm FDSOI technology with RRAM synapses
- Ultra low power scalable and reconfigurable architecture
- 50x lower dissipation than digital equivalent
- TFT based scalable multichip architecture platform
- A technology to implement on-chip learning, using native adaptive characteristics of electronic synaptic elements

A NEW EU COLLABORATIVE PROJECT: NEURAM3

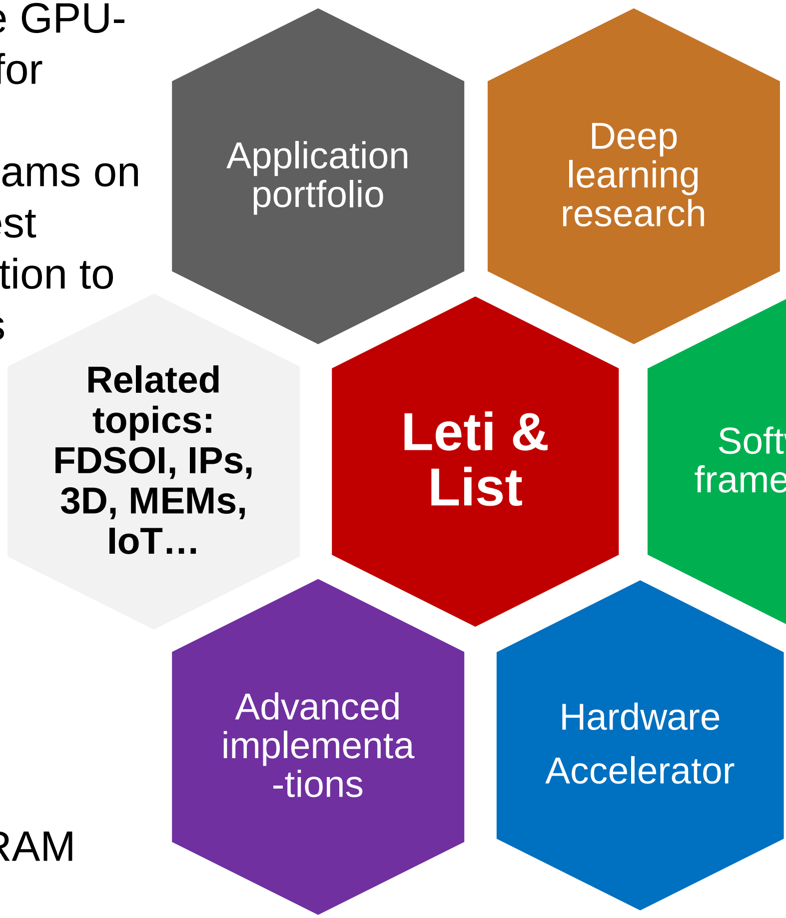
NeuRAM3

Participant no.	Organization name	Short name	Country
1 (Coordinator)	Commisariat a l'energie atomique et aux energies alternatives	CEA	France
2	Interuniversitair Micro-Electronica Centrum IMEC VZW	IMEC	Belgium
3	Stichting IMEC Nederland	IMEC-NL	Netherlands
4	IBM Research Gmbh	IBM	Switzerland
5	University of Zurich, Institute of Neuroinformatics	UZH	Switzerland
6	Agencia Estatal Consejo Superior de Investigaciones Cientificas, Instituto de Microelectronica de Sevilla	CSIC	Spain
7	Consiglio Nazionale delle Ricerche	CNR	Italy
8	Jacobs University Bremen	JAC	Germany
9	ST-Microelectronics S.A.	STM	France

LETI AND LIST ASSETS IN DEEP LEARNING

Summary of key points

- Large-scale database GPU-accelerated learning for CNN
- Among the leading teams on ImageClef2015 contest
- From scratch exploration to industrial applications



- Lead in bio-inspired STDP learning (IEDM’11,12,14)
- Formalized spike-coding for CNN, complete tool flow for co-simulation

- Complete framework with C, OpenCL, CUDA and HLS exports
- Complete tool flow for spike-coding DSP

- 2-PCMs synapse (patented) scheme (IEDM’15)
- Lead in SNN with RRAM devices (IEDM’14)

- Competitive reconfigurable architecture with P-Neuro
- Spike-coding DSP architecture
- Increased efficiency with 3D

Thank you for your attention

Leti, technology research institute
Commissariat à l'énergie atomique et aux énergies alternatives
Minatec Campus | 17 rue des Martyrs | 38054 Grenoble Cedex | France
www.leti.fr

INSTITUT CARNOT CEA LETI