LETI IS YOUR PARTNER FOR ARTIFICIAL INTELLIGENCE BASED SYSTEMS

ENABLING ARTIFICIAL INTELLIGENCE TECHNOLOGIES

Leti Innovation Days | June 28-29, 2017

Entering in
Human and machine collaboration era

Atos Codex

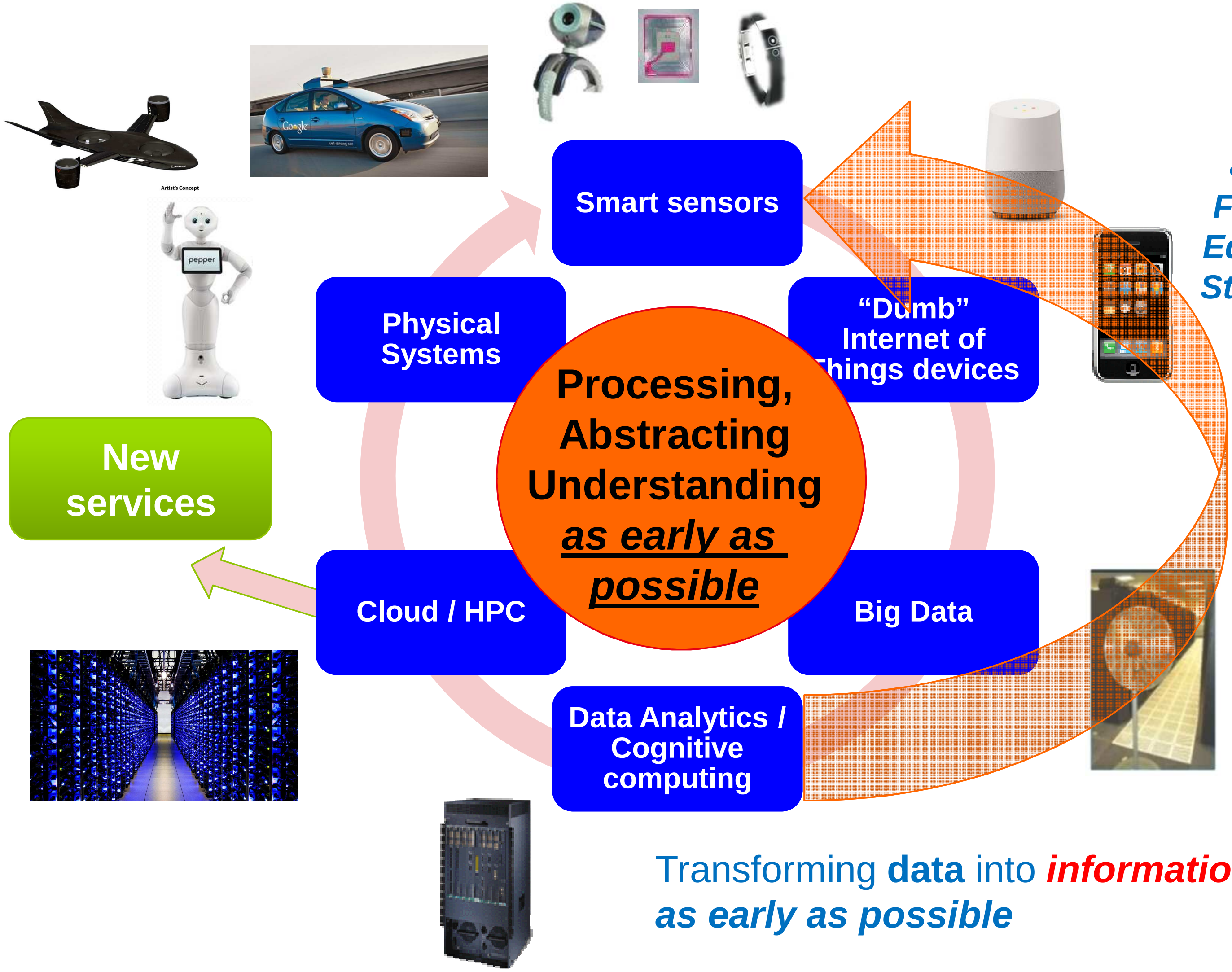
ENABLED BY ARTIFICIAL INTELLIGENCE (AND DEEP LEARNING)

- ***Artificial Intelligence*** **is changing the man-machine interaction – natural interfaces, ”intelligent” behavior**
  - Image and situation understanding
  - Voice recognition and synthesis
  - Data analysis
  - Decision taking
  - ….

AtoS Codex

IBM WATSON

COMPUTING DISTRIBUTION FOR ”COGNITIVE” SYSTEMS

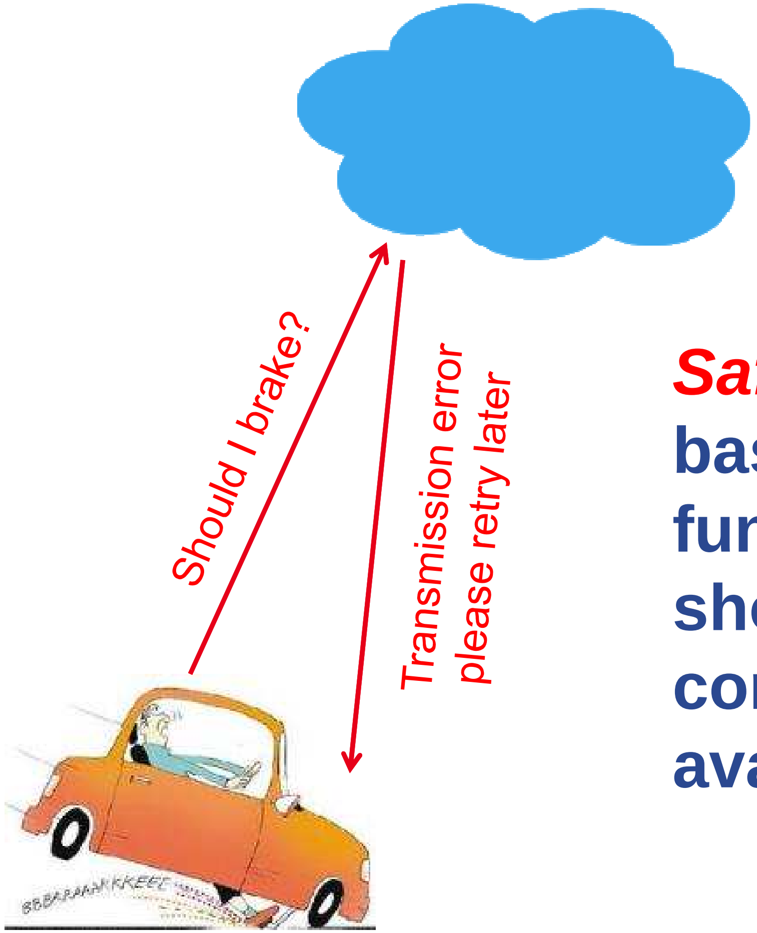


Artist's Concept
Smart sensors
Physical Systems
“Dumb” Internet of Things devices
Processing, Abstracting Understanding *as early as possible*
Cloud / HPC
Big Data
Data Analytics / Cognitive computing
New services

Real-time Embedded Intelligence at the edge: Fog computing Edge computing Stream analytics Fast data…

Transforming **data** into ***information***
as early as possible

EMBEDDED INTELLIGENCE NEEDS LOCAL HIGH-END COMPUTING



Should I brake?

Transmission error
please retry later

**System should be autonomous to
make good decisions in all conditions**

Safety **will impose that basic autonomous functions
should not rely on “always connected” or “always available”**

EMBEDDED INTELLIGENCE NEEDS LOCAL HIGH-END COMPUTING

Dumb sensors

Smart sensors: Streaming and distributed data analytics

Bandwidth will require more **local processing**

EMBEDDED INTELLIGENCE NEEDS LOCAL HIGH-END COMPUTING

Example: detecting elderly people falling in their home

Privacy will impose that some **processing should be done locally** and not be sent to the cloud.

CURRENT CONTRIBUTIONS OF LETI TO DEEP LEARNING

1) Provide tools and IPs (Hardware and software) for *fast and efficient development* of deep-learning techniques (mainly inference and data fusion) *at the edge* under constraints of:

- Performance
- Speed
- Cost
- Power consumption
- Choice of hardware
- Size

2) Provide *innovative technologies* for tomorrow’s unsupervised learning systems

N2D2: NEURAL NETWORK DESIGN & DEPLOYMENT

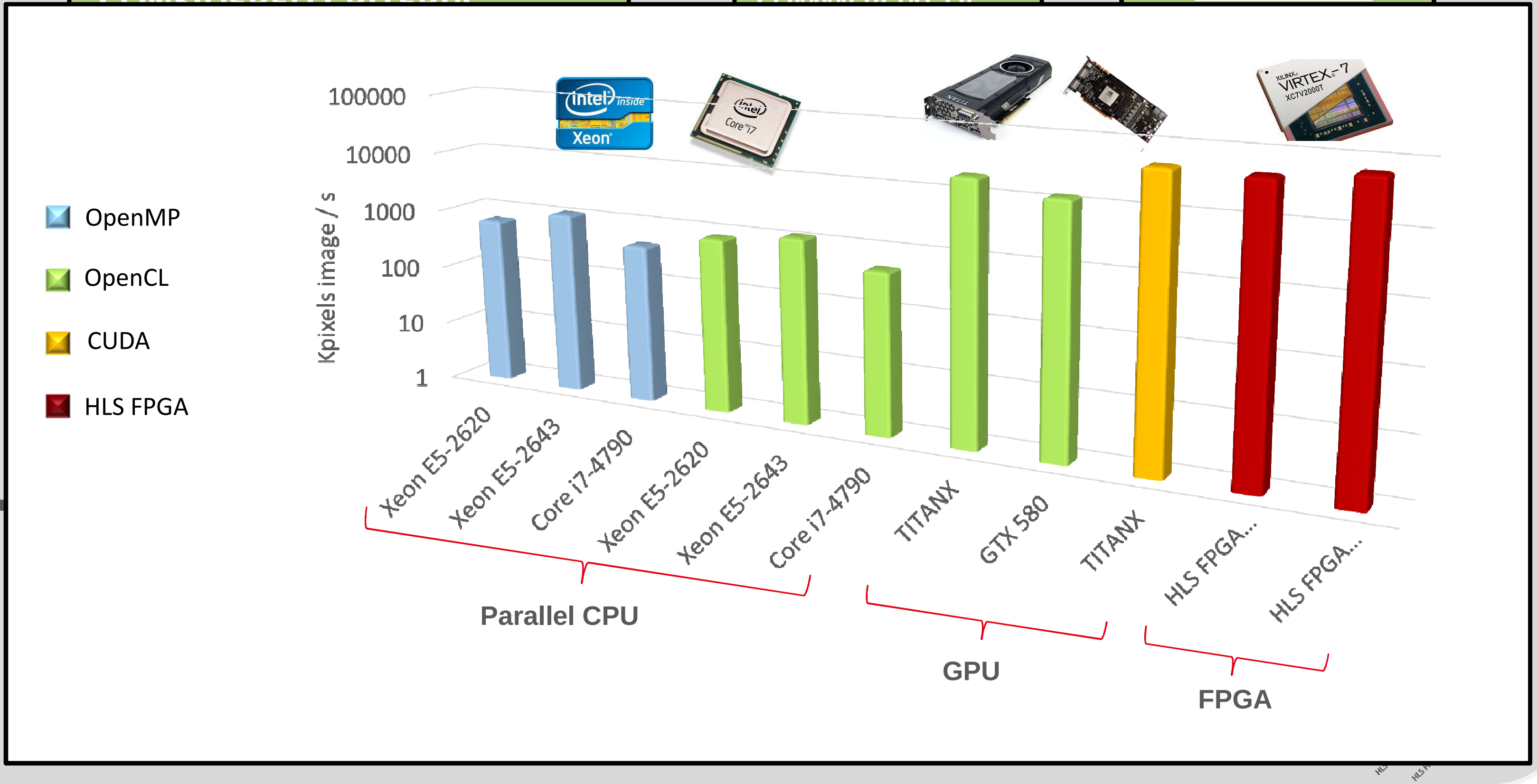
Learning & Test databases

Data conditioning

CONSIDERED CRITERIA

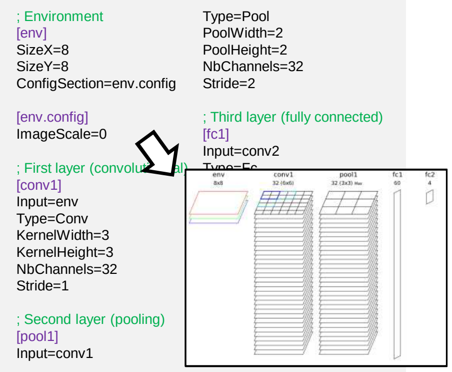
COMMERCIAL OF

HARDWARE



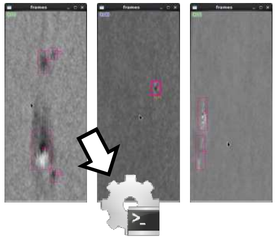
FAST AND ACCURATE DNN EXPLORATION

1 Deep Network builder

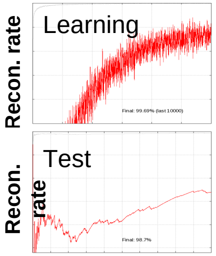


```
; Environment
[env]
SizeX=8
SizeY=8
ConfigSection=env.config

[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
```

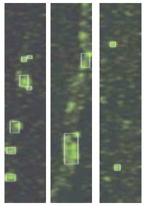
2 Learning a database



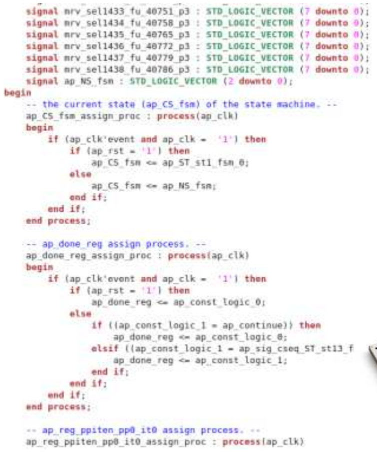
3 Analysis of network Performance



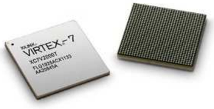
Output categories and localization

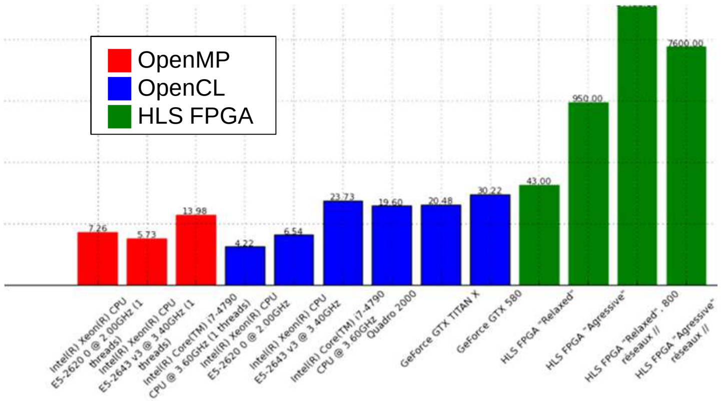


4 CPU, GPU and FPGA-based Real-time implementation



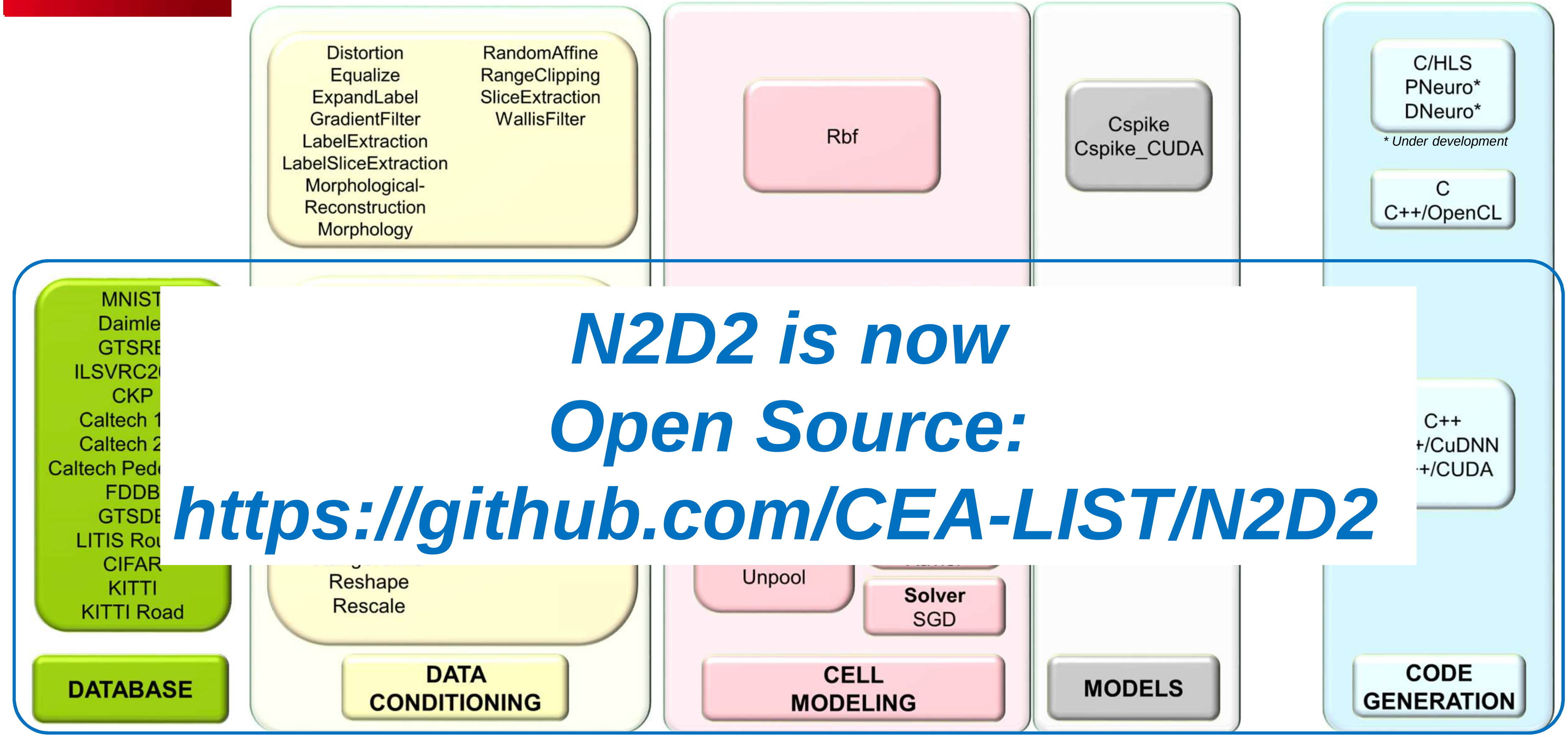
```
    signal mrv_sel1433_fu_40751_p3 : STD_LOGIC_VECTOR (7 downto 0);
    signal mrv_sel1434_fu_40758_p3 : STD_LOGIC_VECTOR (7 downto 0);
    signal mrv_sel1435_fu_40765_p3 : STD_LOGIC_VECTOR (7 downto 0);
    signal mrv_sel1436_fu_40772_p3 : STD_LOGIC_VECTOR (7 downto 0);
    signal mrv_sel1437_fu_40779_p3 : STD_LOGIC_VECTOR (7 downto 0);
    signal mrv_sel1438_fu_40786_p3 : STD_LOGIC_VECTOR (7 downto 0);
    signal ap_NS_fsm : STD_LOGIC_VECTOR (2 downto 0);
begin
    -- the current state (ap_CS_fsm) of the state machine. --
    ap_CS_fsm_assign_proc : process(ap_clk)
    begin
        if (ap_clk'event and ap_clk =  '1') then
            if (ap_rst = '1') then
                ap_CS_fsm <= ap_ST_st1_fsm_0;
            else
                ap_CS_fsm <= ap_NS_fsm;
            end if;
        end if;
    end process;

    -- ap_done_reg assign process. --
    ap_done_reg_assign_proc : process(ap_clk)
    begin
        if (ap_clk'event and ap_clk =  '1') then
            if (ap_rst = '1') then
                ap_done_reg <= ap_const_logic_0;
            else
                if ((ap_const_logic_1 = ap_continue)) then
                    ap_done_reg <= ap_const_logic_0;
                elsif ((ap_const_logic_1 = ap_sig_cseq_ST_st13_f
                    ap_done_reg <= ap_const_logic_1;
                end if;
            end if;
        end if;
    end process;

    -- ap_reg_ppiten_pp0_it0 assign process. --
    ap_reg_ppiten_pp0_it0_assign_proc : process(ap_clk)
```





➔ **Wide targets range, perfs and power metrics**

OVERVIEW OF N2D2 MODULES



Distortion
Equalize
ExpandLabel
GradientFilter
LabelExtraction
LabelSliceExtraction
Morphological-
Reconstruction
Morphology

RandomAffine
RangeClipping
SliceExtraction
WallisFilter

Rbf

Cspike
Cspike_CUDA

C/HLS
PNeuro*
DNeuro*

** Under development*

C
C++/OpenCL

MNIST
Daimle
GTSRE
ILSVRC2
CKP
Caltech 1
Caltech 2
Caltech Pede
FDDB
GTSDE
LITIS Rou
CIFAR
KITTI
KITTI Road

Reshape
Rescale

Unpool

Solver
SGD

C++
+/CuDNN
+/CUDA

DATABASE

DATA CONDITIONING

CELL MODELING

MODELS

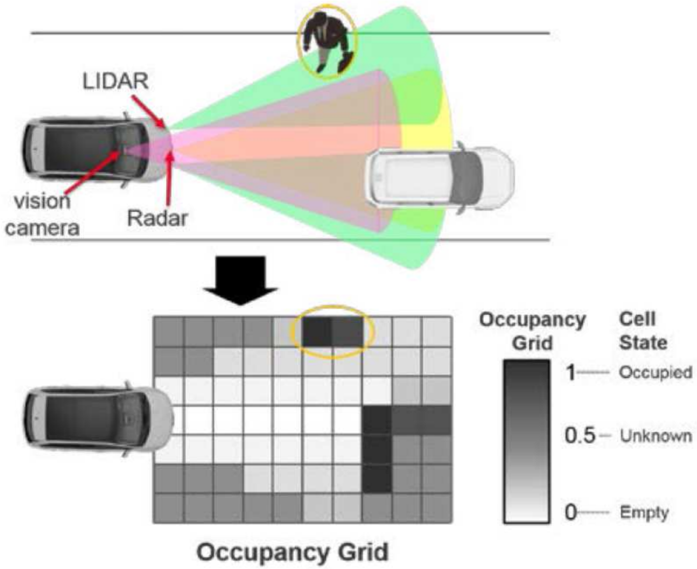
CODE GENERATION

N2D2 is now
Open Source:
https://github.com/CEA-LIST/N2D2

ΣFUSION: SENSOR DATA PREPROCESSING AND FUSION

ΣFusion technology:

- Bayesian Fusion with **only integer arithmetic**
- Compatible with **ASIL-D processors**
- **Real-time performance** on µC **(Cortex M7),** 200MHz
- **Fully certifiable** (deterministic and predictable)
- **Power efficiency** increased by a factor of **100x**
- Suitable for **multi-modal sensor fusion**



Occupancy Grid

- **Perception for autonomous vehicles**
  - 2x Velodyne VLP 16 Lidars
    - 600000 data points per sec
    - 16 Mbits/sec data bandwith (Ethernet)
  - 1x Tara stereo system
    - Disparity map computed on a Nvidia TX1
    - ~150000 data points per sec
    - ~ 4Mbits /sec
  - Environment model
    - 272x480 cells (130560 cells)
    - computed in real-time (40 ms)
    - Spatial accuracy of ~4cm
    - Scanning horizon ~8m
  - Hardware used:
    - ***STM32F7 @200 MHz***

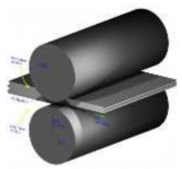
See the stand on the right

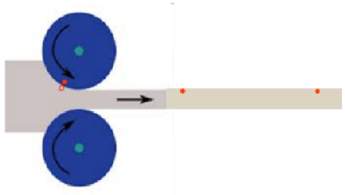
- Obstacle detection
- Vision glare with external light source

- **Featured in EEtimes, Embedded Computing, Eenews, …**
  - http://www.eenews.net/stories/1060048190
  - http://embedded-computing.com/31082-ces-2017-leti-the-biggest-little-organization-you-never-should-have-heard-of/
  - http://www.eetimes.com/document.asp?doc_id=1331148&page_number=7

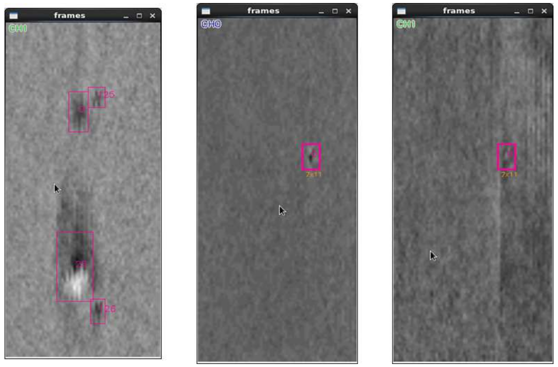
APPLICATION: DEFECTS DETECTION

- **Defects identification on metal after rolling**
  - Constrains:
    - Real-time with extremely high throughput
    - Tiny and low contrasted defects
  - Solutions:
    - Database labeling and pre-processing
    - Fast NN topology exploration
    - Performances vs complexity analysis

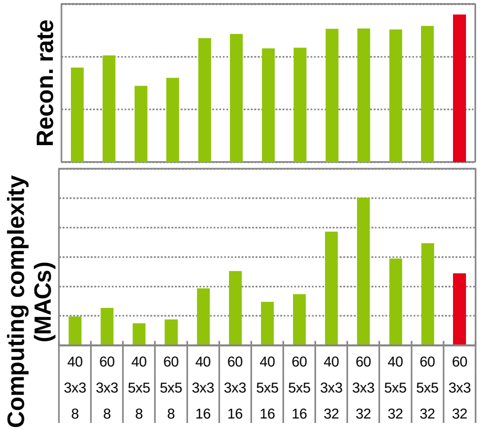




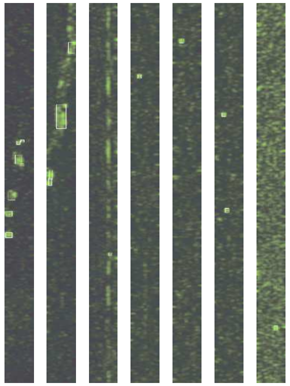
1) Defects labeling and visualization



2) NN Exploration and benchmarking



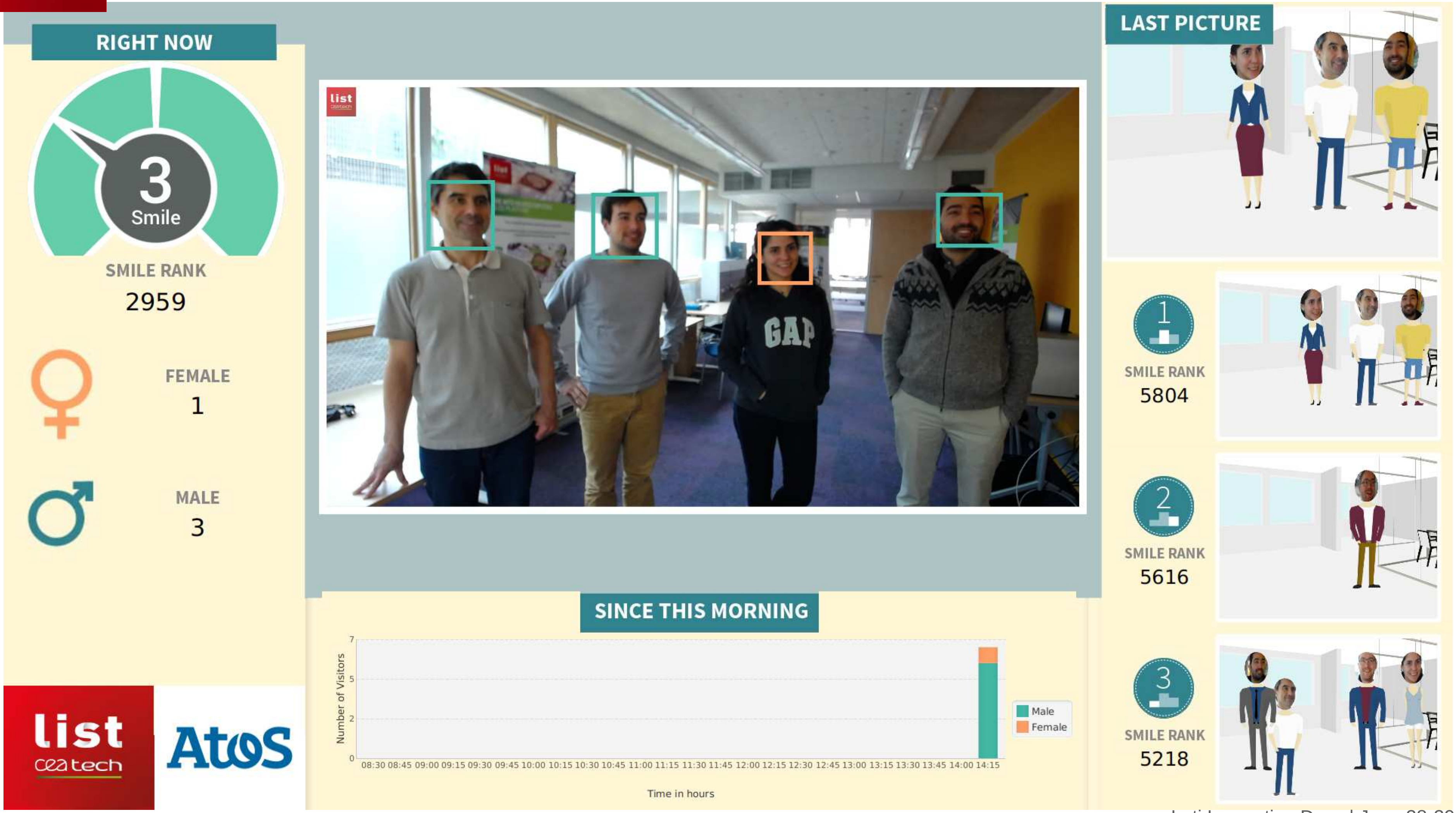
3) Defects identifications after NN learning



➔ From scratch exploration (database and NN construction) to industrial application

➔ 50,000 MACs NN synthetized in 100 cycles on FPGA @ 100 MHz (500 MACs/cycle)

APPLICATION: REAL-TIME FACES DETECTION WITH GENDER & EMOTION



RIGHT NOW

3 Smile

SMILE RANK 2959

FEMALE 1

MALE 3

LAST PICTURE

1 SMILE RANK 5804

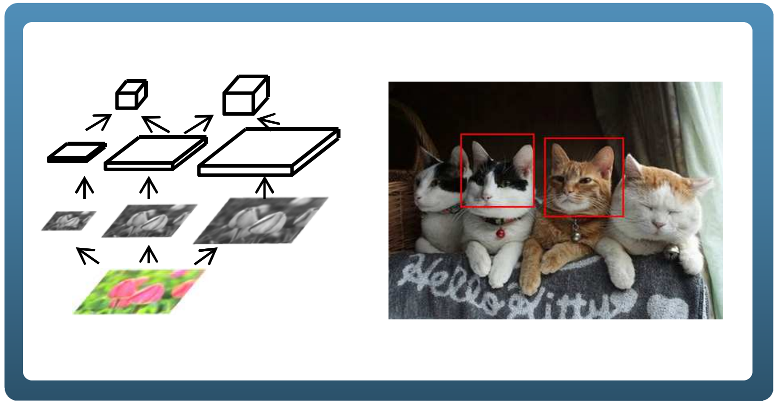
2 SMILE RANK 5616

3 SMILE RANK 5218

SINCE THIS MORNING

APPLICATION: Q-LEARNING BASED SOC ENERGY MANAGEMENT

Energy saving reinforcement learning



- Dynamic software applications with performance constraints, e.g., throughput

eLinux

android

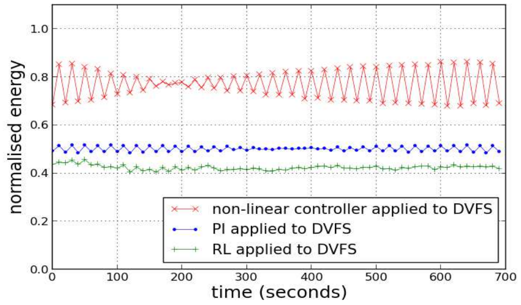
- Standard Linux-based operating system

- Multi/many core SoCs

Source: NXP i.MX6

Source: ST/CEA

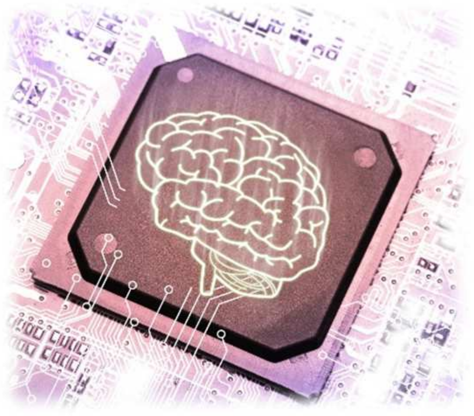
- Q-learning energy manager
  - On-line, gradually learn the SoC operating points such that performance constraints are respected and **energy consumption is reduced**
  - No need to model the dynamics of the system



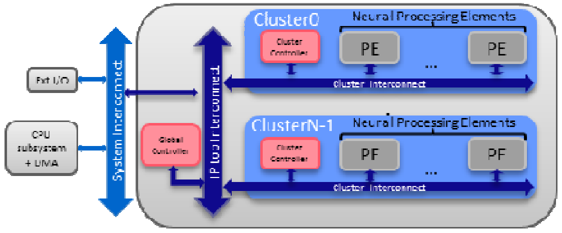
Up to 44% energy reduction, wrt. state-of-the-art (proportional-integral and non-linear controllers)

HARDWARE ACCELERATION FOR DEEP NEURAL NETWORKS

Dedicated computing IPs with high TOPS/Watt performance

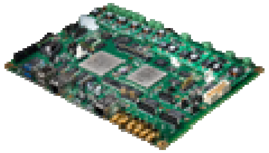


PNeuro programmable

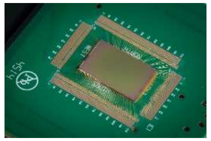


PNeuro Eng'ne

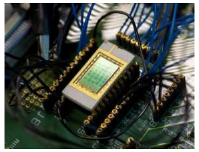
DNeuro dedicated IP /
High level synthesis



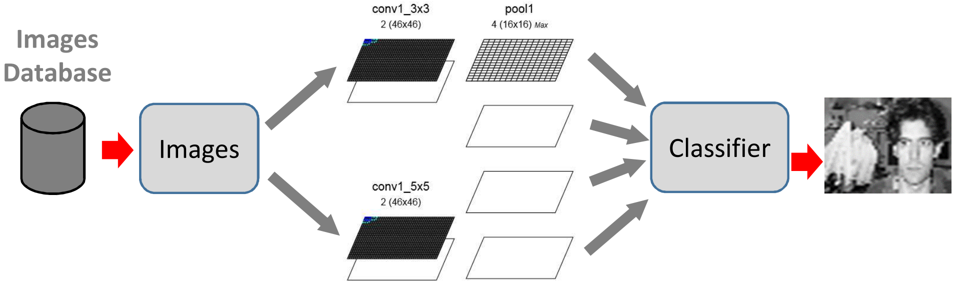
3D stacked architectures



Advanced architectures
(Spike-based, mixed-signal, NVMs…)

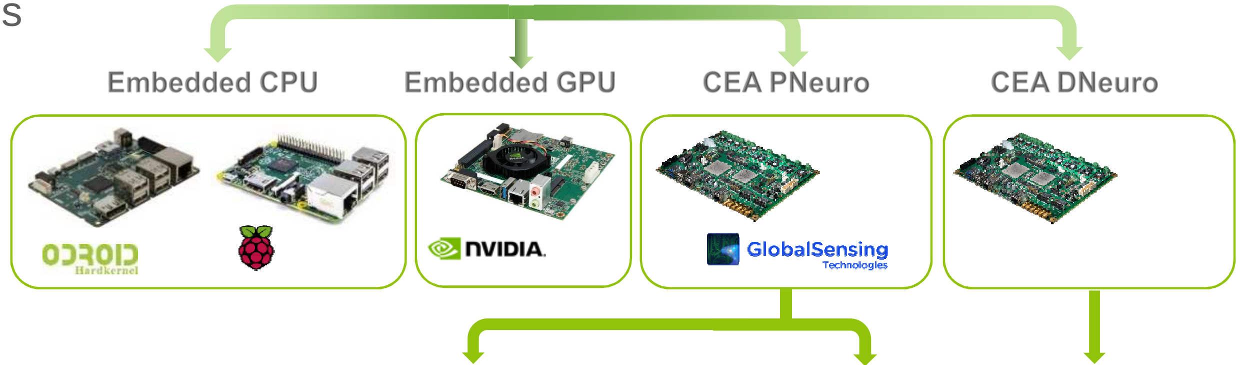


PNEURO BENCHMARKING



- **Benchmark application:**
  - Face extraction on a database of 18,000 images
  - 60 neurons on the hidden layer, 450 Kops
  - Recognition rate 97%

- **Optimized code for 5 architectures:**
  - Embedded CPU: Quad Arm A7 & A15
  - Embedded GPU: NVidia Tegra K1
  - PNeuro Quad Neuro-Cores / DNeuro

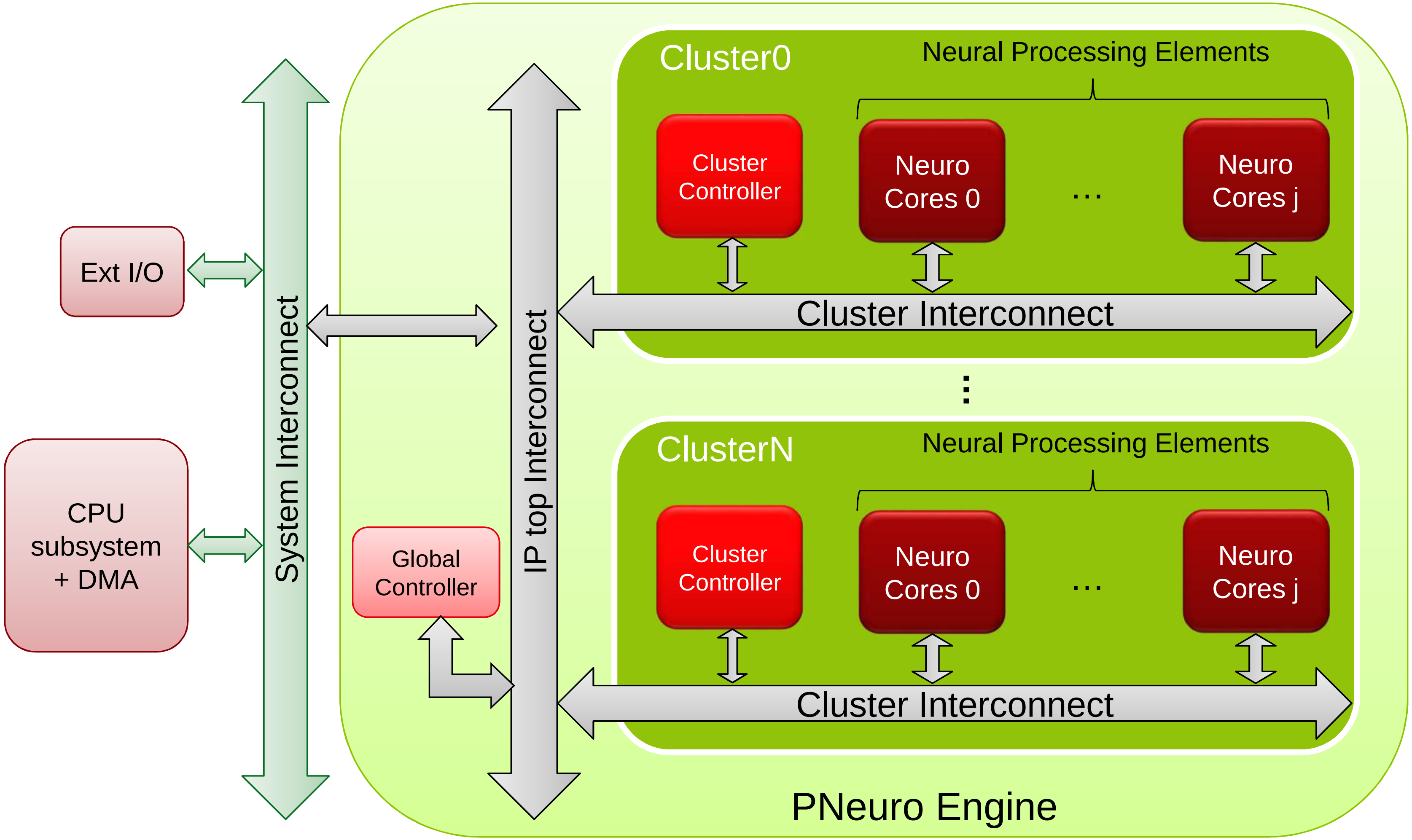


Target	Quad ARM A7 900 MHz	Quad ARM A15 2 GHz	Tegra K1 850 MHz	PNeuroV2 (FPGA) 100 MHz	PNeuroV2 (ASIC) 500 MHz	DNeuro (FPGA) 100 MHz
Performance	480 images/s	870 images/s	3 550 images/s	7 000 images/s	25 000 images/s	45 000 images/s
Energy Efficiency	380 images/s/W	350 images/s/W	600 images/s/W	2 800 images/s/W	125 000 images/s/W	18 000 images/s/W

- **PNeuro and DNeuro performance comparison vs Tegra K1 with N2D2:**

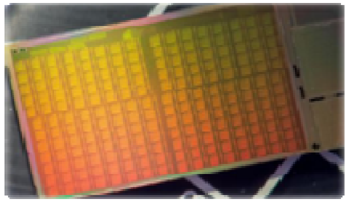
– Faster	x 2	x 7	x 12.5
– More Energy Efficient	x 4.5	x 200	x 30

PNEURO OVERVIEW

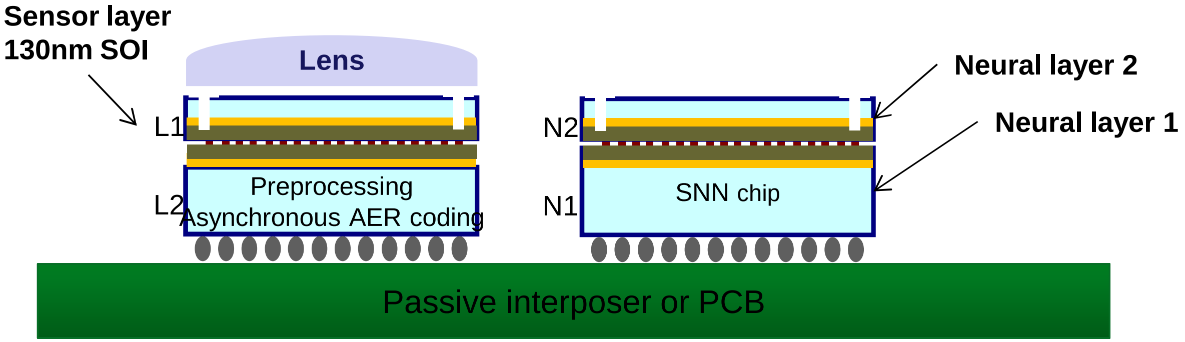


3D STACKED RETINA WITH SPIKING NEURAL NETWORKS

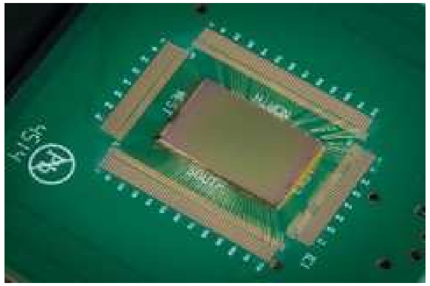
- **RETINE: image sensor + 3D stacked SIMD processors**
  - Image sensor: 70% fill factor, 12 µm pixel, >1000 fps
  - SIMD processors: 3072 units, distributed memory, 11.7 MOPS/mW
  - Feed SNN with Asynchronous Event Representation (AER) after pre-processing



Processor array die



Retine Chip
ALTIS 130nm, CuCu bonding

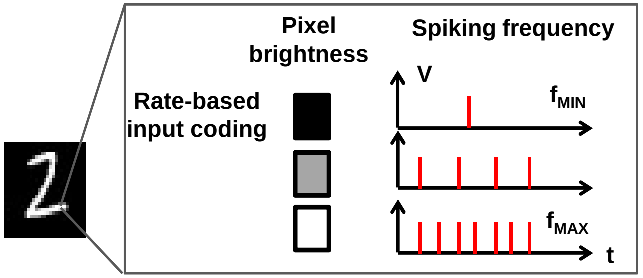


- **Pre-processing performances: (L1+L2 stacked retina)**

	RETINE	ARM cortex A9 +NEON	STxP70
Frequency (Mhz)	150	400	350
Performance (GOPS)	**72**	0,67	0,28
Power consumption (W)	4,8	0,25	0,08
Energy / frame (mJ)	2,74	0,68	5,6
Energy efficiency (normalized, GOPS/W)	**45**	2,68	5,25

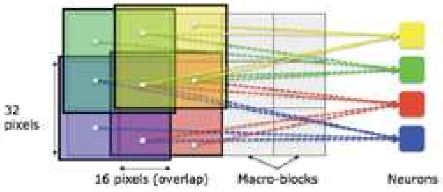
➔ x100 computing power, x10 energy efficiency, /15 processing latency vs competition

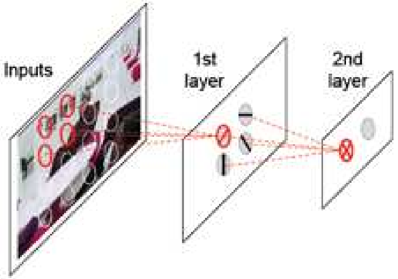
3D SPIKING NEURAL NETWORK

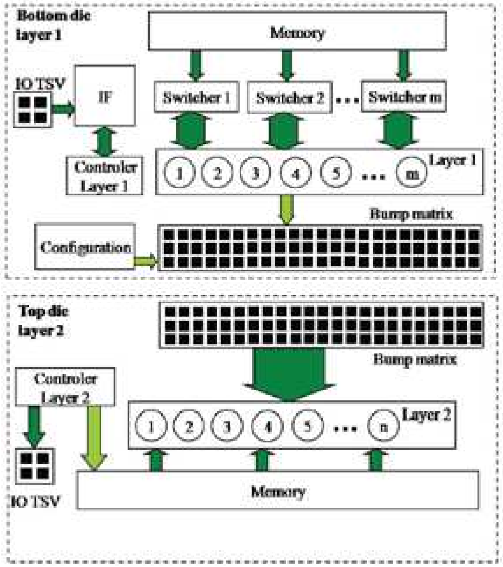


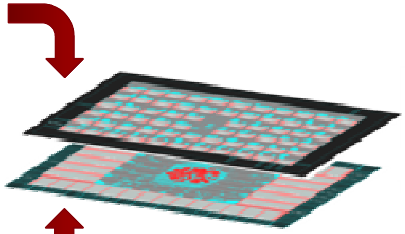
NEMESIS 3D two-layers SNN test chip

- 1st layer: 48 macro-block neurons, 1024 synapses per neuron (49 152 total)
- 2nd layer: 50 fully connected neurons, 2 400 synapses

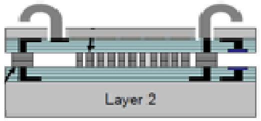








Nemesis Test Chip
ALTIS 130nm
CuCu bonding

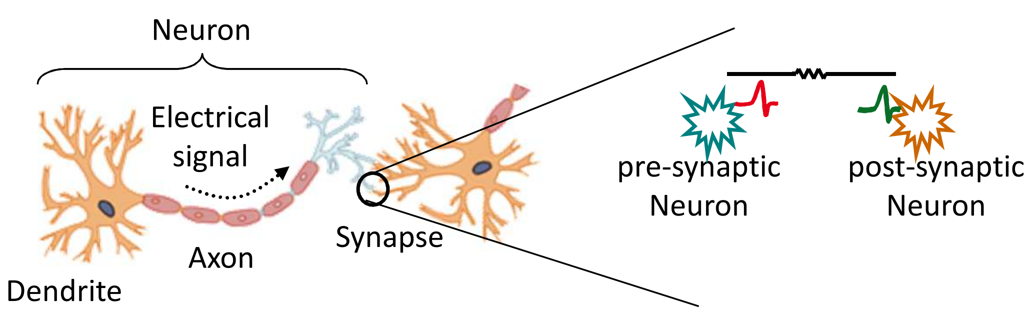


Two-layers SNN circuit	2D	3D
Total area (mm²)	7,97	3,63 (-54%)
Power (mW)	428	354 (-17%)
Critical path (ns)	9,00	6,63 (-26%)

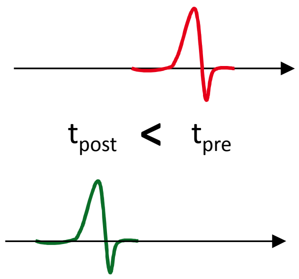
[B. Belhadj, R. Heliot, P. Vivet, CASSES'2014]

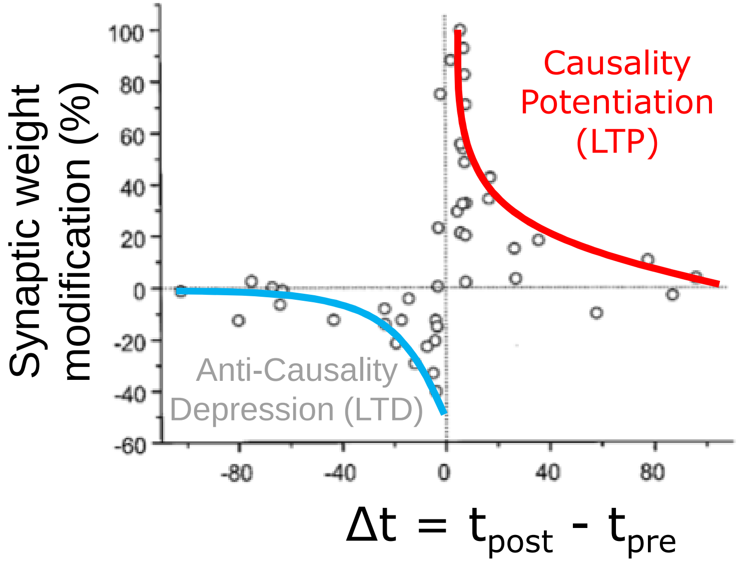
➔ 3D offers 2x better total area and 25% better power efficiency vs 2D

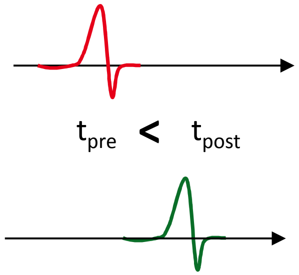
LEARNING FROM NEUROSCIENCE: A STDP (SPIKE TIMING DEPENDENT PLASTICITY) PRIMER



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

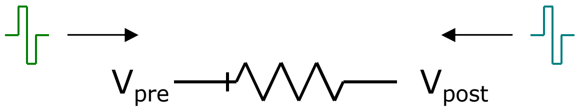


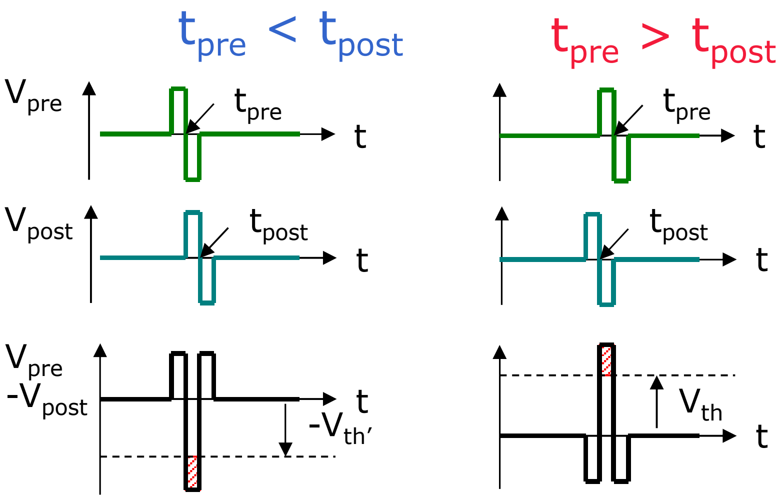




PRINCIPLE CROSSBARS OF MEMRISTORS

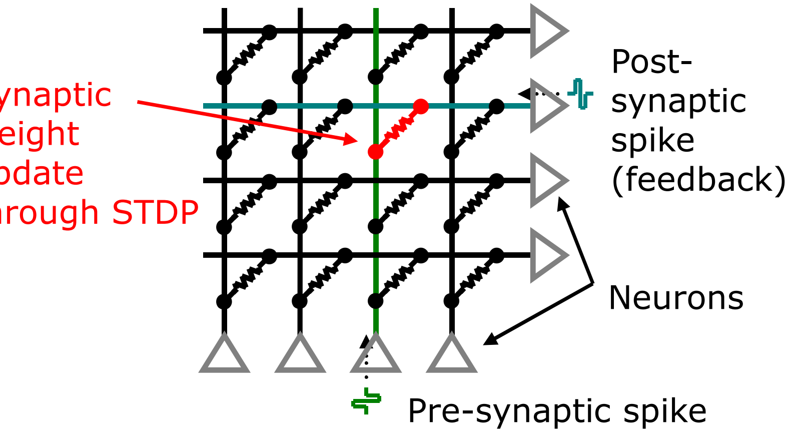
First Proposed by Snider(1)



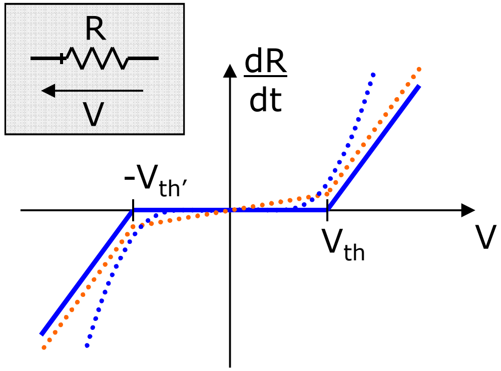


R decreases

R increases



Synaptic weight update through STDP
Post-synaptic spike (feedback)
Neurons
Pre-synaptic spike

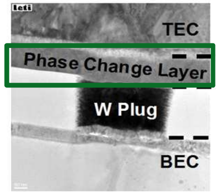


1. G. Snider, *Nanoscale Architectures*, 2008
2. B. Linares-Barranco et al, *Nature Precedings*, 2009

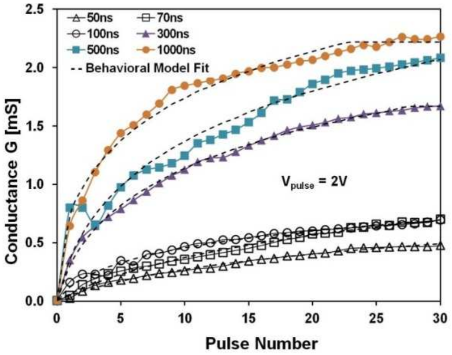
NVM SYNAPSES IMPLEMENTATIONS

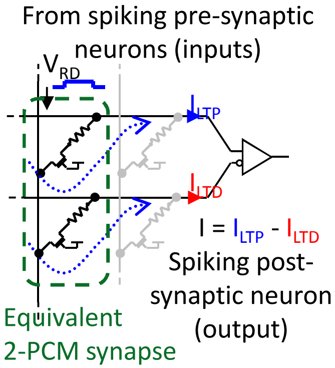
- **2-PCM synapses for unsupervised cars trajectories extraction**

PCM



Crystallization/ Amorphization



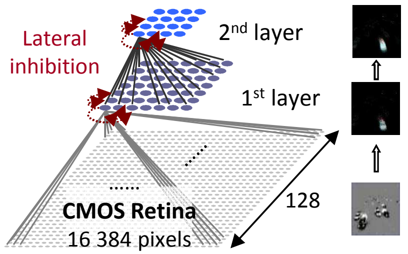


From spiking pre-synaptic neurons (inputs)

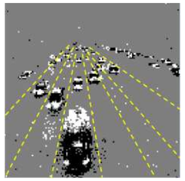
Spiking post-synaptic neuron (output)

Equivalent 2-PCM synapse

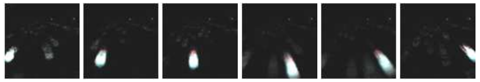
[O. Bichler et al., Electron Devices, IEEE Transactions on, 2012]



Traffic lanes visualization



Neuron sensitivity maps

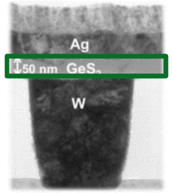


[O. Bichler et al., Neural Networks, 2012]

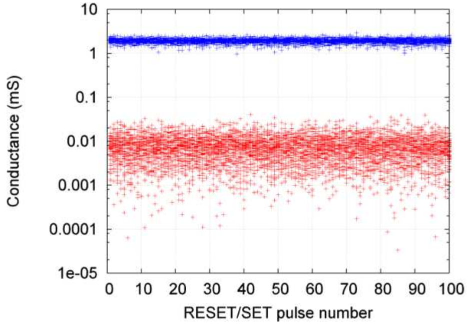
First real world application of STDP on spiking artificial retina

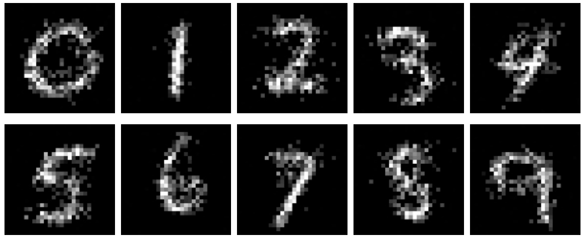
- **CBRAM binary synapses for unsupervised MNIST handwritten digits classification with stochastic learning**

CBRAM



Forming/Dissolution of conductive filament





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

PUTTING IT ALL TOGETHER: NEURAM3

NEURAL COMPUTING ARCHITECTURES IN ADVANCED MONOLITHIC 3D-VLSI NANO-TECHNOLOGIES

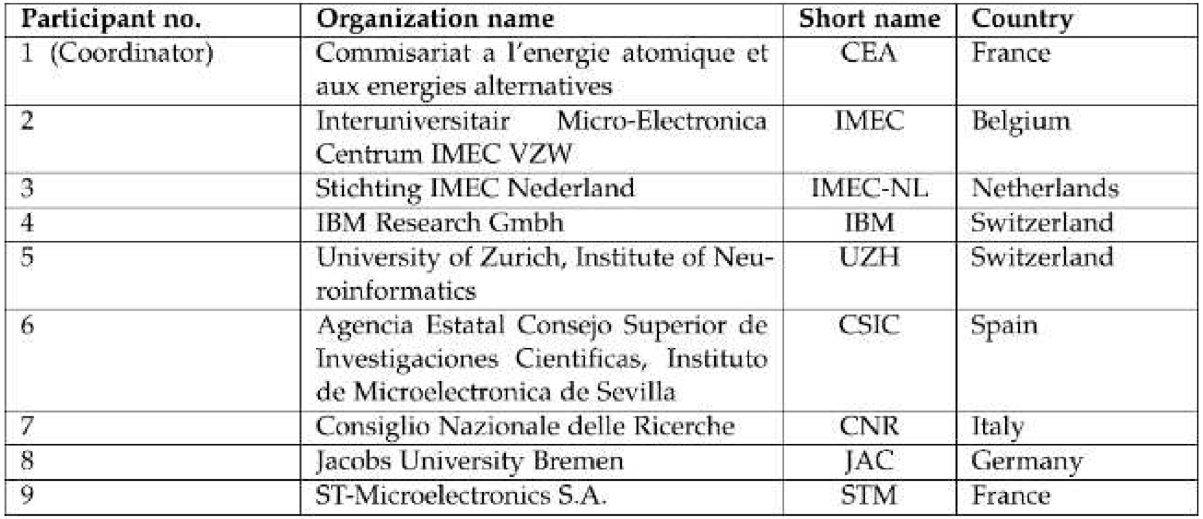
- **EU collaborative project (ICT)**

- **Objective**
  - Fabricate a chip implementing neuromorphi architecture with state-of-the-art machine and spike-based learning

- **Features**
  - Ultra low power, scalable and configurable NN architecture
  - Gain x50 in power consumption vs conventional digital solutions
  - 3D FDSOI at 28nm integrating RRAM synaptic elements
  - TFT device technology to interconnect multiple processor chips

- **Consortium**

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

1ST DIGITAL CHIP EXPECTED FOR SUMMER 2017:

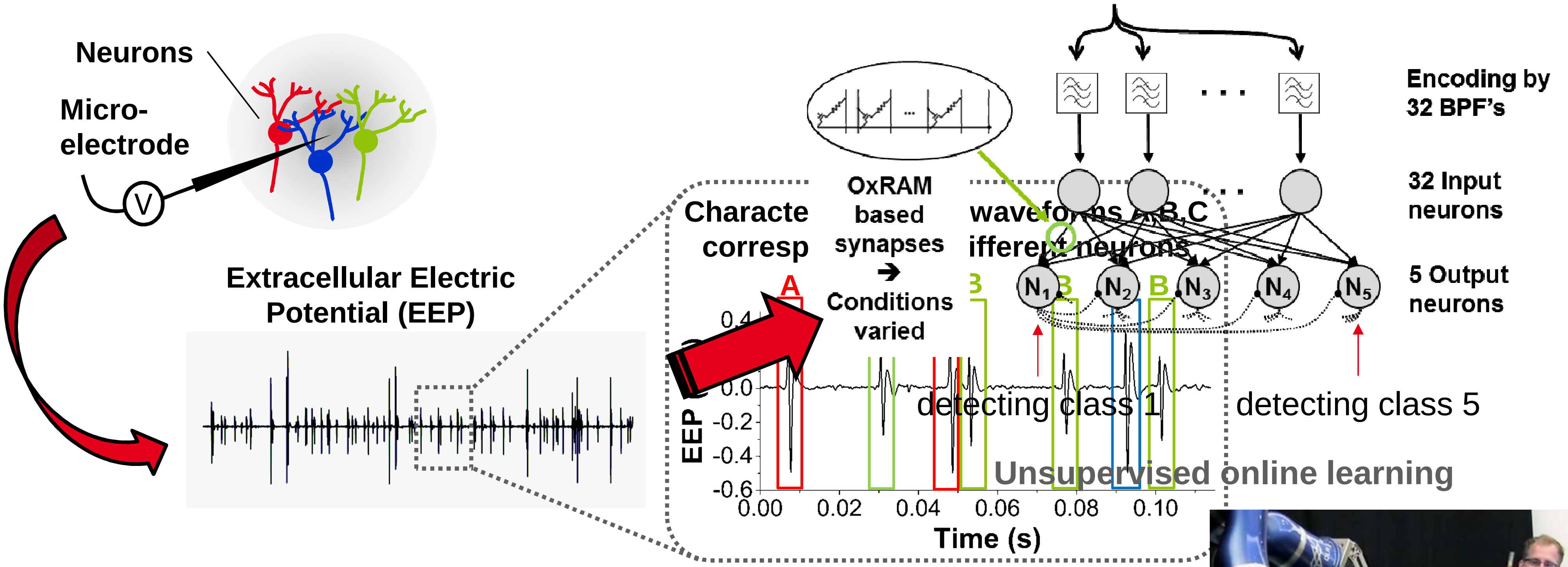
	Neuram3 1st chip	IBM True North
Technology	28 nm FDSOI	28nm CMOS
Supply Voltage	1 V	0.7V
Neuron Type	Analog	Digital
Neurons per core	256	256
Core Area	0.36 mm^2	0.094 mm^2
Computation	Parallel processing	Time multiplexing
Fan In/Out	2k/8k	256/256
Synaptic Operation per Second per Watt	300 GSOPS/W[*1]	46 GSOPS/W
Energy per synaptic event	<2 pJ[*2]	10 pJ
Energy per spike	<0.375 nJ[*3]	3.9 nJ

∗ 1 At 100Hz mean firing rate, by appending 4 local-core destinations per spike, 400 k events will be broadcast to 4 cores with 25% connectivity per event. 400 k x 1 k x 25% / 300 μ W = 300 GSOPS/W

∗ 2 In case of 25% match in each core, energy per synaptic event = energy per broadcast / (256*25%) =120pJ/64 = 2 pJ

∗ 3 Energy per spike = total power consumption / spikes numbers = 300 uW/800 k = 0.375 nJ

NEUROMORPHIC APPROACH FOR SPIKE SORTING



- Identification of single neurons by characteristic of spike shapes
- **Output neurons allow to classify spikes after sufficient learning period**
- **Extraction of spike enables decoding the brain activity (BCI)…**
- **…. Opening new applications, like brain controlled prosthetics..**

LETI IS YOUR PARTNER FOR ARTIFICIAL INTELLIGENCE BASED SYSTEMS

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-cea.com

INSTITUT CARNOT
Leti