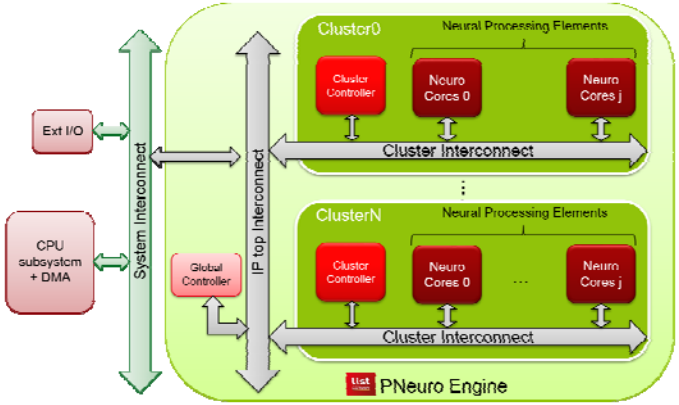
GlobalSensing Technologies



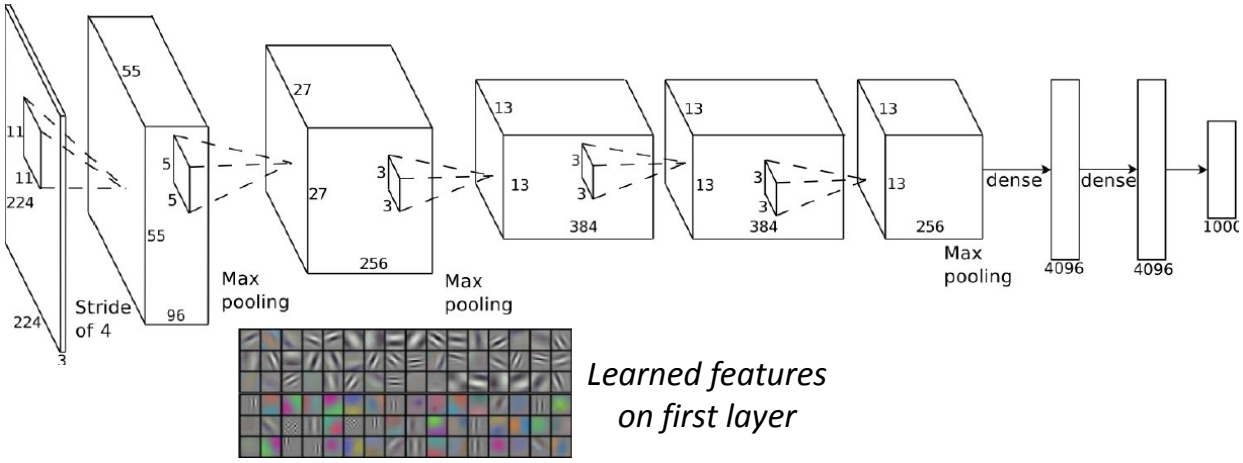
Low-Power Neural Processor for Embedded Vision Applications.

Michel Paindavoine[1]

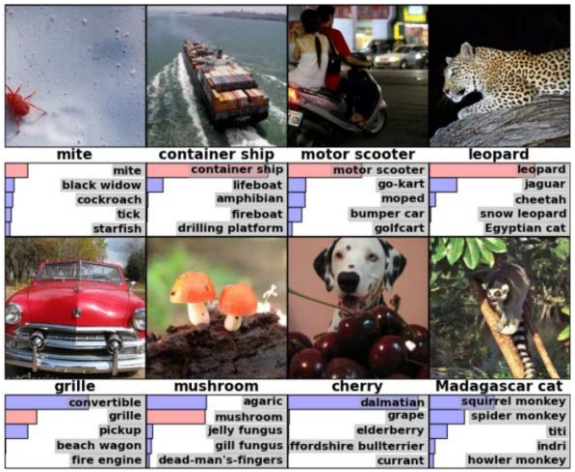
(1) GlobalSensing Technologies (GST) – Dijon, France www.gsensing.eu

Deep Neural Network Models

- ImageNet classification (Hinton's team, hired by Google)
  - 1.2 million high res images, 1,000 different classes
  - Top-5 17% error rate (huge improvement)

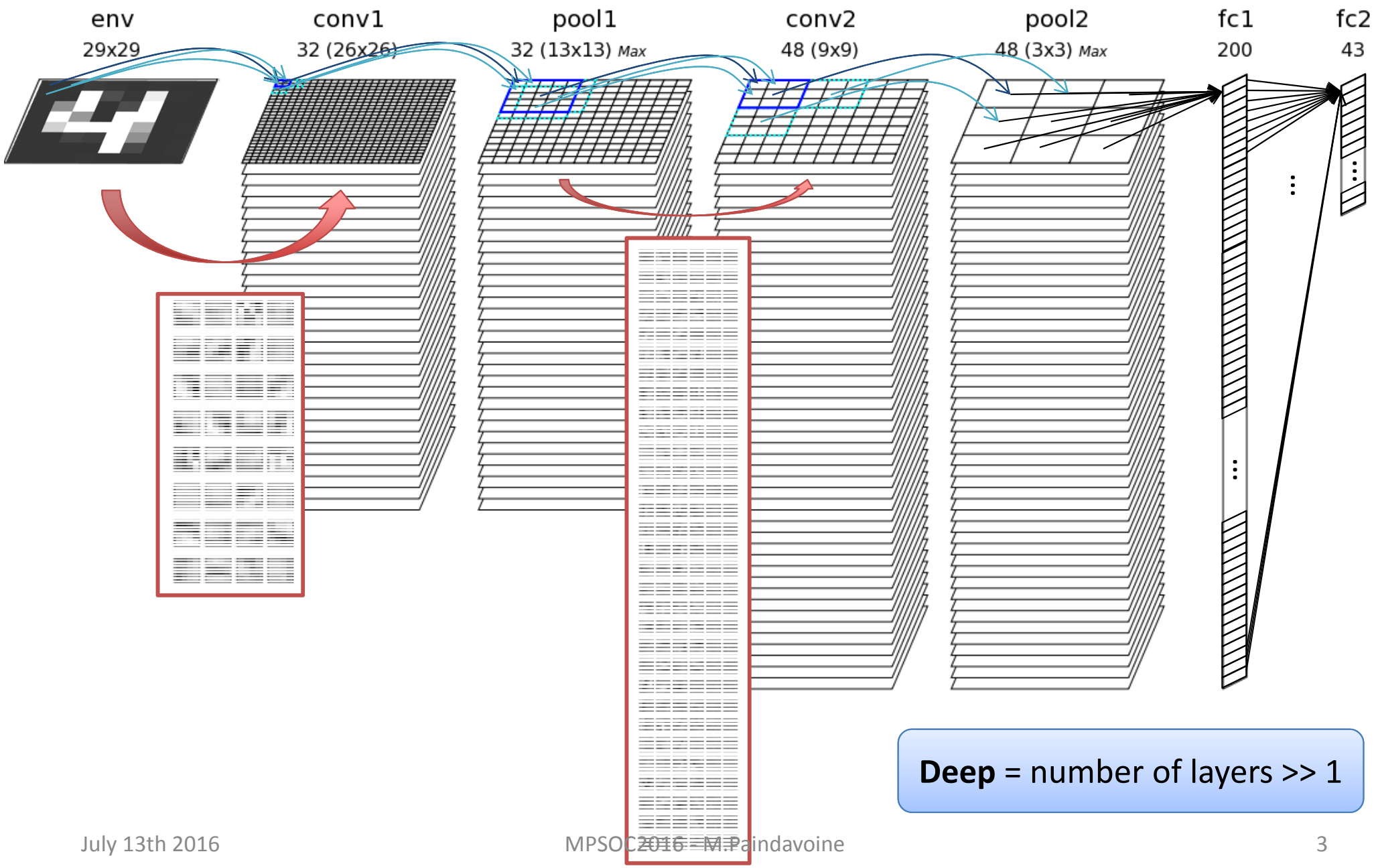


Learned features on first layer



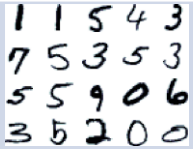
- Facebook's 'DeepFace' Program (labs head: Y. LeCun)
  - 4 million images, 4,000 identities
  - 97.25% accuracy, vs. 97.53% human performance

CNNs Organization



Deep = number of layers >> 1

State-of-the-art in Recognition

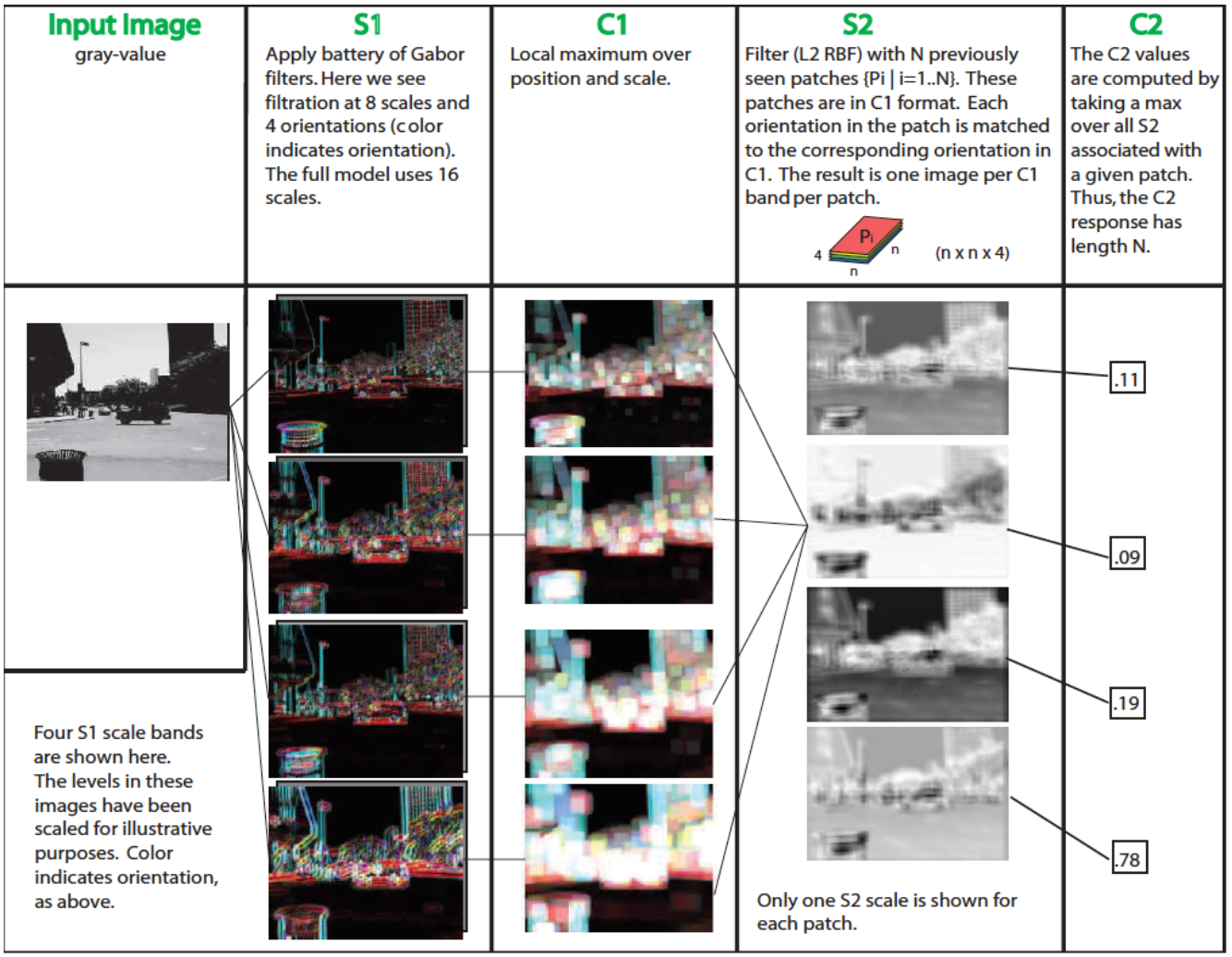
Database	# Images	# Classes	Best score
MNIST *Handwritten digits* 	60,000 + 10,000	10	99.79% [3]
GTSRB *Traffic sign* 	~ 50,000	43	99.46% [4]
CIFAR-10 *airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck* 	50,000 + 10,000	10	91.2% [5]
Caltech-101 	~ 50,000	101	86.5% [6]
ImageNet 	~ 1,000,000	1,000	Top-5 83% [1]
DeepFace 	~ 4,000,000	4,000	97.25% [2]

INCREASING COMPLEXITY

- **State-of-the-art are Deep Neural Networks *every time***

An other Neuro-Inspired Model: The Hmax

(a NeuroScience Approach)



Serre et al . Robust Object Recognition with Cortex-like Mechanisms IEEE PAMI 2007

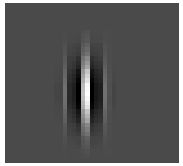
Hmax : S1 and C1 layers

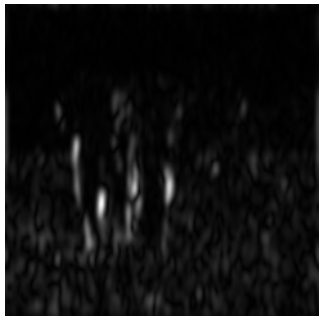
C_1 layer			S_1 layer		
Scale band S	Spatial pooling grid ($N_S \times N_S$)	Overlap Δ_S	filter size s	Gabor σ	Gabor λ
Band 1	8×8	4	7×7 9×9	2.8 3.6	3.5 4.6
Band 2	10×10	5	11×11 13×13	4.5 5.4	5.6 6.8
Band 3	12×12	6	15×15 17×17	6.3 7.3	7.9 9.1
Band 4	14×14	7	19×19 21×21	8.2 9.2	10.3 11.5
Band 5	16×16	8	23×23 25×25	10.2 11.3	12.7 14.1
Band 6	18×18	9	27×27 29×29	12.3 13.4	15.4 16.8
Band 7	20×20	10	31×31 33×33	14.6 15.8	18.2 19.7
Band 8	22×22	11	35×35 37×37	17.0 18.2	21.2 22.8

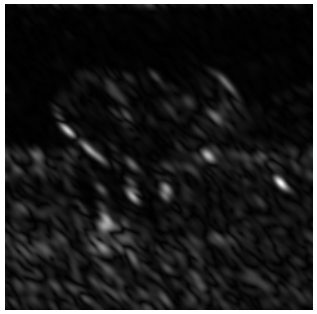
Serre et al . Robust Object Recognition with Cortex-like Mechanisms IEEE PAMI 2007

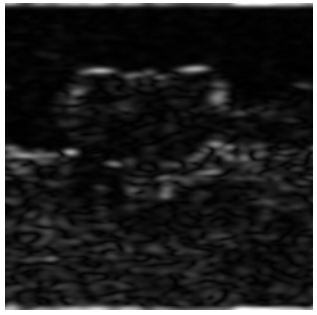
Gabor Filters

Original Image









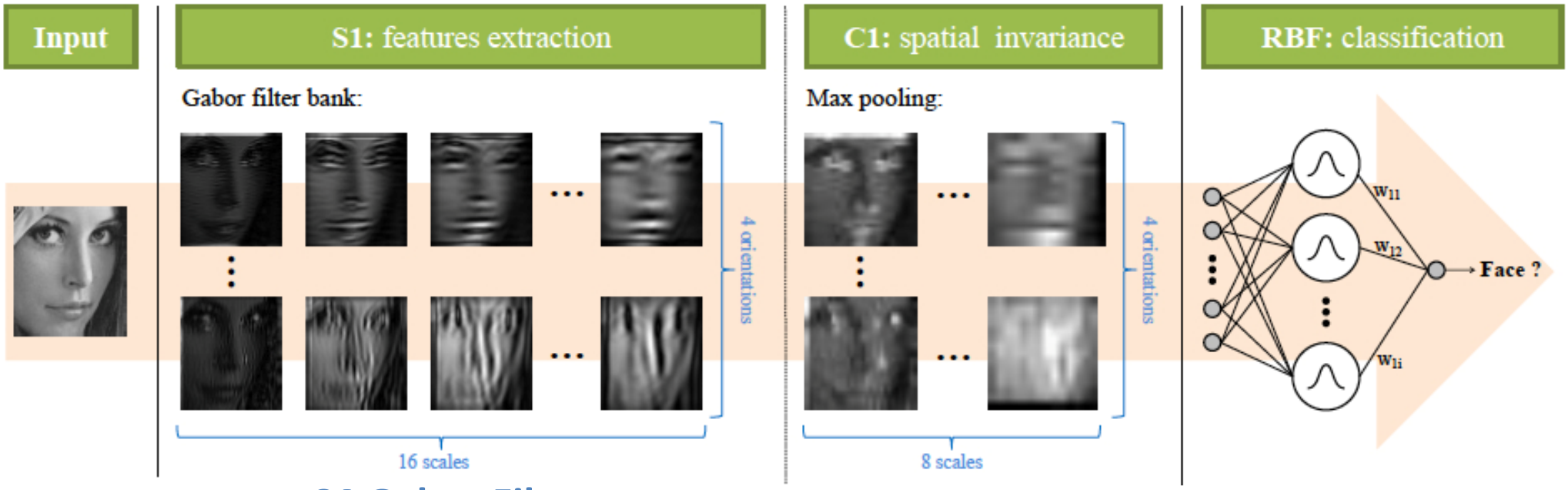
Neuro-Inspired Models:

The Hmax

Hmax Model performances

Data sets	Proposed	[Serre 07]	Others
Airplanes	96.0	96.7	94.0 [Fergus 03]
Motorcycles	98.0	98.0	95.0 [Fergus 03]
Cars (Rear)	96.0	99.8	84.8 [Fergus 03]
Leaves	92.0	97.0	84.0 [Weber 00]

Hmax accelerator: Complexity



64 Gabor Filters

1 Mpixels Image complexity:

S1: Optimized Gabor Filters: 2.9 GMAC

C1: Max: 0.13 GOP

RBF Neural Network : 0.4 GOP

Total: **3.43 GMAC & OP**

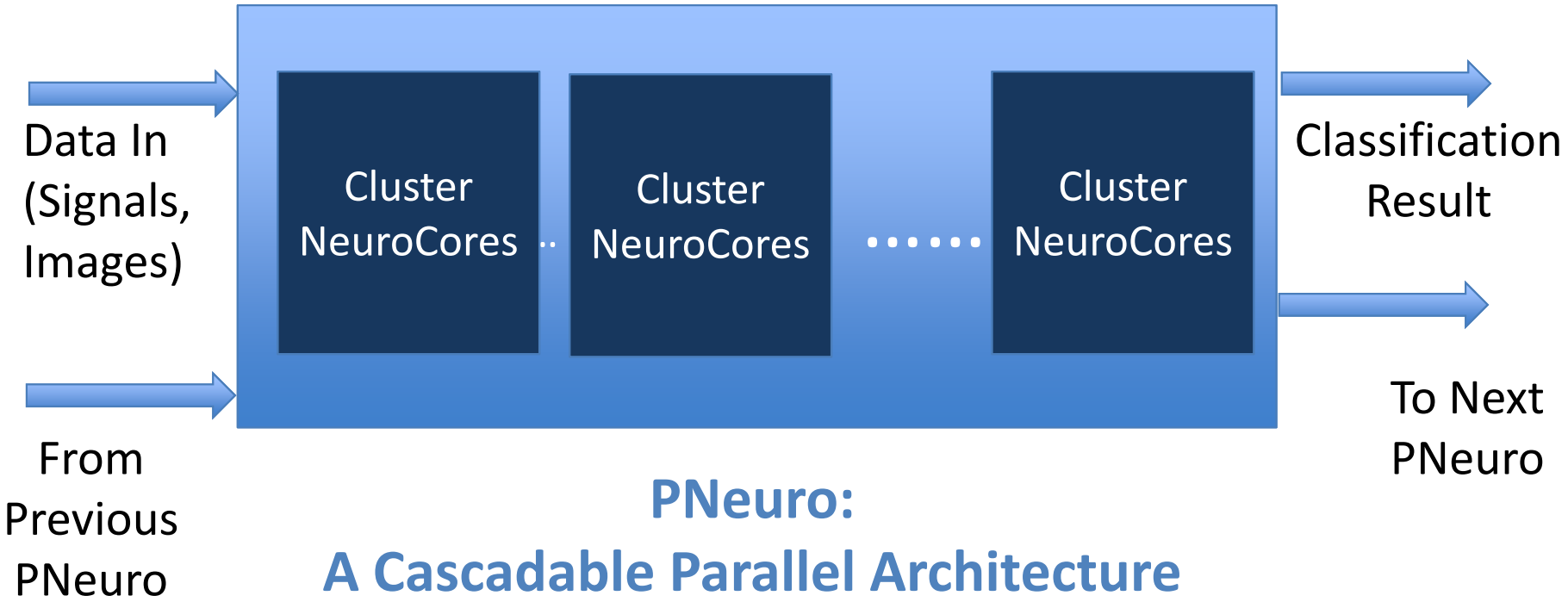
One IP camera 1M pixels @ 30 fps: 103 GOP/sec

Pneuro accelerator

(Joint Laboratory CEA & GST initiated in 2013)

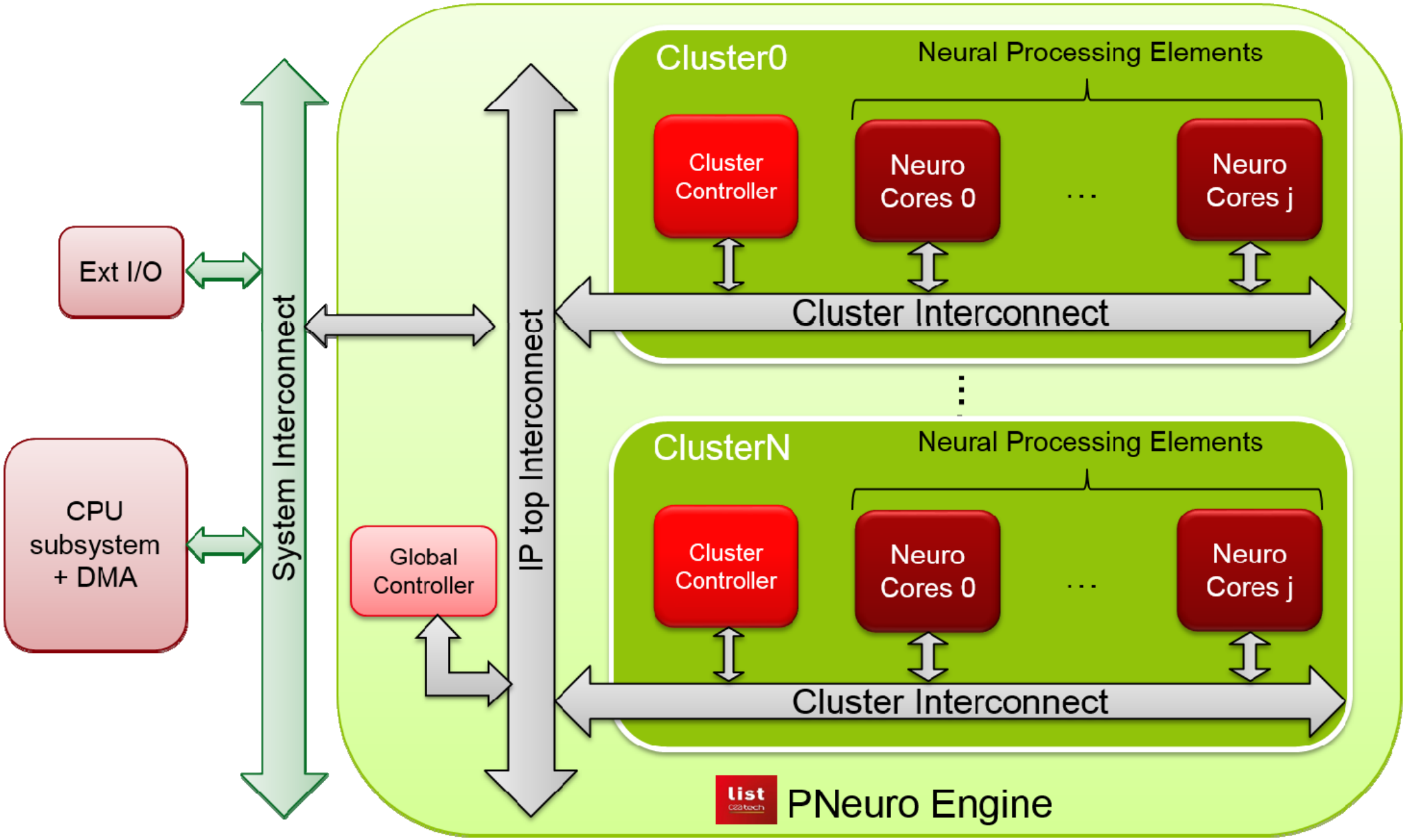
Objective:

Design a processor integrating within the same chip signal processing functions and neuronal functions: Hmax, CNN



PNeuro:
A Cascadable Parallel Architecture

PNeuro accelerator overview



GlobalSensing Technologies

cea

PNeuro accelerator: Main Specifications

- **Fully-programmable energy efficient hardware accelerator**
  - Designed for DNN processing chains
    - CNN (OK), HMax (OK), RNN (under investigation)
  - Supporting traditional image processing chains (filtering, etc.)

- **Clustered SIMD architecture**
  - Optimized operators for MAC & NL-approximation
  - Optimized memory accesses to perform efficient data transfers to operators
  - ISA including ~50 instructions (control + computing)

- **Programming tools under development**
  - Library including most-common kernels with associated parameters (convolution, max pooling, fully-connected layers) to ease programming
  - Based on N2D2 platform with dedicated exports for PNeuro

PNeuro accelerator: Performances

Profiling result: based on FDSOI 28 nm technology

One cluster of 4 Neuro-Cores @ 1GHz: 32 GMAC/sec with 70mW power consumption, including memories and the controller

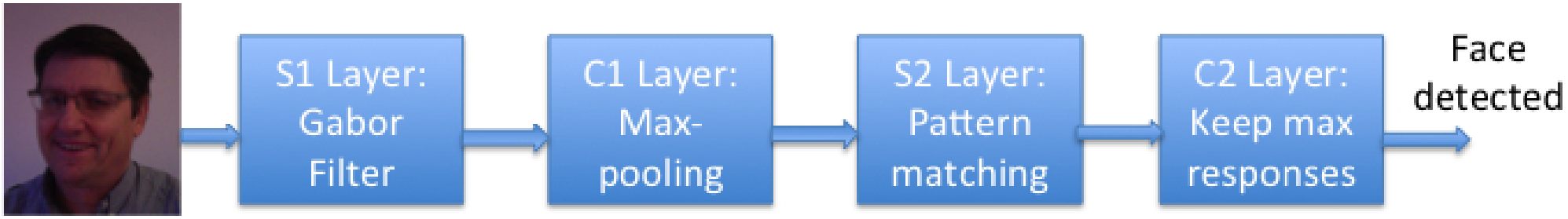
32 Neuro-Cores @ 1GHz: 1024 GMAC/sec – 2.2W

→ **Energy Efficiency: 465 GMAC.s^{-1}/W**

Full Hmax →One IP camera 1M pixels @ 30 fps: 103 GOP/sec

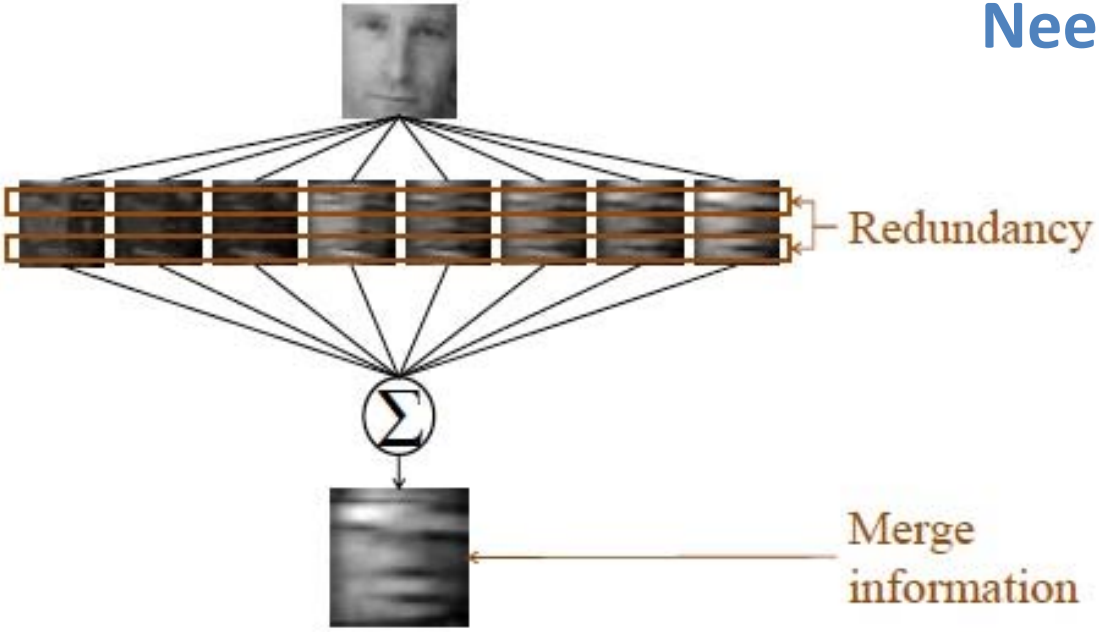
Needs 4 clusters of 4 Neuro-Cores (sup[103/32]) ⟶ **280mW**

Face Detection Application Example with Hmax



Complexity Calculation divided by 8 (merge 8 scales):

One 1M pixels camera @ 30 fps:
12.9 GOP.sec^{-1} (103 GOP.sec^{-1} /8)



Needs One Cluster with 2 NeuroCores:
Power consumption < 35mW

VGA Image @ 30 fps
only 1 NeuroCore: < 20 mW

Pneuro on FPGA

- **First demonstration on a FPGA-based Pneuro using ConvNet (CNN)**
  - Single cluster configuration (4 Neuro-Cores)
  - Embedded CNN application (60 neurons on the hidden layer, 450 KOps)
    - Faces extraction, 18000 images on the database, 96% recognition rate
    - Same application ported on 5 different architectures
      - Embedded CPU: Raspberry PI 2 B, Odroid Xu3
      - Embedded GPU: NVidia Tegra K1 (batch)
      - Desktop CPU: Intel I7
      - PNeuro, Quad Neuro-Cores
    - Using an internal prototyping board

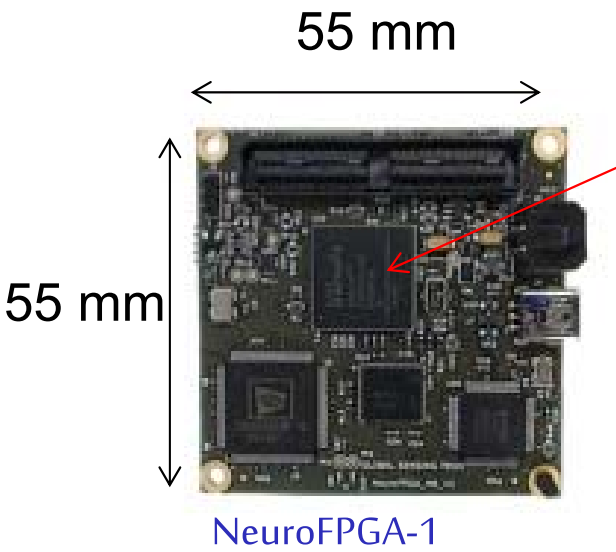
Target	Freq (MHz)	Energy Eff. (Images/W)	Perf (Images/s)
Intel I7	3400	160	5800
Quad ARM A15	2000	350	860
Quad ARM A7	900	380	480
Tegra K1	850	600	3550
PNeuro (FPGA)	**100**	**2000**	**4960**

- **FPGA approach is already competitive with existing CPU & GPU solutions**
  - First FPGA product developed for early 2017 by GST
    - Embedded FPGA: Artix 100 (~1W), 17.6cm² for the board, including one cluster

Pneuro on FPGA: Using NeuroFPGA

GlobalSensing Technologies

→ SmartNeuroCam



NeuroFPGA-1

+

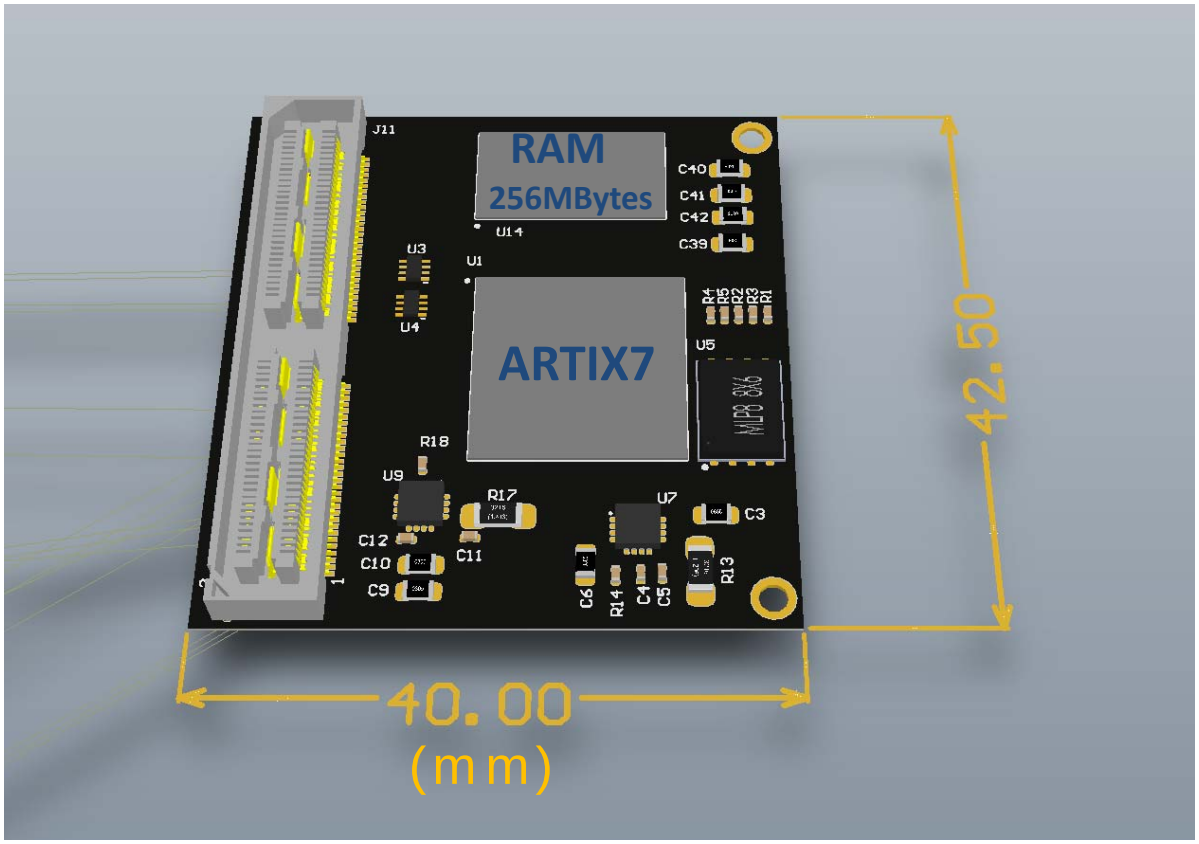
Camera Head

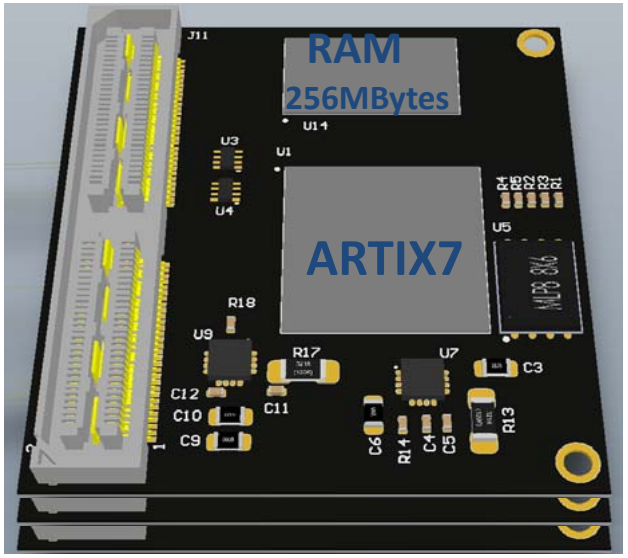
=

Aptina CMOS Image sensor:
752 x 480 pixels @ 60fps

Neuro-FPGA-2

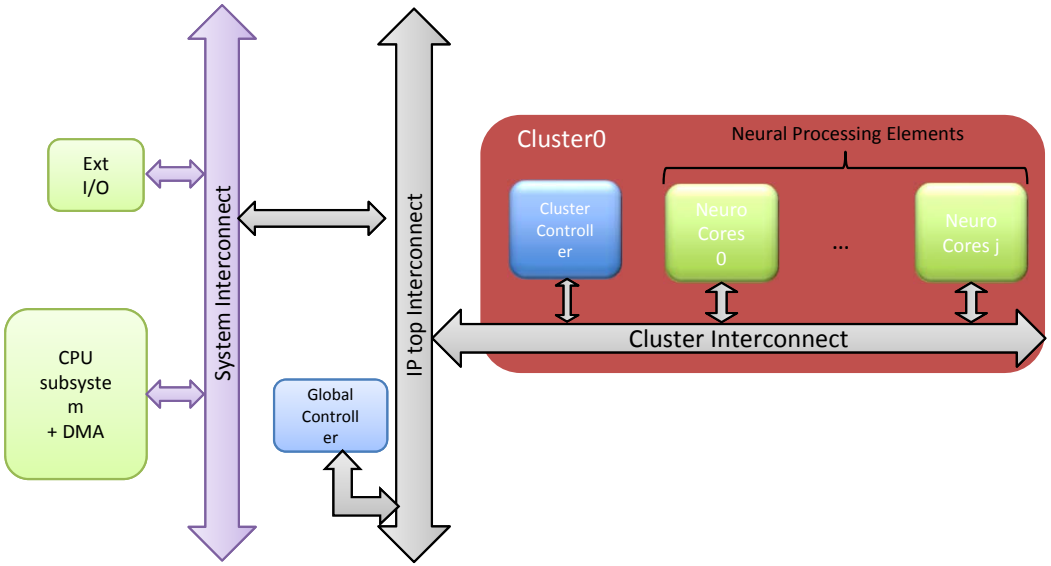
GlobalSensing Technologies

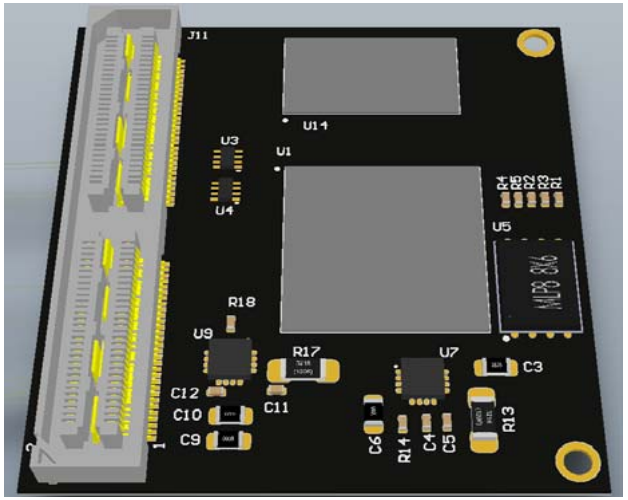




Scalability Capacity

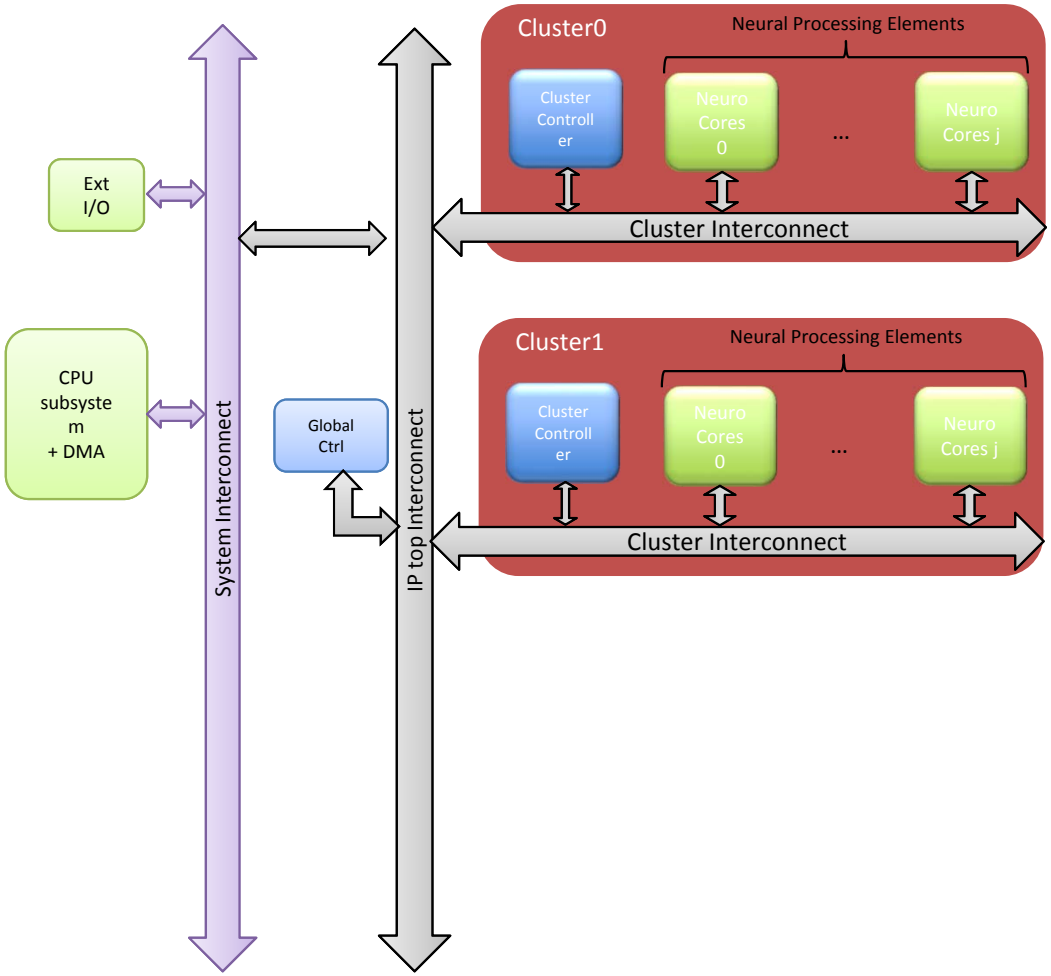
PNEURO LEVERAGING ON NEUROGPGA BOARD SCALABILITY

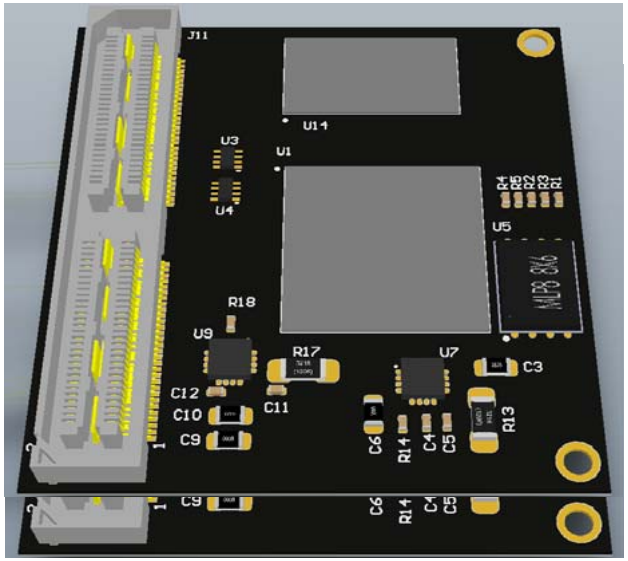




- **1 single-cluster Pneuro fits into one NeuroFPGA-2 board @100MHz**
  - 4 NeuroBlocs included providing 32 operations/cycle

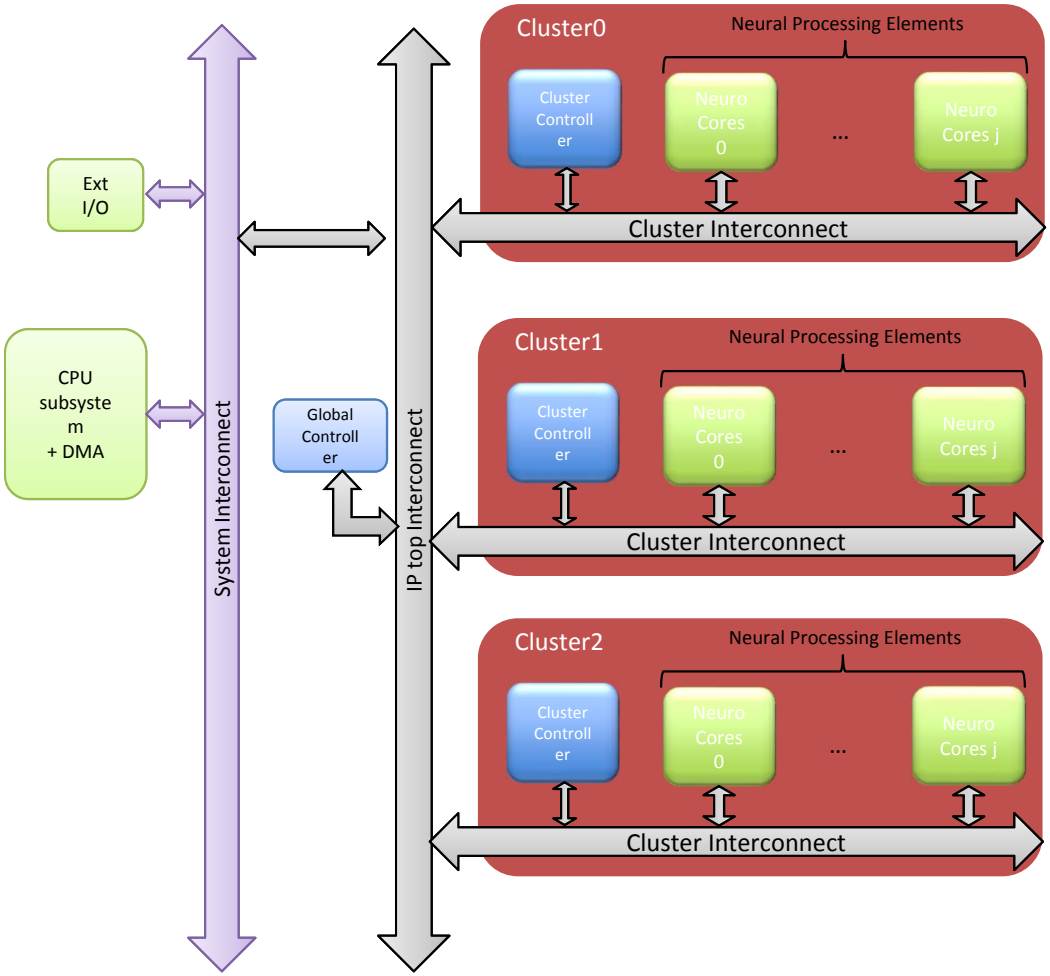
PNEURO LEVERAGING ON NEUROGPGA BOARD SCALABILITY

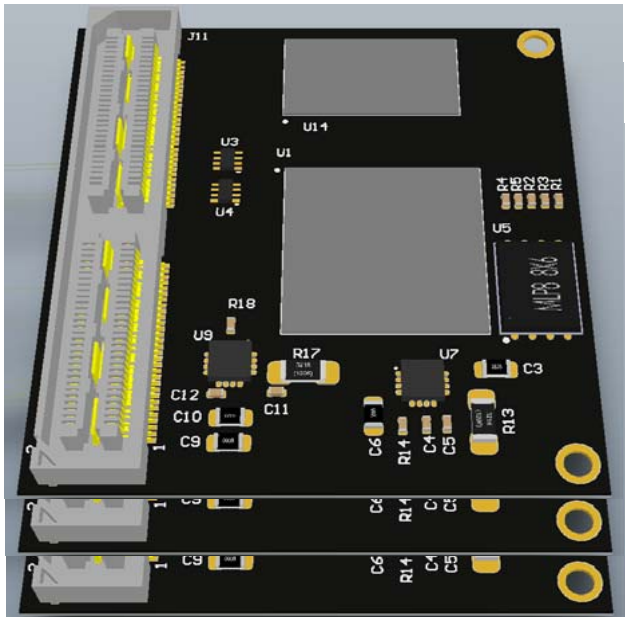




- **Additionnal clusters can fit in daughters and communicates through high bandwitdh multiboard interconnect**
  - Up to 200 high speed links shared betweens daughter boards

PNEURO LEVERAGING ON NEUROGPGA BOARD SCALABILITY





- NN Scalability properties are completely exploited thanks to a Board & IP Codesign between GST & CEA

• ASIC EVALUATION

- Caracterization chip in fabrication (tapeout end of june) in FDSOI 28nm
- Peak performances up to 1.8 TOPS/W @500MHz
- 0.4 mm² for a single cluster and its control, with a power consumption under 35 mW@500 MHz