Embedding Position Sensitivity in the Detector with ANNA: an Analog Neural Network ASIC

M.Amadori, M.Ronchi, G.Borghi, M.Carminati and C.Fiorini

Politecnico di Milano and INFN, Milano, Italy

PSD14: The 14th International Conference on Position Sensitive Detectors
London (UK), 03-09-26

RadLab

INFN

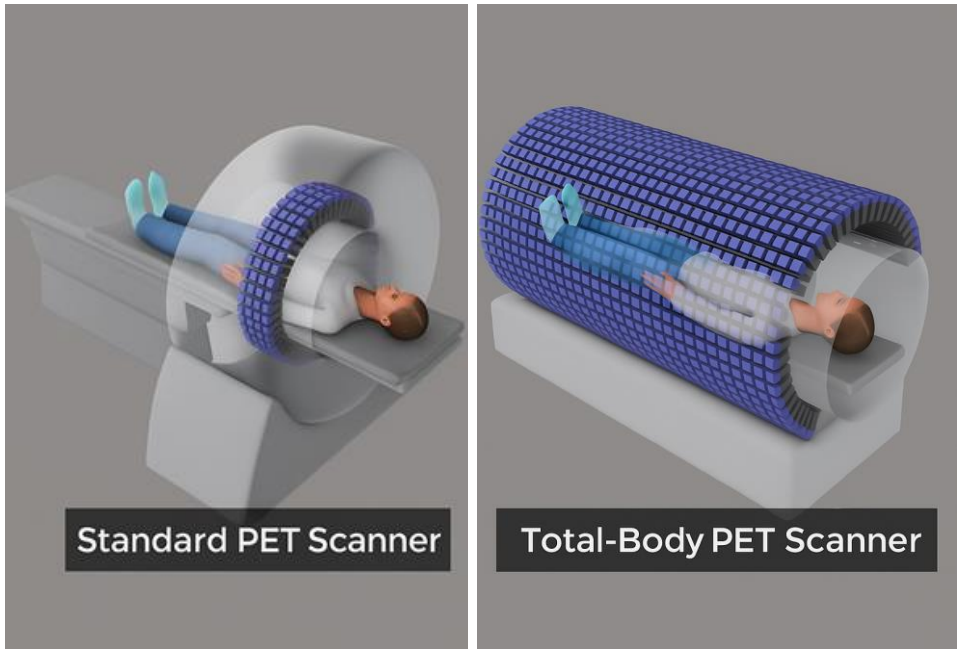
Outline

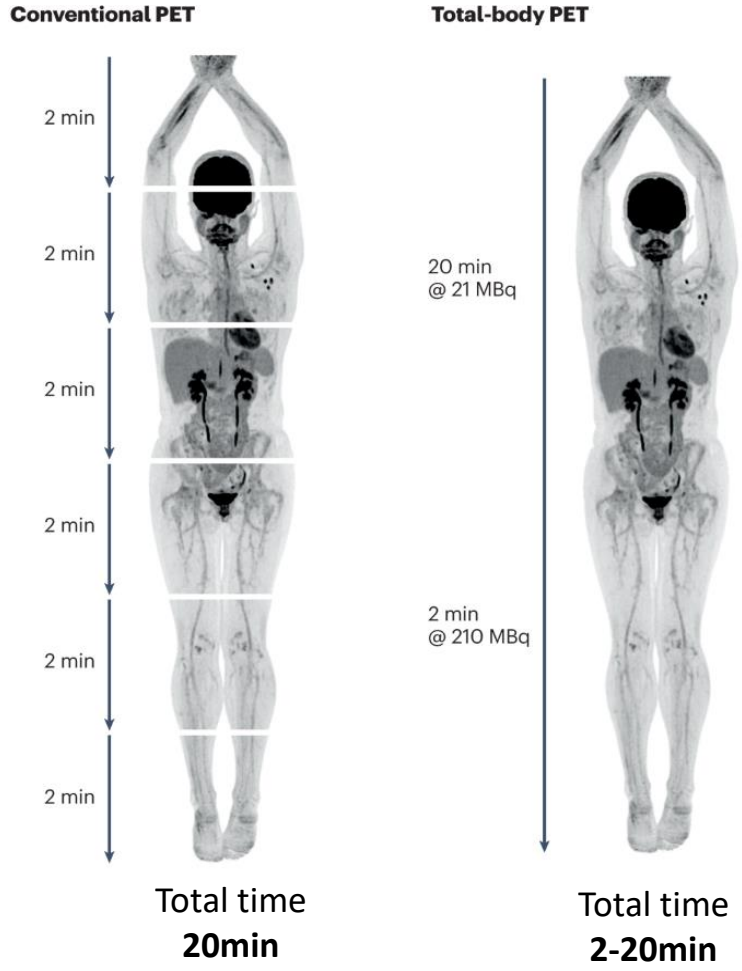
- Motivation: edge-processing in large detection systems (Total-Body PET)
- Neural networks for event reconstruction in monolithic scintillators
- The ANNA ASIC: V1 and V2
- Experimental measurements
- Future application for LGADs readout
- Conclusions

Motivation: edge-processing in large detection systems (TB PET)

Trend towards Total Body PET systems

(Katal S. et al. - Advantages and Applications of Total-Body PET Scanning. Diagnostics, 2022)





Advantages

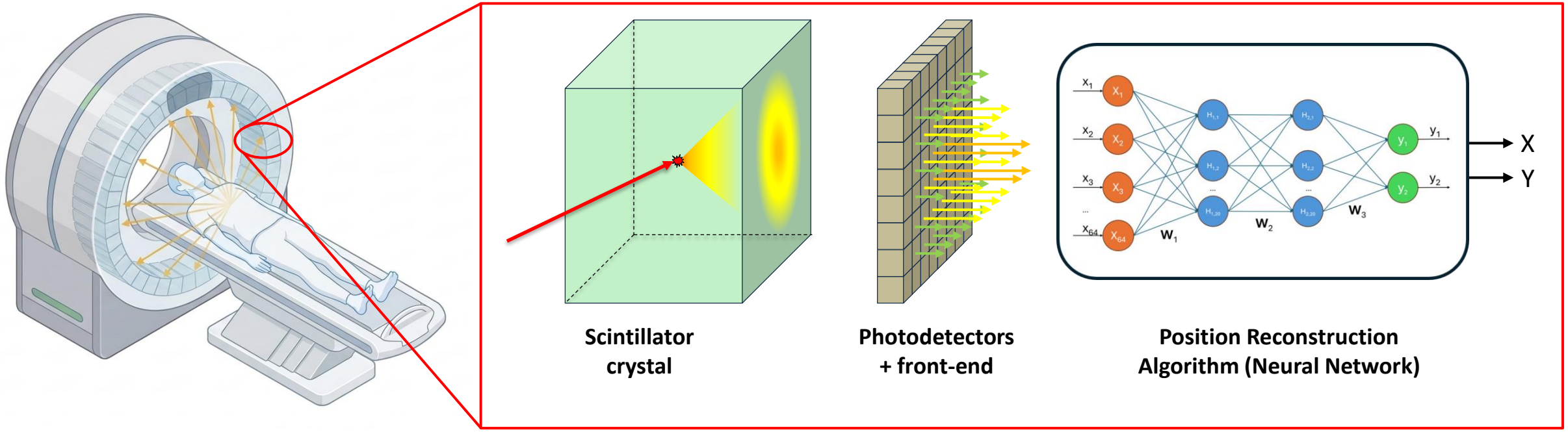
- Full-body coverage
- Higher sensitivity
- Faster scans
- Lower radiation dose

Disadvantages

- High number of channels
- Complex system implementation
- High power

⇒ Low-power, edge-processing of detector signals required
(also for other applications with large number of detector modules)

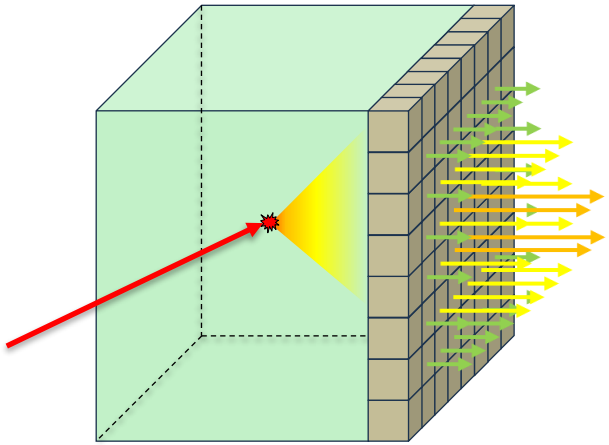
Monolithic Scintillator Detector and Neural Network Reconstruction



- **Monolithic Scintillator**: simplifies the scintillator structure (vs. pixelated), facilitate DOI, at cost of possible pile-up
- **SiPMs Readout**: now quite standard vs. PMTs as more compact and insensitive to magnetic fields
- **Neural Network Algorithm:** better and faster estimation of γ-ray interaction in the crystal, vs. Anger and other methods

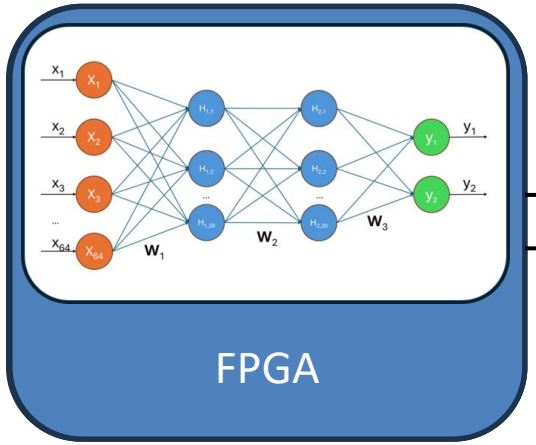
Traditional Digital Processing vs Analog Processing (ANNA)

Traditional Processing

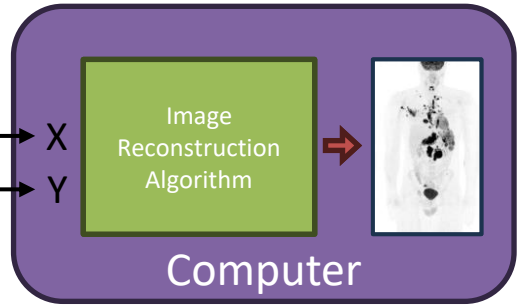


64 ADCs

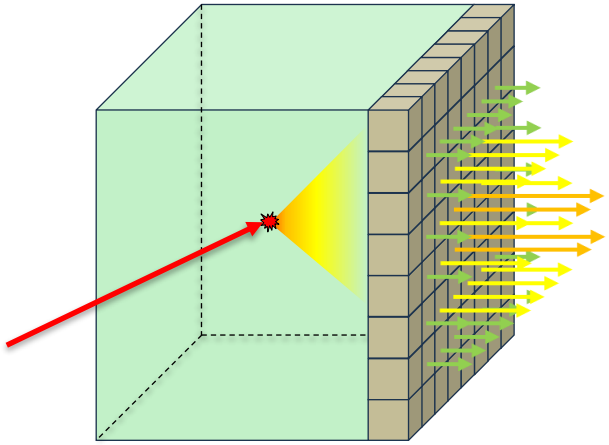
x 64 digital signals



X
Y



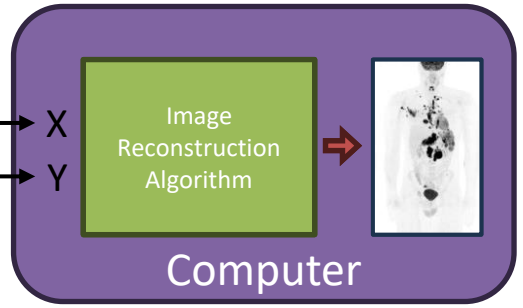
Analog Processing



2 ADCs

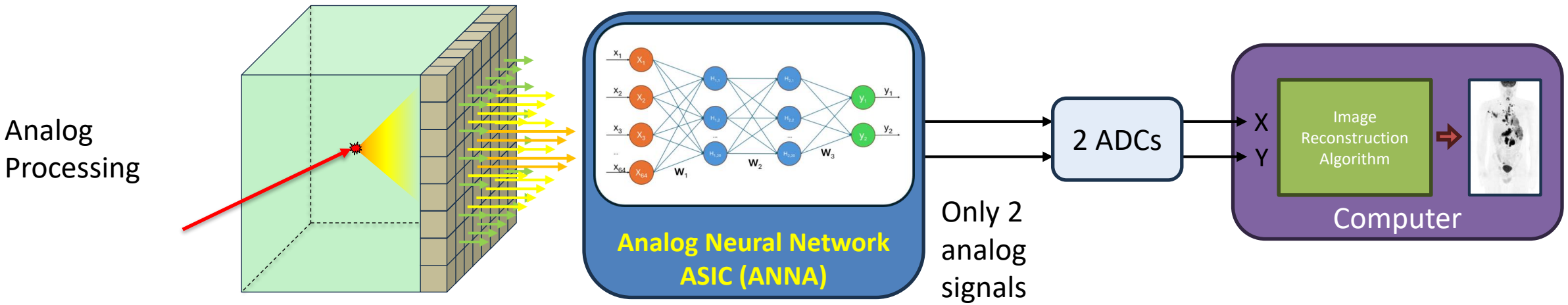
Only 2 analog signals

X
Y



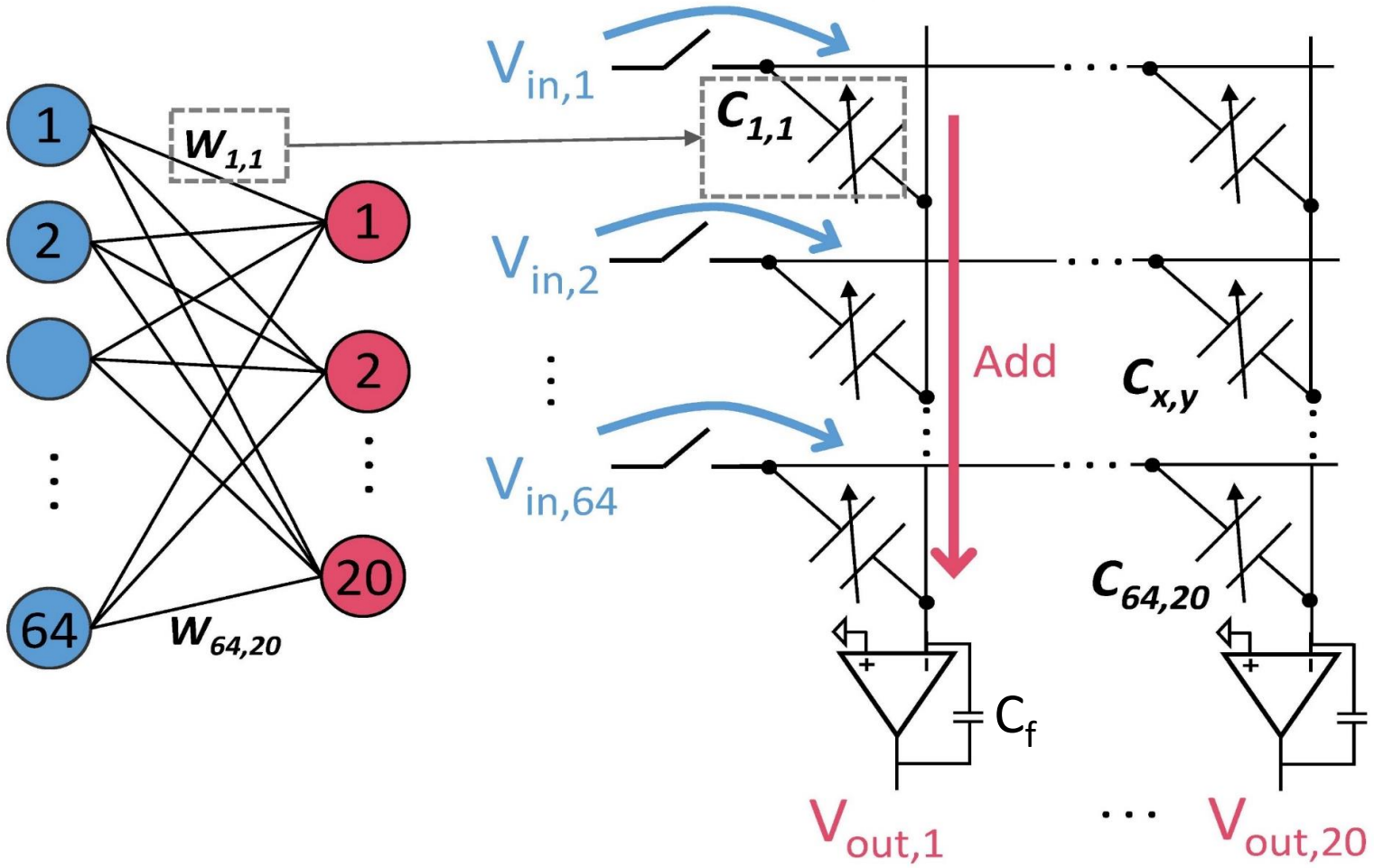
Traditional Digital Processing vs Analog Processing (ANNA)

- **Analog operations** performed directly on **analog signals** coming from photodetectors
- **No need for ADC and FPGA** for embedded processing
- **Interaction coordinates (X,Y)** directly at the output of the **ASIC**
- **Compatibility with Front-end ASIC** for SiPMs current processing (Amplitude, Timing)
- **Same approach possible also for other segmented solid-state detectors**
- **Analog approach (as memristors) to in-memory computing** (vs. digital)



ANNA: Analog Multiply and Accumulate (MAC) principle

Multiplication: $Q_{1,1} = C_{1,1} \cdot V_{in,1}$



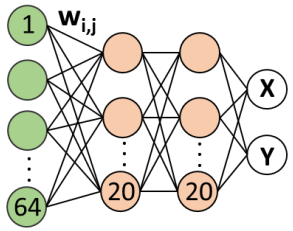
Features:

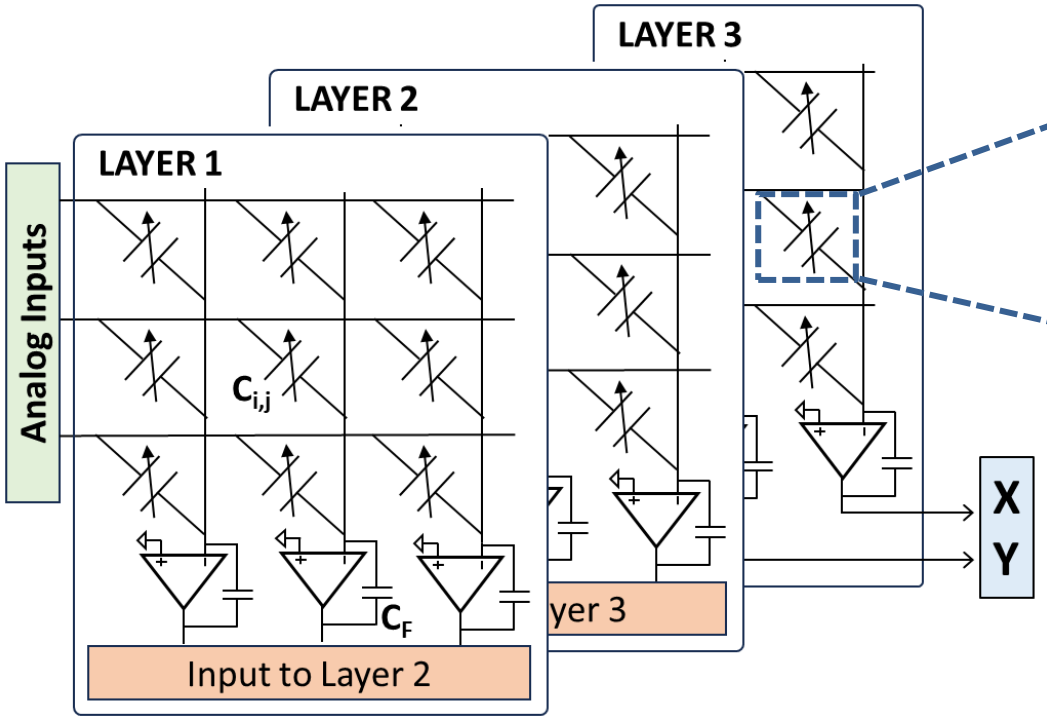
- Capacitive **cross-bar array**
- Charge **sum** by an integrator
- **Activation function** implemented in the integrator

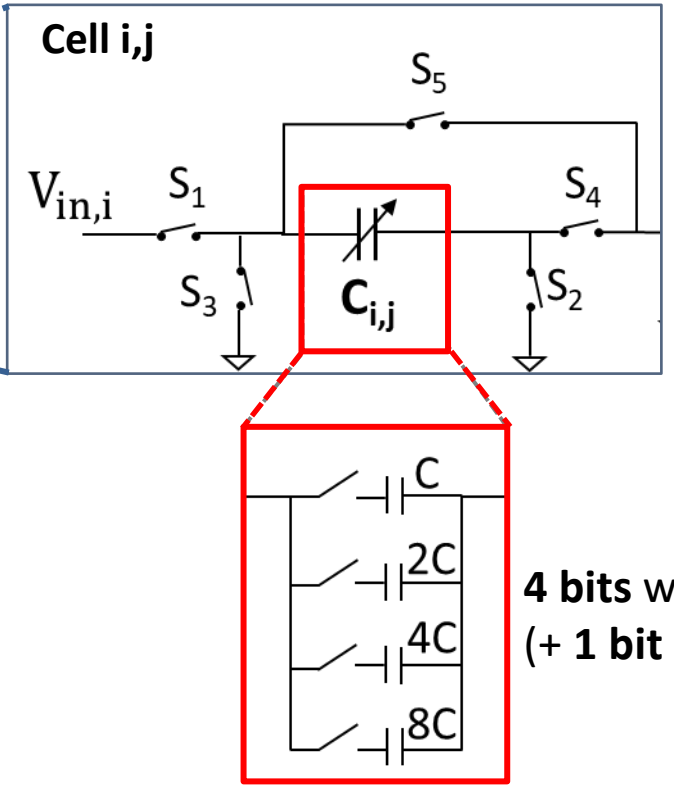
$$V_{out,1} = f\left(\frac{1}{C_f}\sum_{j=0}^{64} C_{1,j} \cdot V_{in,1}\right)$$

f: *activation function*

ANNA (Analog Neural Network ASIC): V1 version (proof-of-concept)





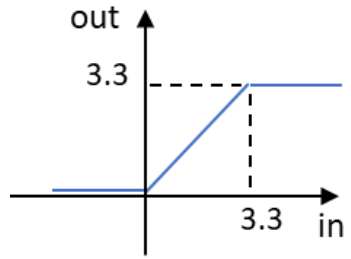


4 bits weight
(+ **1 bit** sign)

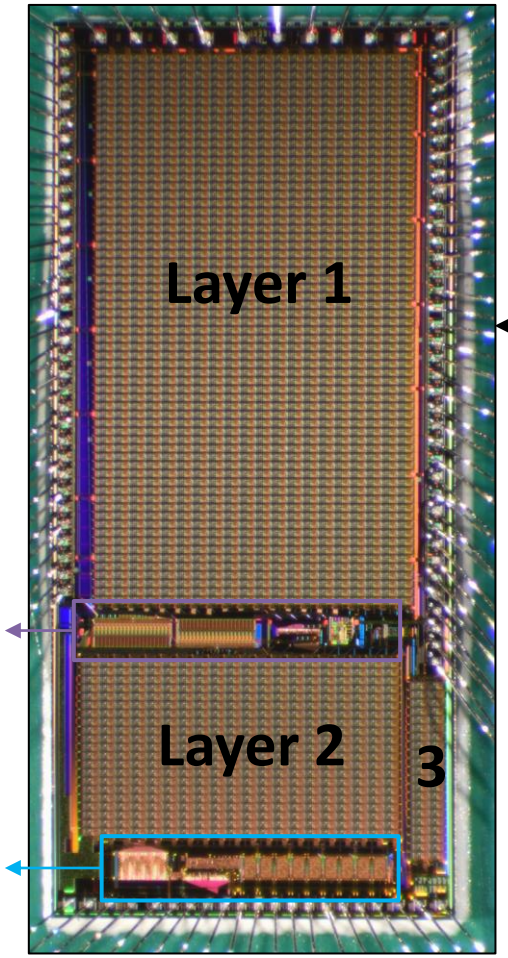
MAC: $V_{out,j} = \sum \frac{C_{i,j}}{C_F} V_{in,i}$

weight

- **Switches**: allows to circulate the charge in two directions (positive and negative weights)
- **Activation function**: clipped ReLU implemented by the op-amp bias limits
- **Quantization-Aware Training**

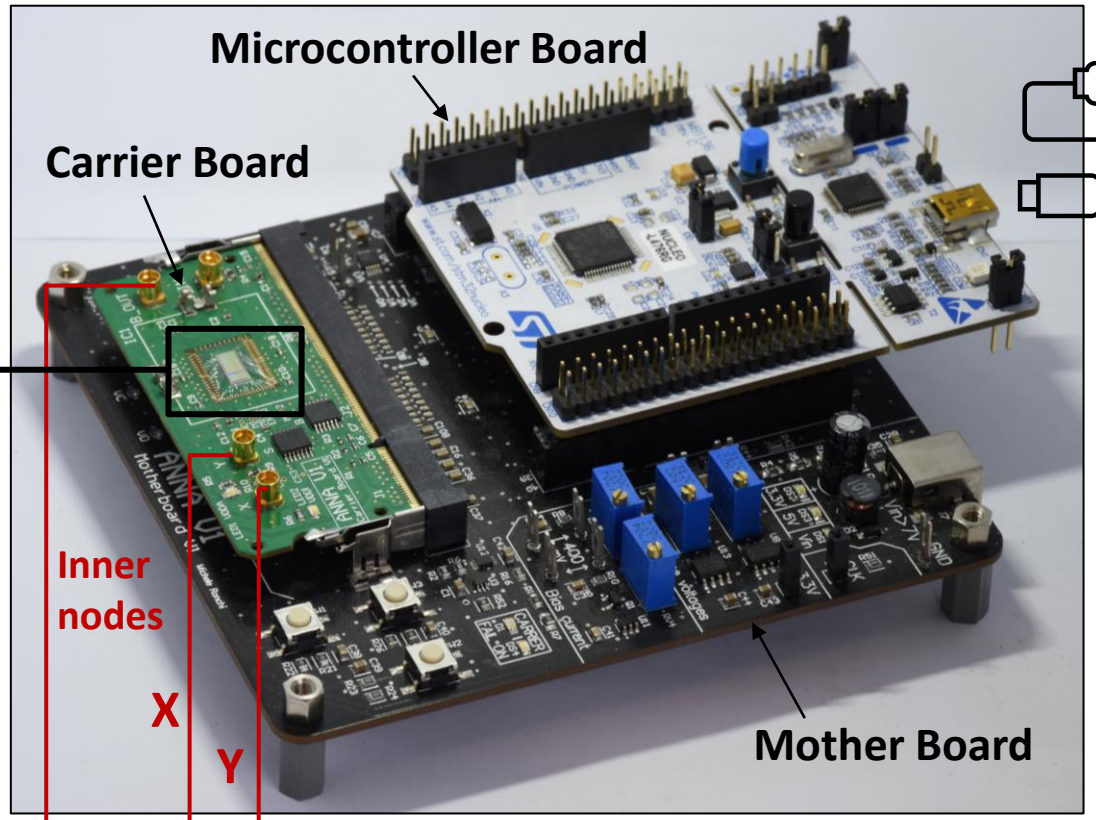


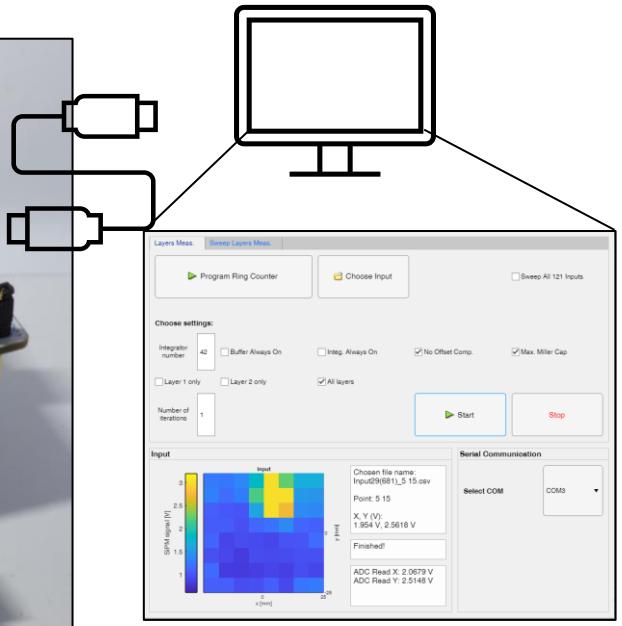
Characterization of ASIC V1 (CMOS 0.35 µm, 24 mm²)



Integrators

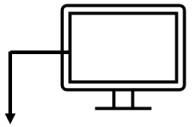
Digital (SPI)





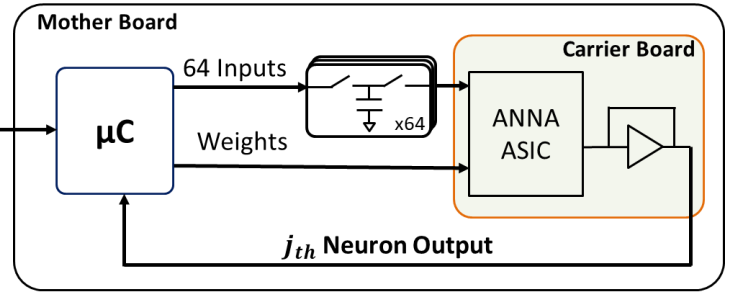
Graphic interface:

- ASIC programming
- Inputs delivery
- Data acquisition



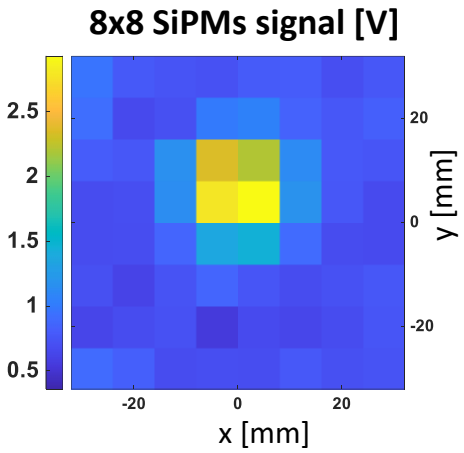
I/O

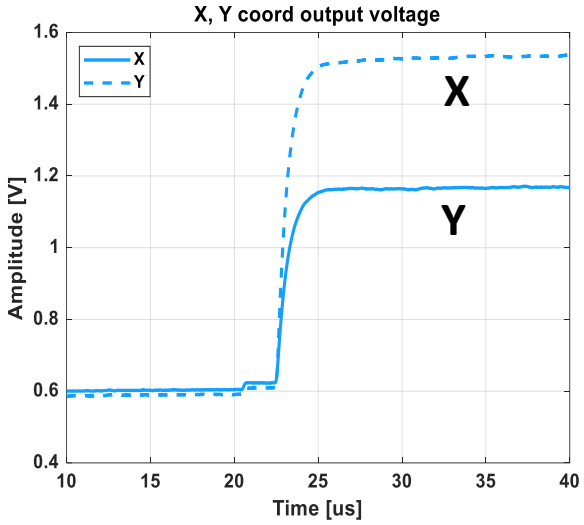
Inputs/weights generation
Data acquisition
Model fitting

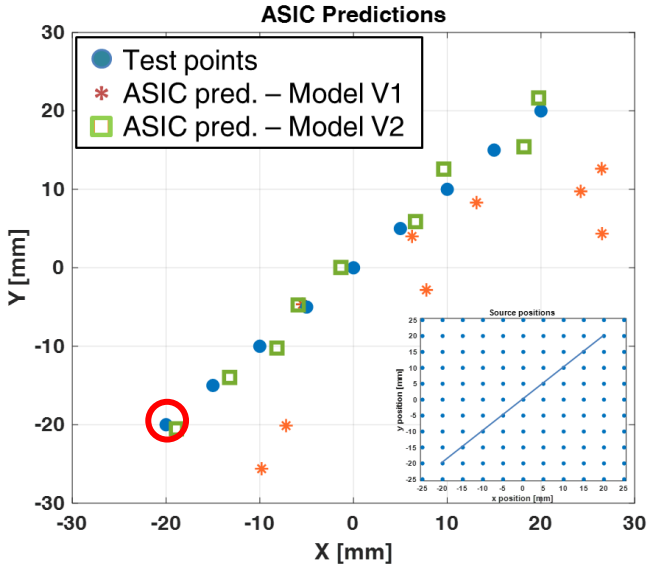


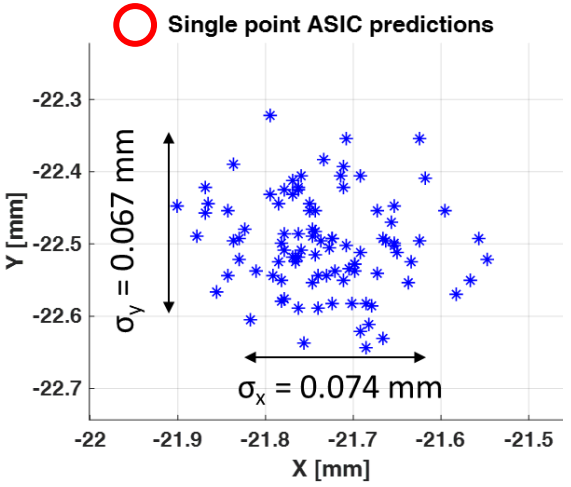
ANNA V1 Performance Characterization

Grid of **121 simulated points** (ANTS2) **provided as ASIC voltage inputs**

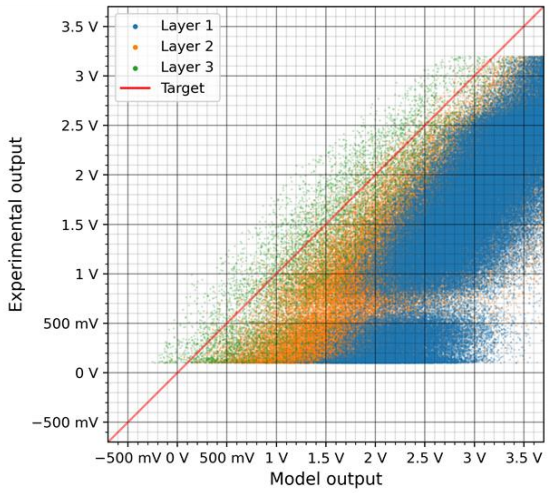




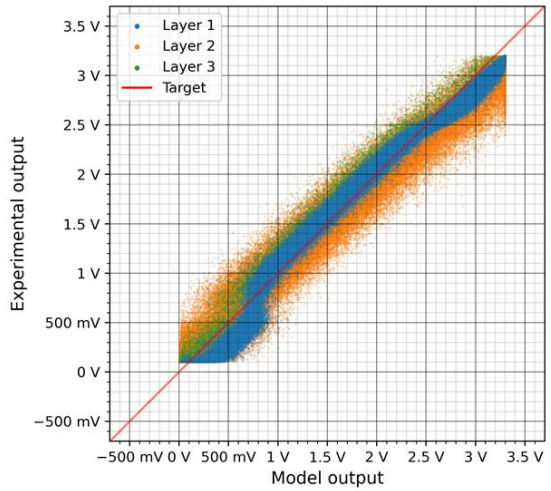


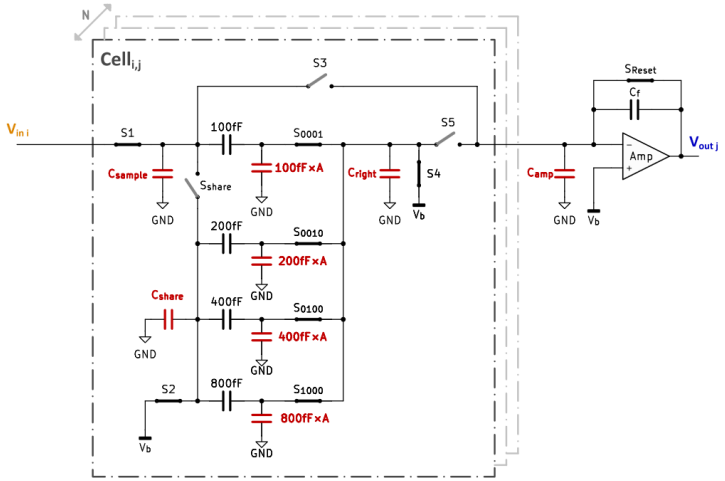


Original, simplest model



Hardware-aware model





Hardware-aware model, used during training, takes into account parasitics, offsets and charge injections, measured during stand-alone test of the ASIC

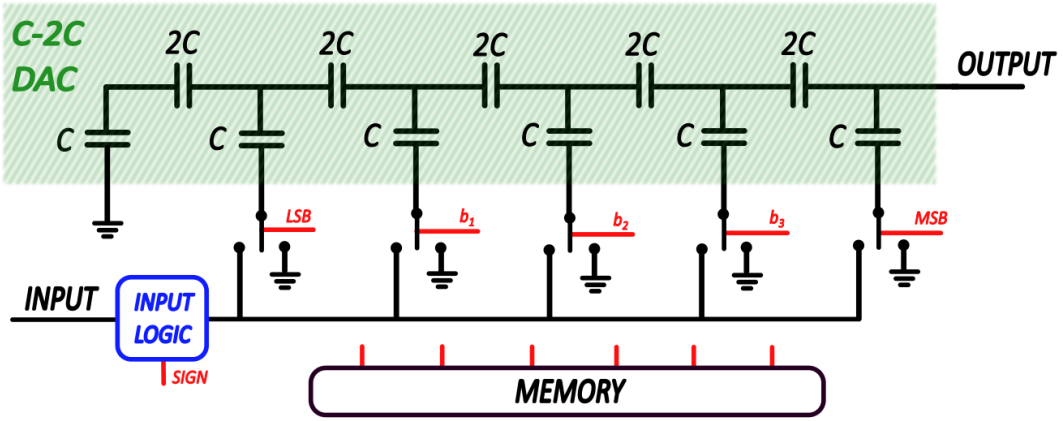
ANNA V2

- **New improved architecture (including recursive network)**
- **Parasitics**: more repeatable and easier to characterize, access to raw analog outputs of each layer
- **Complexity**: simplified neuron design, fewer steps needed for MAC operations

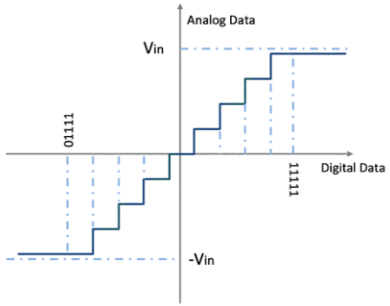
	ANNA V1	ANNA V2
Process node	350 nm	110 nm
Supply voltage	3.3 V	1.2 V
Area	23.5 mm^2	9.5 mm^2
Neural network config	**Fixed**	**Programmable**
Neural network	64 inputs, **2 outputs** **64 x 20 x 20 x 2** (**3 weighted layers**)	69 inputs, **32 outputs** **69 x 32 x 32 x 32 x 32 x 32** (**5 weighted layers**)
Parameters (weights and biases)	1762 (5 bits – 31 analog levels)	6496 (**6 bits – 63 analog levels**)
Inference execution time	**4.5 µs**	**< 1.5 µs**
Throughput	**0.77 GOP/s**	**> 10 GOP/s**
Efficiency	**0.05 TOPs/W**	**> 5 TOPs/W**

ASIC V2 design: new neuron schematic

New Multiplying Element (C-2C DAC)

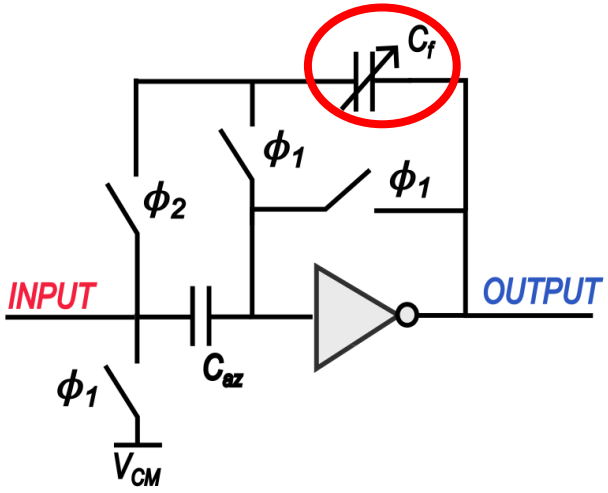


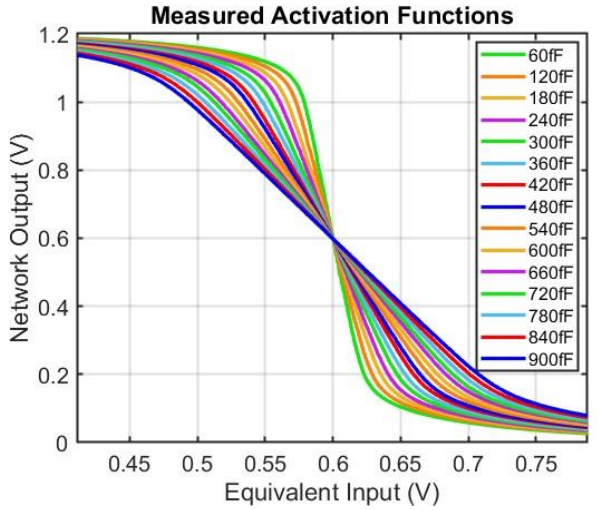
$$-\frac{31}{32} \leq W_i \leq \frac{31}{32}$$



- 6-bit weights precision
- Output capacitance always 2C
- Charge injection reduction by design
- Power reduction due to C reduction (30fF)

New Integrator (inverter-based)

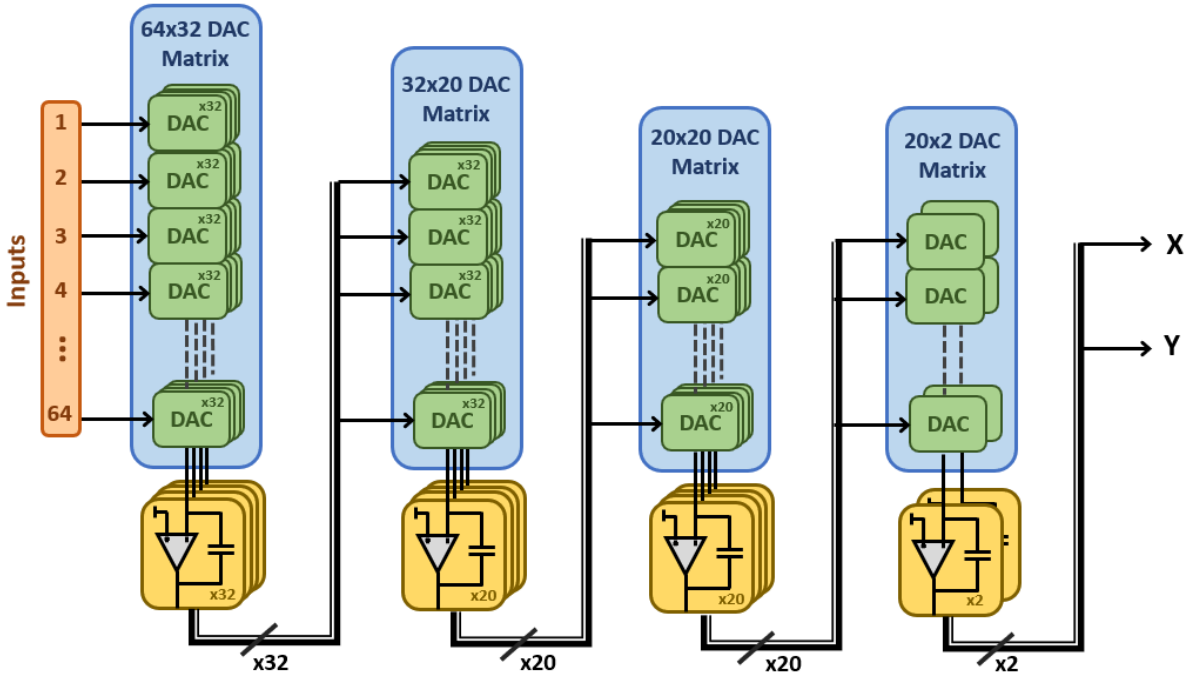




- Tanh-like activation function
- Variable feedback capactitor (15 values)
- Programmable power consumption
- More compact layout

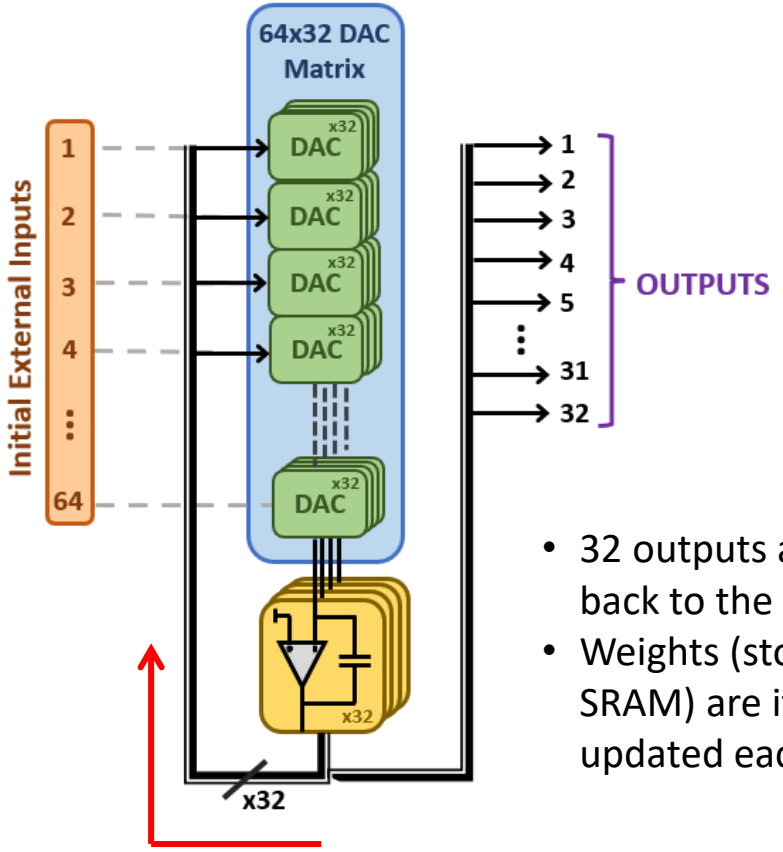
ASIC V2 architecture: recursive network

Standard network (as in V1)



- ✗ High Area occupation
- ✗ No Flexibility
- ✓ Simple Design

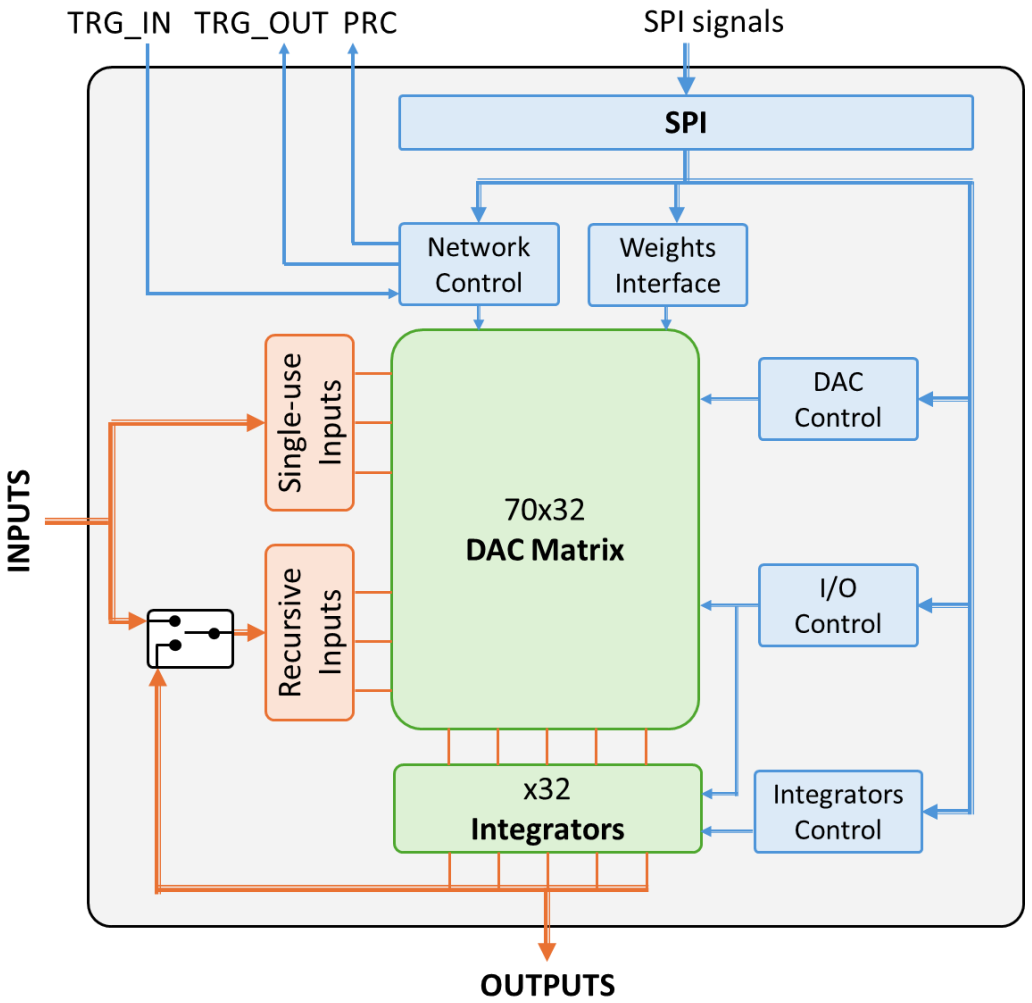
Recursive network



- 32 outputs are feed back to the inputs
- Weights (stored in SRAM) are iteratively updated each layer

- ✓ Low Area
- ✓ High Flexibility
- ✗ Difficult Design

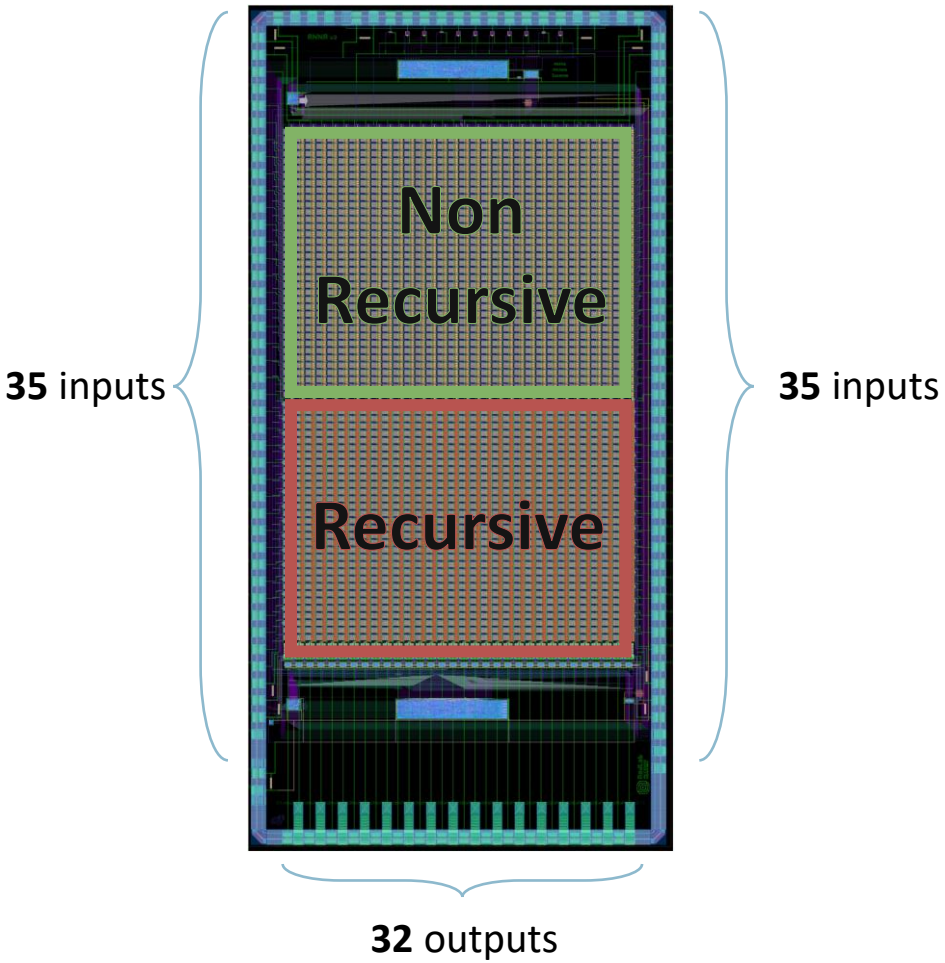
ANNA V2 – Architecture and Layout



- **Fully programmable** neural network (layers and neurons) and parameters (weight, biases, activation functions)
- Configurable **power consumption** and **power cycling**
- Programmable **layer duration** and synchronization
- Trigger based input processing and options to improve throughput

ANNA V2

Single reusable layer

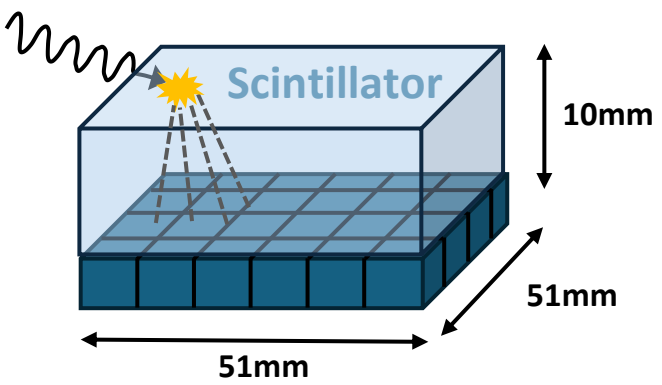


64x20x20x2 Network - Results

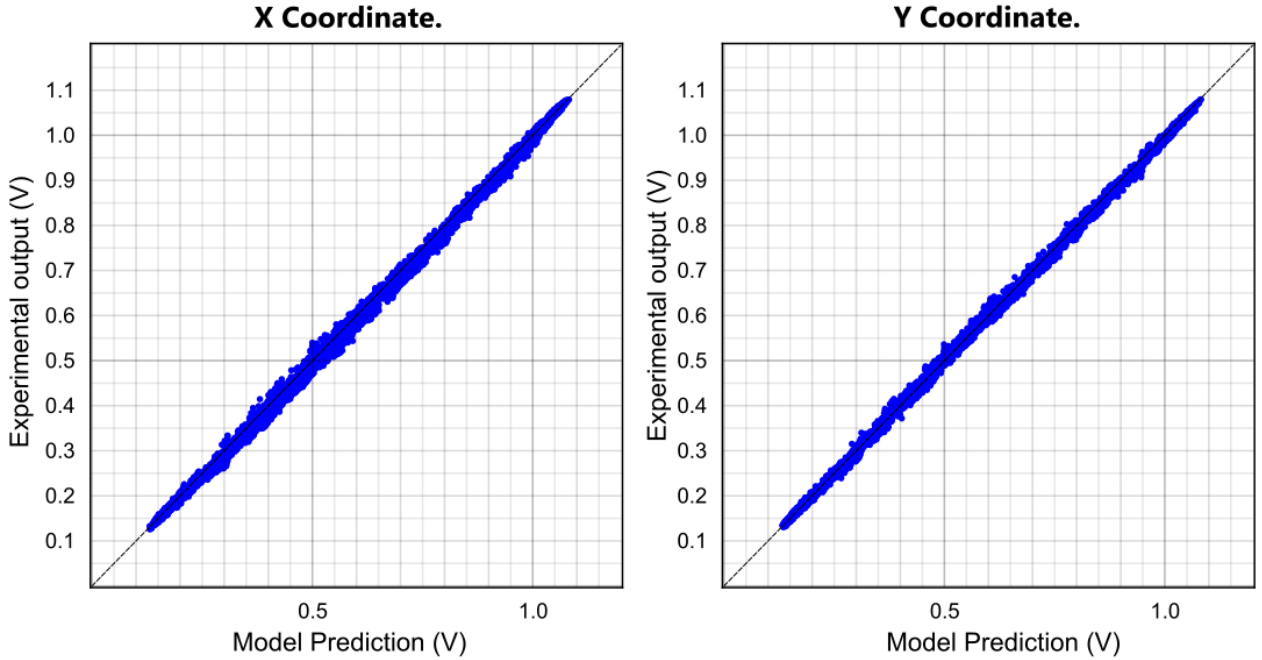
Grid of gamma-ray events emulated with **ANTS3**

Simulated events on a **51mm x 51mm LYSO scintillator**

Linear mapping between mm and Voltages $\mathrm{d_{mm}} = \frac{51_{[\mathrm{mm}]} \cdot V_{OUT\,[V]}}{V_{FSR\,[V]}}$



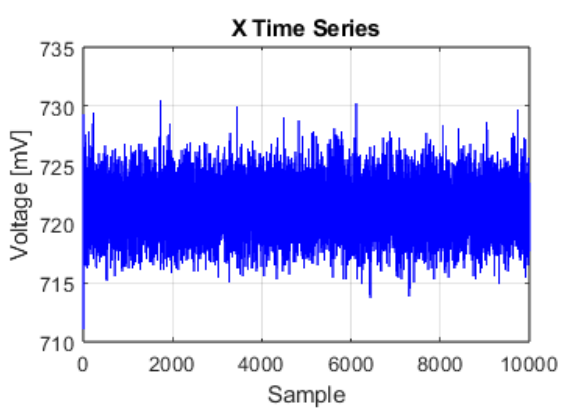
Model vs ASIC Results

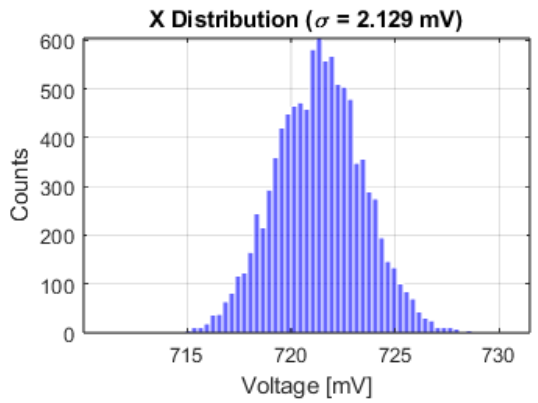


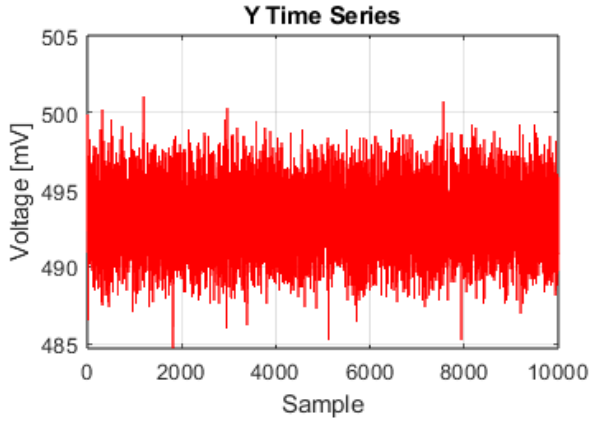
Mean Absolute Error (MAE)

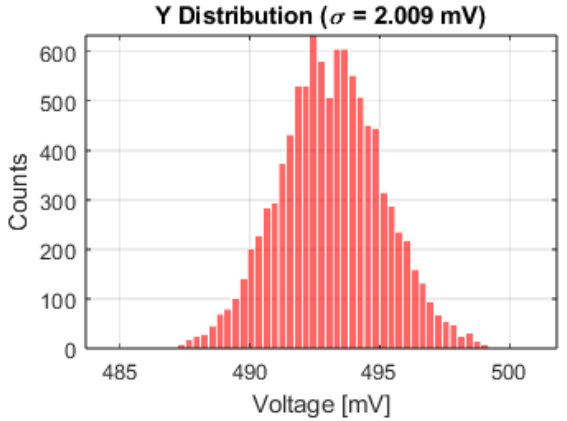
Tensorflow Model	~ 0.9 mm
ANNA	**~ 1.1 mm**

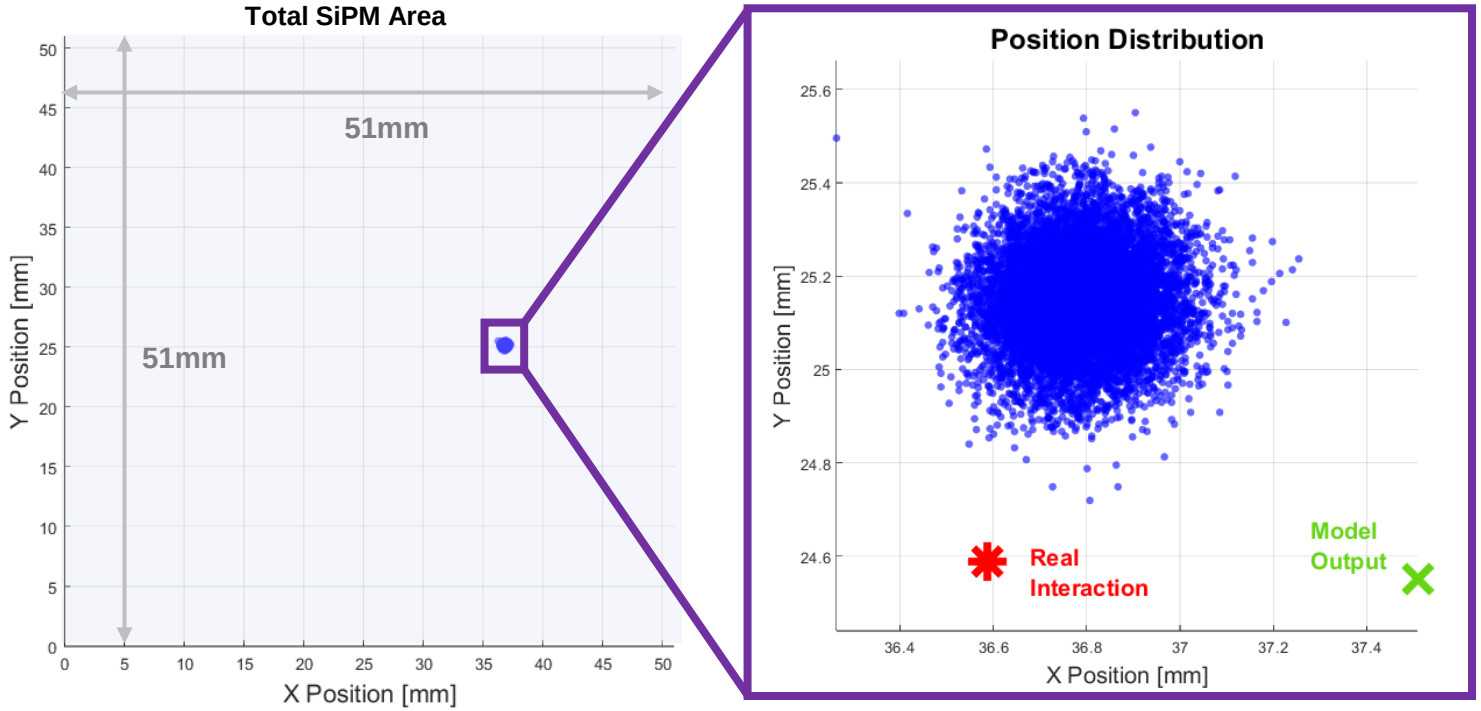
64x20x20x2 Network – Noise Analysis











Noise of **10000 events**

Cf = 120fF for every layer

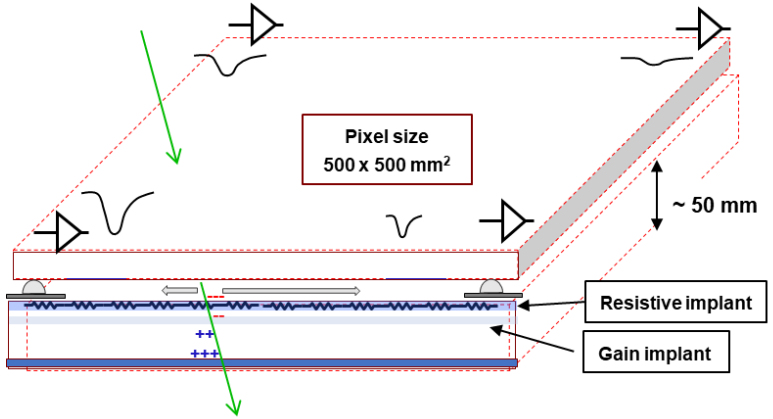
Spatial Precision of

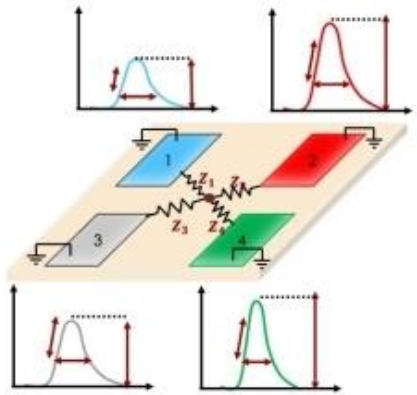
$$\sigma_X = 108\ \mu m$$

$$\sigma_Y = 102\ \mu m$$

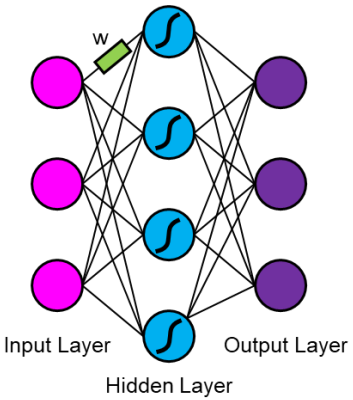
Future Developments – LGADs signals processing with ANNA

1 **Resistive LGADs (DC-RSD)** exploits **charge sharing** to achieve excellent spatial resolution, much smaller than the pixel size, reducing number of readout channels, power and material budget (N.Cartiglia, NIMA, 2022)

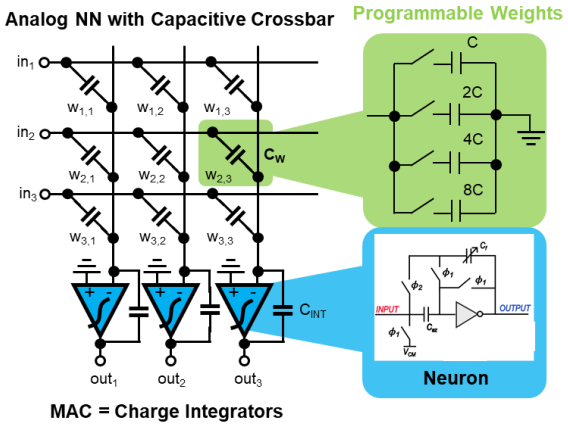


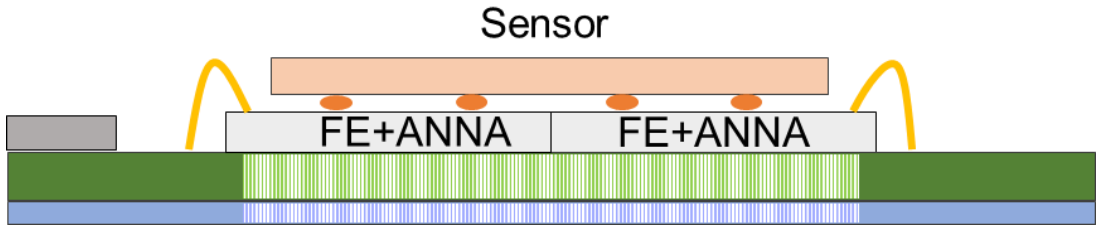


2 **Neural networks (NNs)** allows a fast and energy-efficient reconstruction of events in DC-RSD

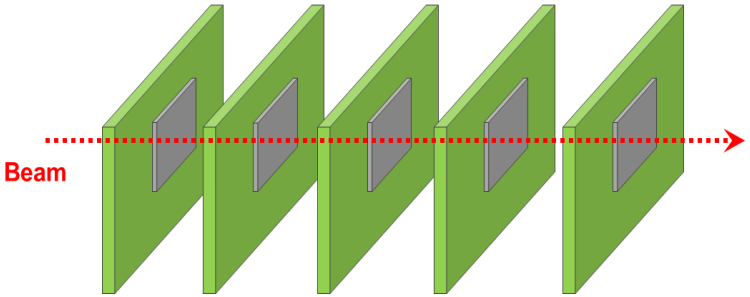


3 **ANNA** implements the NN directly on-chip





We aim to develop a demonstrator for a new, low-power 4D tracker based on **LGADs** with embedded reconstruction of the event features (position, timing) by **ANNA**



Activity in collaboration with INFN-TO/FI, UNI-TO, UNI-Pe

Conclusions and future work

- **ANNA** brings neural-network processing directly into the **analog detector readout chain**, avoiding prior signal digitalization.
- The new ANNA V2 architecture combines high programmability, compact implementation and **ultra-low-power** analog computation.
- Experimental results demonstrate **accurate position reconstruction**, not too far from the TensorFlow model.
- This approach enables true **edge processing in large-scale detector systems**, reducing data throughput and system-level complexity.
- Beyond monolithic scintillators, ANNA opens the way to **“intelligent detectors” with embedded reconstruction**, including future **low-power LGAD trackers**.

Next steps:

- Construction (ongoing) and test of **PETRARCA**: a PET detector with embedded, ANNA-based position recontruction
- Development of a proof-of-concept prototype of **resistive LGAD detector with ANNA processing**

Relevant Publications

- S. Di Giacomo, et al., "Implementing an Integrated Neural Network for Real-Time Position Reconstruction in Emission Tomography with Monolithic Scintillators“, IEEE TRANSACTIONS ON RADIATION AND PLASMA MEDICAL SCIENCES, VOL. 8, NO. 5, MAY 2024.
- S. Di Giacomo, et al., «Experimental Validation of ANNA: Analog Neural Network ASIC for Event Positioning in Monolithic Scintillation Detectors», IEEE Transactions on Radiation and Plasma Medical Sciences, 2025.
- M.Amadori, et al., «A Compact Neuron Implementation for a Capacitive Analog Neural Network for In-Sensor Processing», 2024 31st IEEE International Conference on Electronics, Circuits and Systems (ICECS).
- M. Ronchi, et al., «Design, Implementation and Analysis of an Integrated Switched Capacitor Analog Neuron for Edge Computing AI Accelerators», IEEE Transactions on Circuits and Systems I: Regular Papers, 2025.

Acknowledgements

This work was supported by the IANNET project (Italian **MUR**) and by the ANNA project (Italian **INFN**)

The RADLAB team

Thank you for your attention!

Backup

Analog Neuron Operation

Charge-redistribution approach

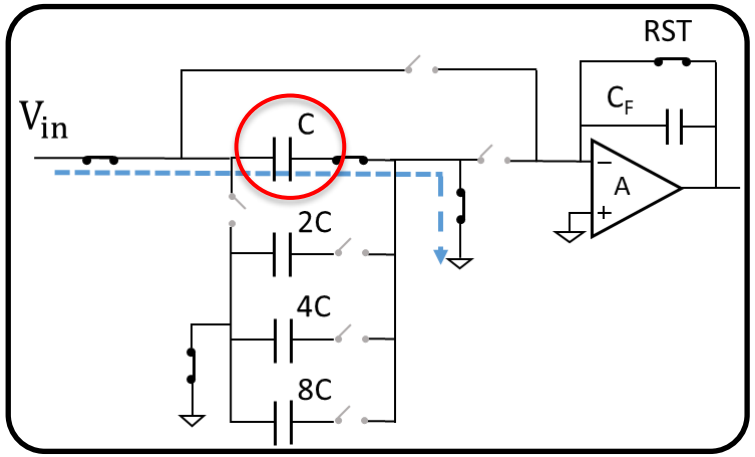
1) Charge **only C** with V_{in} to **minimize energy consumption** ($E = CV^2$)

2) Charge is **redistributed** among all capacitors closing S_s

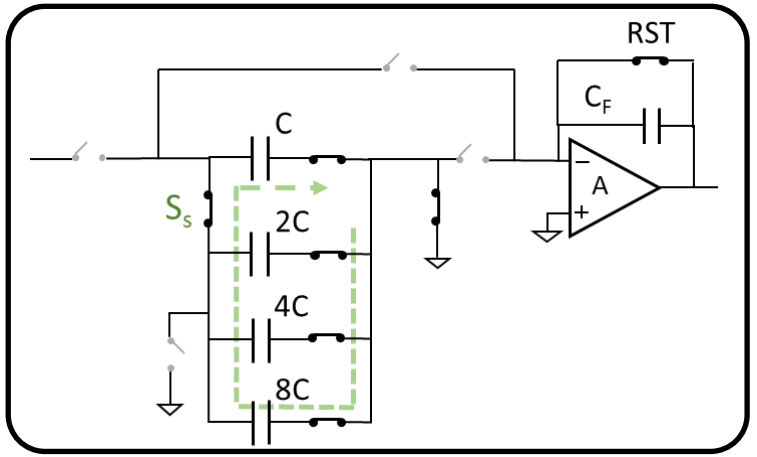
3) **Only the capacitors** corresponding to the **desired weight** are connected to the integrator, while the others are disconnected by means of their respective switch.

The output is $$V_{out,j} = \frac{V_{in,i}}{15}\frac{C_{ij}}{C_F}$$

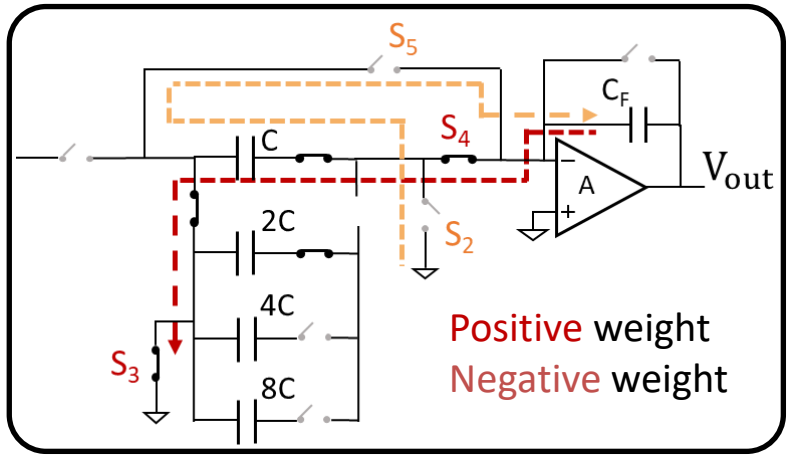
Step 1



Step 2



Step 3

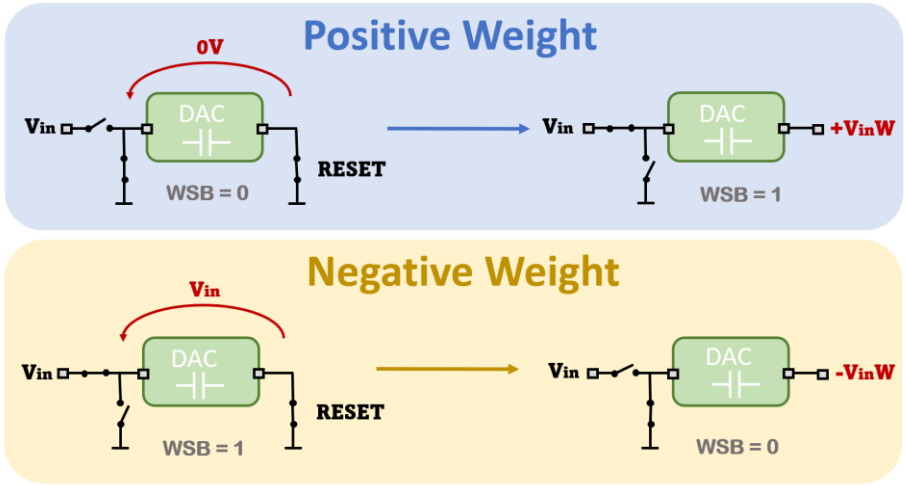


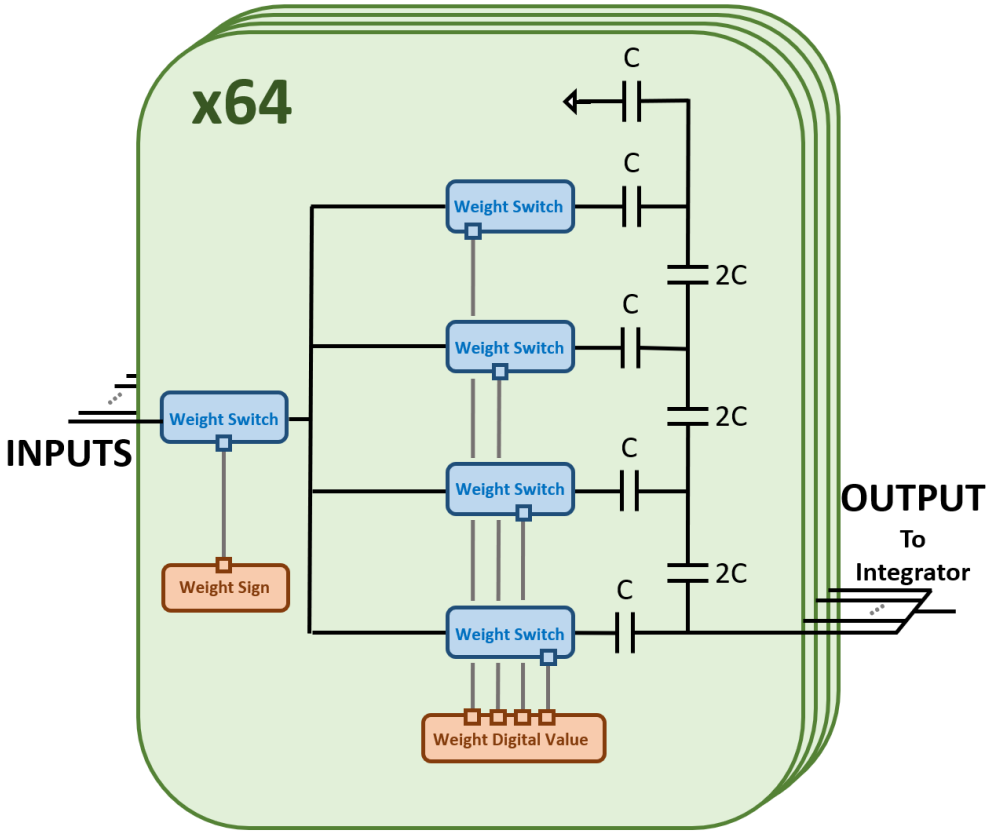
ANNA V2 – Multiply and accumulate

The multiplication phase is implemented through a 2-2C architecture, providing a weighted charge to the integrator.

Compared to the previous implementation:

- **Faster** (less switching steps)
- **Much more compact** (less capacitors: above 4 bits the C-2C is more area-efficient than binary weighted)
- **More accurate** (consistent parasitics, constant output impedance)

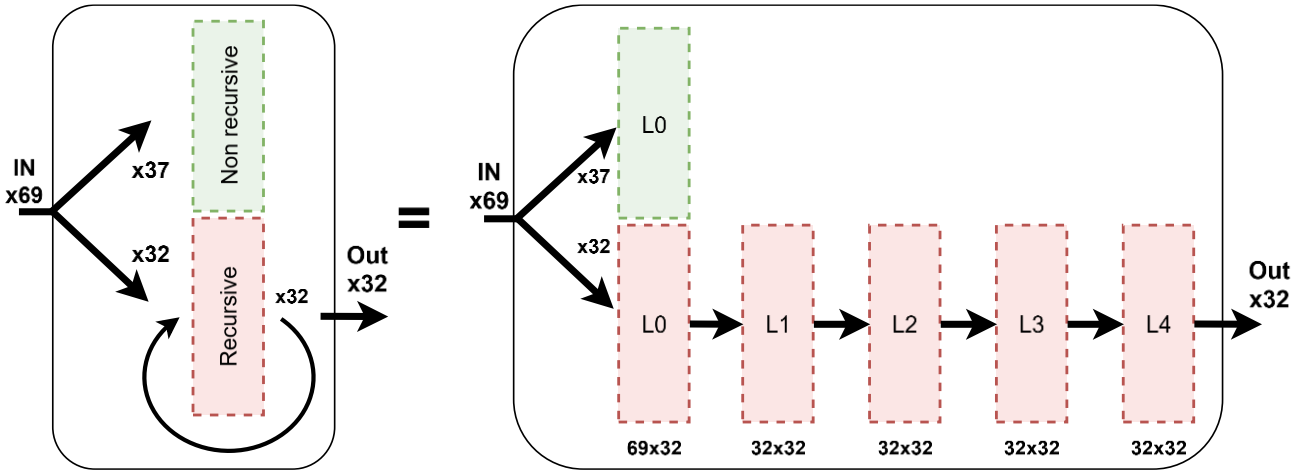


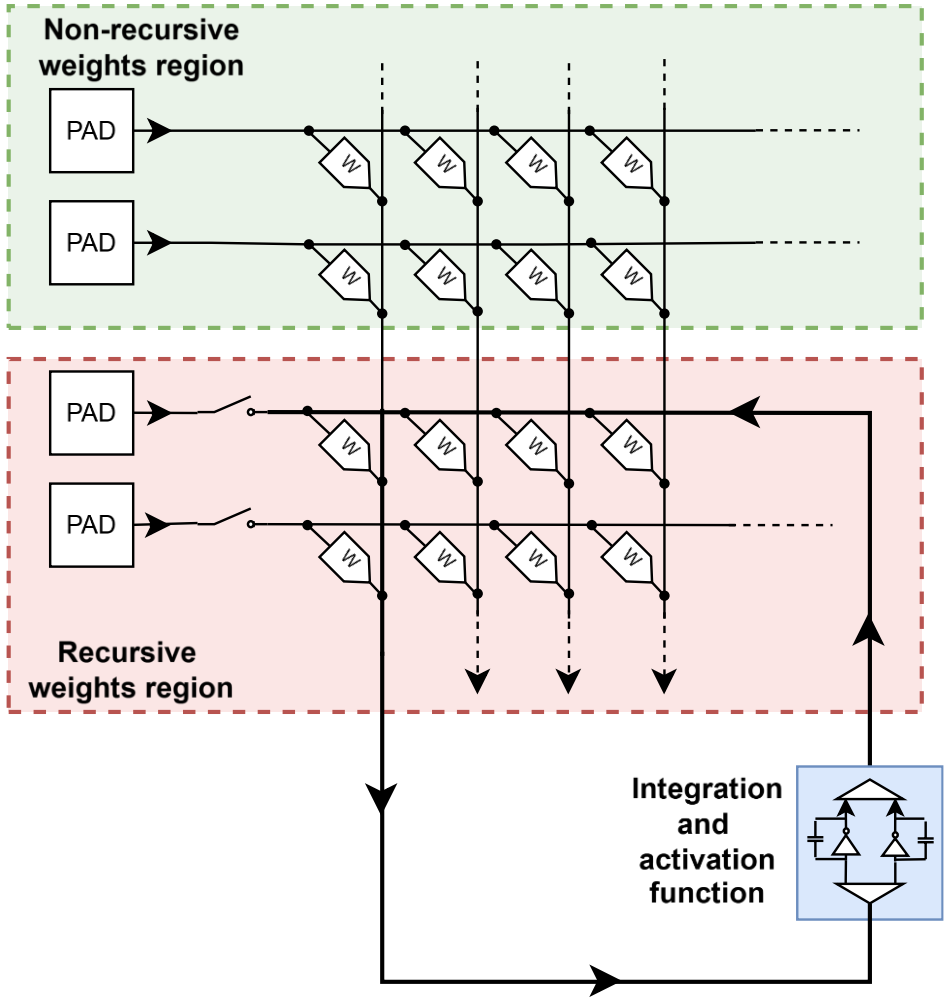


ANNA V2 – new architecture

Two computation regions:

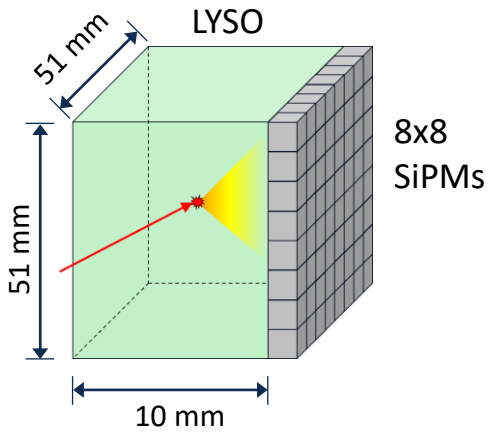
- **Full matrix** (69+bias)×32: **first layer**
- **Recursive region (**(32+bias)×32): **subsequent layers**
- The recursive region is **isolated** from inputs during operation
- Weights are **iteratively updated** each layer
- All weights are stored in SRAMs in the ASIC





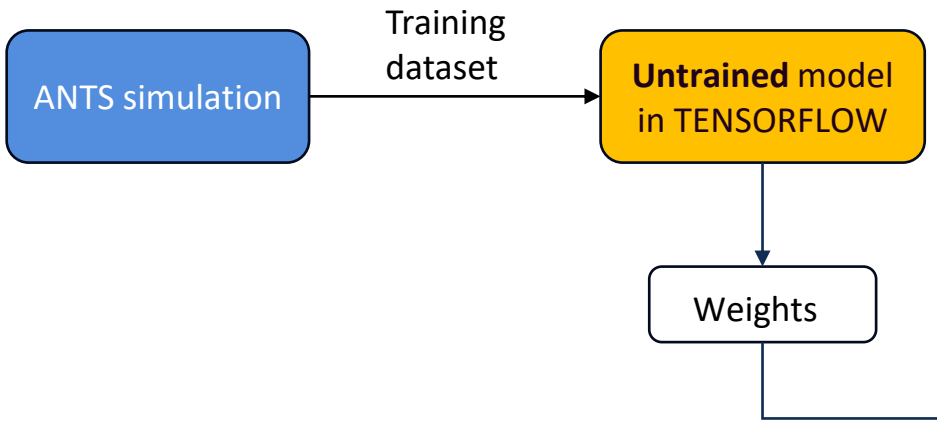
Detector simulation

- Detector simulated in ANTS: LYSO scintillator $51 \times 51 \times 10$ mm^3 coupled to an 8×8 SiPM array.
- Generated a training dataset (27000 events) and test dataset (4500 events).
- Labeled datasets: the input data are the SiPM signals and the associated labels are the coordinates of the interaction point in the crystal.

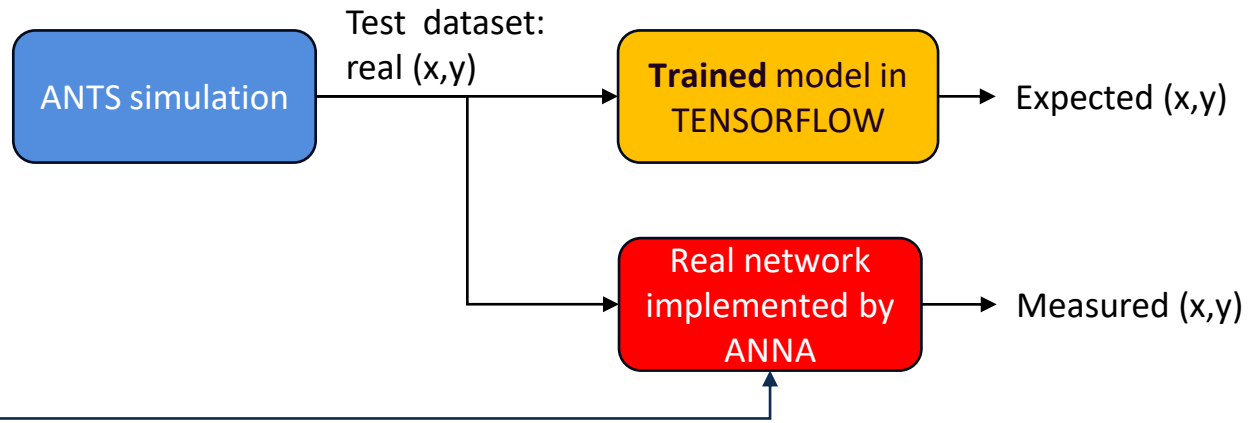


- Chosen neural network topology: $64 \times 20 \times 20 \times 2$.
- Model of the network trained and tested in Tensorflow.

TRAINING PHASE



TEST PHASE



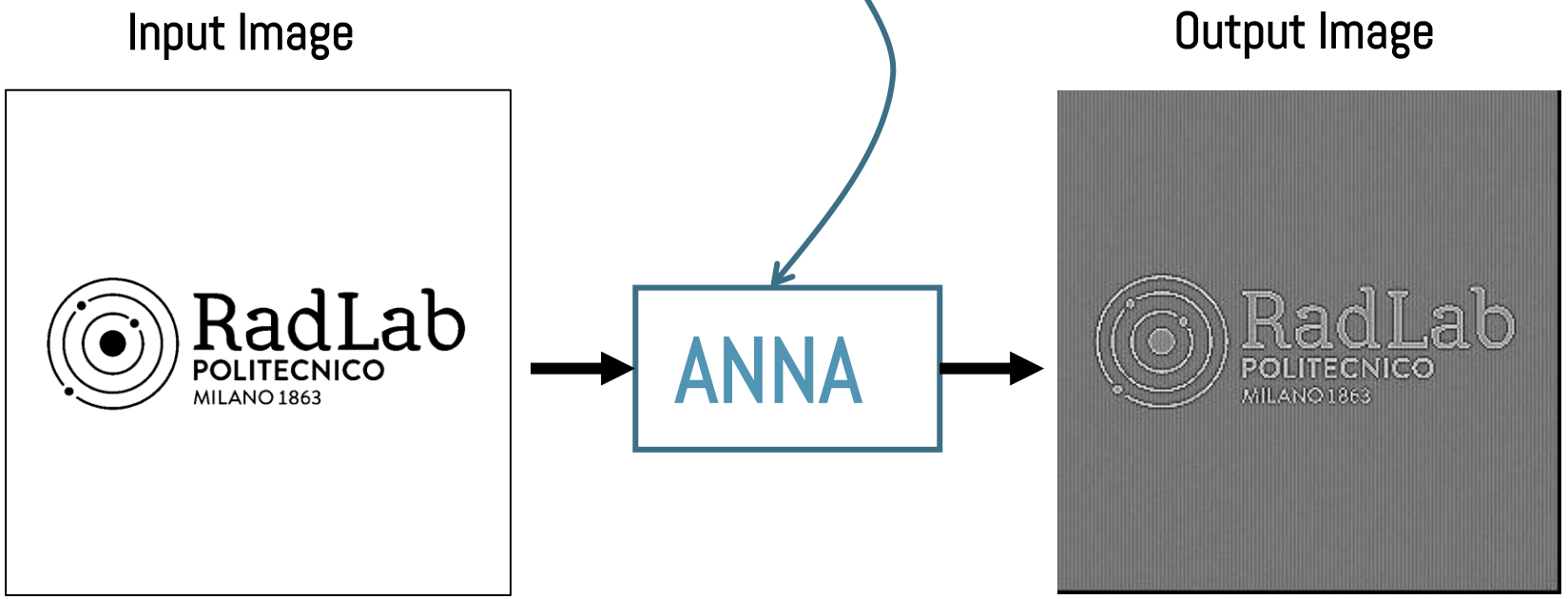
Results: Edge Detection

Edge detection:

- Image provided on the analog domain, 3x3 patches
- Analog edge filter on the ASIC
- Analog output sampled

Convolutional Filter

$$W = \frac{1}{32} \cdot \begin{bmatrix} 0 & -7 & 0 \\ -7 & 28 & -7 \\ 0 & -7 & 0 \end{bmatrix}$$



ANNA V2 Summary