An Analog Neural Network ASIC for Image Reconstruction Embedded in Detectors

S. Di Giacomo, M. Ronchi, G. Borghi, M. Carminati, C. Fiorini

marco1.carminati@polimi.it

International Workshop 24 th iWoRiD on Radiation Imaging Detectors

POLITECNICO MILANO 1863

RadLab POLITECNICO MILANO 1863

INFN

Outline

Motivation

- Edge Computing
  - Conventional vs. In-Memory
- Analog In-Memory Computing for Neural Network (NN)
- Neural Network for event reconstruction

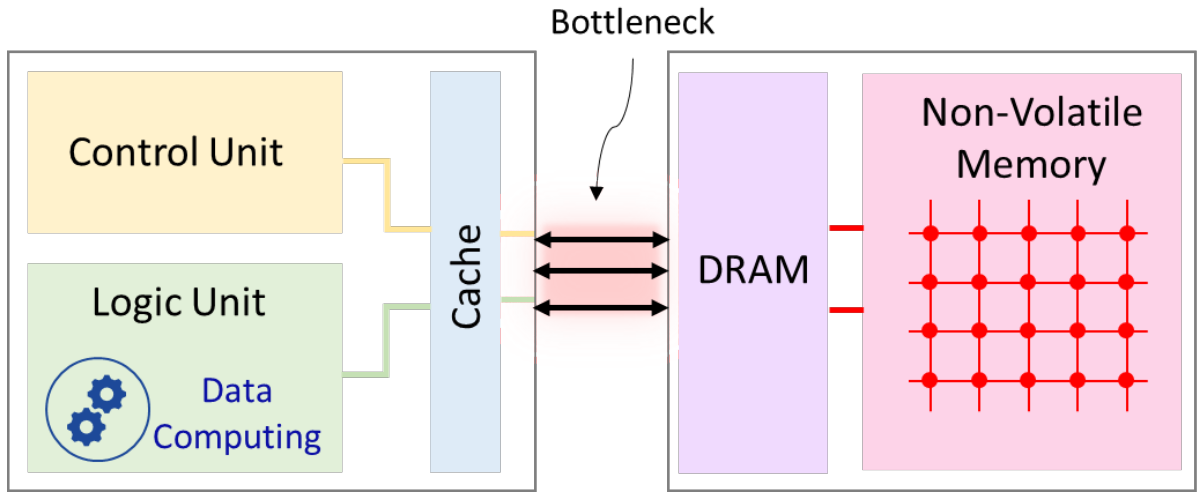
ANNA ASIC

- Neural Network training and quantization
- Analog implementation as capacitive crossbar array
- Circuit challenges and non-idealities
- Energy Efficiency

Edge Computing

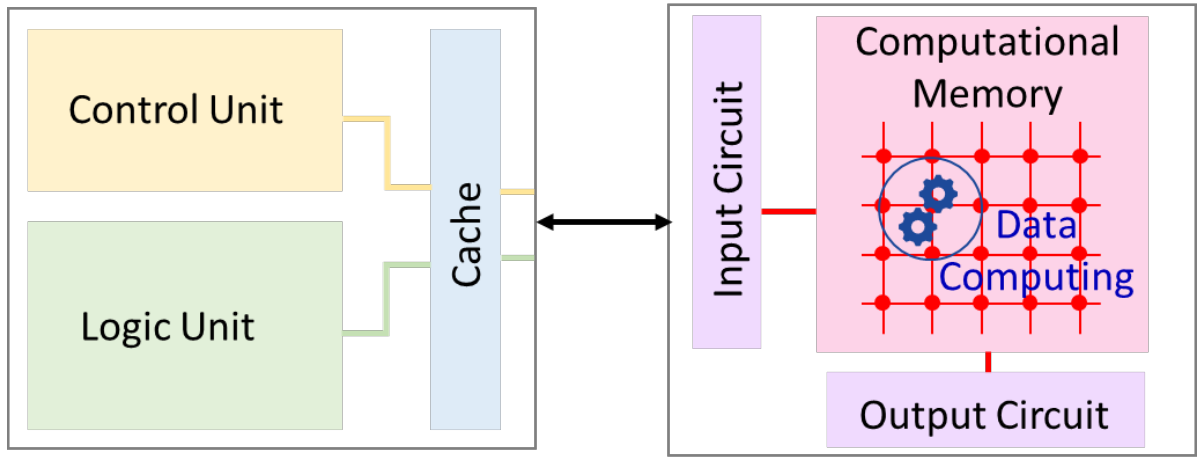
Conventional Von Neumann Computing

- Power hungry data movement
- Long memory access latency
- Limited memory bandwidth



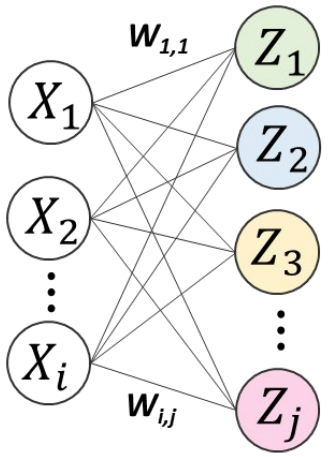
In-Memory Computing (IMC)

- Parallel data processing
- Negligible data movement
- Operations inside memory elements
- Energy efficient
- Programmable



Analog IMC for Neural Network inference

NN basic operation → Multiply-and-Accumulate (**MAC**)



MAC operation

$$\begin{bmatrix} X_1 \\ X_2 \\ X_i \end{bmatrix}^T \begin{bmatrix} w_{1,1} & \cdots & w_{1,j} \\ \vdots & \ddots & \vdots \\ w_{i,1} & \cdots & w_{i,j} \end{bmatrix}$$

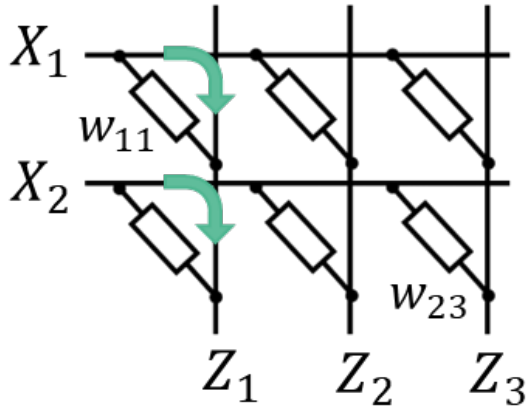
Digital approach:

- Intensive and constant dataflow
- Pipeline multiple full-adders

Analog accelerator exploits **Ohm's law** and **Kirchhoff's laws:**

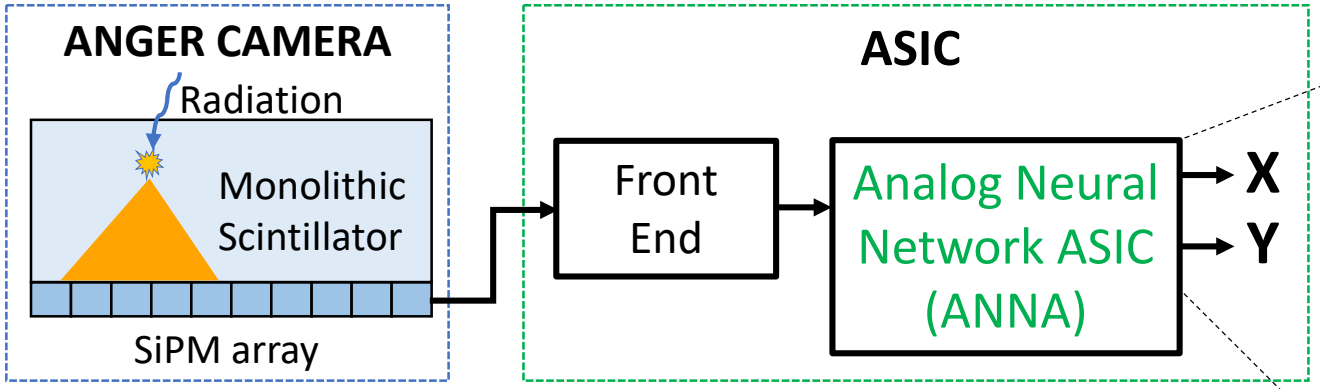
- **Multiplication**: non-volatile memories, used also for **weight** storage
- **Accumulation**: current or charge summation on a wire

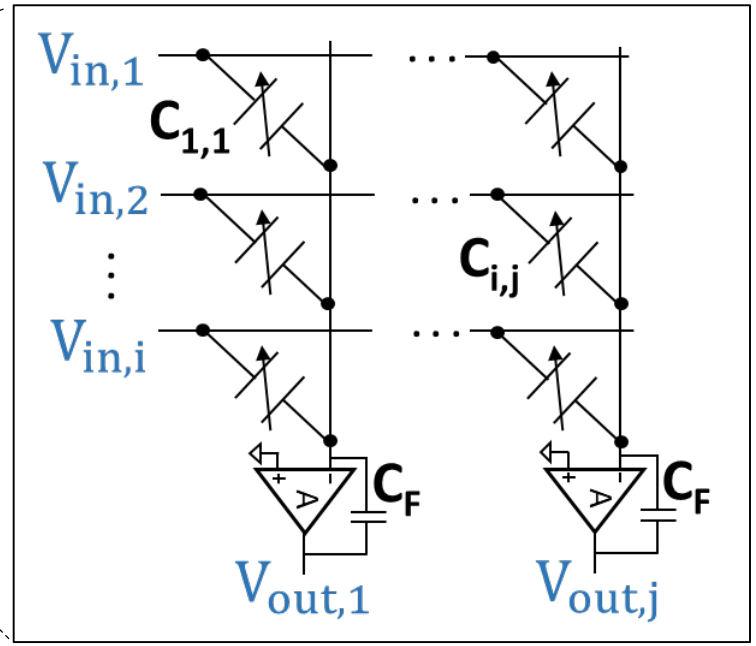
✓ Low power
✓ Throughput and speed improvements (high parallelism)
✓ Monolithic integration with CMOS IC



$$Z_j = \sum w_{i,1} X_i$$

Neural Network for Event Reconstruction





ANNA application

NN for the **localization of the radiation event** from detector signals for nuclear medical imaging (e.g. Anger Camera for PET).

→ Crossbar array of **programmable switched capacitors**

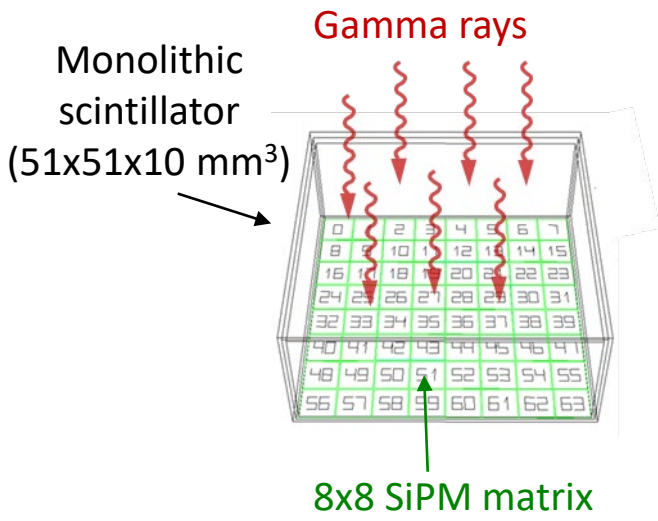
→ **Analog operations** performed directly on **analog signals** coming from photodetectors

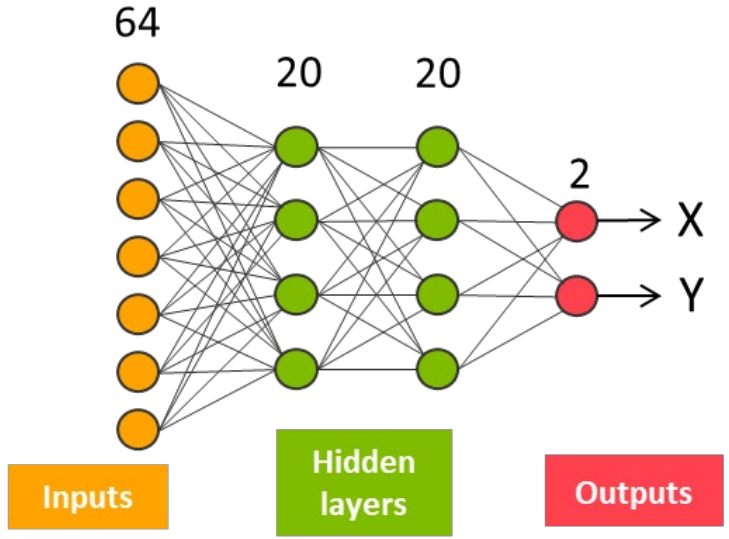
→ No need for signal ADC and FPGA for embedded processing

→ **Interaction coordinates** directly at the output of the **ASIC**

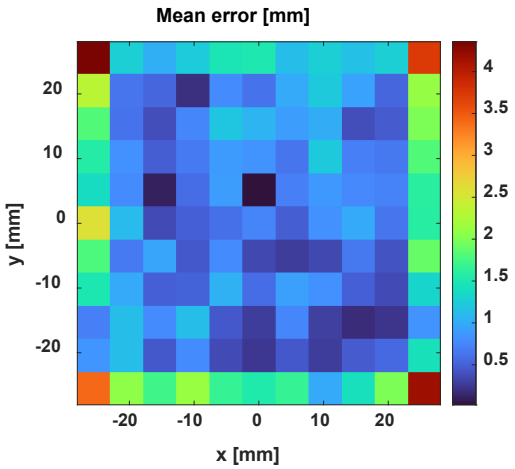
NN software training and performance

- Simulated dataset for training
- NN with **64** inputs, 2 hidden layers of **20** neurons each, **2** outputs
- Training in MATLAB with **weights quantization** (**5-bit** resolution)



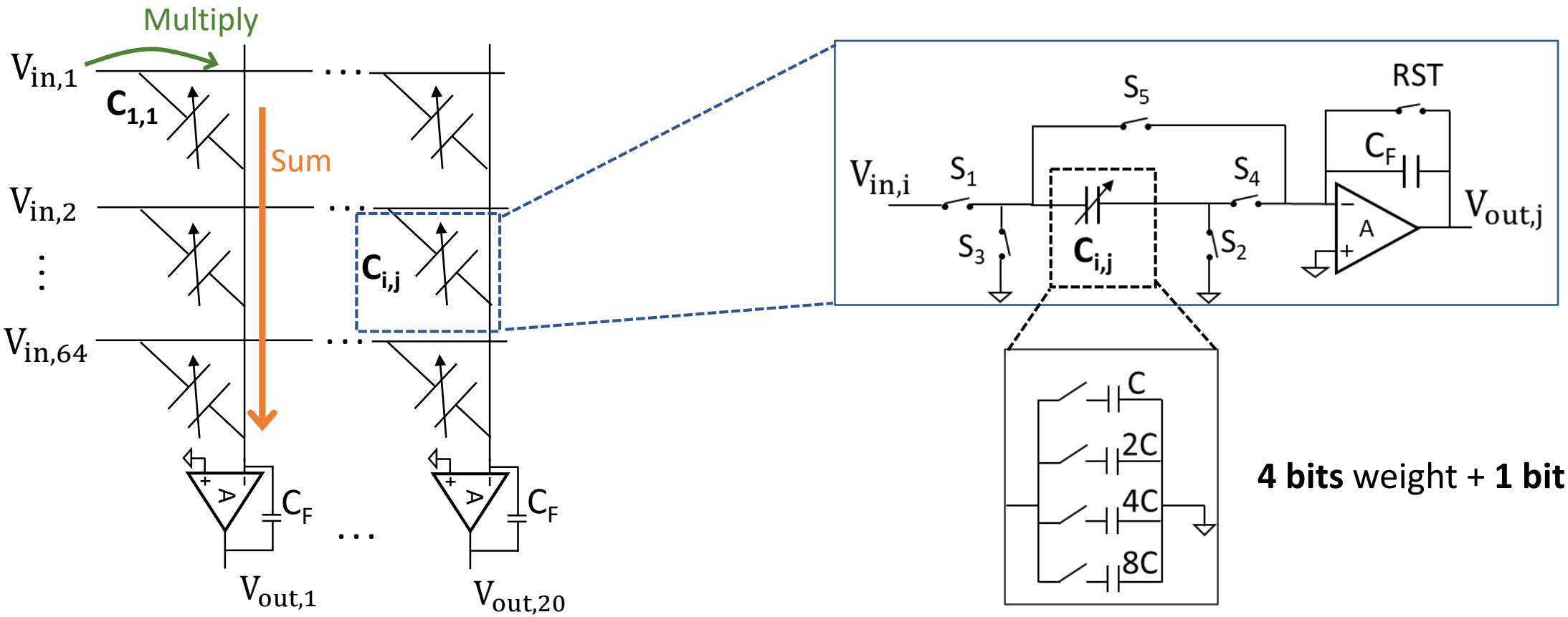


Resolution [mm]	*x*	*y*	Total
FWHM (2D PSF)	1.58	1.81	
$r_{50\%}$	0.79	0.79	1.84
$r_{90\%}$	2.44	2.44	4.83
MAE	1.16	1.13	1.80



- **FWHM**: of the 2D PSF of the reconstruction error
- $\mathbf{r_{50\%}}$ and $\mathbf{r_{90\%}}$: 50% and 90% percentile of the normalized error distribution
- **MAE**: mean absolute error

Analog Neural Network implementation

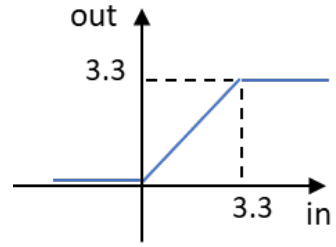


4 bits weight + **1 bit** sign

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

weight

Activation function: clipped ReLU implemented within the op-amp feedback line



Analog Neuron operation

Charge-redistribution approach

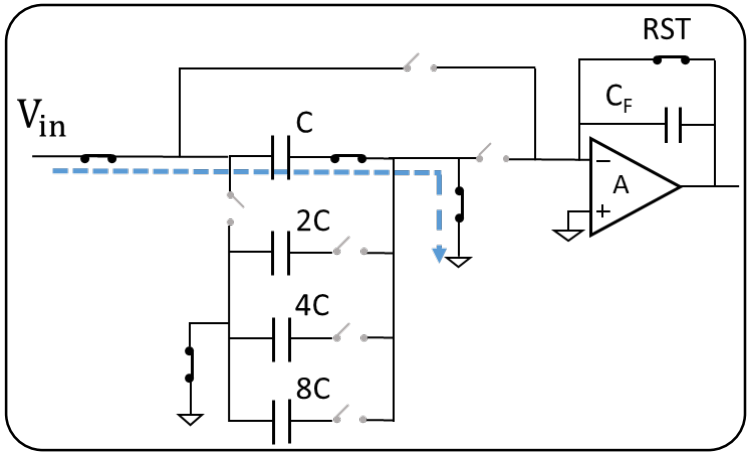
1) Charge **only C** with V_{in} to **minimize energy consumption** $(E = CV^2)$

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

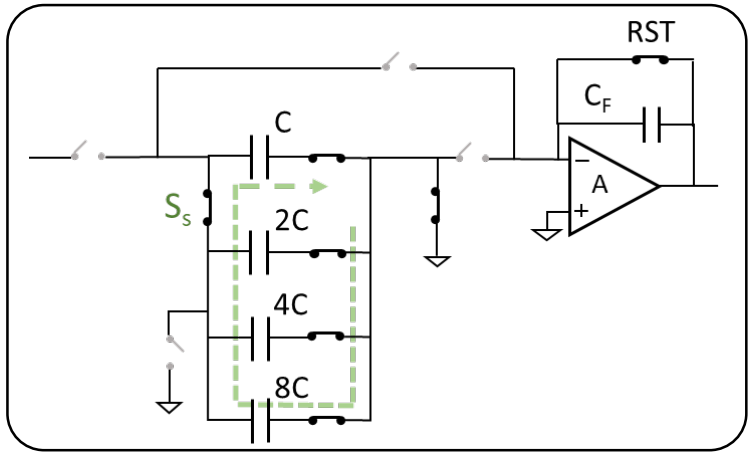
3) **Only the capacitors** corresponding to the **desired weight** are connected to the integrator, while the others are disconnected by means of their respective switch.

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

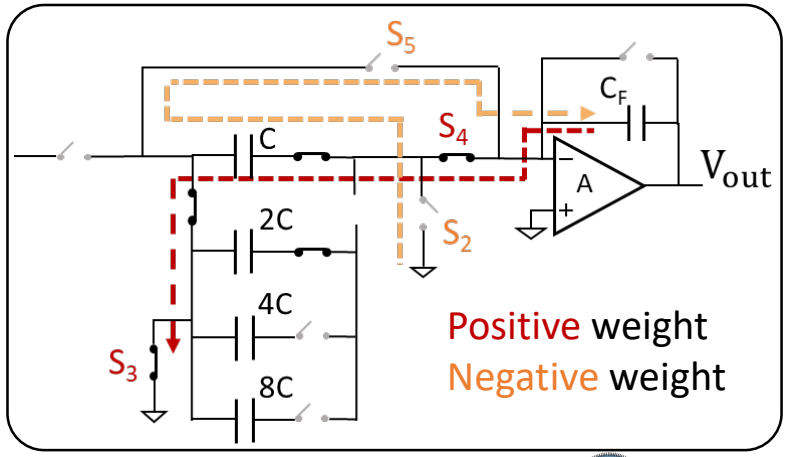
Step 1



Step 2

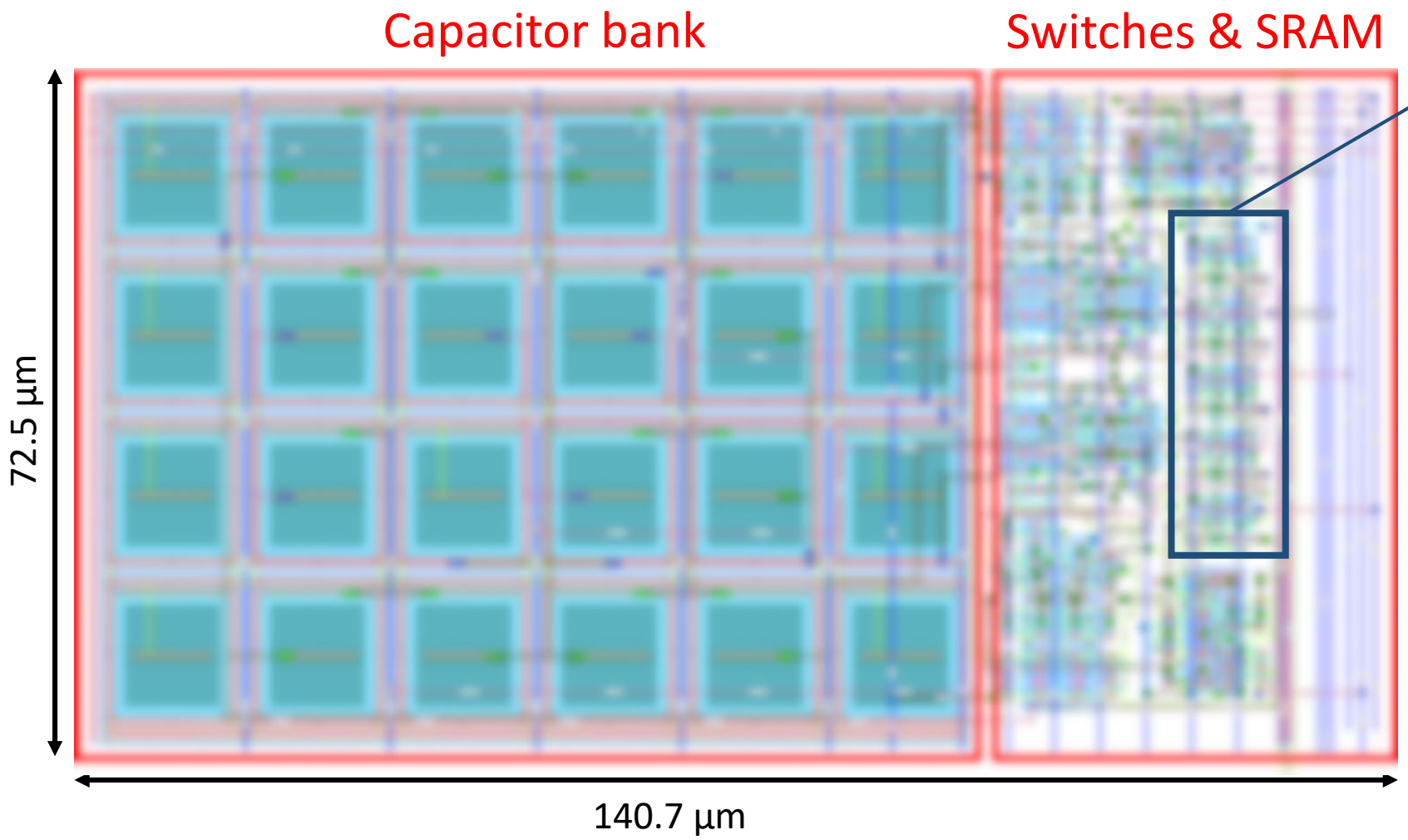


Step 3

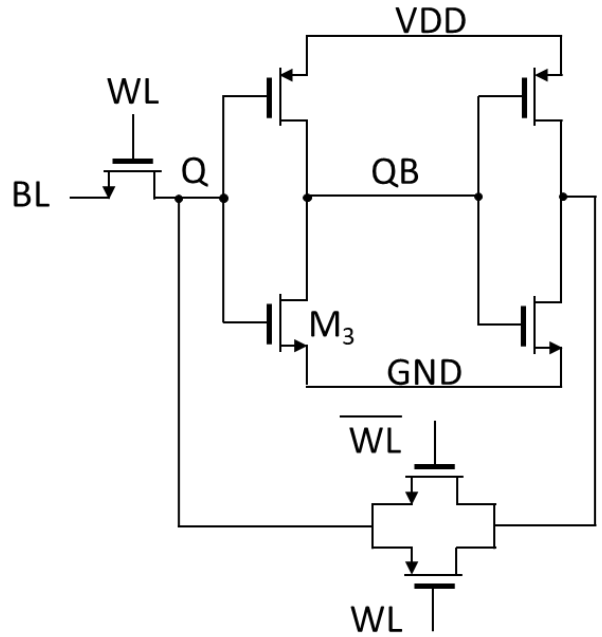


Analog Neuron Layout

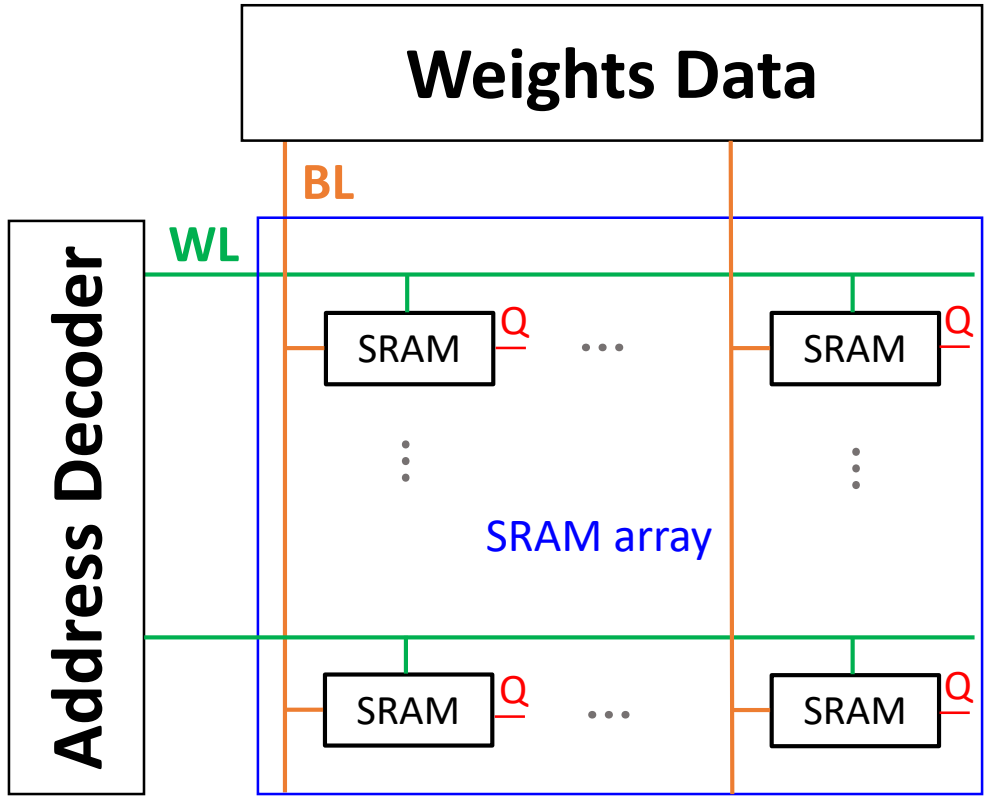
Proof of concept: implementation in CMOS 0.35 µm with C_{LSB} = **100 fF**



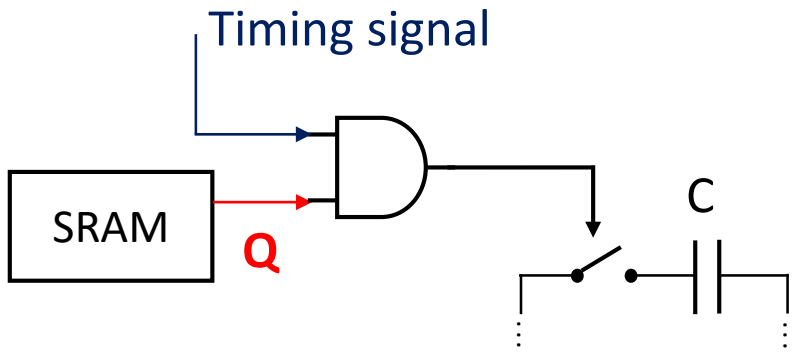
SRAMs to store weight



Weight Programming



- Weight programming by means of **SPI**
- Stored data Q, plus additional logic, to close/open capacitor bank **switches**
- Timing signals provided by programmable **Ring Counters**



Circuit challenges

Analog Switches

- Charge injection and clock feedthrough
- Parasitic capacitances

Integrator

- Large dynamic range (0 – 3.3 V), low power
- Very low offset error
- Stability for different input capacitance (NN weight)
- Electronic noise

Analog Switches (1/2)

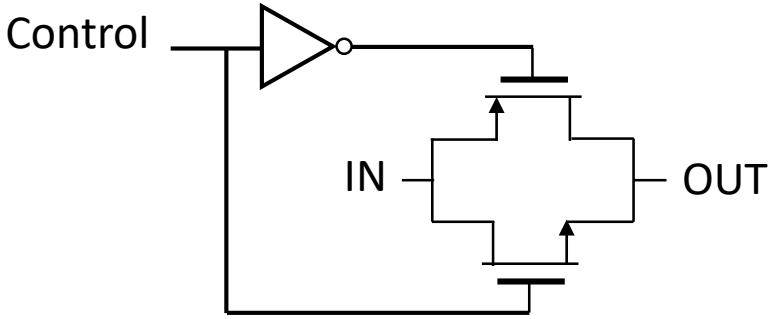
- Analog switches implemented as **transmission gates** (TG)

- Two non-idealities:

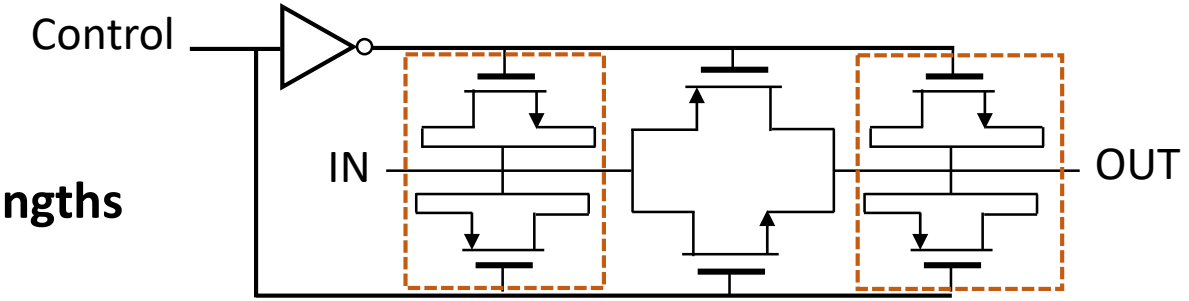
  1) Charge injection: added or subtracted charge from drain/source in an asymmetric way, depending on impedance

$$Q_{ch} = WLC_{ox}(V_{DD} - V_{in} - V_{th})$$

  2) Clock feedthrough: added charge due to gate-source and gate-drain capacitances



➢ Solution
- **Dummy switches** to absorb charge from TG
- Cadence **optimization** tool to set **widths and lengths** that minimize injected charge

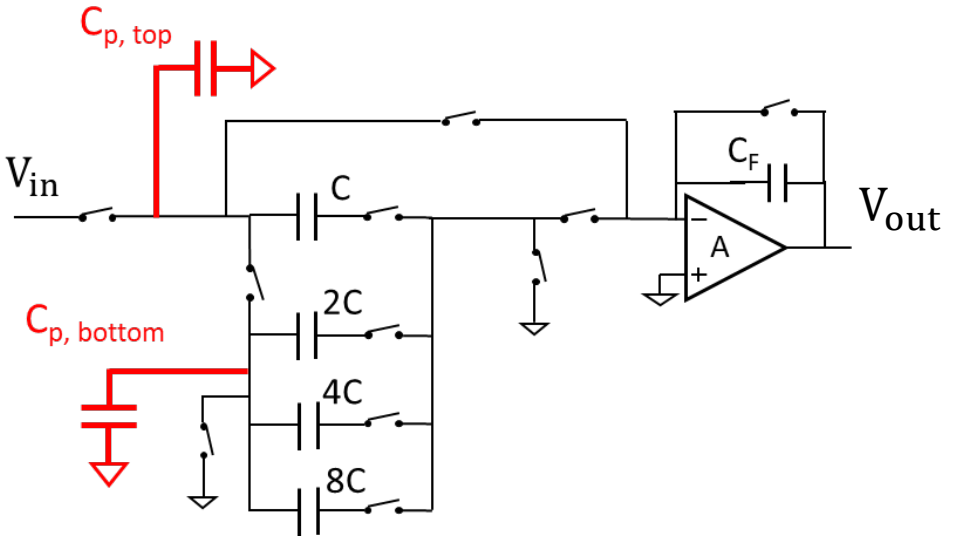


Dummy switches

Analog Switches (2/2)

- Additional parasitic capacitors added by switches
- Errors during input sampling and charge redistribution phases
- Estimated from **post-layout** simulations

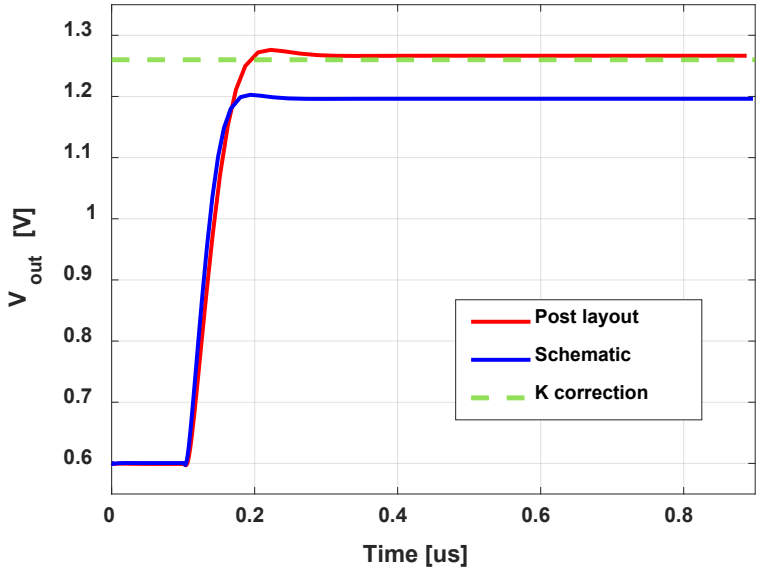
➢ A **corrective factor *K*** can be calculated and added to neural network Matlab model, to take it into account during training



$$V_{out,ideal} = \frac{V_{in}}{15}\frac{C_{ij}}{C_F}$$

$$V_{out,real} = \frac{V_{in}}{15}\frac{C_{ij}}{C_F}\underbrace{\left(\frac{C_{LSB} + C_{p,top}}{C_{LSB} + C_{p,top}/15 + C_{p,bottom}/15}\right)}_{K}$$

$$V_{out,real} = K\, V_{out,ideal}$$



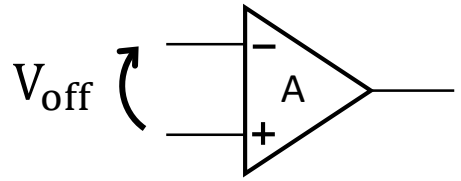
Integrator (1/3)

- Low power, rail-to-rail class A amplifier

Offset error

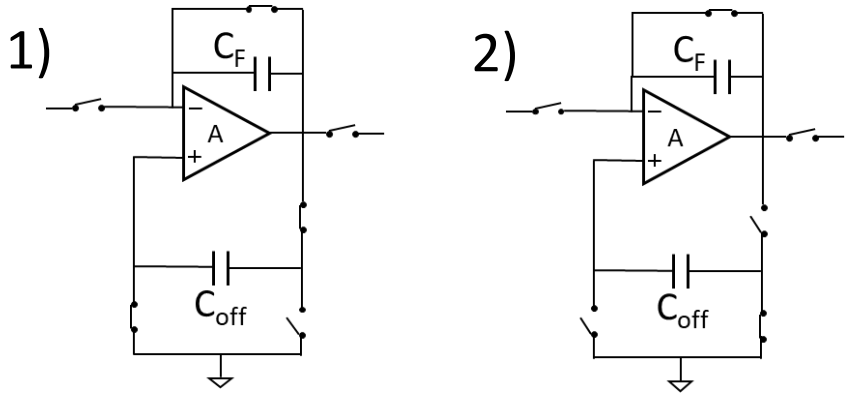
- Offset at integrator differential input subject to a huge amplification

$$V_{out} = \frac{V_{in}}{15}\frac{C_{ij}}{C_F} + V_{off}\left(1 + \frac{|C_{ij}|}{C_F}\right) \qquad \text{with } V_{off} = 96\ \mu V \pm 5 mV$$

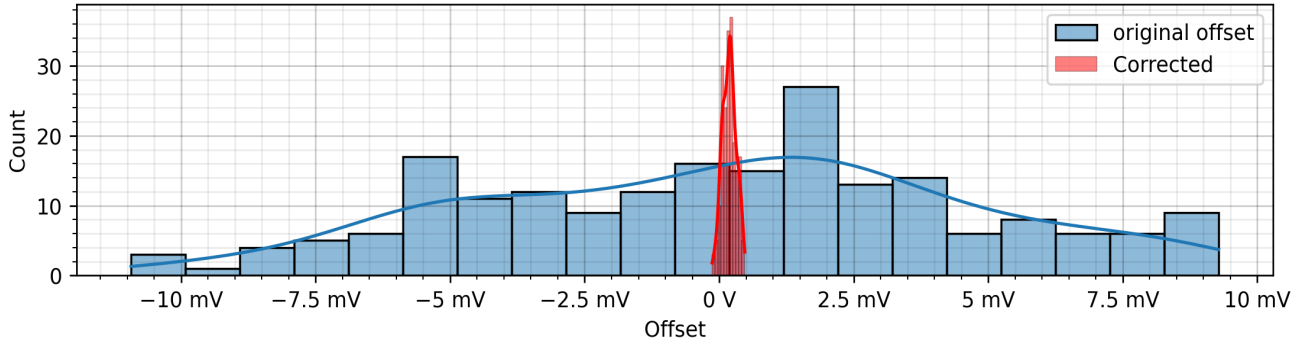


- **Offset compensation** phase to minimize its effect

1) The offset is sampled on a 1.5 pF capacitor
2) The capacitor is flipped and the offset subtracted from V+



MC post-layout offset



$$V_{off} = 187\ \mu V \pm 125 \mu V$$

POLITECNICO MILANO 1863

Integrator (2/3)

Variable input capacitance

- Variable input capacitance C_{ij} (from 0 to 97.5 pF) → large variability of feedback factor

$$\beta = \frac{C_F}{C_{ij} + C_F} \qquad \frac{1}{\beta} \rightarrow \text{from 3.5 dB to 54 dB}$$

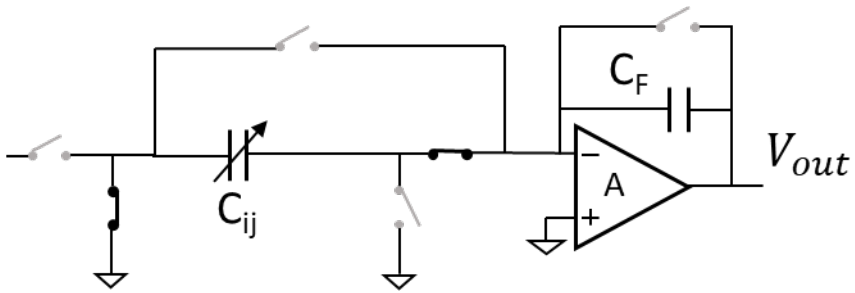
➢ **Programmable Miller compensation** to always ensure fast and stable response

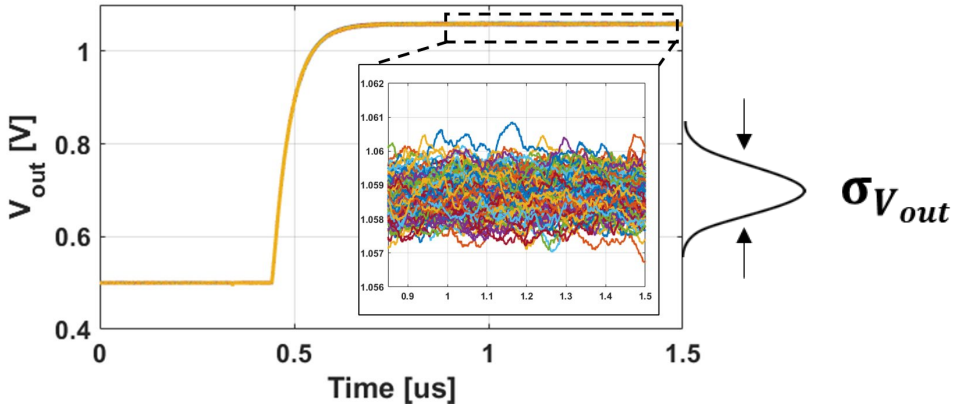
- Can be chosen among four capacitances or a combination of them (25 fF, 50 fF, 100 fF, 400 fF)
- Capacitance settings stored in SRAM cells
- Target phase margin of $67° \pm 7°$

Integrator (3/3)

Electronic noise

Transient noise simulations on a **single-input neuron** to estimate noise contributions to integrator output:





$\sigma_{V_{out,kTC}} \approx 440\ \mu V$ → **kTC** noise contribution

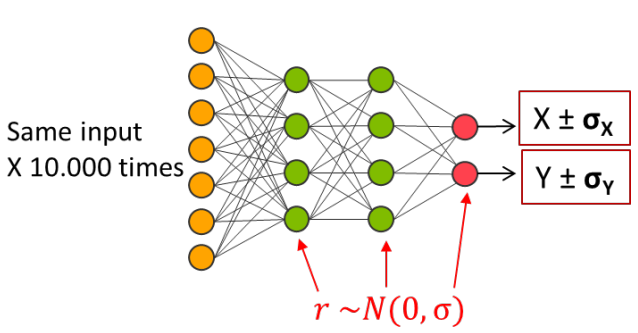
$\sigma_{V_{out,OPA}} \approx 400\ \mu V$ → **op-amp** only noise contribution

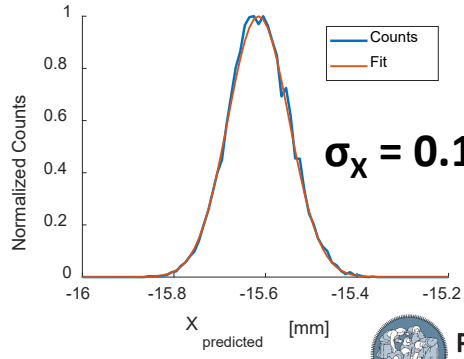
$$\boldsymbol{\sigma_{V_{out,1}}} = \sqrt{\sigma^2_{V_{out,kTC}} + \sigma^2_{V_{out,OPA}}} \approx 595\ \mu V \text{ (1 input)}$$

$$\boldsymbol{\sigma_{V_{out,64}}} = \sqrt{64} * 595\mu V \approx 4.7\ mV \text{ (max 64 inputs)}$$

- **Matlab** simulation: **given set of input signals** and **noise σ = 5 mV** at the output of each neuron, to consider contribution of **all inputs**

  Std. of the **predicted coordinates** to evaluate effect of **noise** on the **NN performances**





σ_X = 0.11 mm

Energy Efficiency

- Energy consumption during NN **inference** at f_{clock} = 10 MHz, estimated from post-layout simulations
- Input buffers and integrators are powered on only when needed to save energy

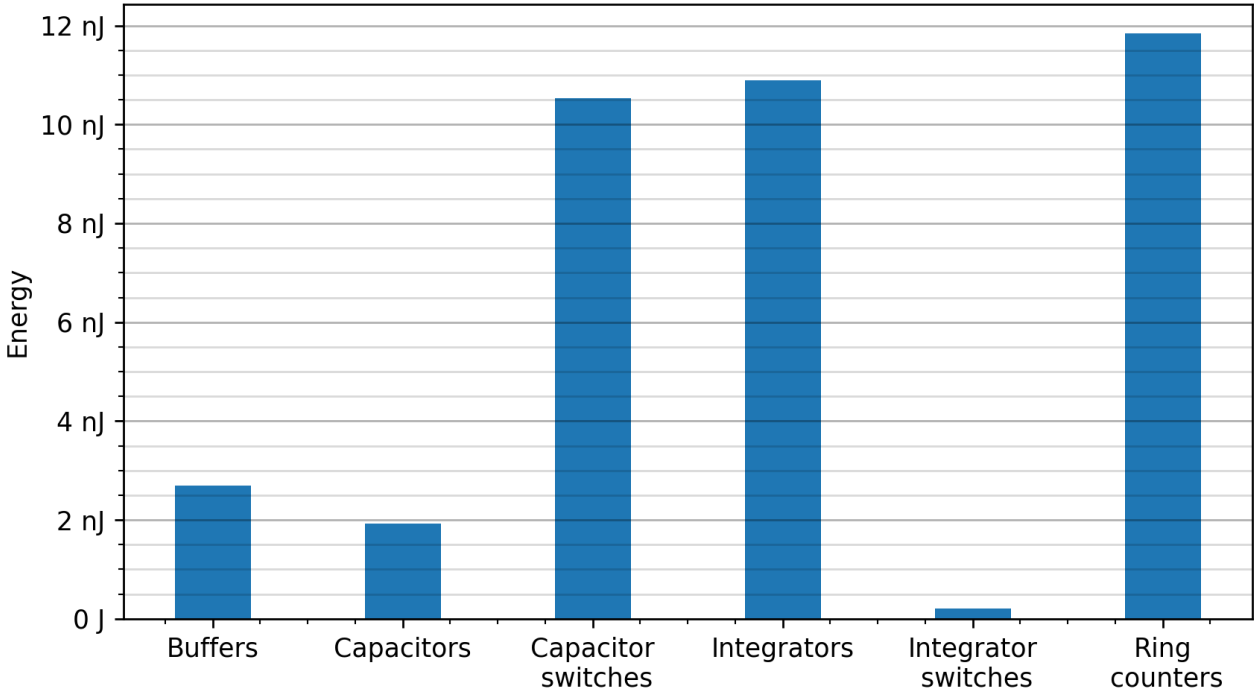
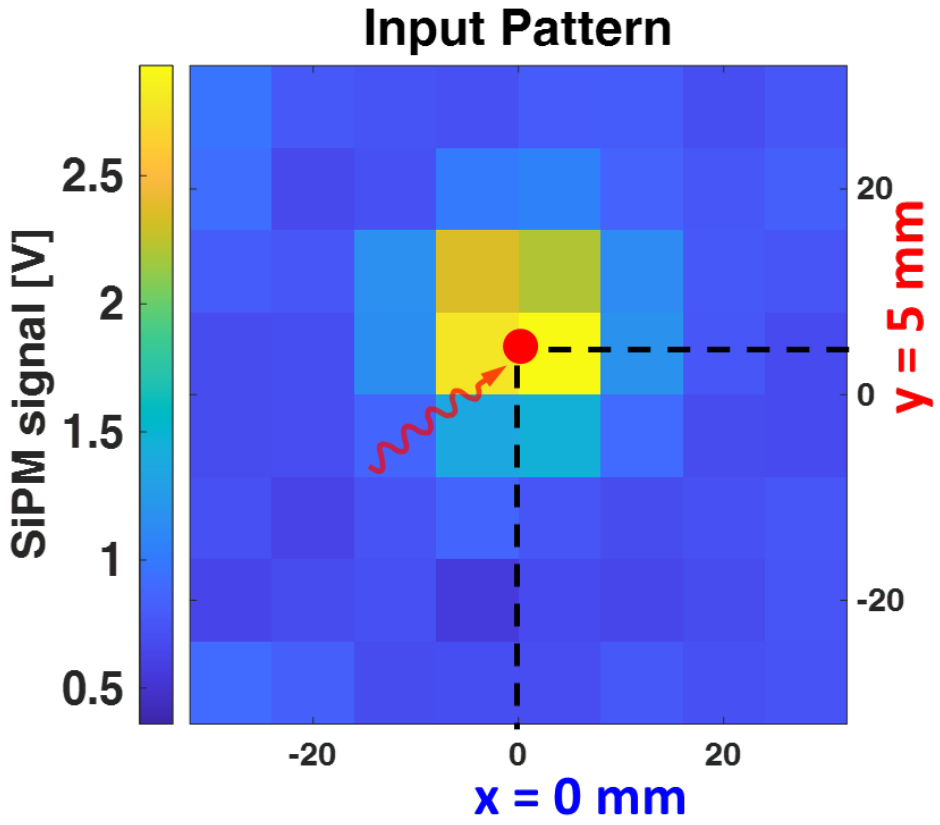
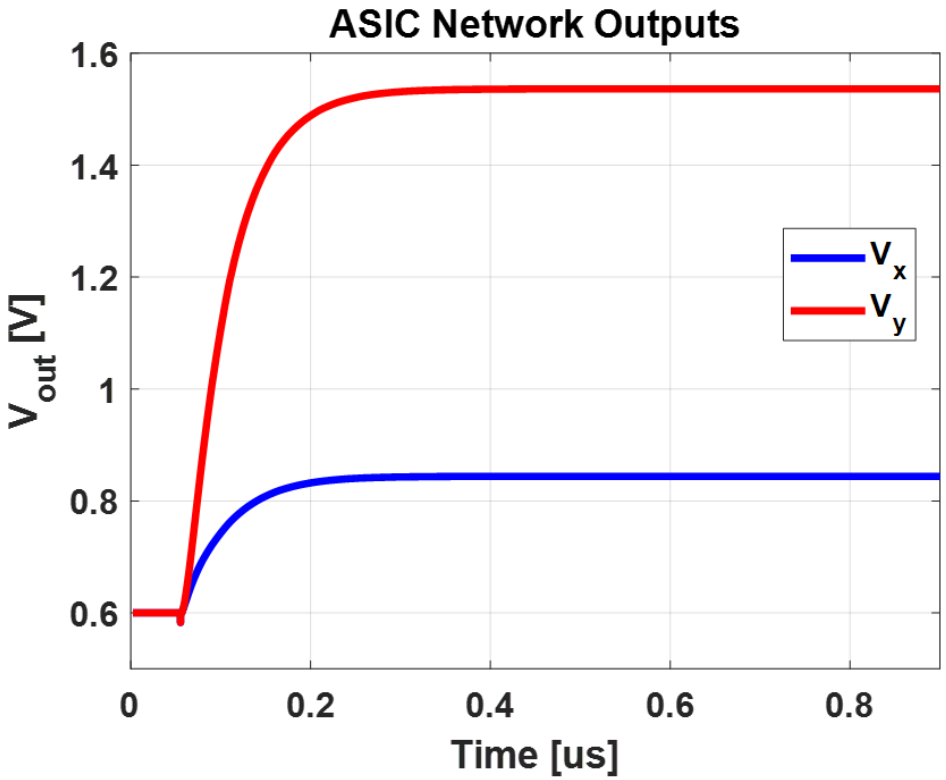


Fig. of merit	Estimate
I/O latency	4.6 µs
I/O total operations	3566
Total consumption	38.12 nJ
Efficiency	775.21 MOP/s
Analog energy efficiency	135.74 GOP/J
Total energy efficiency	**93.55 GOP/J**

Full Neural Network ASIC Simulation

- Preliminary full neural network ASIC **schematic simulation** in CADENCE
- Input event (64 SiPM voltage signals) with position of interaction (0mm, 5mm)
- The **two output voltages** represents the predicted **x, y** coordinates





Conclusions

- ✓ **Fully analog** neural network able to perform 5-bit MAC operations in the **charge domain (ASIC)**
- ✓ Can be monolithically **integrated** in the front-end ASIC
- ✓ Inference in a more **efficient** way in terms of computational cost and energy, compared to a fully-digital implementation
- ✓ ASIC prototype with $\boldsymbol{C_{min} = 100fF}$ in CMOS 0.35 µm node
- ✓ Energy consumption estimated for inference ≈ **38.12 nJ**
- ✓ Energy Efficiency estimated for inference ≈ **93.55 GOP/J**

- Will be submitted soon for fabrication.
- Can be applied to several detector challenges where NN are adopted (charge sharing correction etc…)

POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA

RadLab
POLITECNICO
MILANO 1863

INFN
MILANO
Istituto Nazionale di Fisica Nucleare
Sezione di Milano

Thank you!