



ANNA: Analog Neural Network ASIC for radiation detectors

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

XII Front-end Electronics Workshop
Torino, June 16th 2023



POLITECNICO
MILANO 1863



RadLab
POLITECNICO
MILANO 1863



Motivation

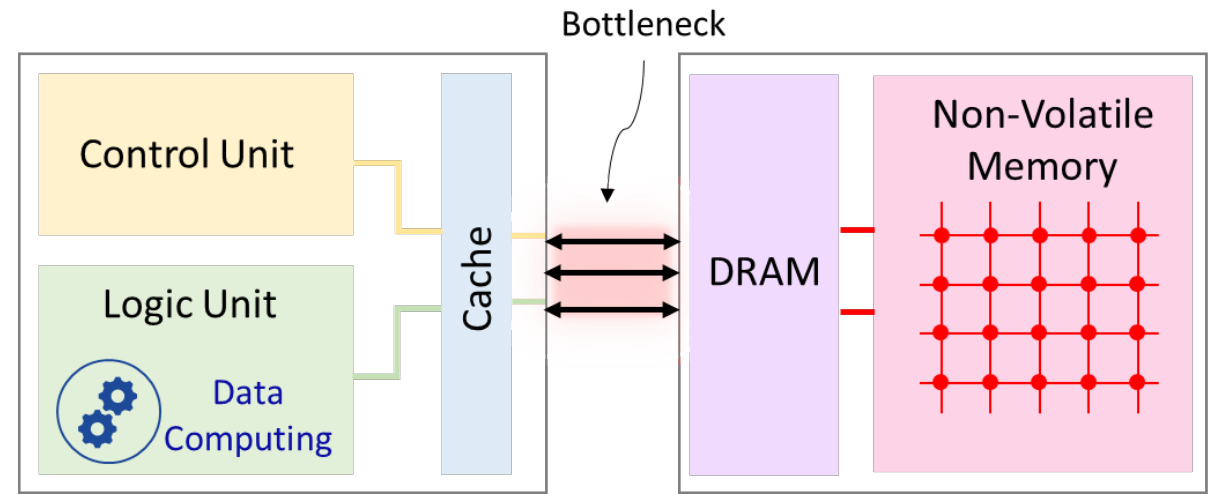
- Edge Computing
 - Conventional vs. In-Memory
- Analog In-Memory Computing for Neural Network (NN)
- Neural Network for event reconstruction

ANNA ASIC

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

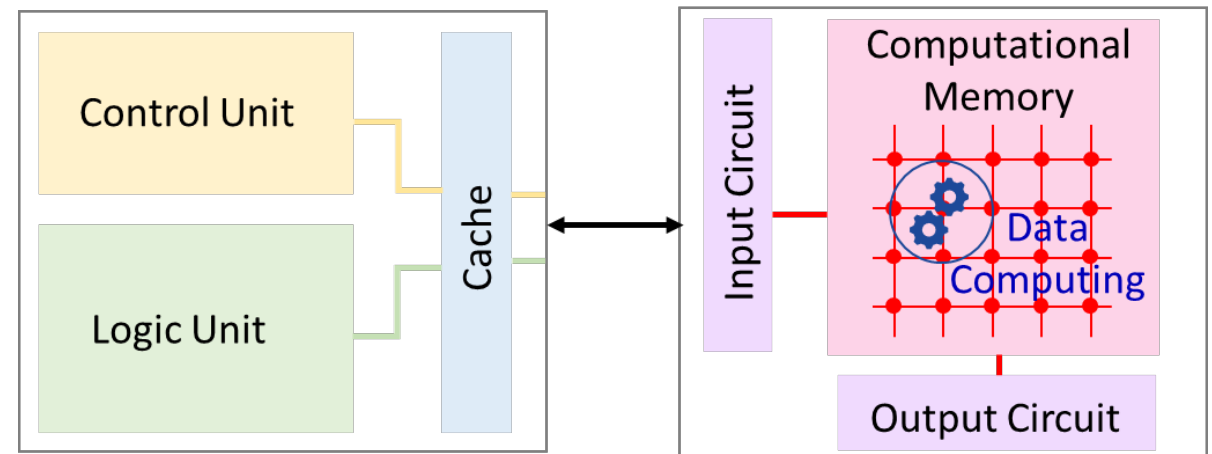
Conventional Von Neumann Computing

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



In-Memory Computing (IMC)

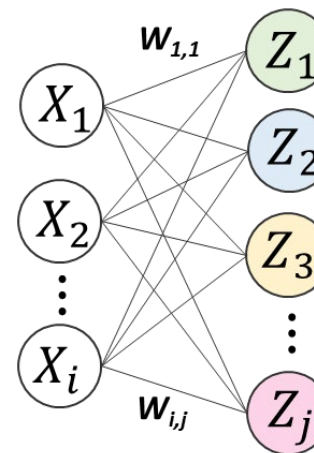
- Parallel data processing
- Negligible data movement
- Operations inside memory elements
- Energy efficient
- Programmable



NN basic operation \rightarrow Multiply-and-Accumulate (**MAC**)

Digital approach:

- Intensive and constant dataflow
- Pipeline multiple full-adders

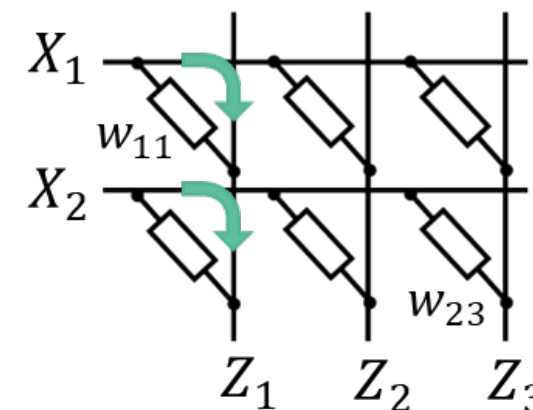


MAC operation

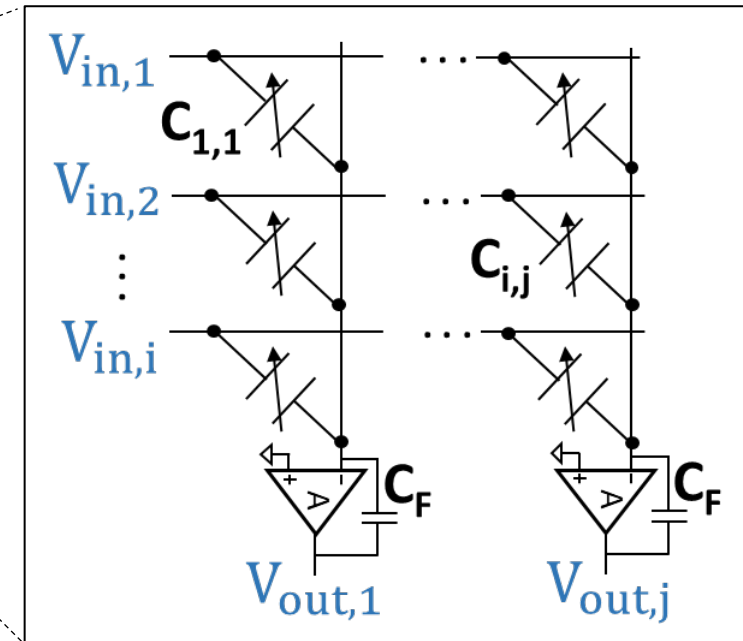
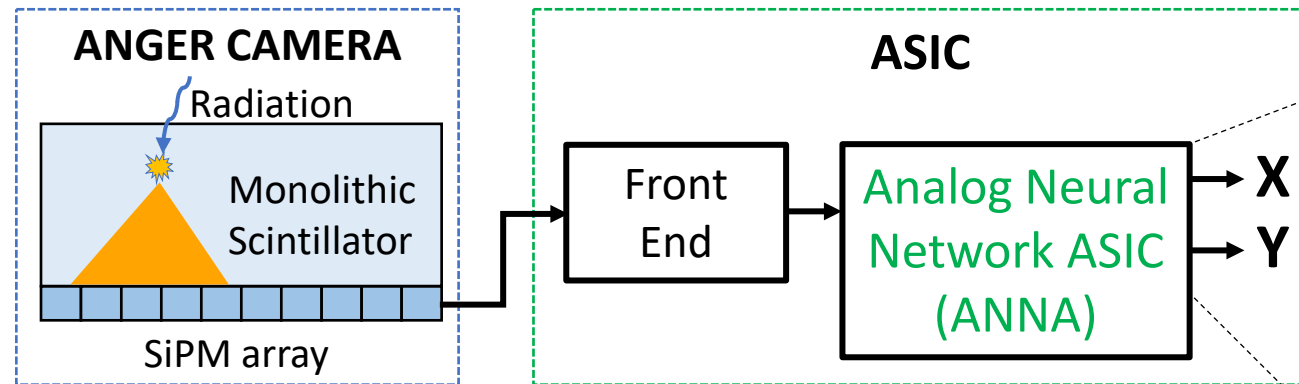
$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_i \end{bmatrix}^T \begin{bmatrix} w_{1,1} & \cdots & w_{1,j} \\ \vdots & \ddots & \vdots \\ w_{i,1} & \cdots & w_{i,j} \end{bmatrix}$$

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



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

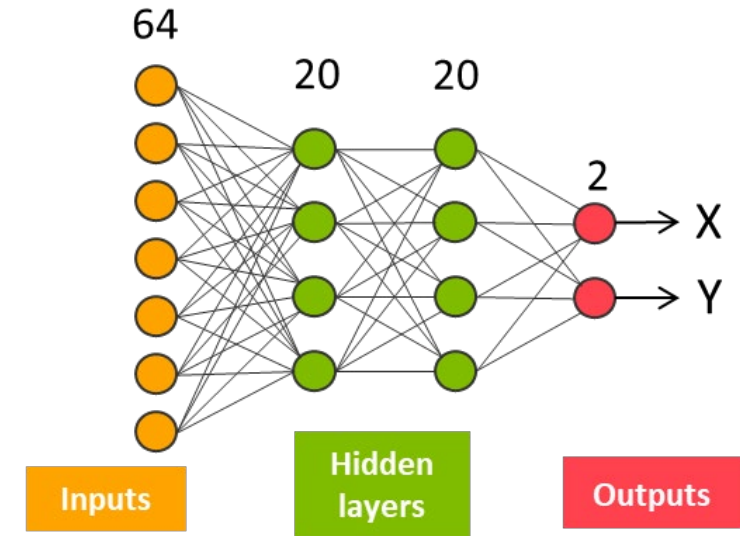
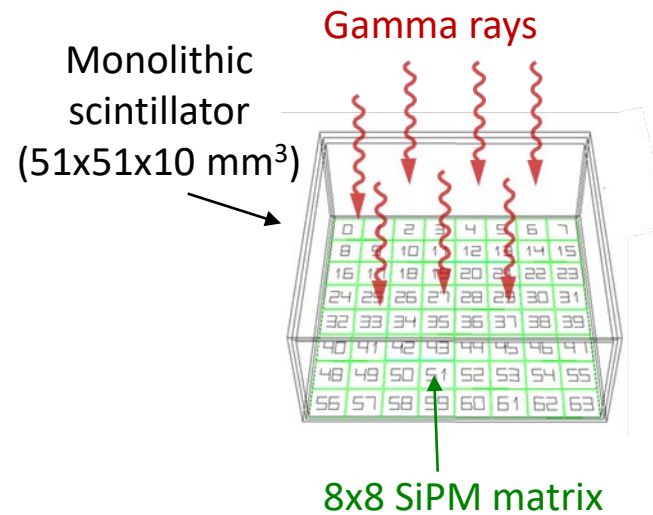


ANNA application

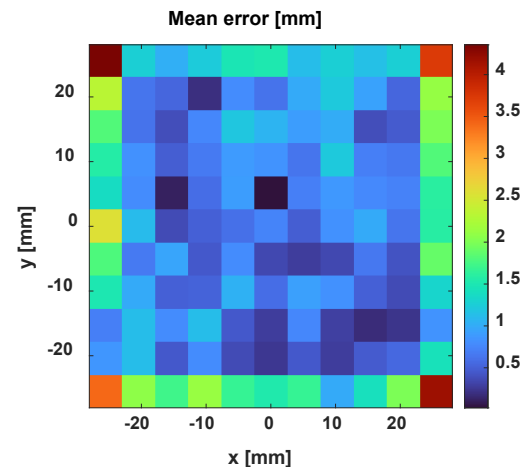
NN for the **localization of the radiation event** from detector signals for nuclear medical imaging (e.g. Anger Camera for PET).

- 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**

- 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)**

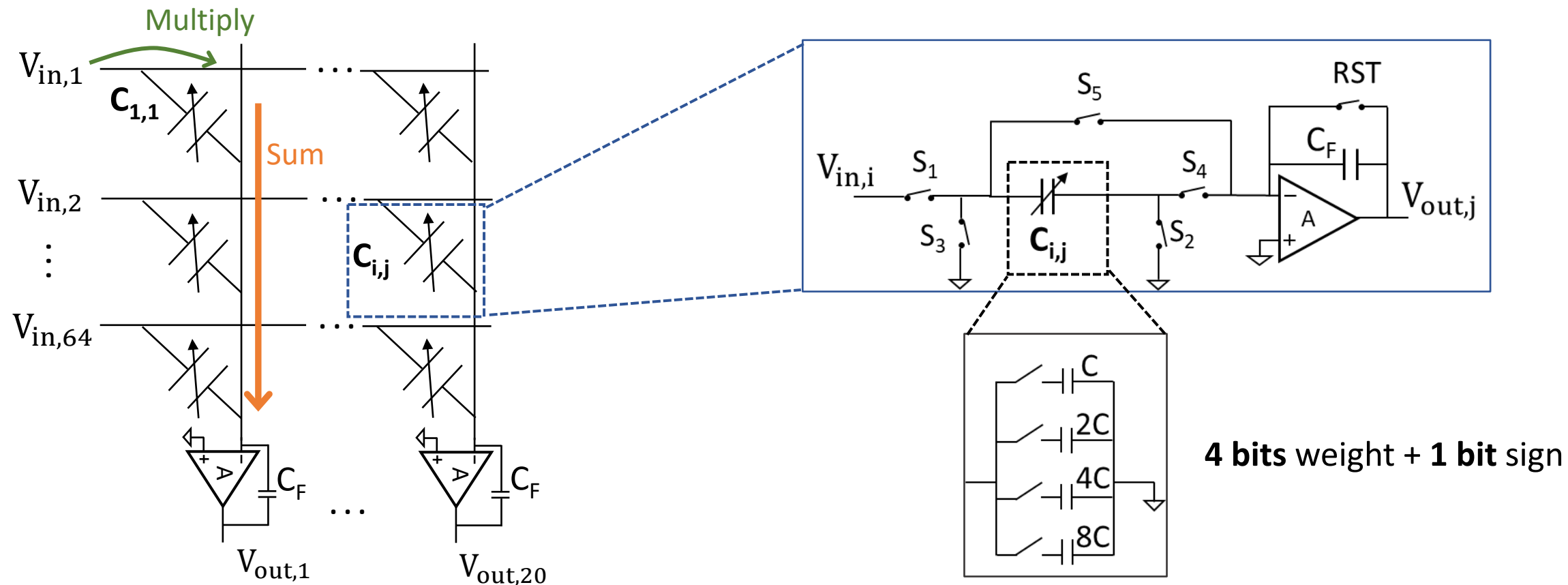


Resolution [mm]	x	y	Total
FWHM (2D PSF)	1.58	1.81	
$r_{50\%}$	0.79	0.79	1.84
$r_{90\%}$	2.44	2.44	4.83
MAE	1.16	1.13	1.80



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

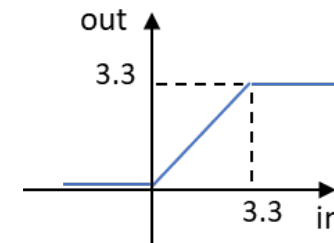
Analog Neural Network implementation



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

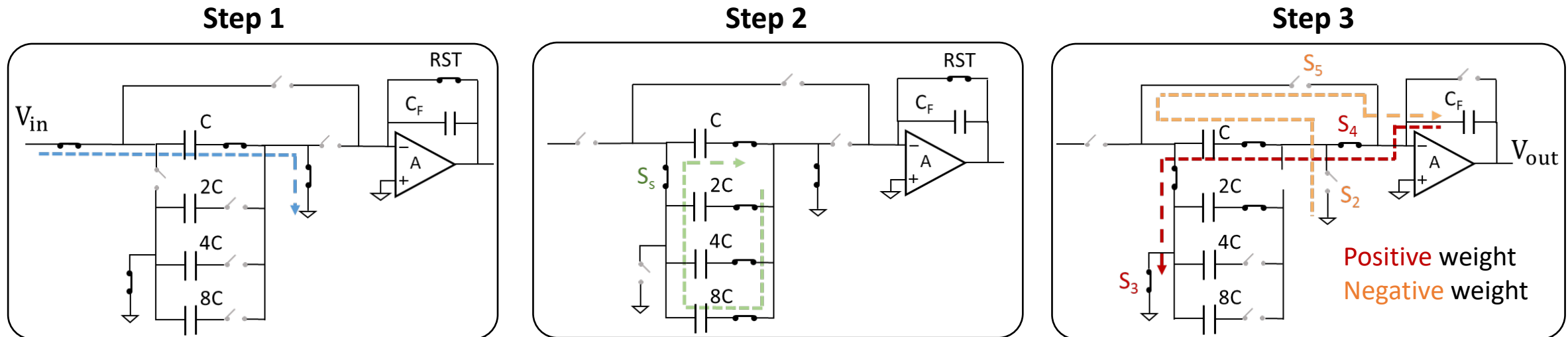


Analog Neuron operation

Charge-redistribution approach

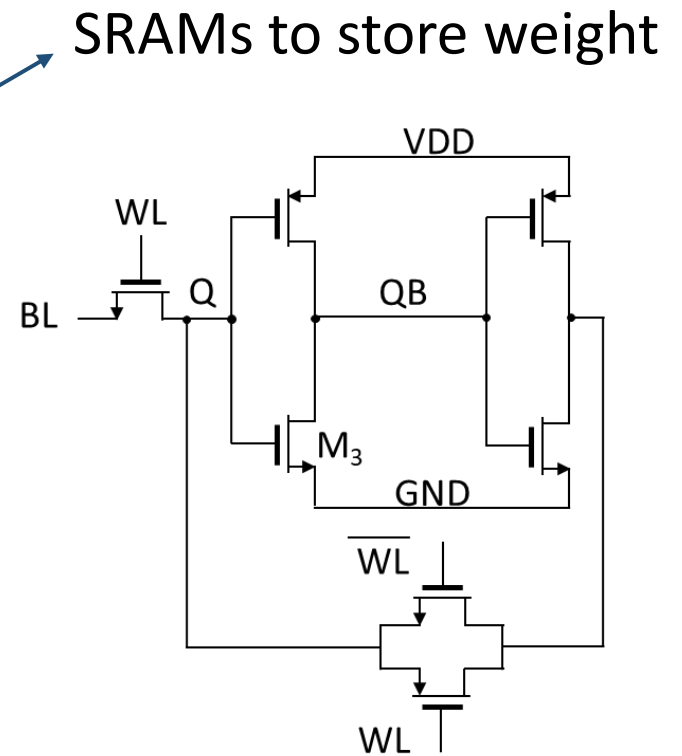
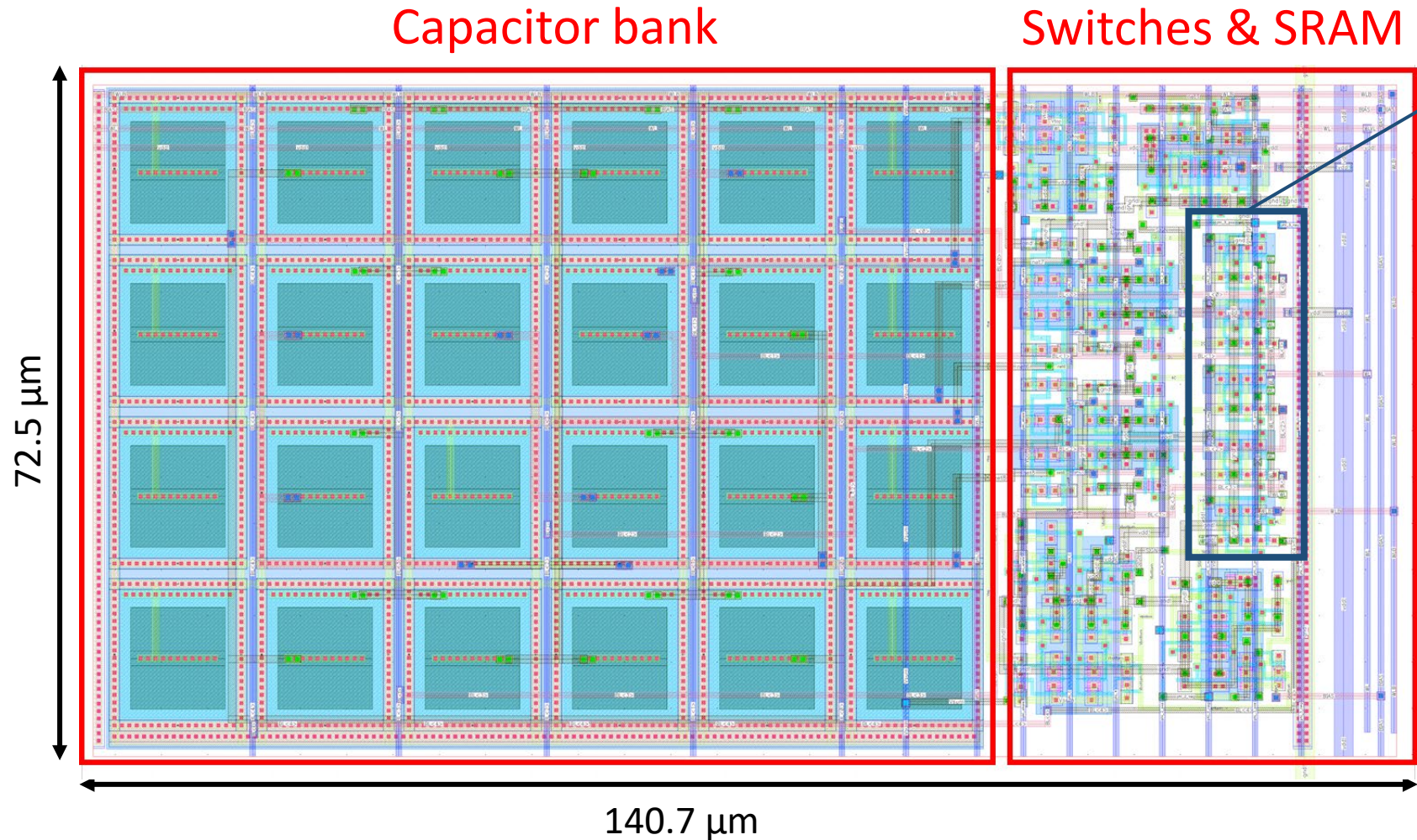
- 1) Charge **only C** with V_{in} to **minimize energy consumption** ($E = CV^2$)
- 2) Charge is **redistributed** among all capacitors closing S_s
- 3) **Only the capacitors** corresponding to the **desired weight** are connected to the integrator, while the others are disconnected by means of their respective switch.

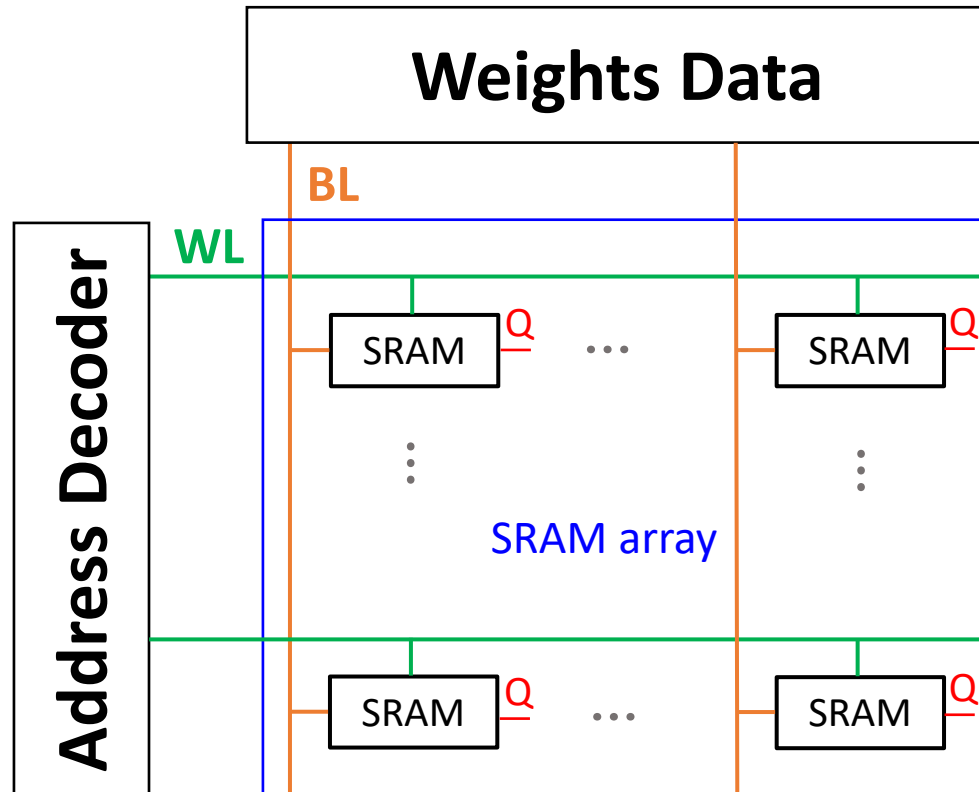
The output is
$$V_{out,j} = \frac{V_{in,i} C_{ij}}{15 C_F}$$



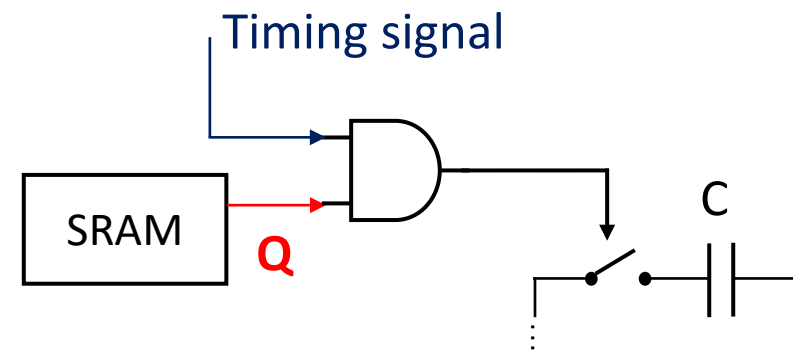
Analog Neuron Layout

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





- 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**



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 implemented as **transmission gates** (TG)
- Two non-idealities:

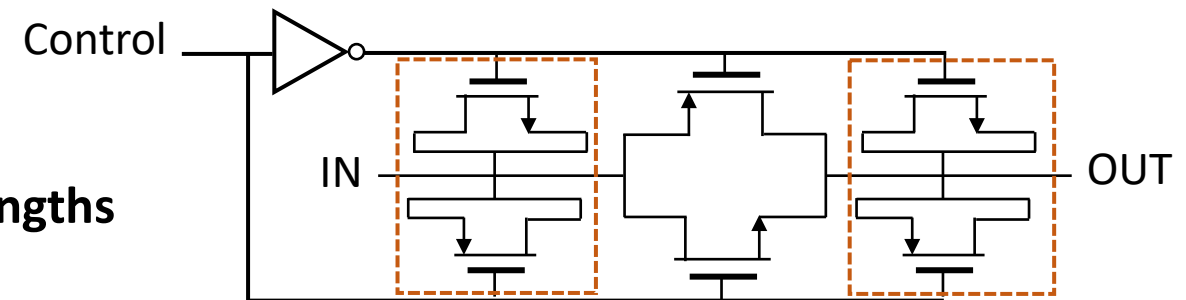
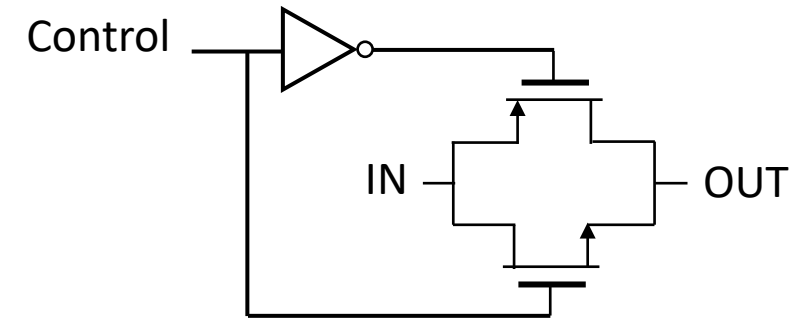
1) **Charge injection**: added or subtracted charge from drain/source in an asymmetric way, depending on impedance

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

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

➤ Solution

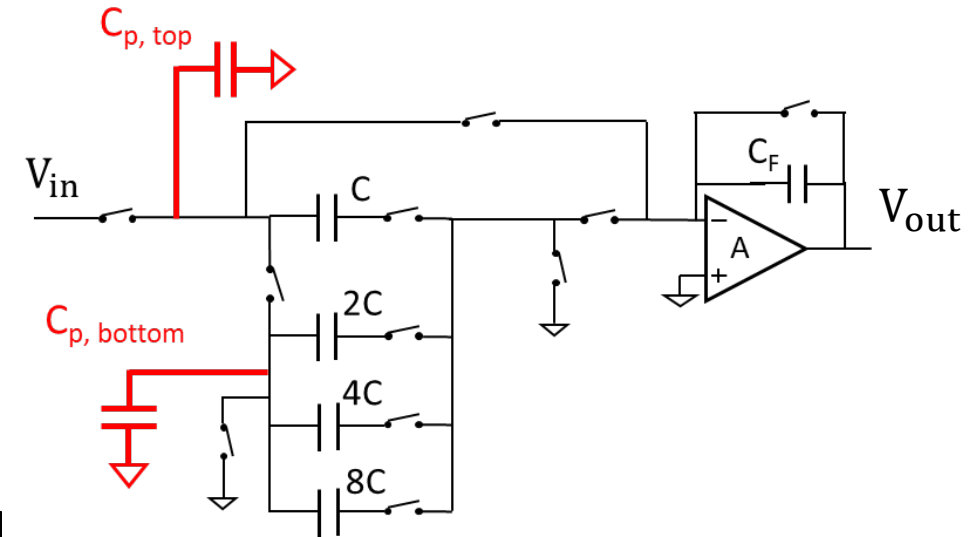
- **Dummy switches** to absorb charge from TG
- Cadence **optimization** tool to set **widths and lengths** that minimize injected charge



Dummy switches

Analog Switches (2/2)

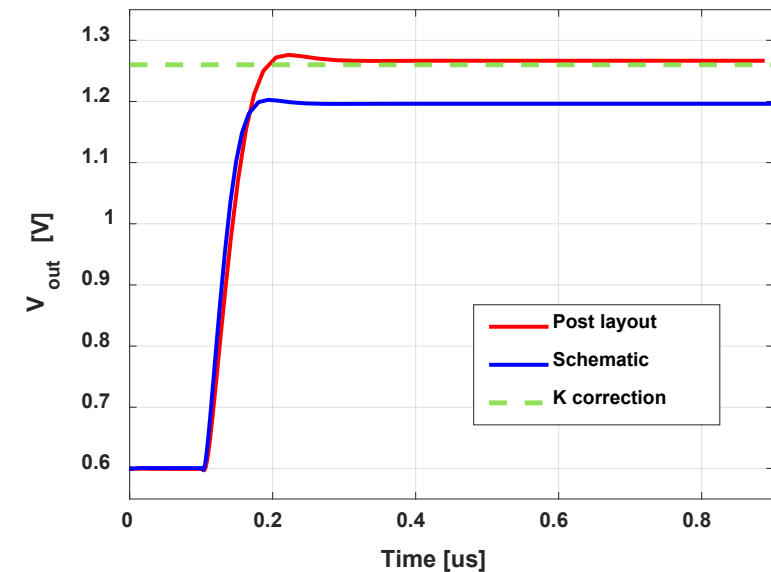
- Additional **parasitic capacitors** added by switches
- Errors during input sampling and charge redistribution phases
- Estimated from **post-layout** simulations
- A **corrective factor K** can be calculated and added to neural network Matlab model, to take it into account during training



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

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

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

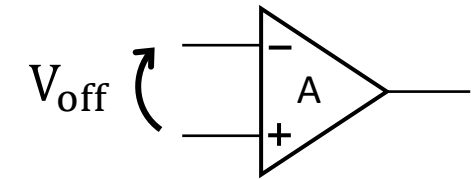


- Low power, rail-to-rail class A amplifier

Offset error

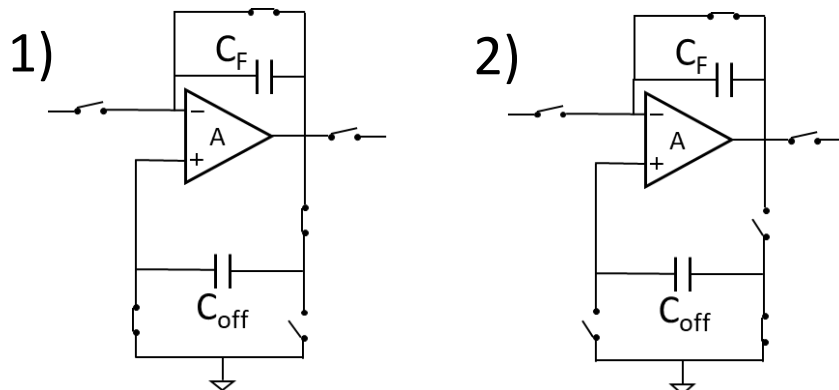
- Offset at integrator differential input subject to a huge amplification

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

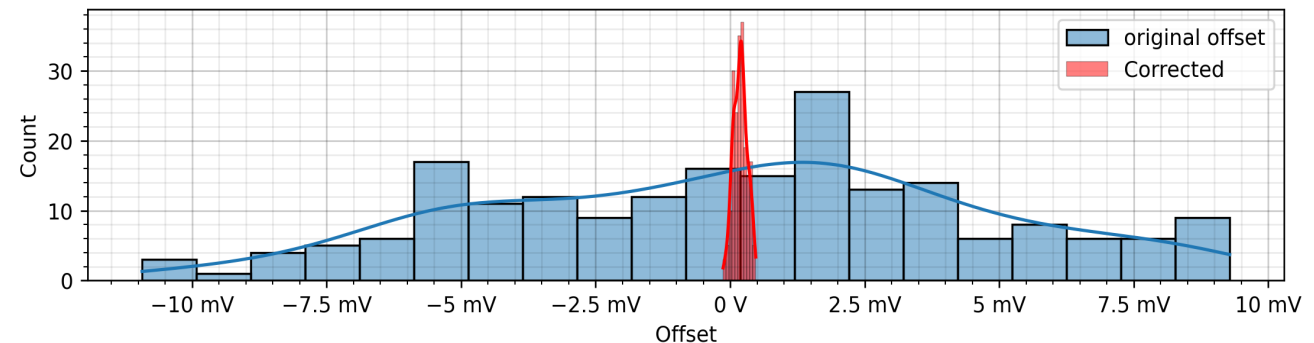


- **Offset compensation** phase to minimize its effect

- 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



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

Variable input capacitance

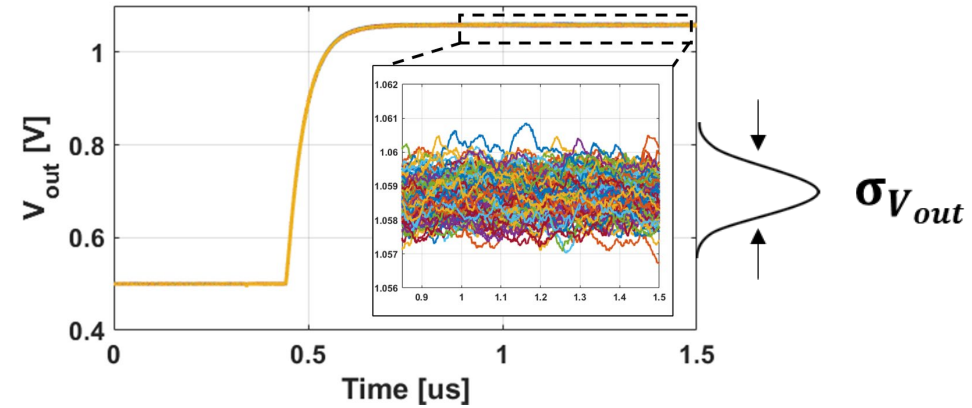
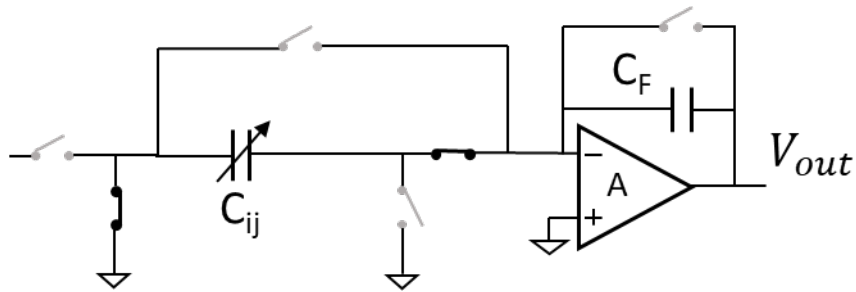
- Variable input capacitance C_{ij} (from 0 to 97.5 pF) \rightarrow large variability of feedback factor

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

- **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 $67^\circ \pm 7^\circ$

Electronic noise

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



$$\sigma_{V_{out,kTC}} \approx 440 \mu V \rightarrow \text{kTC noise contribution}$$

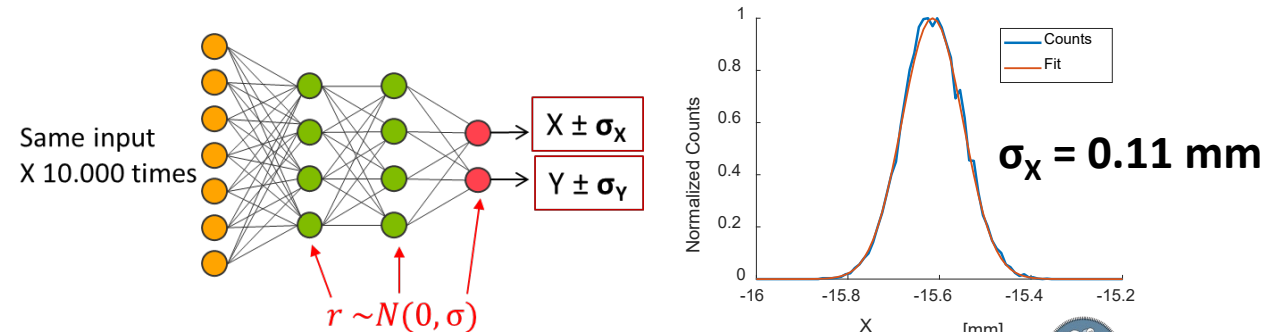
$$\sigma_{V_{out,OPA}} \approx 400 \mu V \rightarrow \text{op-amp only noise contribution}$$

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

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

➤ **Matlab** simulation: given set of input signals and **noise $\sigma = 5 \text{ mV}$** 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**



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

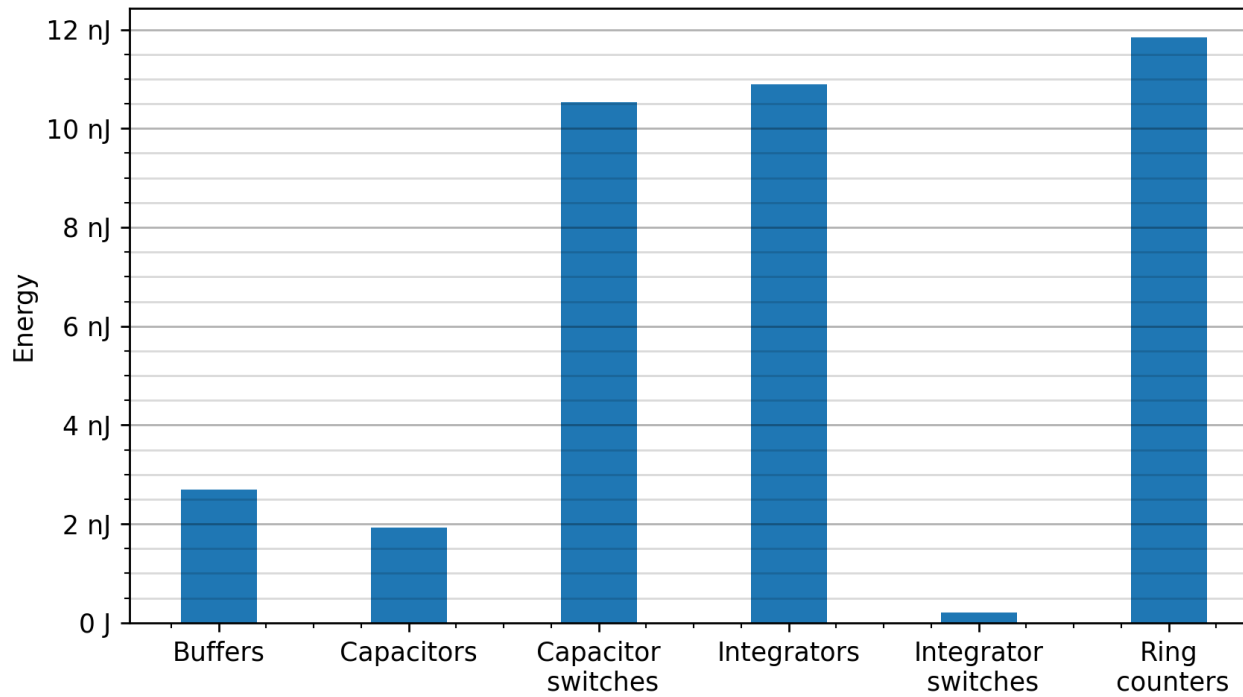
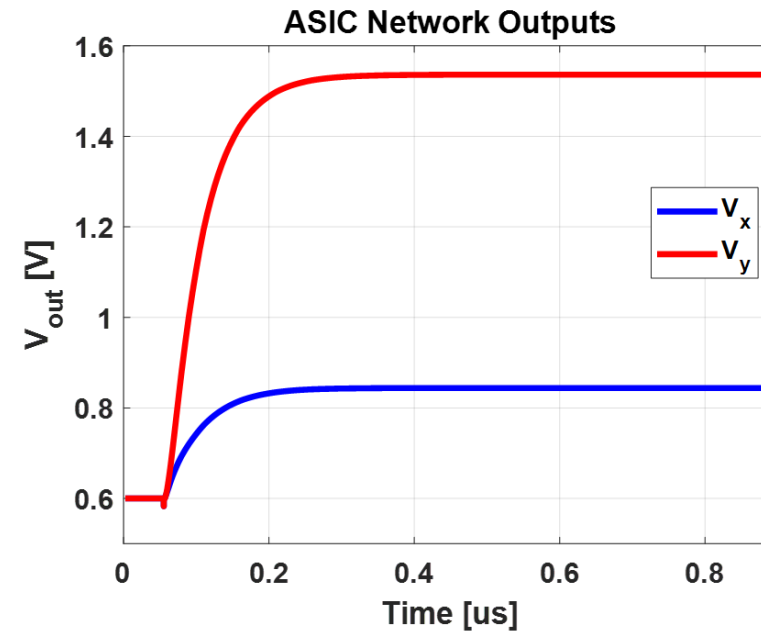
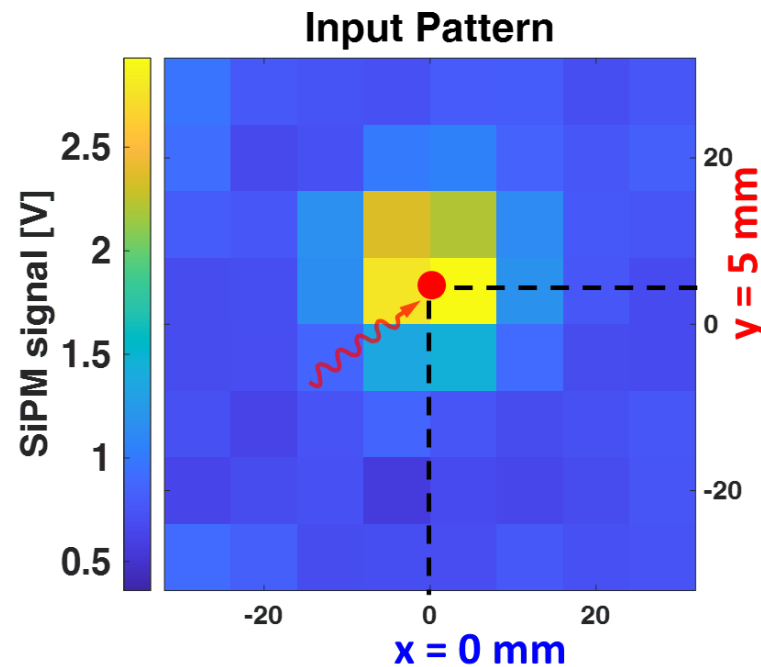


Fig. of merit	Estimate
I/O latency	4.6 μ s
I/O total operations	3566
Total consumption	38.12 nJ
Efficiency	775.21 MOP/s
Analog energy efficiency	135.74 GOP/J
Total energy efficiency	93.55 GOP/J

- 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

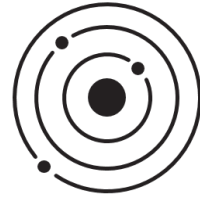


- ✓ **Fully analog** neural network able to perform 5-bit MAC operations in the **charge domain (ASIC)**
- ✓ Can be **integrated** in the front-end ASIC
- ✓ Inference in a more **efficient** way in terms of computational cost and energy, compared to a fully-digital implementation
- ✓ ASIC prototype with $C_{min} = 100fF$ 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.



POLITECNICO
MILANO 1863

DIPARTIMENTO DI ELETTRONICA
INFORMAZIONE E BIOINGEGNERIA



RadLab
POLITECNICO
MILANO 1863



Istituto Nazionale di Fisica Nucleare
Sezione di Milano

Thank you!

susanna.digiacomopolimi.it