Beyond-CMOS Systems for Fast Machine Learning in Physics

a PRIN 2022 project funded in the PE2 ERC sector

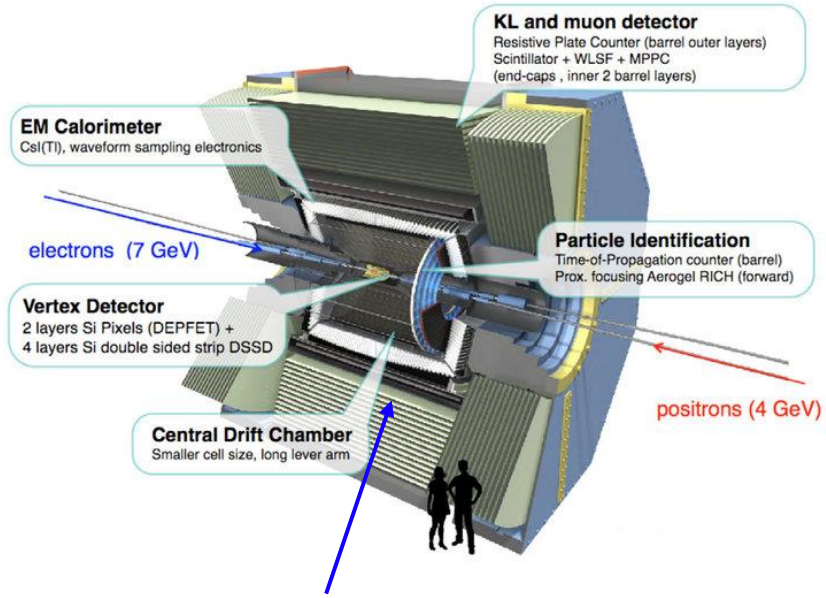
University of Naples “Federico II” and Istituto Nazionale di Fisica Nucleare

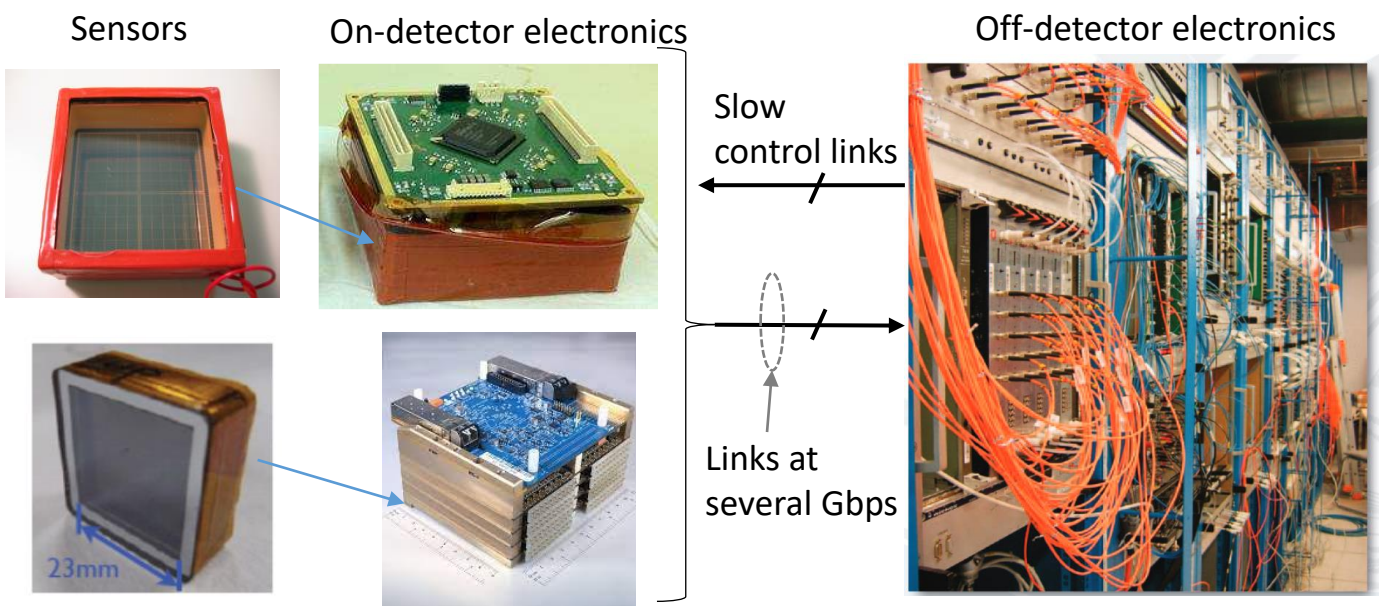
Outline

- Motivation
- Project goal
- People, timing and funds

Motivation

- Trigger and data acquisition systems in HEP are complex
  - Many heterogenous (fast) sensors, ~10-100 millions of electronic channels
  - ~1-10 Petabyte/year raw data rate
- Complexity will grow even more, power consumption too
- on/near-sensor processing for reducing data rates and triggering
  - ECFA recommendation [1]
- Reconfigurability required
  - e.g. search for different physics processes, change of collider conditions





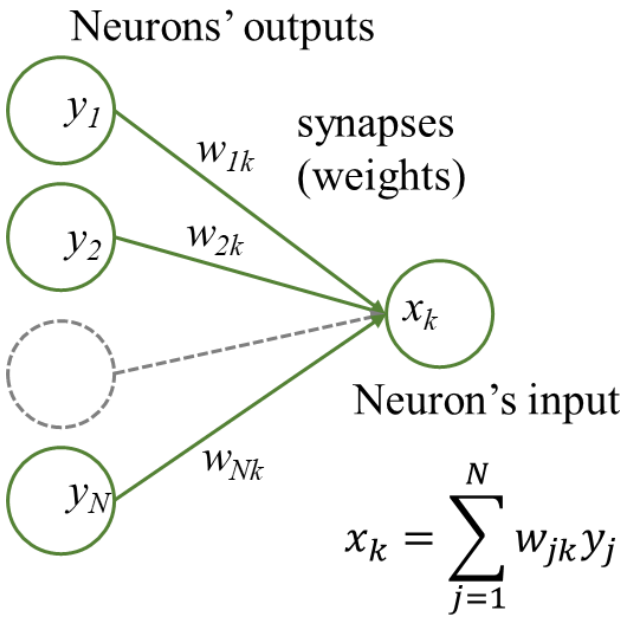
[1] www.doi.org/10.17181/CERN.XDPL.W2EX

Digital-Circuit-based Machine Learning

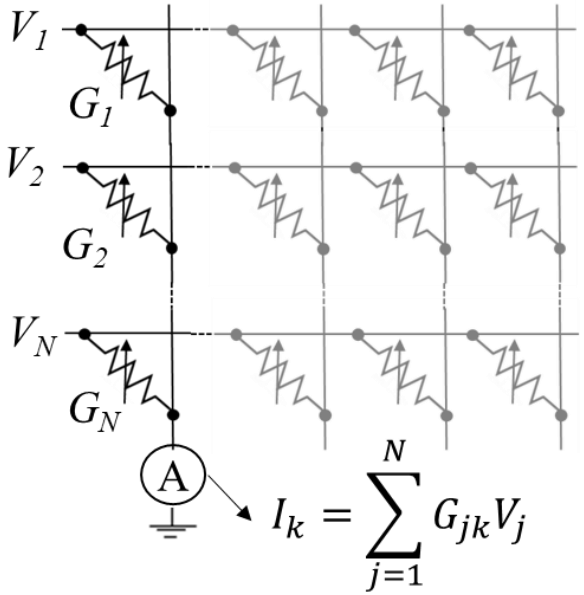
- ML today widely used in off-line data processing
- Fast (i.e. on-line) increasingly being used, even on detector
  - Mostly FPGA and ASIC-based implementations
  - Examples of trigger [1,2] (off-detector) and data compression [3] (on-detector)
  - Latency: tens of ns (ASIC) to microseconds (FPGA)
- CMOS digital hardware designed for logic and arithmetic
  - not optimized for Neural Networks

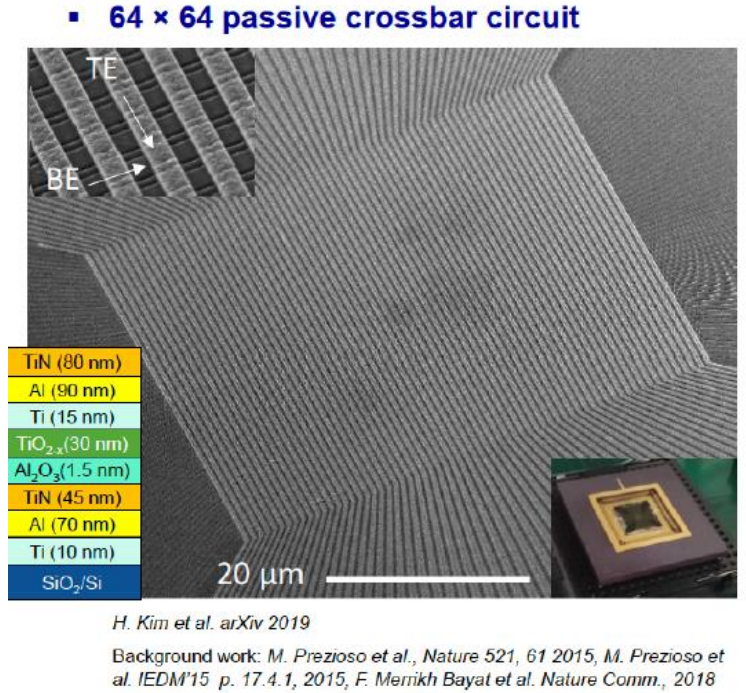
[1] www.doi.org/10.48550/ARXIV.1910.13679
[2] www.doi.org/10.1051/epjconf/202024501021
[3] www.doi.org/10.1109/TNS.2021.3087100

Beyond-CMOS Hardware for Machine Learning



Memristor crossbar





H. Kim et al. arXiv 2019

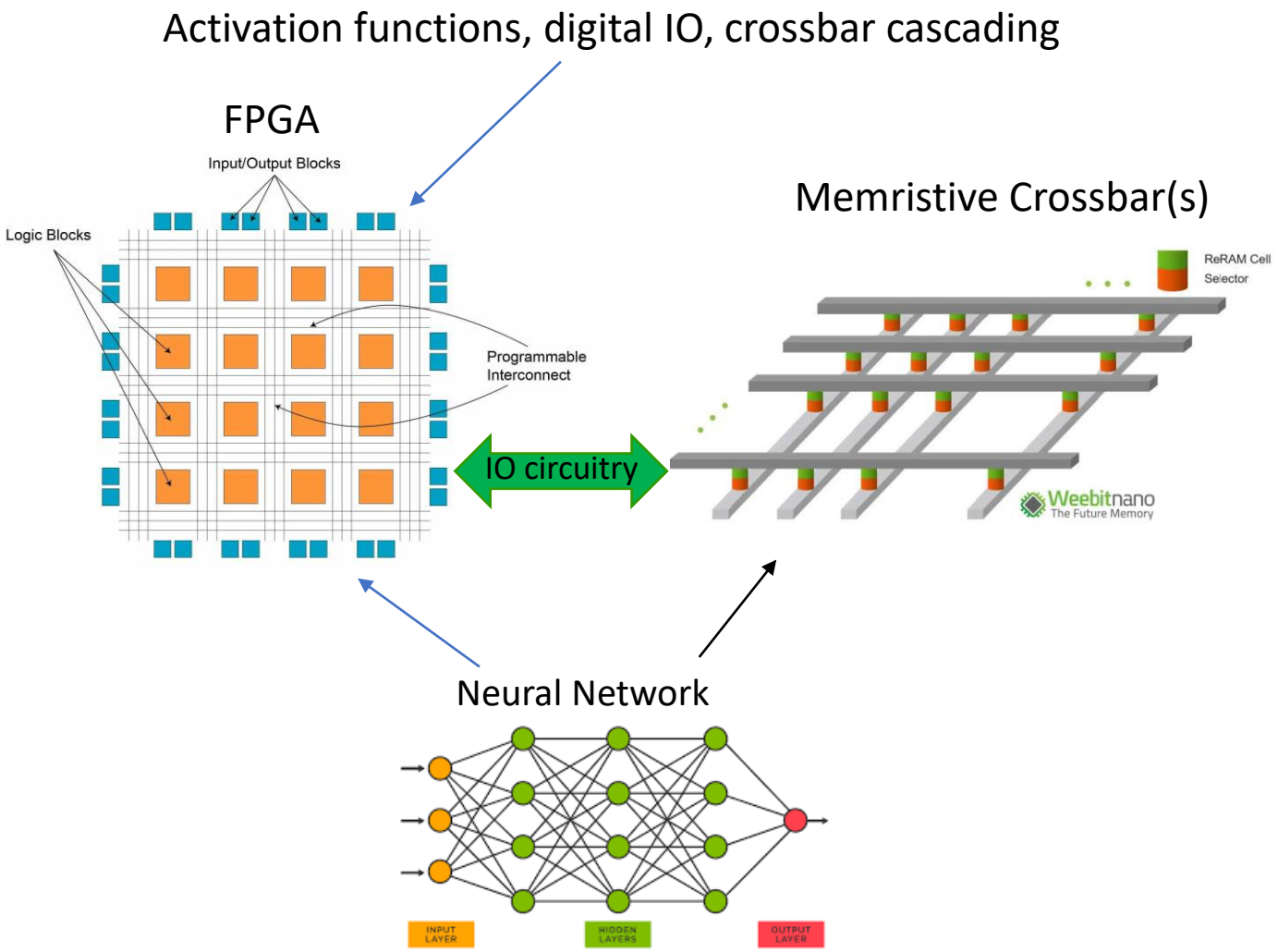
Background work: *M. Prezioso et al., Nature 521, 61 2015, M. Prezioso et al. IEDM'15 p. 17.4.1, 2015, F. Merrikh Bayat et al. Nature Comm., 2018*

https://doi.org/10.1038/s41467-021-25455-0

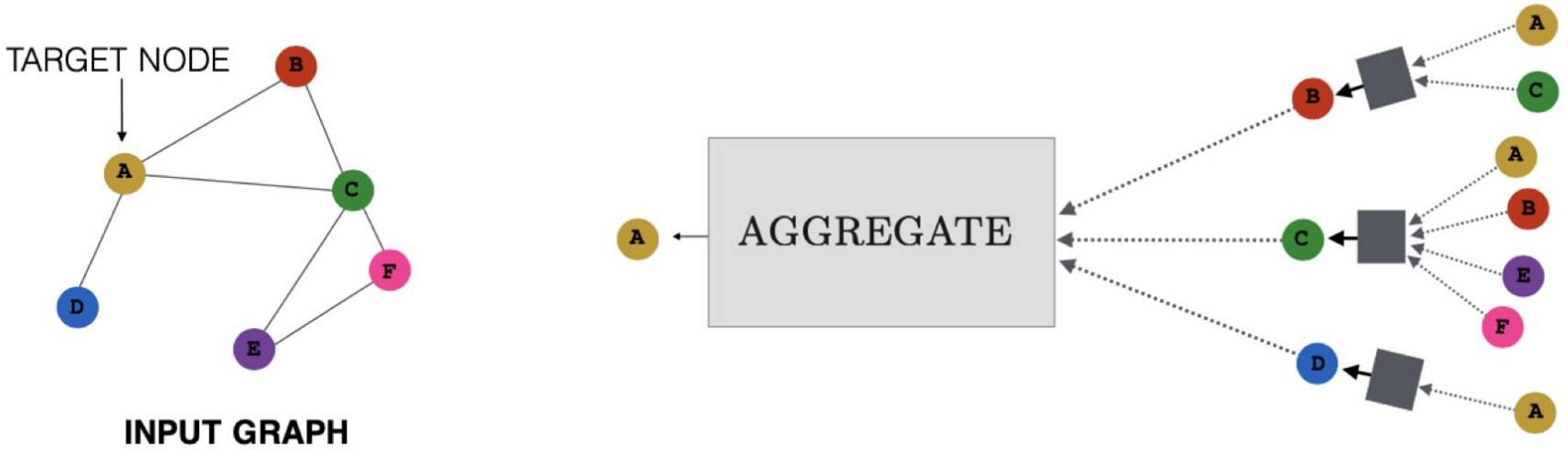
- Emerging hardware represents directly neurons and synapses
  - Memristors: CMOS-compatible, high circuital density (Tb/cm^2) and low power consumption (fJ 100x100 VMM at 6b)
  - Computing in memory
- R&D required to leverage new hardware => this project

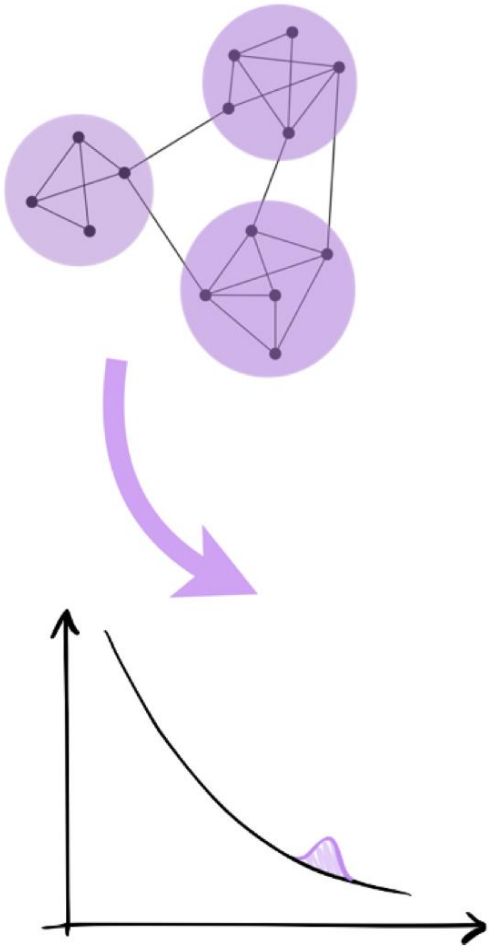
Goal

- Demonstrate fast hybrid digital-neuromorphic computing with FPGAs interfaced to memristive crossbars
  - Perform analog-grade IO with compact circuitry
  - Study of performance and limiting factors (e.g. noise)
  - Compare with full digital ML implementations
- Case Study
  - Graph-based Neural Networks for anomaly detection



Machine Learning with Graphs

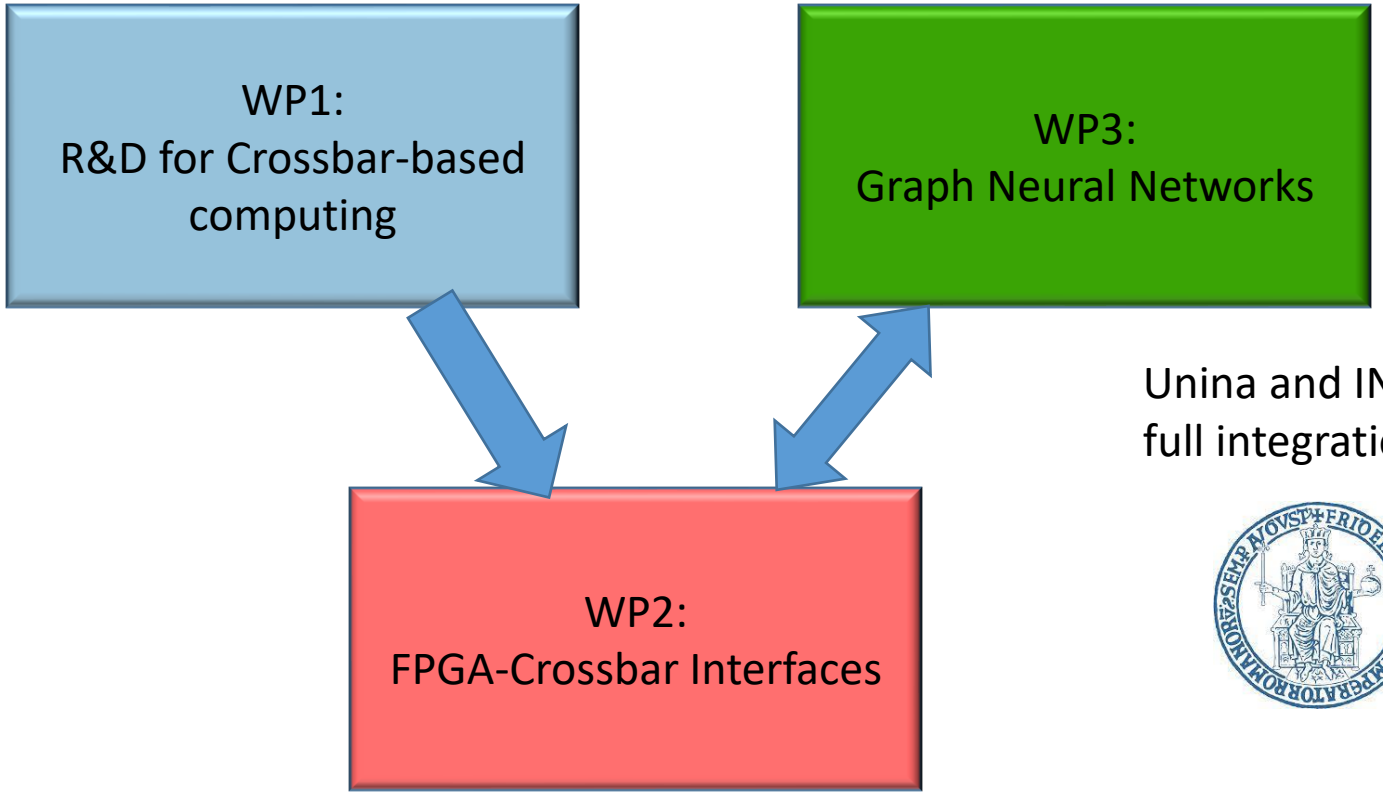




- detector output as “graph”
  - Nodes (e.g. calorimeter clusters) + edges (e.g. distance)
- generalization of “images”
  - aggregate properties of connected objects
  - determine representation of the event in an abstract space
- assign anomaly score to event
  - e.g. autoencoder, distance in abstract space…

V. Ippolito – INFN Sezione di Roma

Work Packages



Unina and INFN participate to all WPs, full integration of activities

Team, Times and Budget

- Presently 9 members (7 staff + 2 post docs)
  - Unina unit (Unina + INFN NA): F. Conventi, R. Giordano (PI), V. Izzo, E. Rossi + 1 post Doc
  - INFN unit (INFN Roma1): V. Bocci, F. Iacoangeli, V. Ippolito (coPI) + 1 post Doc
  - Both units looking for talented young post docs for hardware and software activities
- Official time frame (2 years): Sept. 2023 – Aug. 2025
- Funds – 260 kEUR
  - 20% cut with respect to request
  - rescaling expenses to fit

	Contributo MUR (kEUR)	Confinanziamento (kEUR)
Unina	109.6	24
INFN	90.4	36
Totale	200	60