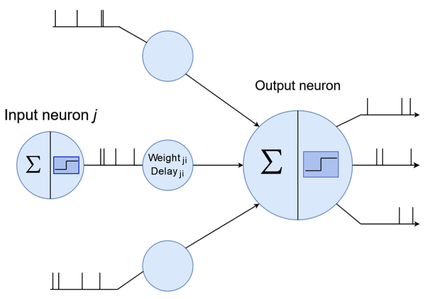
AIGOR: Novel Neuromorphic Computing Architectures with a Multi-Node FPGA system

INFN Sezioni di Roma, Roma 2, APE Lab

AIGOR is a neuromorphic computer architecture *prototype built on a multi-node FPGA (Field Programmable Gate Array) system. It's main objective is to serve as an accelerator platform for efficiently executing Spiking Neural Networks on edge devices. Brain-inspired Spiking Neural Networks represent a promising frontier in computational models, offering potential advantages over traditional computing paradigms in terms of energy efficiency, temporal information processing, and adaptability to dynamic data. This can benefit numerous applications, such as real-time signal processing and pattern recognition in resource-constrained environments. Neuromorphic computing is an approach to hardware architecture design to efficiently implement these biologically-inspired networks, balancing biological plausibility against computational efficiency. Leveraging on a proprietary framework for flexible, low latency communication we aim to deploy our architecture prototype on a multi-FPGA system, adopting a software-hardware codesign workflow that relies on the High Level Synthesis (HLS) programming paradigm for relatively fast and simple translation from a high level simulator of the architecture to the hardware design.*

Neuromorphic Computing



Spiking Neural Networks (SNNs):
Artificial neural networks that communicate through "spikes" or pulses, similarly to biological neurons

Neuromorphic computing
A brain-inspired approach to hardware and algorithm design that efficiently realizes Spiking Neural Networks

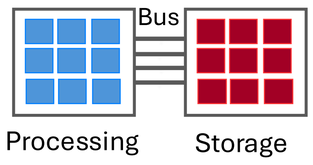
Why mimic the brain?

- **In-Memory Computing**. Co-location of processing and storage

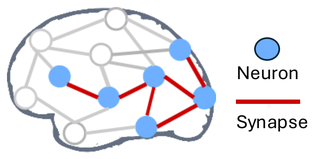
Traditional Von Neuman computing: The cost of data movement

Operation	Energy consumption
Addition of data	1x
Access data (onchip cache)	60x
Access data (offchip RAM)	3500x

Sze et al, IEEE Custom Integrated Circuits Conference (2017)



Brain: processing-storage closely tied, highly parallel 3D stacked architecture



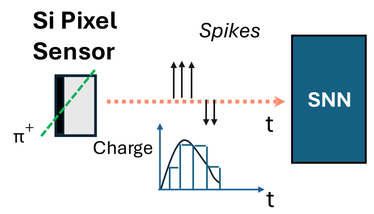
- Sparse, distributed information encoding through spikes
- Asynchronous, event-based computation relying on spikes time
- Local learning rules
  - No costly back-propagation

High energy efficiency, real-time processing, incremental learning

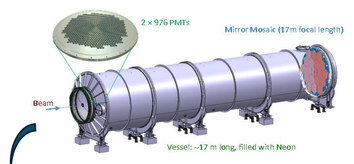
Neuromorphic Computing - Applications

- Any input can be encoded as spikes
- Particularly effective when input is already sparse, event-based or is acquired as a time sequence

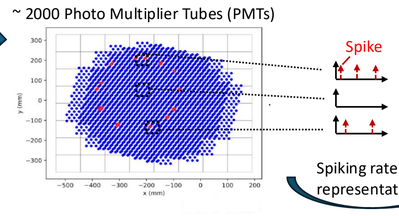
- Identify interesting applications in physics
  (particle physics sensors, anomaly detection, event cameras, time series data …)

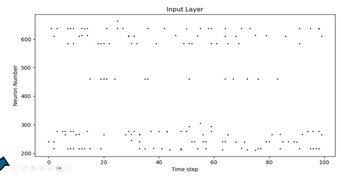


- **Test-case**: NA62 Ring Imaging Cherenkov (RICH) detector



~ 2000 Photo Multiplier Tubes (PMTs)



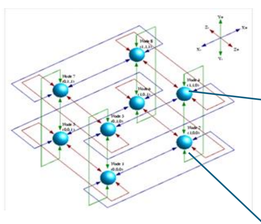


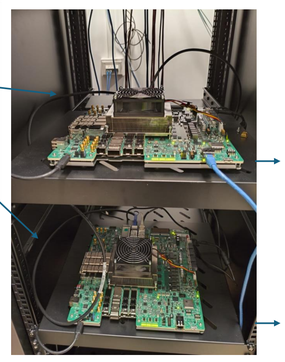
Input layer 784 neurons
Hidden layer 1000
Output layer 4 (**Counting Rings**: 0,1,2, 2+)
10000 samples (7000 train, 2000 test)
Test Set Accuracy: ~ 80%
Trained in *snnTorch*

Ideal for **edge Computing, on-sensor processing, streaming** applications

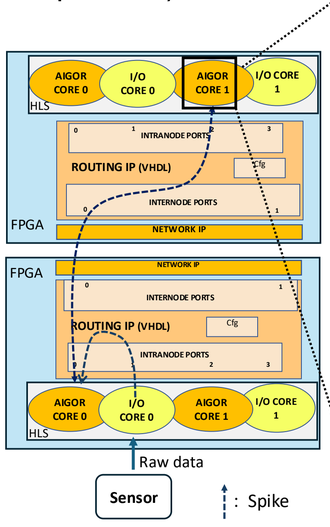
Multi-FPGA architecture prototype

- Direct network of interconnected FPGAs. 3D Torus.
- INFN Routing IP for low-latency inter-FPGA and intra-FPGA communication (< 1 us for up to 1 kB packets)

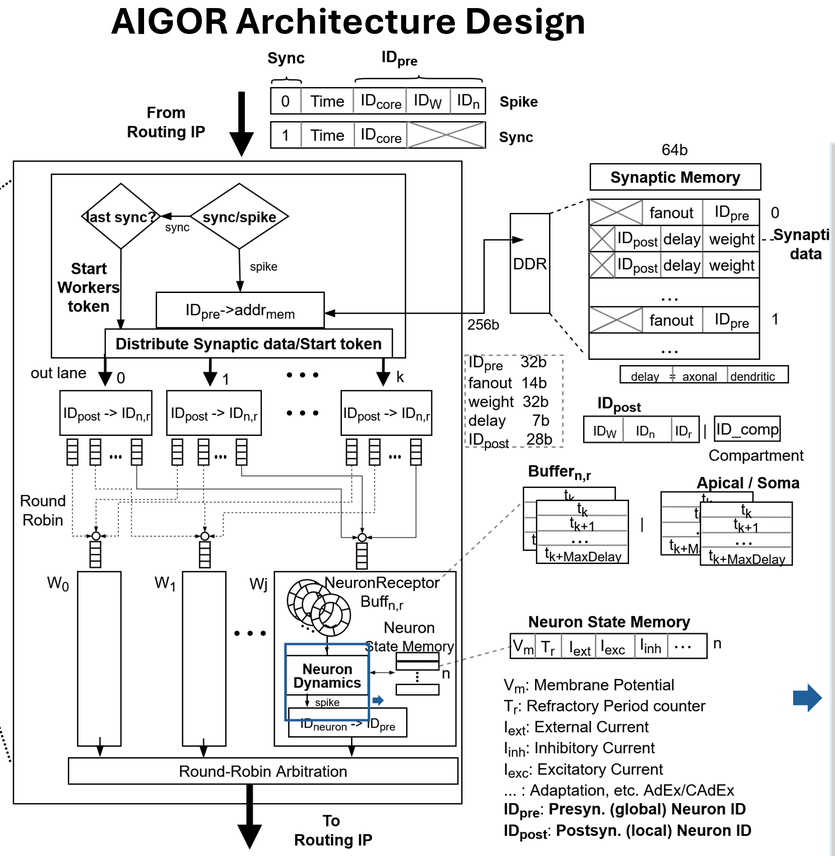




AMD Versal Premium VPK180

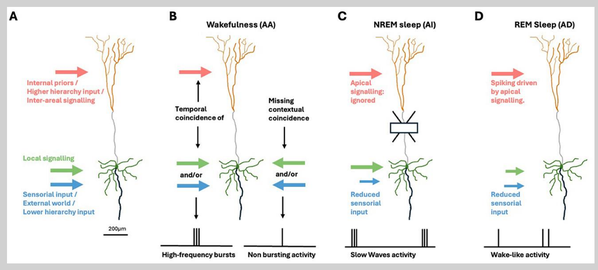


AIGOR Architecture Design



V$_m$: Membrane Potential
T$_r$: Refractory Period counter
I$_{ext}$: External Current
I$_{inh}$: Inhibitory Current
I$_{exc}$: Excitatory Current
… : Adaptation, etc. AdEx/CAdEx
ID$_{pre}$: Presyn. (global) Neuron ID
ID$_{post}$: Postsyn. (local) Neuron ID

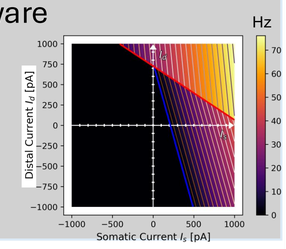
- **Adaptable neuron dynamics module** to explore usage of biologically realistic neuron models
- e.g. multi-compartment neuron [1]



- Specialized in **incremental learning**.
- ThetaPlanes: piece-wise linear approximation of its transfer function for hardware implementation

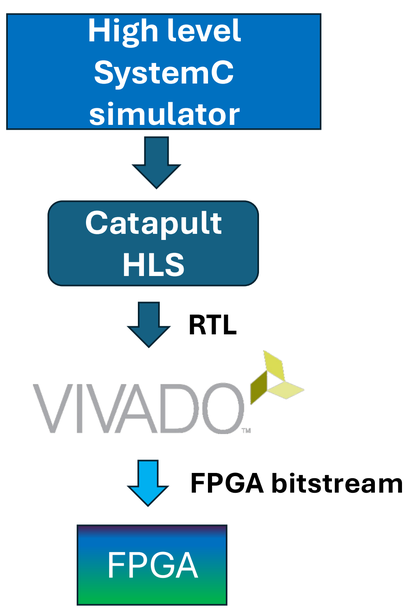
$$\nu_F(I_s, I_d; \nu) = \Theta_\rho(1 - \Theta_H) \cdot \nu_- + \Theta_H \cdot \nu_+$$

[1] Pastorelli et al.)2025) doi: 10.3389/fncom.2025.1566196



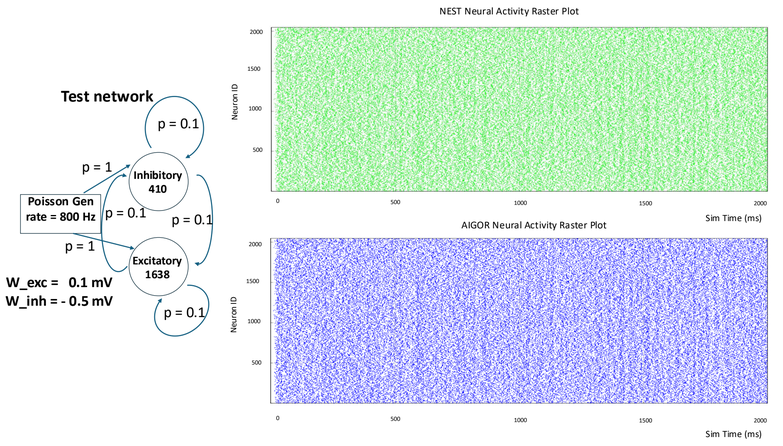
- **Modular, scalable, reconfigurable** architecture prototype with multiple levels of parallelism
- Configurable I/O cores for diverse input encoding pipelines
- Authomatized configurable synthesis workflow. Config » number of nodes, router ports (Aigor-I/O Core), neuron model, bit-field sizes, sim parameters, etc.

Flexible development Workflow Based on High Level Synthesis



- **SystemC (C++ based event-driven) architecture simulator**
  - Validate and test architecture functionality
  - Modelling multi-FPGAs communication delays
  - SystemC allows higher or lower levels of abstraction (TLM, RTL-like)
- **Catapult High Level Synthesis (HLS)**
  - Direct transition from high-level simulator to RTL
  - Benefits: high level language, software emulation mode, early estimates on latency/throughput and FPGA-resources consumption → relatively easy validation, reprogrammability, and debug

Funcional Validation against the nest software simulator



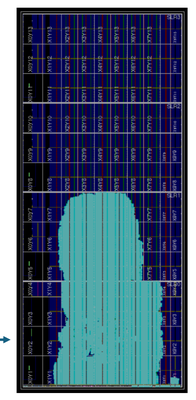
C++ simulator, with a Python interface, implementing a large number of models of biological neurons and synapses

- **FPGA Occupancy** single AIGOR-Core (ON VPK180)

Single Aigor-Core	LUT	Flip Flop	DSP	BRAM	Synaptic Mem requires
512 N (4 W, 128 n x W)	75.5k **2.25%**	30.6k **0.45%**	0 **0.00%**	16 **0.32%**	4 Mb
2048 N (32 W, 64 n x W)	764.0k **22.7%**	197.3k **2.93%**	2 **0.01%**	96 **1.94%**	84 Mb
2048 N (16 W, 128 n x W)	708.5k **21.1%**	106.9k **1.59%**	34 **0.24%**	64 **1.30%**	84 Mb

VPK-180 Total PL mem 994 Mb

N: tot. neurons, W: workers
n x W: neurons per worker



Funded by the European Union NextGenerationEU · Italiadomani · Ministero dell'Università e della Ricerca
FAIR Future Artificial Intelligence Research
EBRAINS
EBRAINS Italy

Contacts:
APE Lab website: https://apegate.roma1.infn.it
Presenter contact: pierpaolo.perticaroli@roma1.infn.it