

Exploring Spiking Neural Networks

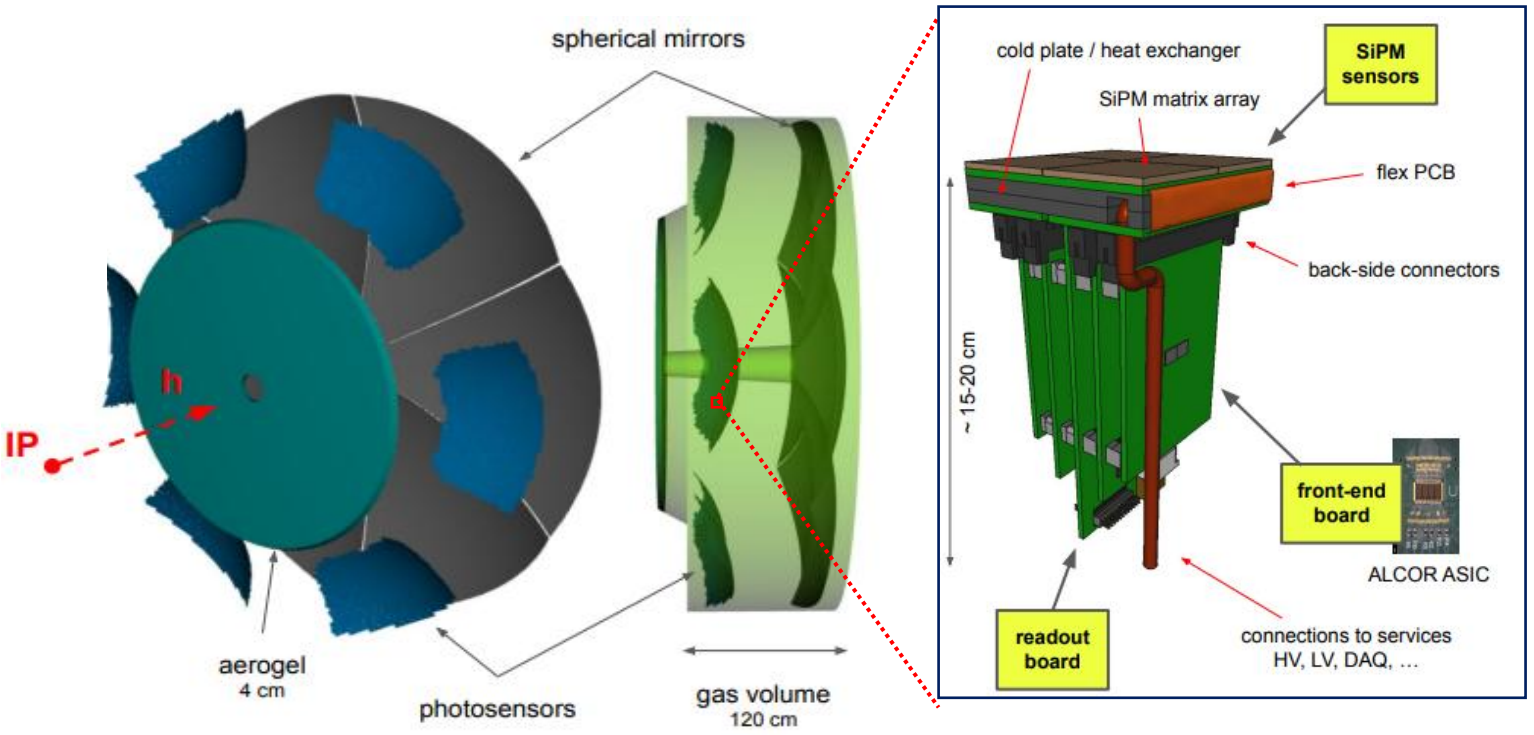
for Online Data Reduction at the ePIC dRICH Detector

Pierpaolo Perticaroli
INFN Sezione di Roma · APE Lab

Digital Twins for Nuclear and Particle Physics — May 11 – 15, 2026

The ePIC dRICH: Detector and Readout Chain

Dual-radiator RICH for hadron PID in the EIC forward region



Aerogel + C_2F_6 gas radiators, spherical mirrors, segmented photosensors

Photodetector Unit · SiPM matrix + ALCOR ASIC

~300k SiPM → 1248 PDUs → 30x DAM FPGAs → ePIC processing & storage

PID & detector

- π / K / p separation from 3 to 50 GeV/c.
- ~300 000 SiPM channels, 6 sectors.
- 1248 Photo Detection Units **(PDU) .**
  - PDU = 4x64 SiPM matrix + 4 FEB + 1 RDO board.

Readout chain

- ALCOR: ASIC on FEBs, handles TDC.
- RDO: buffers and packetizes data from 4xFEBs.
- **Data Aggregation Module DAM:** (FELIX / FLX-155 FPGA), 42 links with RDOs. Connected to ePIC farm through PCIe (100 GbE). 5 x DAMs x Sector.

dRICH: Output Bandwidth

bandwidth/throughput issue

- Sensors DCR: 3-300 kHz (**increasing with radiation damage** ⇒with experiment lifetime).
- Considering planned techinques to manage SiPMs irradiation (e.g. annealing):
  - **worst DCR case: 300 kHz**

➔ Full detector throughput (FE): **6.792,19 Gbps**

➔ a **reduction is needed** to cope with 30 channel (30x100GbE) bandwidth availability

➔ Single Dam output bandwidth **226,41 Gbps**

dRICH DAQ parameters	
RDO boards	1248
ALCOR64 x RDO	4
dRICH channels (total)	319488
Number of DAM	30
Input link in DAM	42
Output links from DAM to TP	1
Number of DAM Trigger Processor	1
Input link to DAM Trigger Processor	30
RDO-DAM Link Bandwidth (VTRX+) [Gb/s]	10
DAM to Echelon-0 Switch Bandwidth [Gb/s]	100
dRICH Interaction tagger reduction factor	1
Interaction tagger latency [s]	1,00E-04
EIC parameters	
EIC Clock [MHz]	98,522
Orbit efficiency (takes into account gap)	0,92

dRICH data stream analysis		Limit
Sensor rate per channel [kHz]	300,00	4.000,00
Rate post-shutter [kHz]	276,00	800,00
Throughput to serializer [Mb/s]	172,50	788,16
Throughput from ALCOR64 [Mb/s]	1.380,00	
Throughput from RDO [Gb/s]	5,39	10,00
Input at each DAM [Gbps]	226,41	420,00
Buffering capacity at DAM [Mb]	23,18	
Output from each DAM [Gbps]	226,41	100,00
Aggregated dRICH data throughput		
Total input at DAM [Gb/s]	6.792,19	
Total output from DAM [Gb/s] to Echelon	6.792,19	

dRICH: Output Bandwidth

data reduction system

- Sensors DCR: 3-300 kHz (**increasing with radiation damage** ⇒with experiment lifetime).
- Considering planned techinques to manage SiPMs irradiation (e.g. annealing):
  - **worst DCR case: 300 kHz**

➔ Full detector throughput (FE): **6.792,19 Gbps**

➔ a **reduction is needed** to cope with 30 channel (30x100GbE) bandwidth availability

- EIC beams bunch spacing: ~10 ns ⇒ bunch crossing rate of 100 MHz
- For the low interaction cross-section (DIS) ⇒ one interaction every ~200 bunches ⇒ interaction rate of ~500KHz.

➔ A system tagging dark current noise-only events can solve the throughput issue (reducing down to 1/5 the data throughput)

➔ Single DAM output bandwidth : **45,64 Gbps**

dRICH DAQ parameters	
RDO boards	1248
ALCOR64 x RDO	4
dRICH channels (total)	319488
Number of DAM	30
Input link in DAM	42
Output links from DAM to TP	1
Number of DAM Trigger Processor	1
Input link to DAM Trigger Processor	30
RDO-DAM Link Bandwidth (VTRX+) [Gb/s]	10
DAM to Echelon-0 Switch Bandwidth [Gb/s]	100
dRICH Interaction tagger reduction factor	5
Interaction tagger latency [s]	1,00E-04
EIC parameters	
EIC Clock [MHz]	98,522
Orbit efficiency (takes into account gap)	0,92

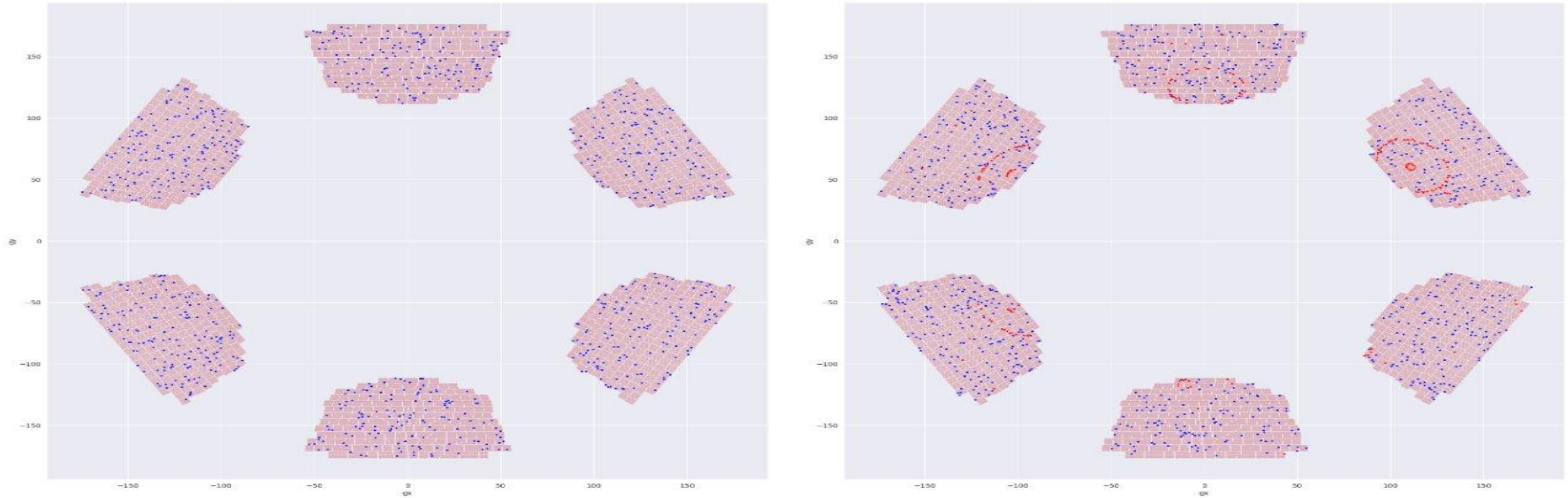
dRICH data stream analysis		Limit
Sensor rate per channel [kHz]	300,00	4.000,00
Rate post-shutter [kHz]	276,00	800,00
Throughput to serializer [Mb/s]	172,50	788,16
Throughput from ALCOR64 [Mb/s]	1.380,00	
Throughput from RDO [Gb/s]	5,39	10,00
Input at each DAM [Gbps]	226,41	420,00
Buffering capacity at DAM [Mb]	23,18	
Output from each DAM [Gbps]	45,28	100,00
Aggregated dRICH data throughput		
Total input at DAM [Gb/s]	6.792,19	
Total output from DAM [Gb/s] to Echelon	1.358,44	

Signal vs. Noise: a Spatio-Temporal Discrimination Problem

The physics is encoded in spatial and timing correlations

Noise-only event

Physics Signal + Physics Background + Noise

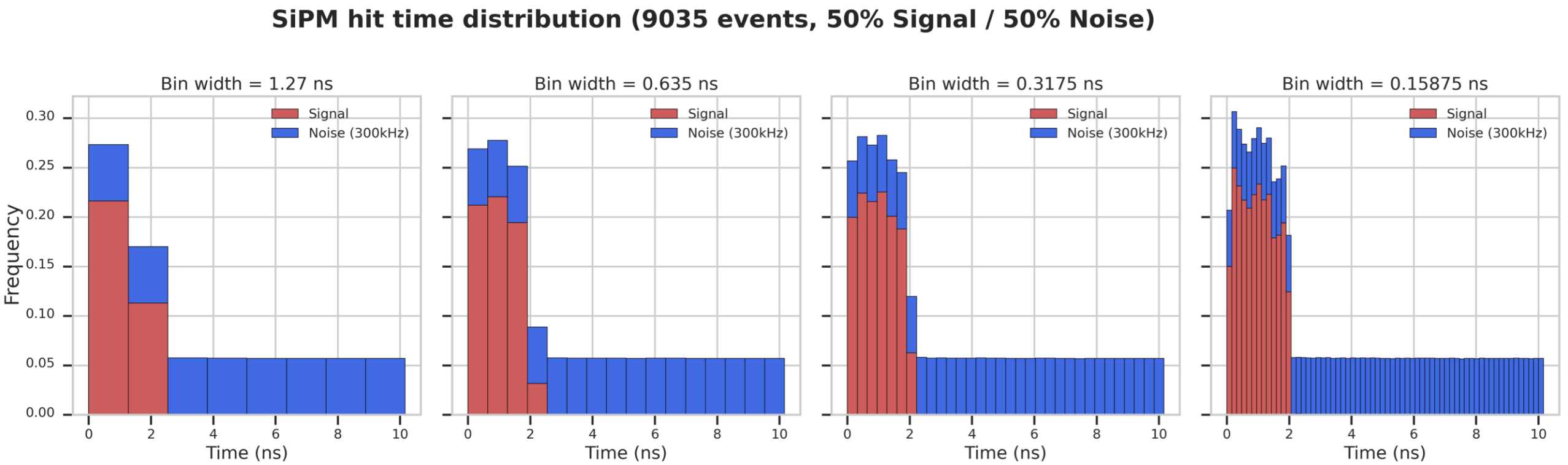


Aggregated hits over a full Bunch Crossing acquisition window (~ 10 ns)

Cherenkov hits are correlated in space (rings); Dark Current hits are uniformly random.

Signal vs. Noise: a Spatio-Temporal Discrimination Problem

The physics is encoded in spatial and timing correlations

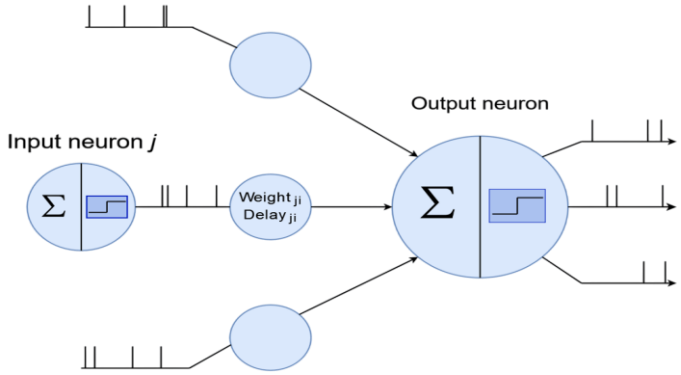


Cherenkov photons arrive bunched in time within the ~10 ns BC window; noise photons are uniformly random.

→ A temporal pattern-recognition problem. Most BCs contain only noise — a smart front-end could suppress them locally.

Why Consider Spiking Neural Networks?

SNNs are natively temporal and event-driven — potentially a great fit



Why they fit the problem

Temporal encoding is native: information lives in precise spike timing, not in data aggregated over a fixed time window

Event-driven processing: no input → no computation. Sparse inputs → sparse compute, small power consumption.

Near-sensor potential: compact, low-power neuromorphic hardware could sit close to the SiPM.

Plasticity and bio-inspired multi-compartment neurons: potential for brain-like sample efficient online learning.

Frame-based DNNs pay for every data frame regardless of content — SNNs do not.

Some interesting works in recent years in HEP

- On-Sensor Data Filtering using Neuromorphic Computing for High Energy Physics Experiments, *Shruti R. Kulkarni et al. (2023)* *doi.org/10.1145/3589737.3605976*
- Unsupervised Particle Tracking with Neuromorphic Computing, *E.Coradin et al. (2025)* *doi.org/10.48550/arXiv.2307.11242*
- *No SNN deployed in a full DAQ pipeline yet (to our knowledge)*

dRICH offers a clear temporal asymmetry, and very high throughput → a clean testbed for SNN benefits & limitations.

SNNs Are Not Plug-and-Play

Fundamental throughput challenges for a 10 ns world

1. The Data Format Mismatch: Packed Buffers vs. Event Streams

- HEP readout systems typically maximize bandwidth by grouping many (near-)simultaneous hits into dense digital words. SNNs natively expect a continuous, high-speed stream of individual events.
- Unpacking dense buffers into a serial spike stream can add latency upfront. The algorithm must artificially "unfold" the 10 ns window into sequential timesteps. Processing these discrete bins sequentially can burn multiple clock cycles.

2. The Communication Bottleneck: One Spike at a Time

- Neuromorphic hardware routes spikes internally using Address-Event Representation (AER)—essentially sending a tiny "Neuron X fired at this time" packet over a shared bus.
- While physics signals are sparse, the continuous background noise (300 kHz DCR) creates a flood of hits. Serializing these onto a shared bus creates a traffic jam within the brief 10 ns window.

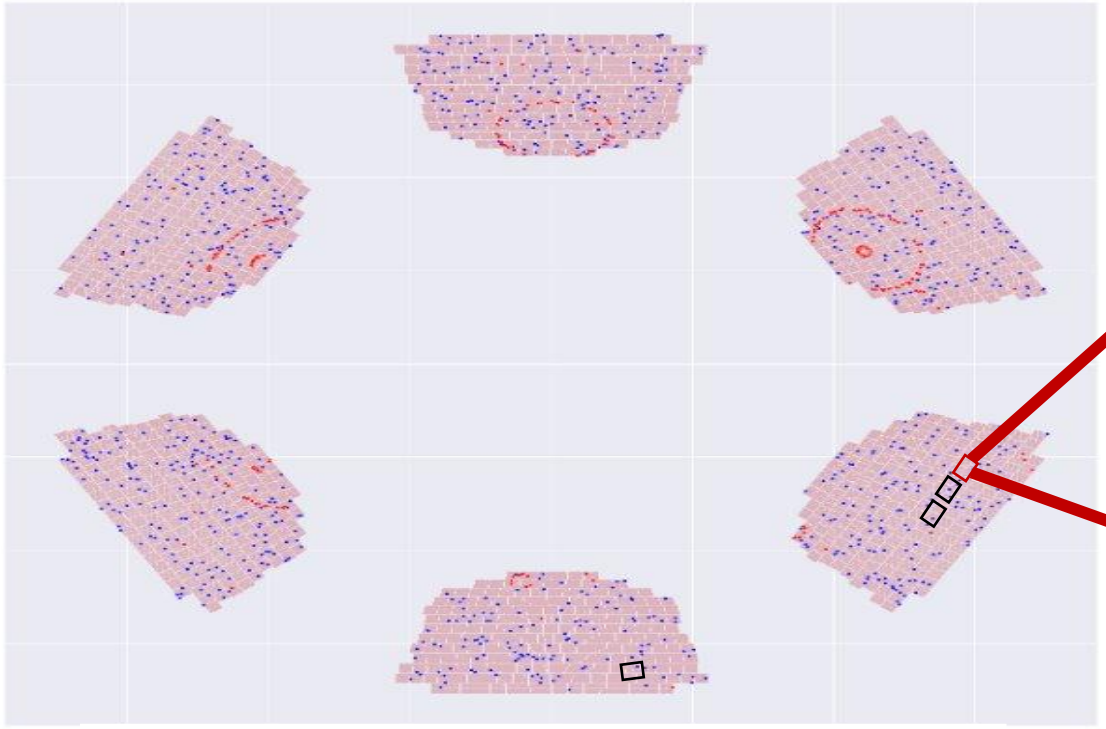
3. The Overhead of Determinism

- To ensure reproducible results, typical multi-core digital SNN chips use "barrier synchronization"—all cores pause and wait until everyone finishes the current timestep.
- This rigid waiting guarantees correct dynamics across the network, but it strips away the theoretical speed benefits of asynchronous, event-driven processing.

Not showstoppers — they demand architectural thinking. That is what we are exploring.

System Overview: Encoder + Classifier

Two stages — smart encoding near the sensors, SNN classification downstream



Stage 1 - LIF Encoder

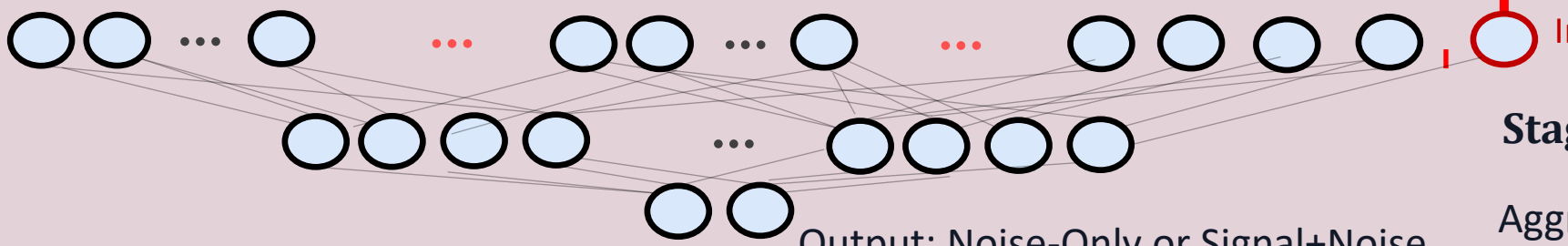
Local To Each **PDU** (256 SiPM matrix)

Input: Raw Hit [timestamp, SiPM id]

Output: Spike [timestep, PDU id]

1 Encoder PDU j =

= 1 SNN Input neuron j

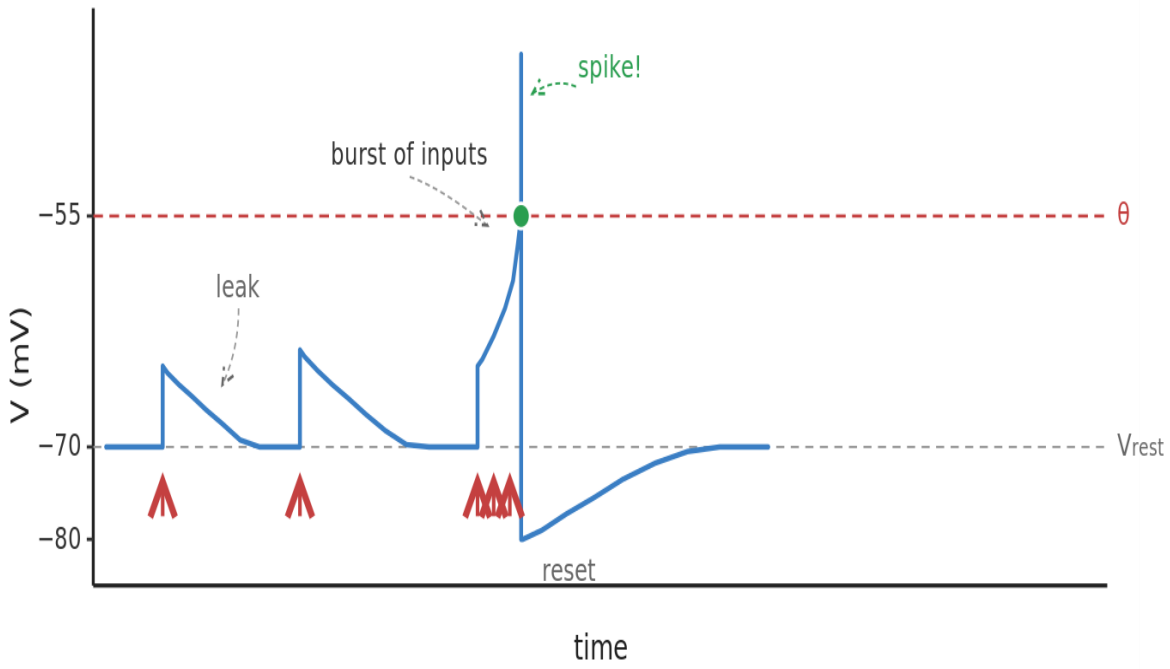


Stage 2 - SNN Classifier on FPGA

Aggregates spatially all dRICH PDUs

The LIF Encoder: Temporal Coincidence Detection

The encoder itself is a data-reduction stage



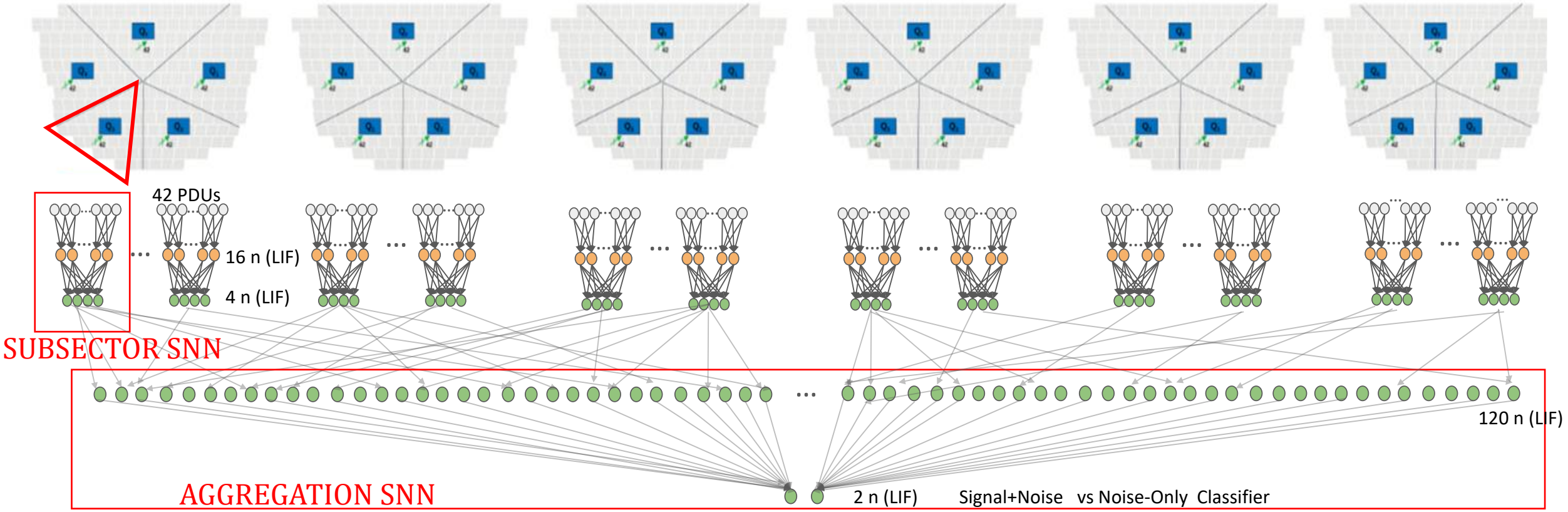
How it works

- Incoming hits increment the membrane potential.
- Leak every timestep.
- Threshold crossing → emit spike (Stop at first spike – Time to First Spike mode).
- **Cherenkov hits cluster in time → rapid charge-up → fires.**
- **Noise hits arrive sporadically → leak dissipates charge → silent.**

Tunable: threshold, leak (shift), time bin, TTFS vs. multi-spike.

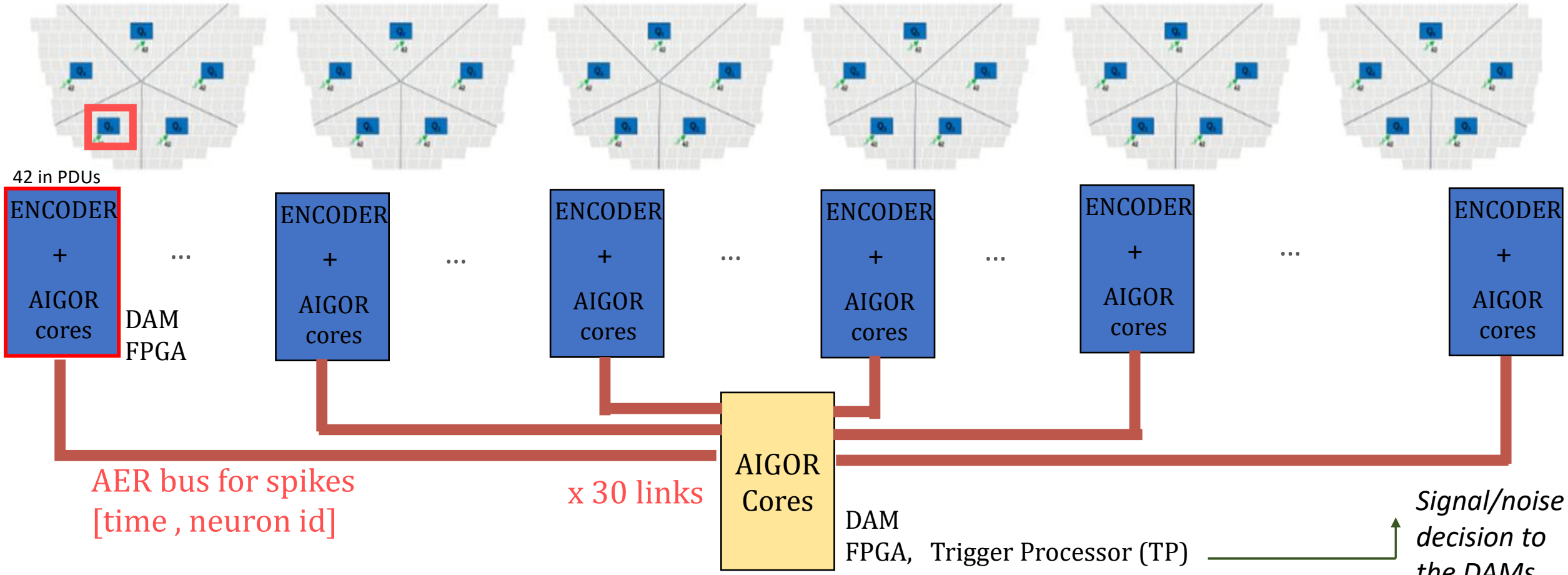
Pre-filters noise before AER Neuromorphic Chip, harnessing time structure.
Reduces the event load (spikes) for the SNN stage

The SNN Classifier: Distributed Spiking Neural Network



- 6 Sectors , 5 DAM FPGAs x Sector -> 30 Subsectors SNNs
- 1 more DAM FPGA (Trigger Processor) -> Aggregation SNN (2 Layers), concatenating features from Subsector SNNs
- Final decision: Classify Bunch Crossing as Signal+Noise or Noise-Only

System Overview: Design of distributed SNN on FPGAs

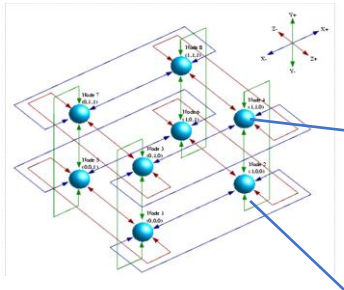


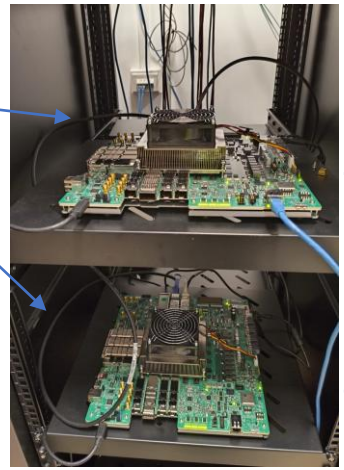
- **AIGOR: multi-core Neuromorphic Computing Architecture prototype → deploy SNN layers.**
- Subsector DAM FPGA: LIF Encoder for RDO hits + AIGOR Cores for Subsector SNN layers. Buffers RDO data.
- Aggregation DAM FPGA: AIGOR Cores for Aggregation SNN Layers. Triggers RDO data forwarding to Farms.

Online AI-based distributed data reduction for the dual-radiator RICH detector in the ePIC experiment, C.Rossi et al. https://doi.org/10.1016/j.nima.2026.171416

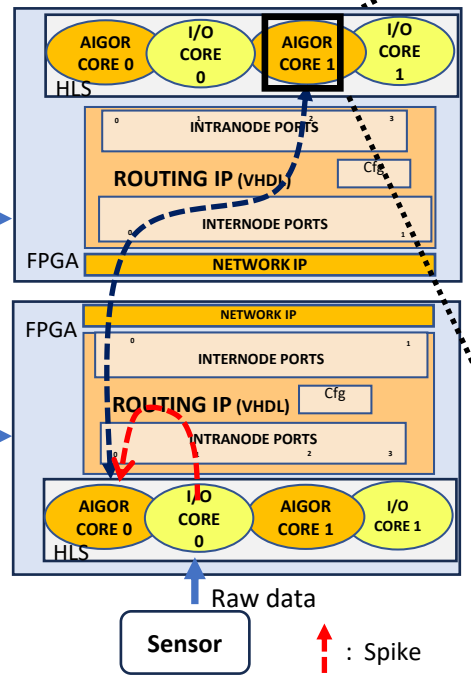
AIGOR: multi-node architecture for SNN applications

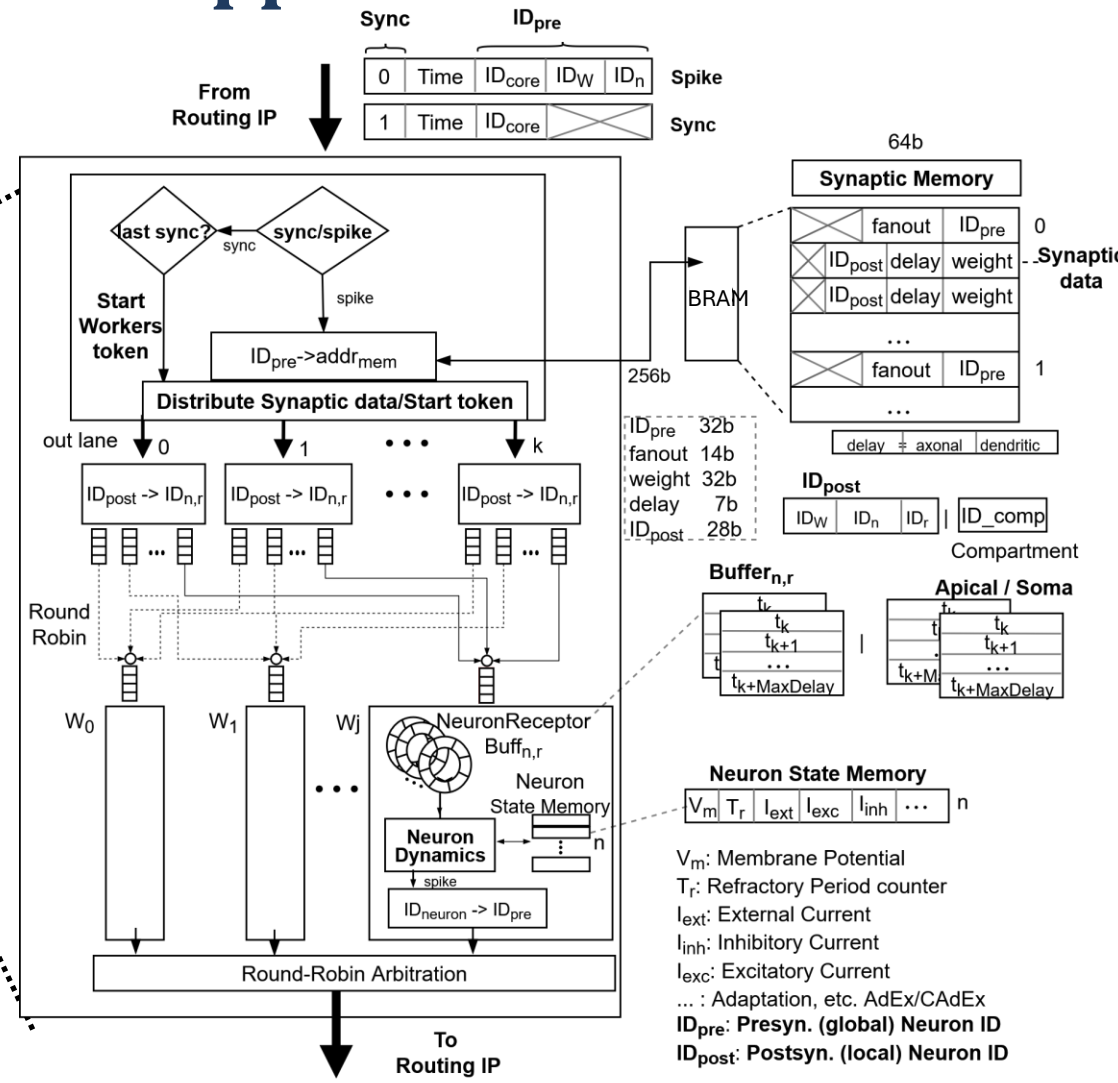
- **Modular, scalable, reconfigurable** architecture prototype for accelerating SNNs.
- Direct network of interconnected FPGAs. 3D Torus.
- INFN Routing IP for low-latency inter-FPGA and intra-FPGA communication (< 1 us for up to 1 kB packets)





AMD Versal Premium
VPK180





- Multiple levels of parallelism (Cores , Workers, Neurons).
- Configurable I/O Cores and Neuron Model for diverse input and processing pipelines.
- **AER based, optimized for sparse inputs. → Spike Load per Bunch Crossing must be low.**

Data Sparsity: Pre vs. Post Spike-Encoding

Quantifying the data-reduction potential

Pre-Encoding

At 300 DCR = 300 kHz

≈ 1687 hits raw hits per 10 ns BC, over the entire dRICH.

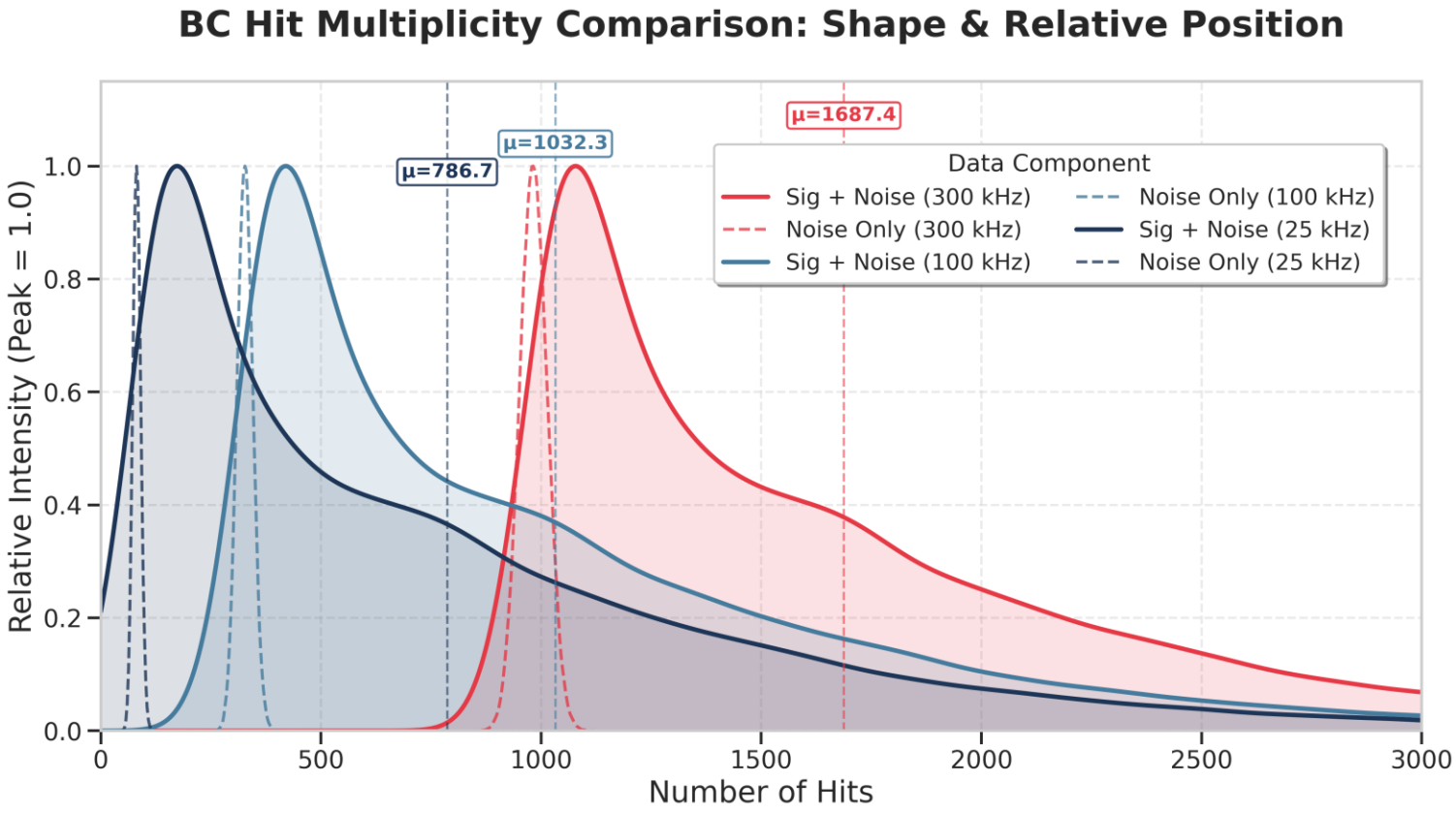
≈ 1000 hits per BC are Noise-Only! (59 %)

1.36 Avg. Number of hits per PDU per BC

However Signal is Bursty

Avg hits per Active PDU (>= 1 hit) per BC:

4.56 Signal vs 1.73 Noise-only

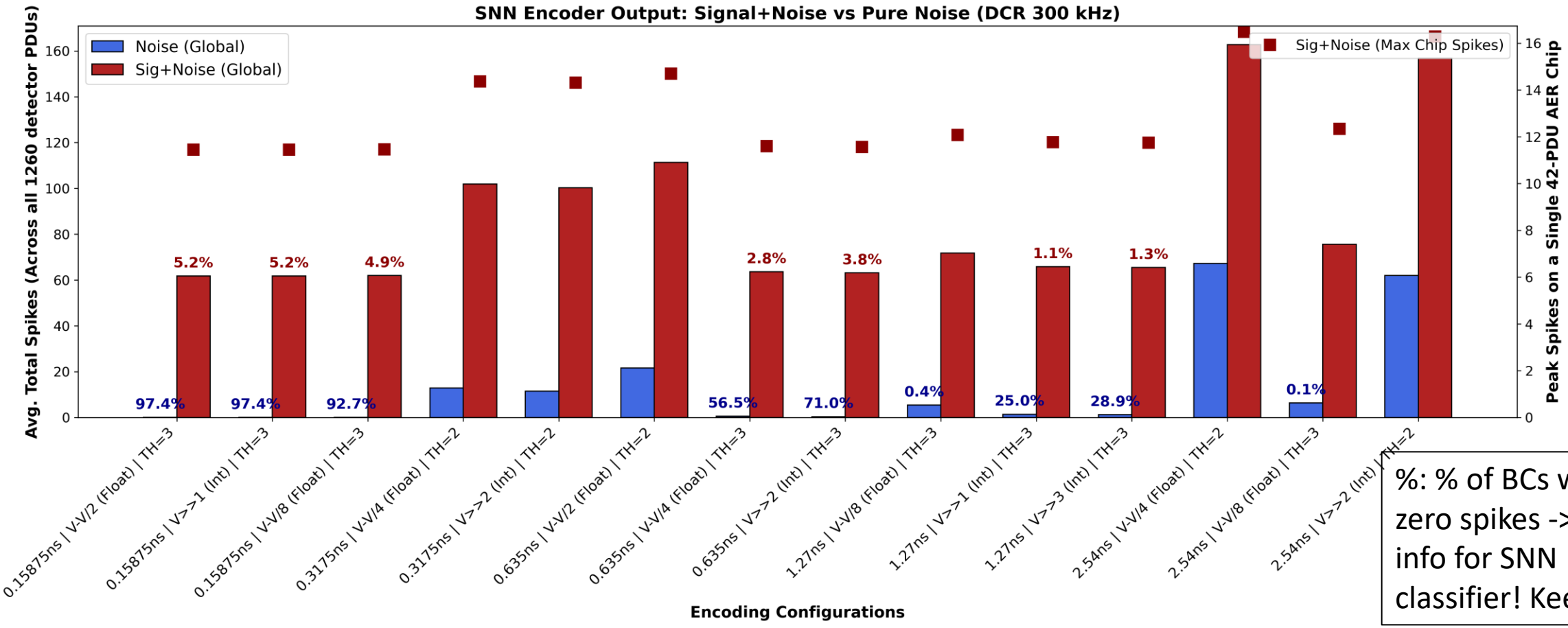


Strong enough signature for the LIF Encoder local to PDU to spike only for Signal hits, with the right threshold settings

Data Sparsity: Pre vs. Post Spike-Encoding

Quantifying the data-reduction potential

Post-Encoding – Multiple LIF Encoder Configurations (Time bin, Leak Strength, Threshold to Spike)



%: % of BCs with zero spikes -> no info for SNN classifier! Keep low!

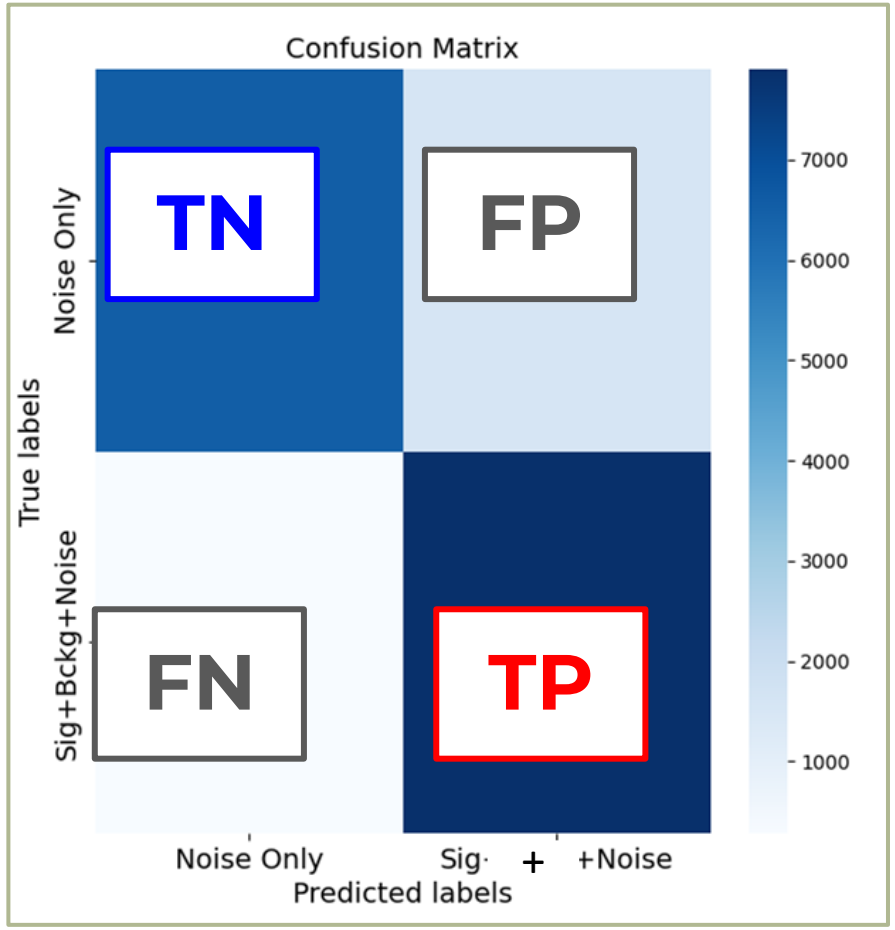
Total Spikes over dRICH ≈ 60-120 → <u>94 % sparsification over raw hits!</u>

Peak # of Spikes per Subdetector ≈ 14

Software: Encoder + SNN Results

Performance Metrics and Target

- To evaluate the model performance, we use:
  - **True Positive Rate (TPR, or *Sensitivity*):** percentage of correctly classified true-labeled Signal+Noise events
  - **True Negative Rate(TNR, or *Specificity*):** percentage of correctly classified true-labeled Noise-Only events



$$TPR = \frac{TP}{TP + FN}$$ **as high as possible**

$$TNR = \frac{TN}{TN + FP}$$ **>= 80% [⇒ reduction factor >=5]**

Software: Encoder + SNN Results

Performance Metrics and Target – Early Exit Strategy for Low Latency

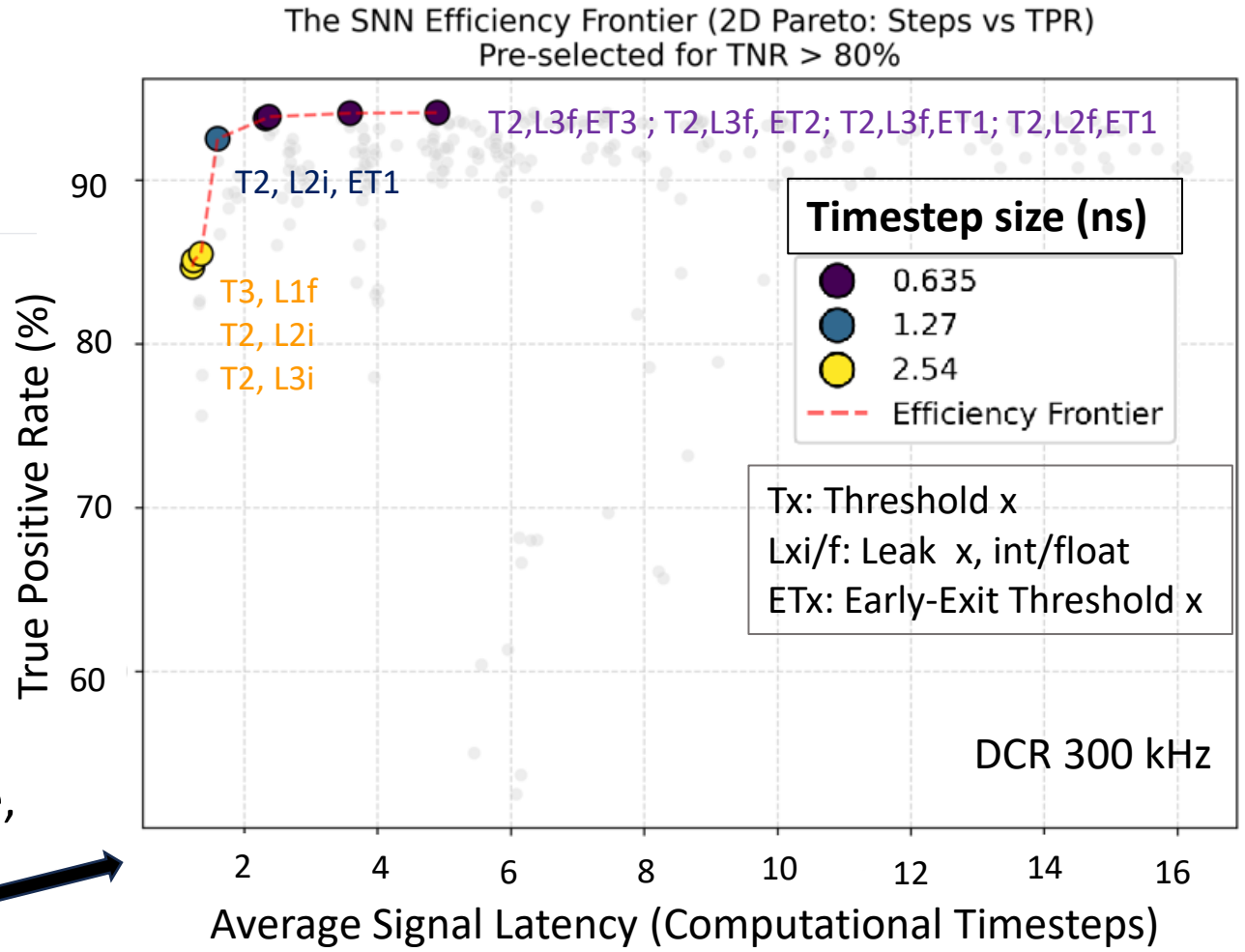
- Average Signal Latency – an hardware oriented metric

  Avg. Nr of Timesteps the SNN needs to be confident that an event is Signal
  - Based on SNN output Sig. neuron activity (configurable Early Exit spikes Threshold)

- Nr of Timesteps is an algorithmic variable
  - Small → Better for hardware throughput

- Physics depends on actual size of timestep
  - Small timestep (e.g. 0.635 ns) → Better TPR

- Select optimal Encoder Configuration (Ts size, threshold, leak) and Early-Exit threshold
  - TPR vs Average Signal Latency



Software: Encoder + SNN Results

Performance Metrics and Target – Early Exit Strategy for Low Latency

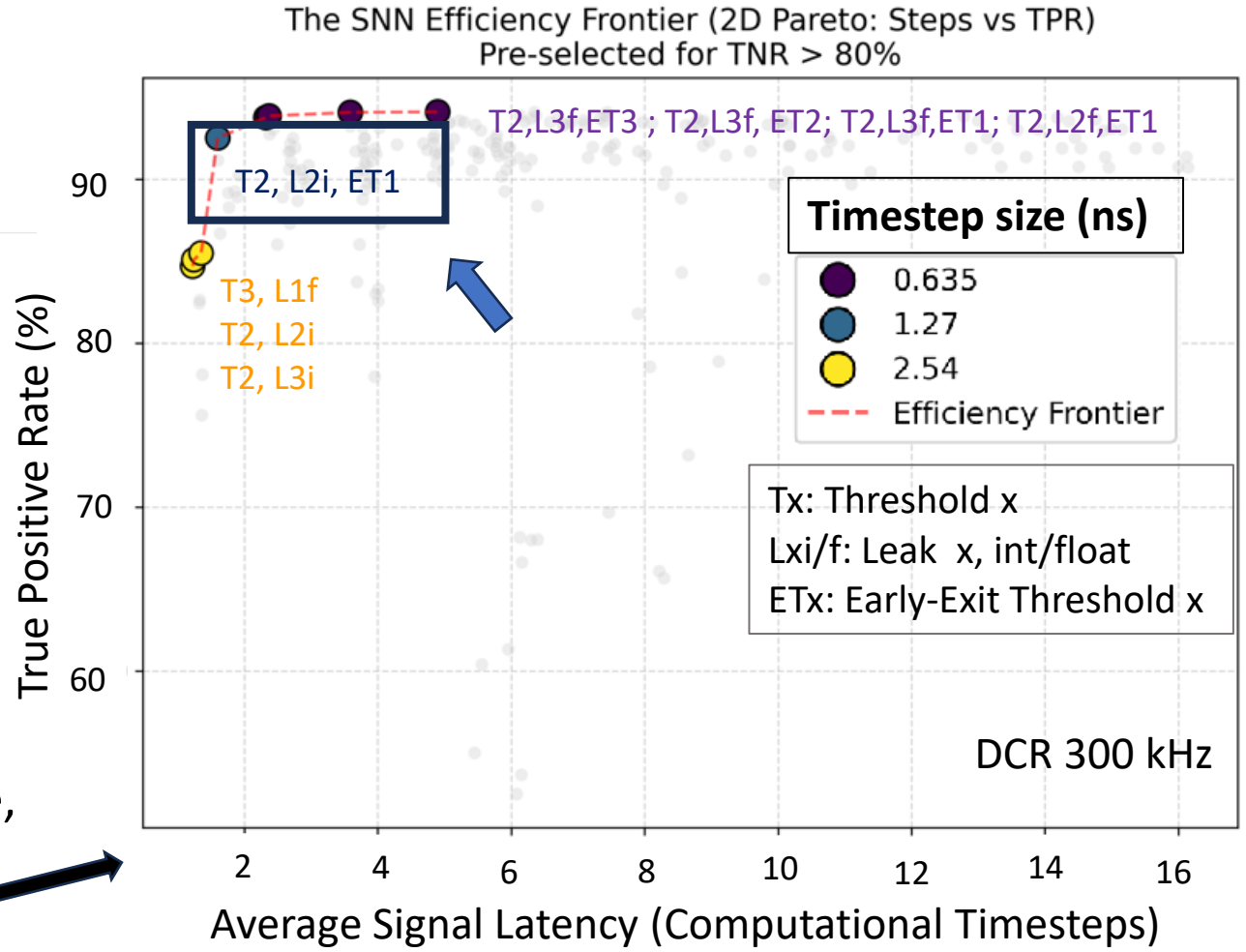
- Average Signal Latency – an hardware oriented metric

  Avg. Nr of Timesteps the SNN needs to be confident that an event is Signal
  - Based on SNN output Sig. neuron activity (configurable Early Exit spikes Threshold)

- Nr of Timesteps is an algorithmic variable
  - Small → Better for hardware throughput

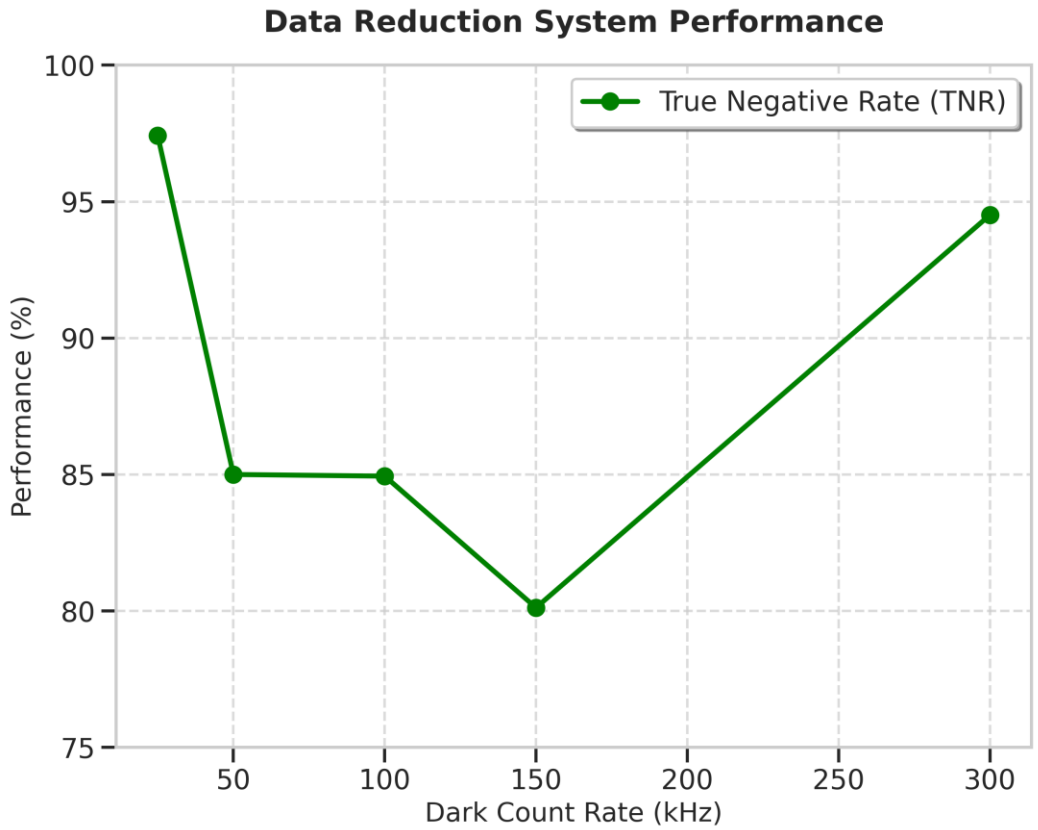
- Physics depends on actual size of timestep
  - Small timestep (e.g. 0.635 ns) → Better TPR

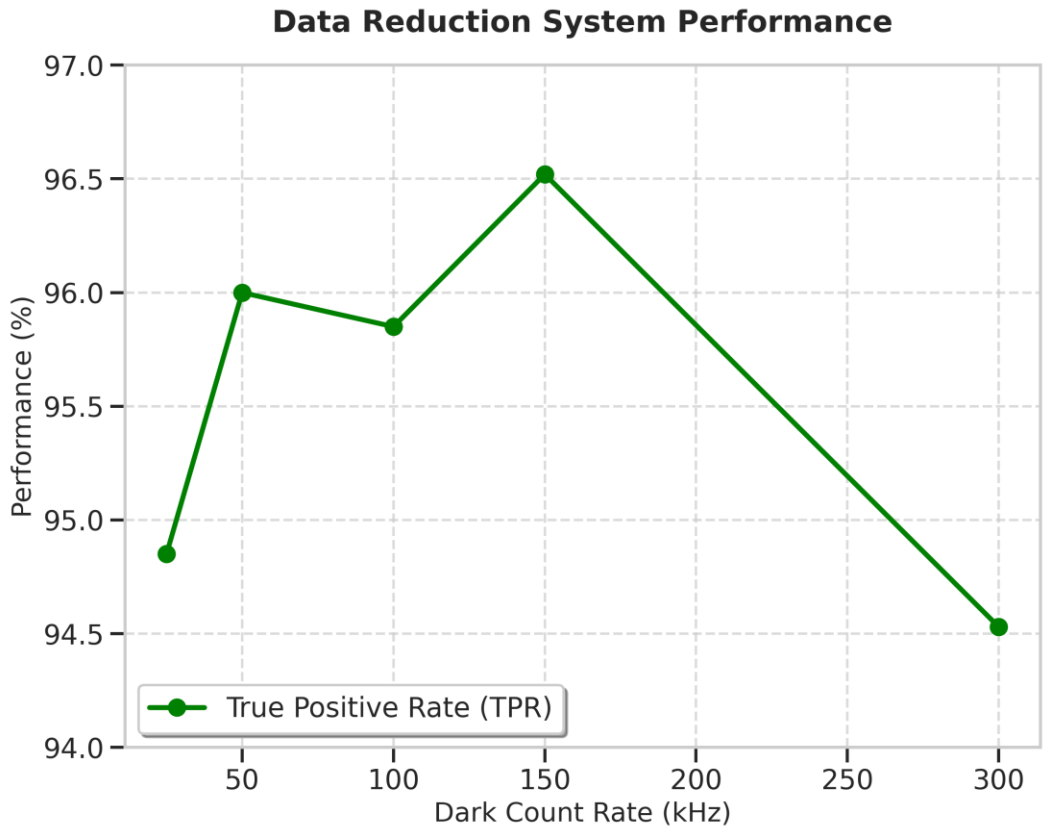
- Select optimal Encoder Configuration (Ts size, threshold, leak) and Early-Exit threshold
  - TPR vs Average Signal Latency



Software: Encoder + SNN Results

Trained on 80k events from ePIC simulation & reconstruction campaigns

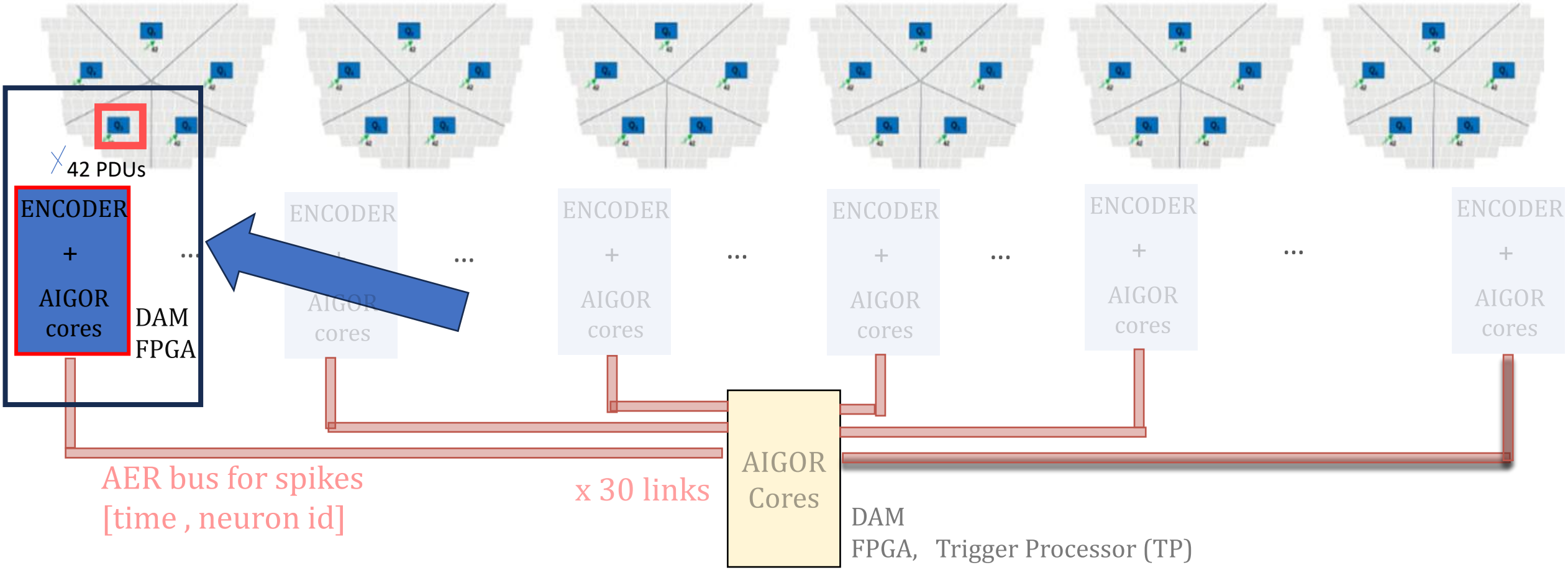




Time Bin 1.27 ns , Encoder Threshold: 2, Leak shift 1 int, No Early exit applied

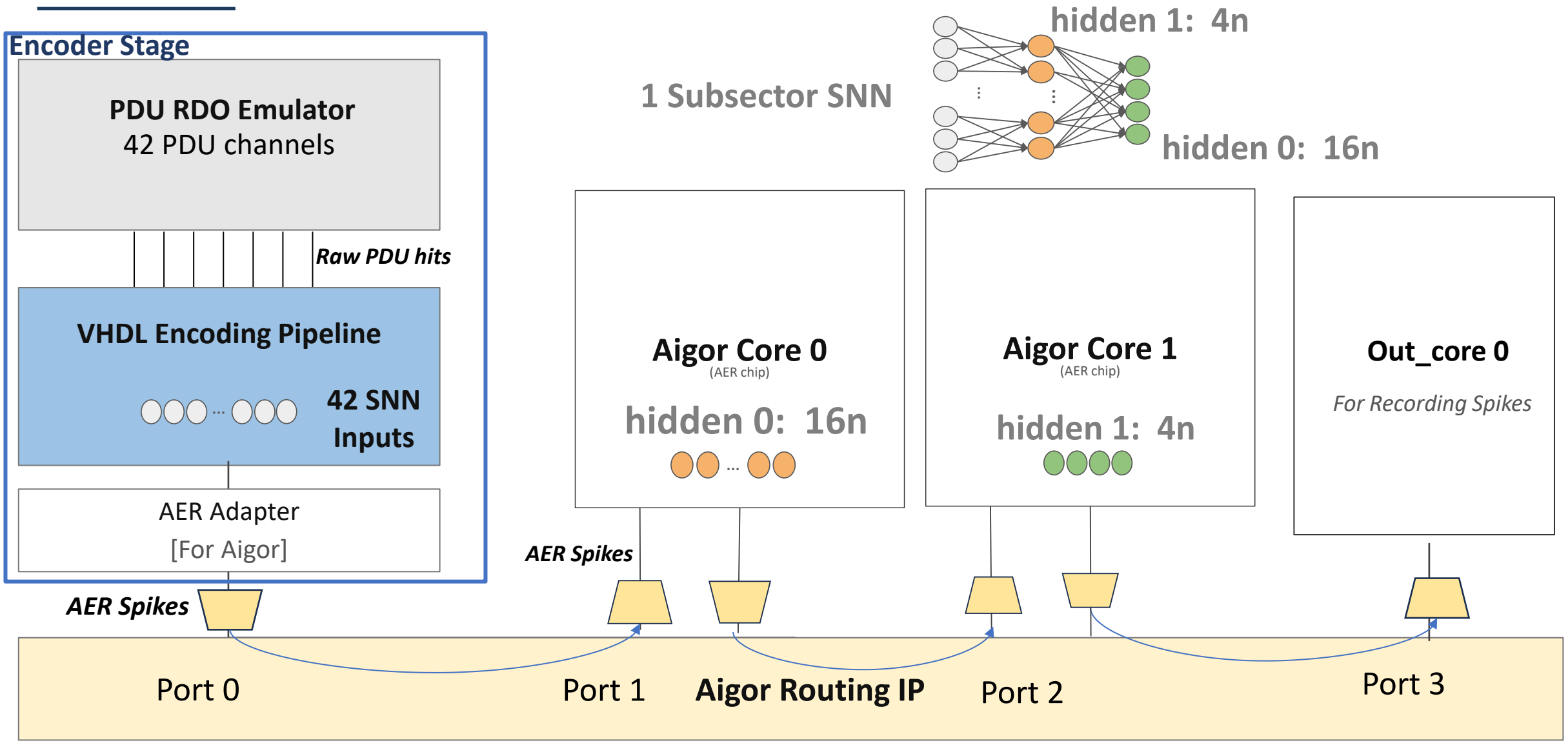
With Early exit Threshold 1 → *lose ~ 3 % TPR/TNR performance*, Avg sig Latency: *2 Timesteps*, at 99th perct.: *5 Timesteps*

Preliminary Hardware Testbed: Single Subsector

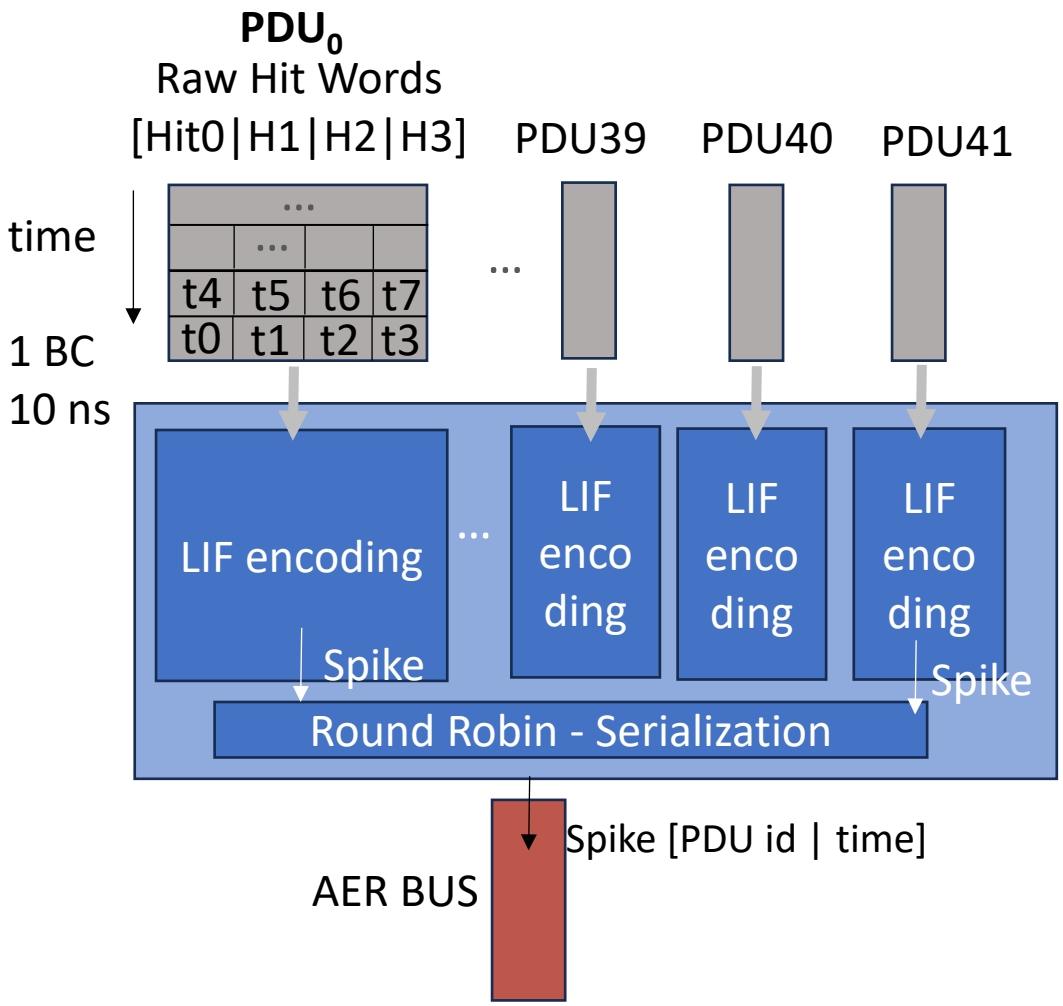


- **AIGOR: multi-core Neuromorphic Computing Architecture prototype → deploy SNN layers.**
- Subsector DAM FPGA: LIF Encoder for RDO hits + AIGOR Cores for Subsector SNN layers.
- Aggregation DAM FPGA: AIGOR Cores for Aggregation SNN Layers.

Preliminary Hardware Testbed: Single Subsector



The Throughput Challenge & Preliminary Results



Encoder Optimization Strategies

- Only One Spike allowed per PDU per BC (Time To First Spike Encoder), flush the remaining BC hits
  - Allows unfolding a word of hits - multiple timesteps- through the LIF integration Chain in a single Clock Cycle and selecting the only spike to emit in the same Clock Cycle.
- Hardware optimized leaky operation (fast bit-shifts).
  Use Int or Fixed-point membrane potential V

$$V_{leak} = V - (V >> k) = V\,(1 - 2^{-k}) \qquad k = leak_shift$$

- **The Serialization Stage is the bottleneck!** But it is event dependent → Few hits ≈ 1.7 per PDU in the BC → no Spike.

Preliminary throughput measurement Encoder-only

- Synthetic dataset reproducing Signal+Noise distribution @300 kHz DCR, ~ 1.7 hit per PDU per BC

≈ 1-10 MHz

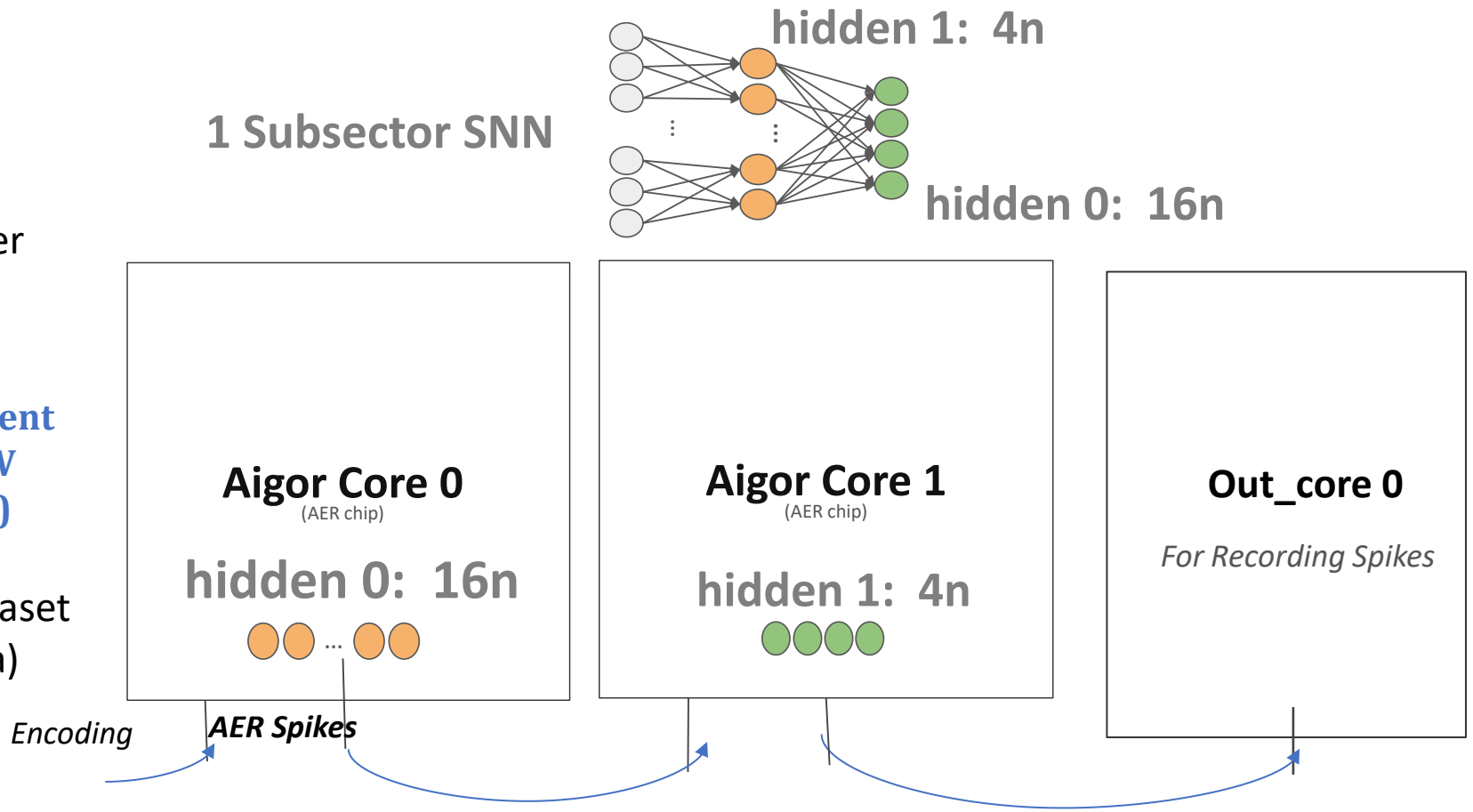
Future optimizations: higher clock frequency (>100 MHz)
Early exit → flush hits after few Timesteps (e.g. 5).
Real dataset → throughput is lower ~1/200 BCs is Only-Noise → less hits

The Throughput Challenge & Preliminary Results

AIGOR Optimization Strategies

- AIGOR Cores deployed in a feedforward pipeline → One Layer per Core → minimize recurrent communication on the same bus.

Preliminary Throughput Measurement
Encoder + Subsector SNN in Vitis HW Emulation (Cycle-accurate RTL emu)
≈ 1.5 MHz
with a semi-realistic throughput dataset with Signal every 100 BCs (ePIC data)



Lots of optimization potential

- Multiple spikes AER packet → but it requires many parallel accesses to the synaptic memory → trade-off between parallelization and sparse event-based efficiency
- Remove barrier messages between timestamps → feasible in feedforward mode, without losing determinism
- Optimize SNN → SNNs are efficient. Even smaller network might work for this task
- Raise Clock Frequency > 100 MHz
- Early exit strategy

Summary & Outlook

Shown

- LIF encoder exploits temporal structure to reduce dRICH hit rates > 90%.
- Encoded spikes → SNN reaches required signal / noise discrimination.
- Preliminary deployment on a configurable neuromorphic processor AIGOR, with a feedforward streaming mode under development.

In progress

- Full hardware throughput characterization.
- Energy efficiency characterization.
- Early-exit and other optimizations.
- Integration with the ePIC readout chain (30-FPGA distributed pipeline).

Open questions

- Can AER-based processing reach the needed throughput at 100 MHz, even with all optimizations?
- Deployment closer to the source of data is ideal. Would in-ASIC near-sensor encoding be feasible and benificial?

Extra

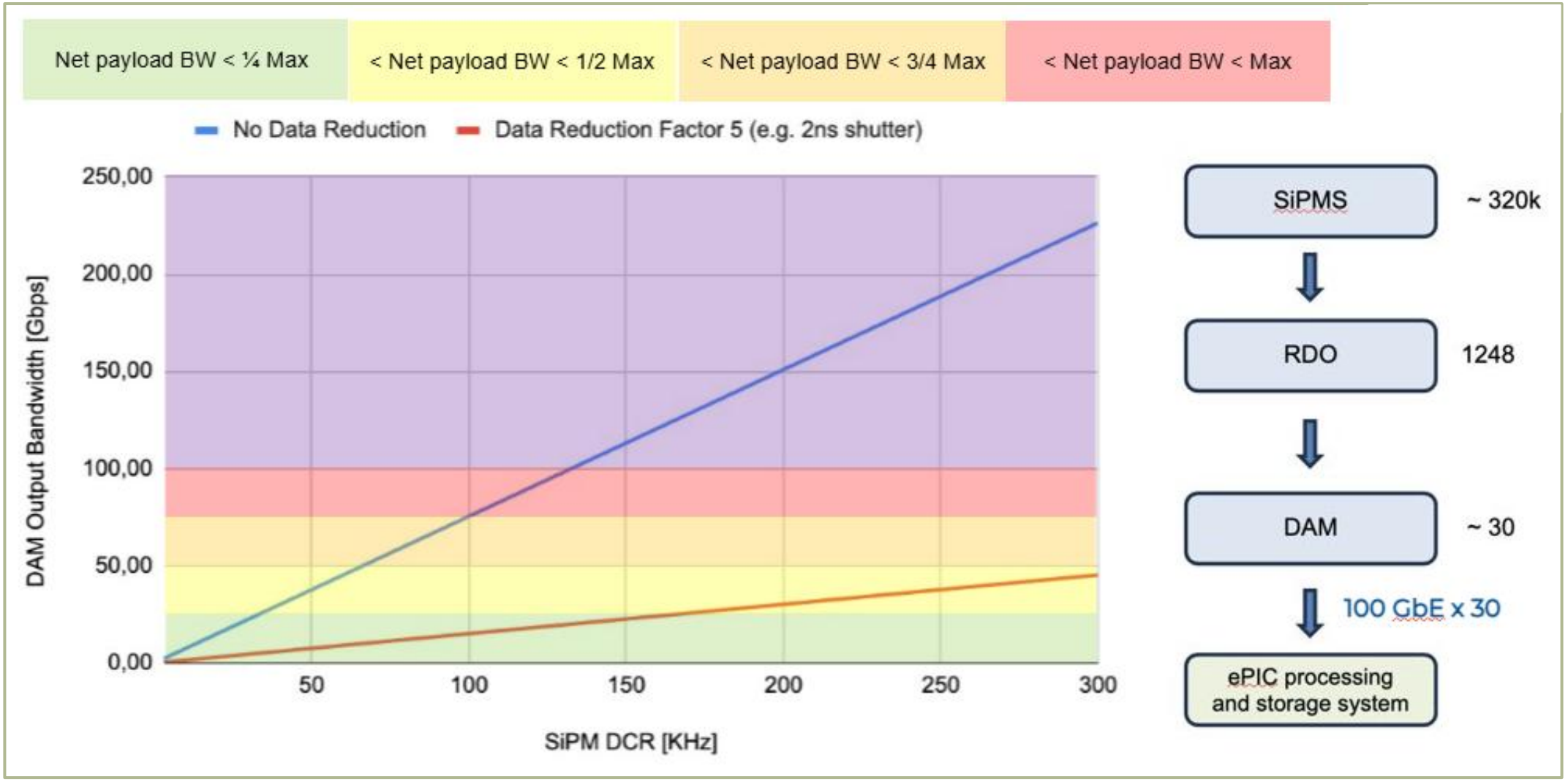
Data reduction with ML Classifier

Online **Signal/Noise discrimination** using ML

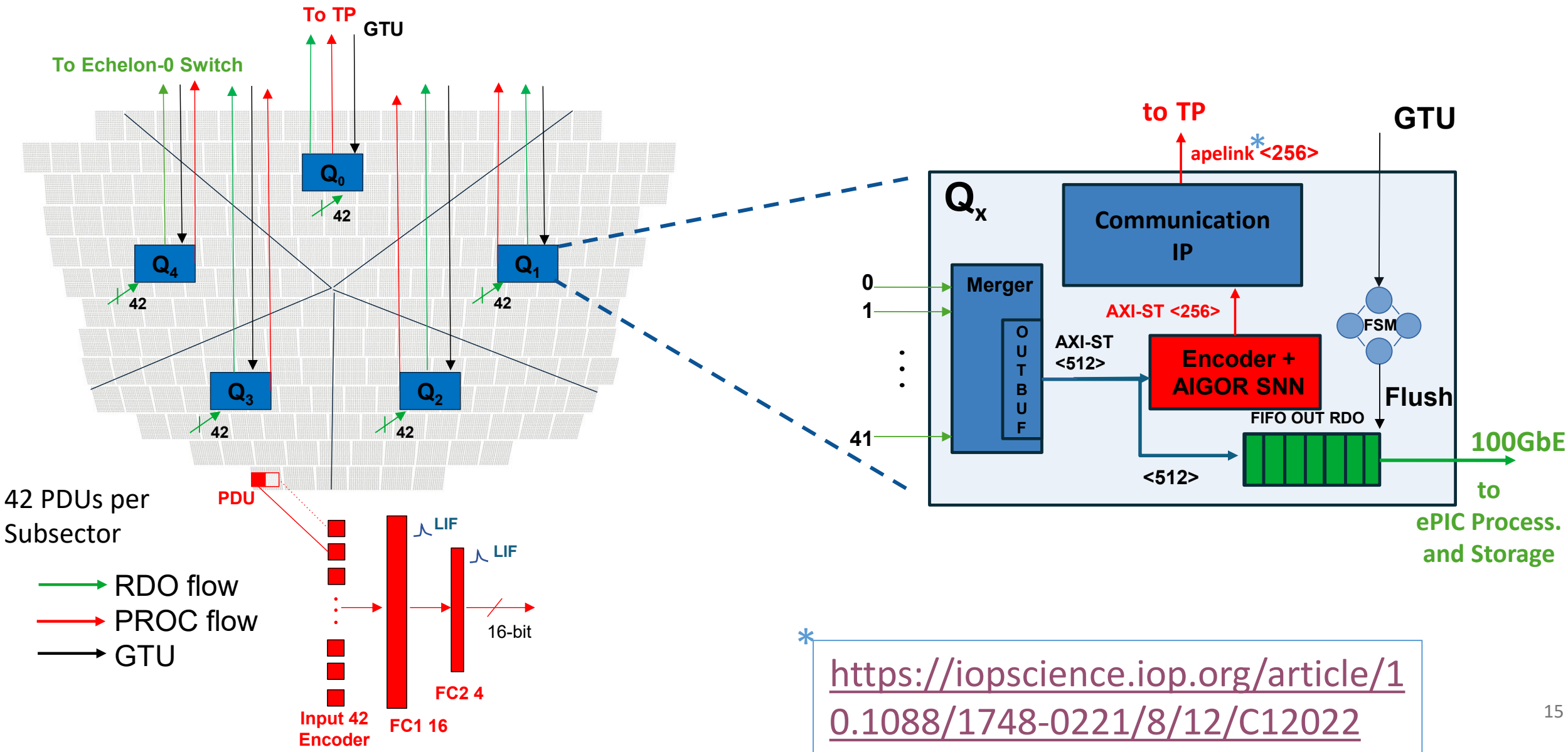
- **Signal**
  (i.e Merged Phys Signal+ Bkg)
  - **Physics Signal:**
    - ➔ e.g. DIS, SIDIS, ...
  - **Physics Background:**
    - ➔ e/p interaction with beam pipe
    - ➔ synchrotron radiation
- **SiPM Noise:**
  - Dark Count Rate (DCR) modeled on the reconstructed dataset

Discriminate between **Noise-Only** and **Signal+Noise** events

dRICH: Output Bandwidth

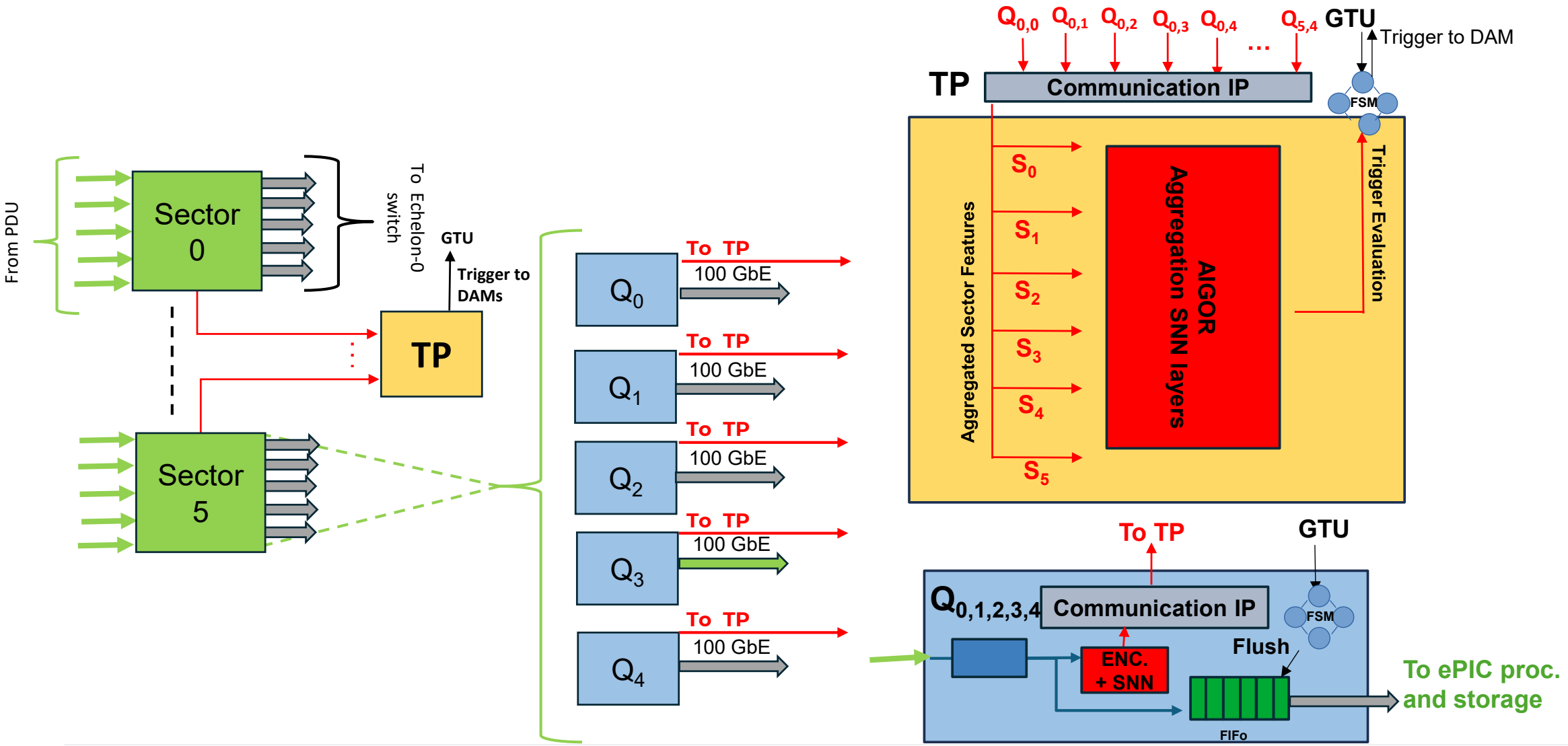


Detailed View: Design of data-reduction stage on multi-FPGAs



* https://iopscience.iop.org/article/10.1088/1748-0221/8/12/C12022

Detailed View: Design of data-reduction stage on multi-FPGAs



- **FPGA Occupancy** single AIGOR-Core

  (on Versal 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

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