European Strategy Update

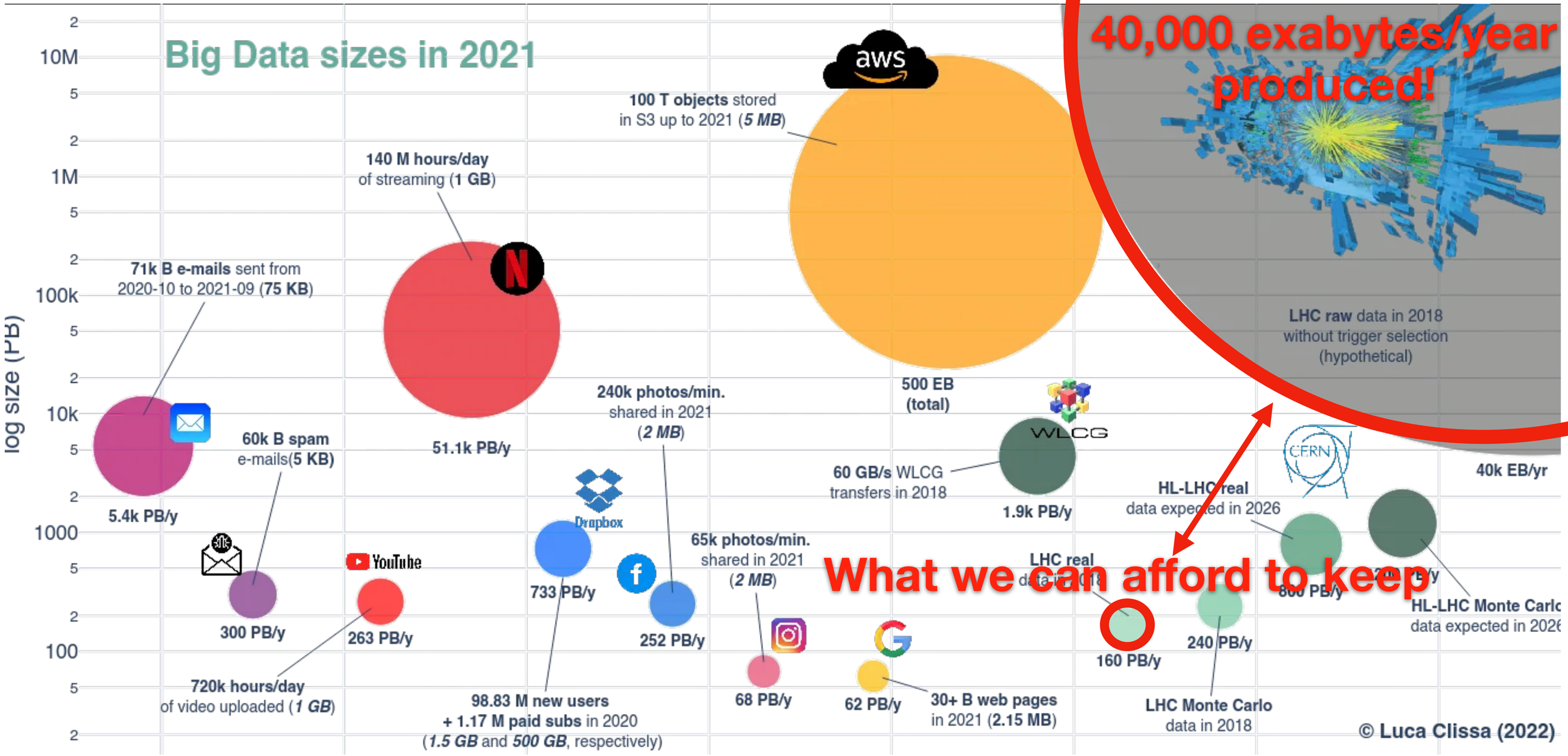
ETH zürich

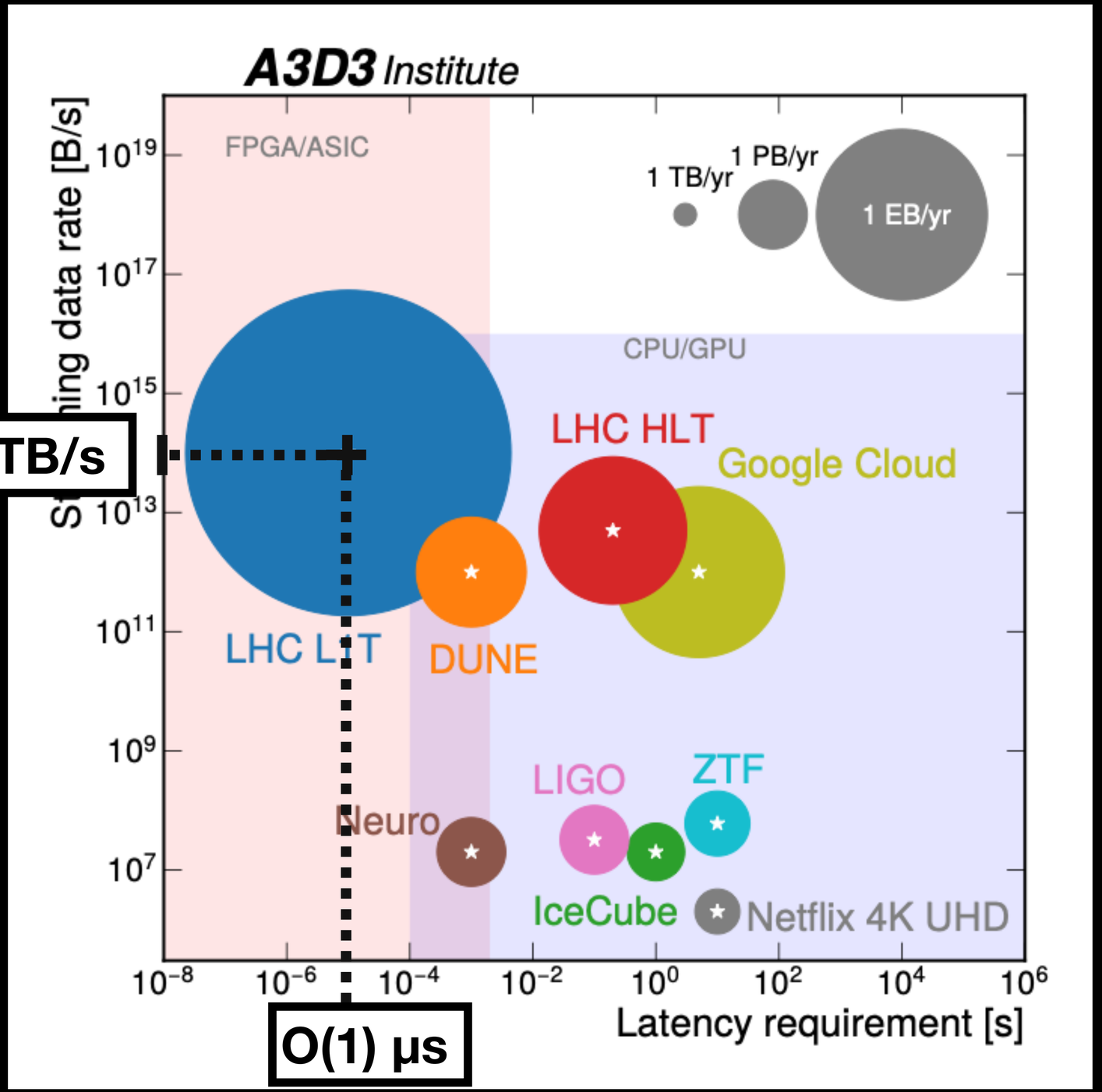
# AI for fast inference in real-time systems: status and future perspectives

## 2026 Update: European Strategy for Particle Physics

Jun 23–27 2025, Venice Lido

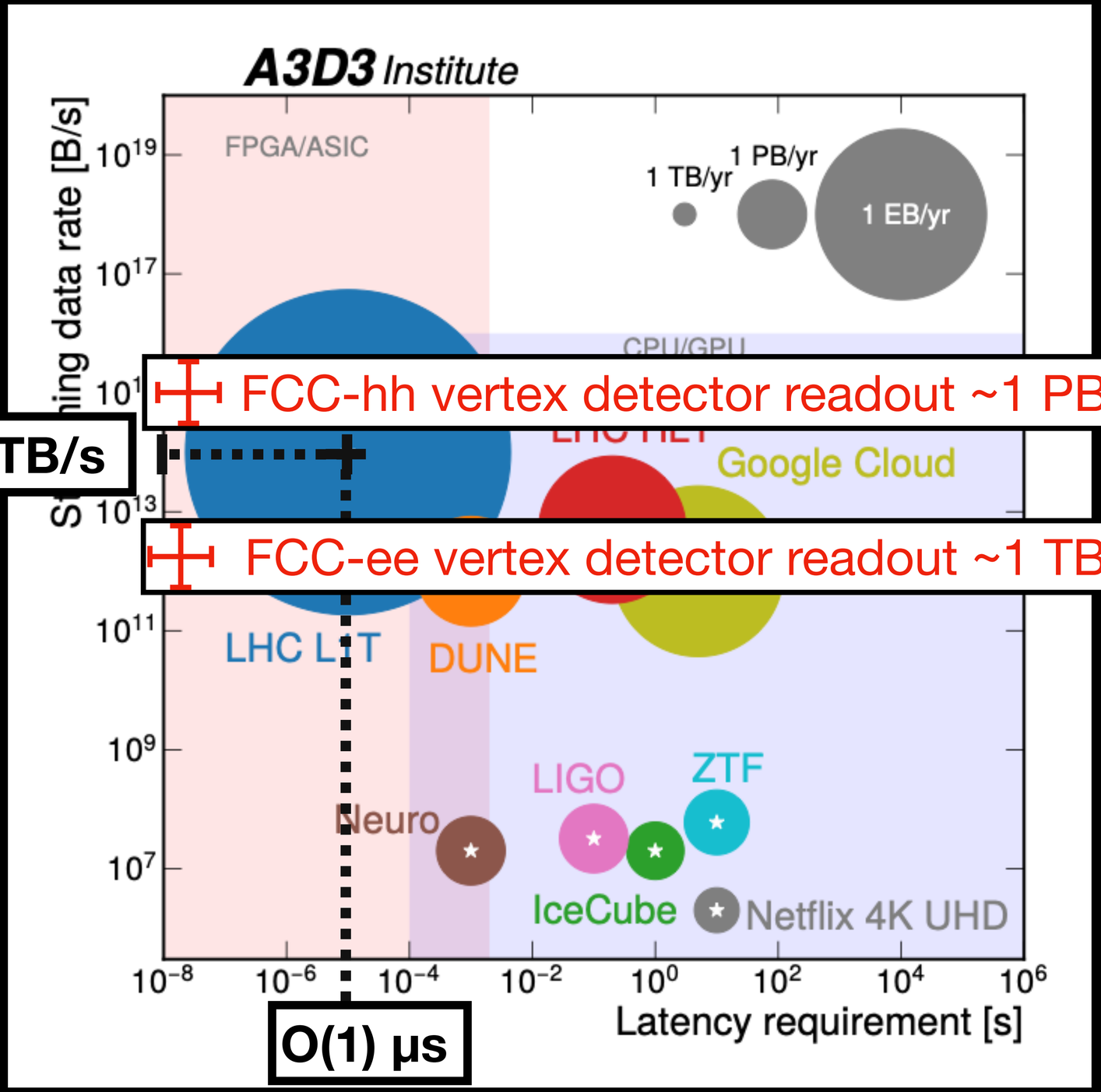
Thea Klæboe Årrestad (ETH Zürich)


Big Data sizes in 2021
log size (PB)
2
10M
5
2
1M
5
2
100k
5
2
10k
5
2
1000
5
2
100
5
2
aws
100 T objects stored in S3 up to 2021 (5 MB)
500 EB (total)
140 M hours/day of streaming (1 GB)
51.1k PB/y
71k B e-mails sent from 2020-10 to 2021-09 (75 KB)
5.4k PB/y
60k B spam e-mails(5 KB)
300 PB/y
720k hours/day of video uploaded (1 GB)
263 PB/y
240k photos/min. shared in 2021 (2 MB)
Dropbox
733 PB/y
252 PB/y
65k photos/min. shared in 2021 (2 MB)
68 PB/y
98.83 M new users + 1.17 M paid subs in 2020 (1.5 GB and 500 GB, respectively)
62 PB/y
30+ B web pages in 2021 (2.15 MB)
WLCG
60 GB/s WLCG transfers in 2018
1.9k PB/y
40,000 exabytes/year produced!
LHC raw data in 2018 without trigger selection (hypothetical)
40k EB/yr
What we can afford to keep
CERN
HL-LHC real data expected in 2026
800 PB/y
HL-LHC Monte Carlo data expected in 2026
LHC real data in 2018
160 PB/y
240 PB/y
LHC Monte Carlo data in 2018
© Luca Clissa (2022)




How much data goes through system

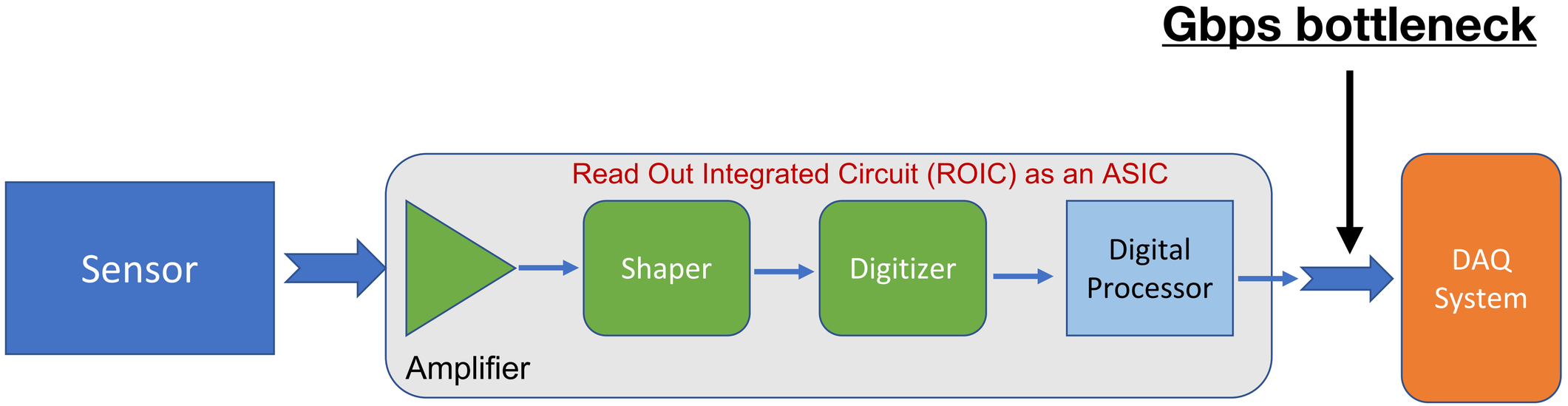
Time to process


A3D3 Institute
FPGA/ASIC
CPU/GPU
1 TB/yr
1 PB/yr
1 EB/yr
Google Cloud
LHC L1T
DUNE
LIGO
ZTF
Neuro
IceCube
Netflix 4K UHD
FCC-hh vertex detector readout ~1 PB/s
FCC-ee vertex detector readout ~1 TB/s
O(100) TB/s
O(1) µs
How much data goes through system
Time to process
Latency requirement [s]

# This readout technology doesn’t (yet) exist!

# Ultimate precision at future colliders could be bottlenecked by our data acquisition systems

# Data readout bottleneck



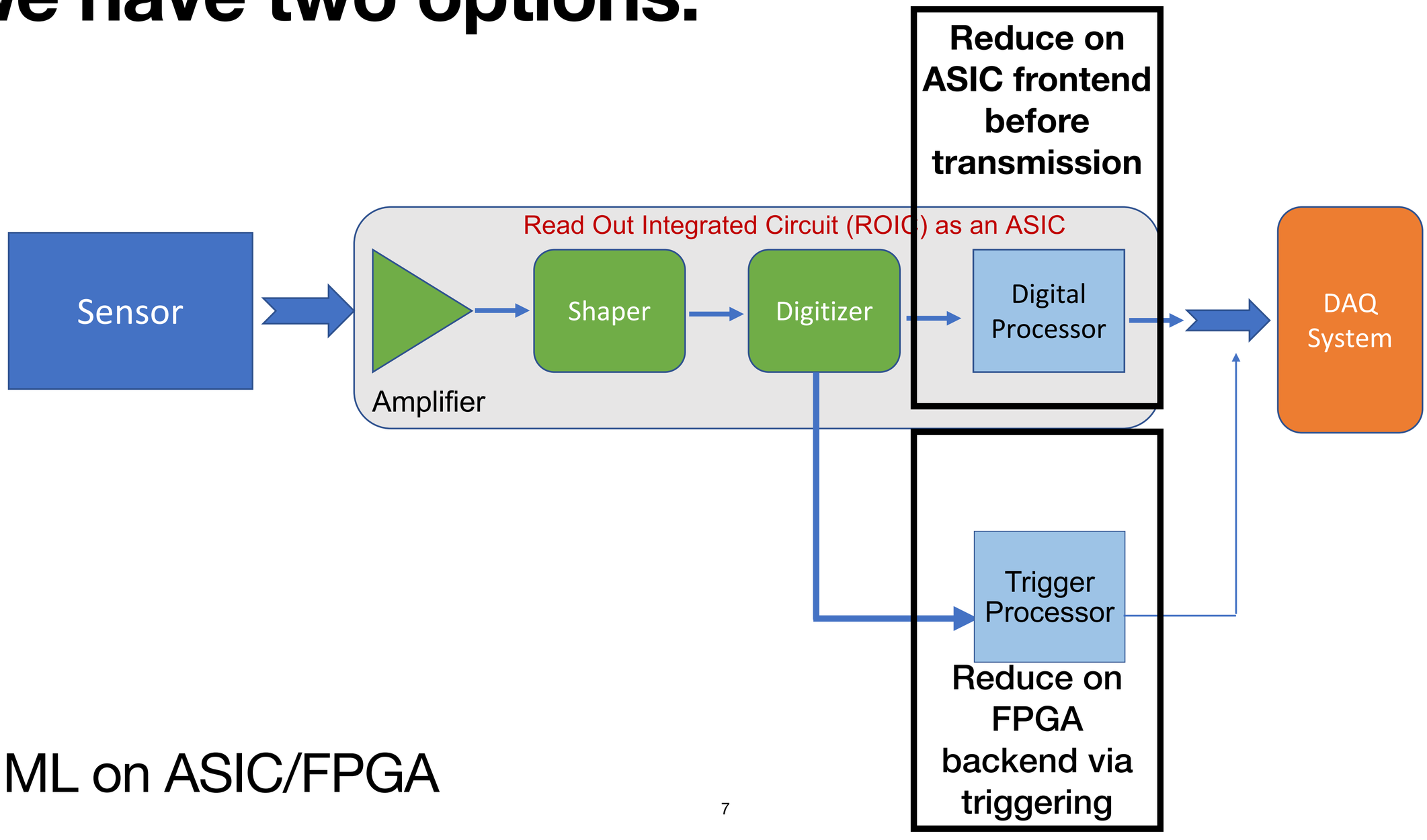

Example vertex detector inner layer from before (#241 FCC-ee):

32 bits pixel data in inner layer @ 200 MHz/cm$^2$

→ **24.4 Gbit/s**

**Current technology 3.2 - 6.4 Gbit/s**

# We have two options:

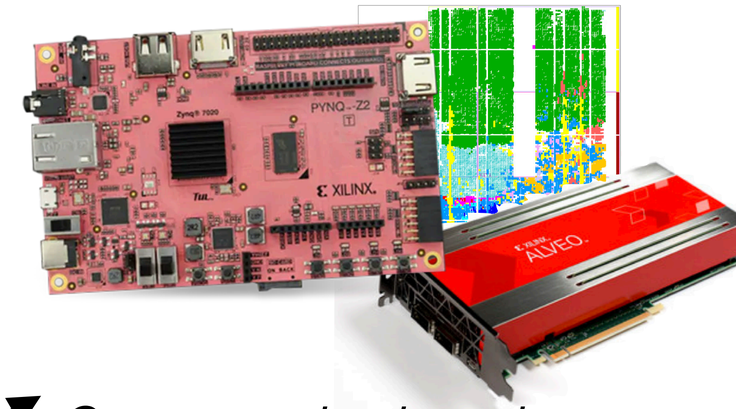



→ML on ASIC/FPGA

#67 #124

# ML inference at low latency

## HEP Tools and Communities




*Co-processing kernel*
*(Xilinx accelerators/SoCs)*

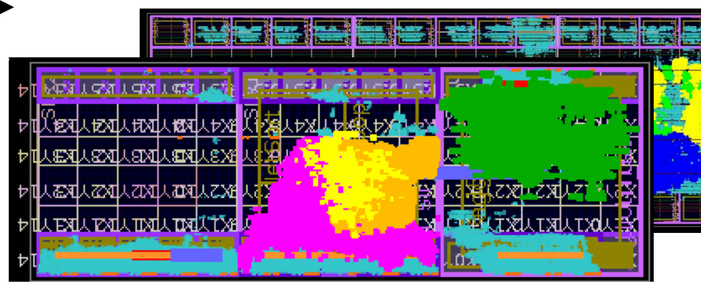
*Model*
*(quantized/pruned)*

***HEP hardware ML libraries:***

hls4ml

Conifer

*FPGA custom designs*
*(eg trigger algorithms)*



***HEP quantization libraries:***

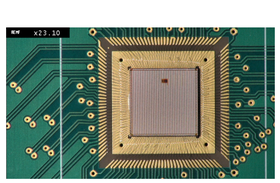
QONNX

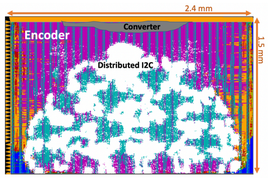
HGQ

QKERAS

( Collaboration with AMD:
Brevitas for PyTorch)

*ASICs*



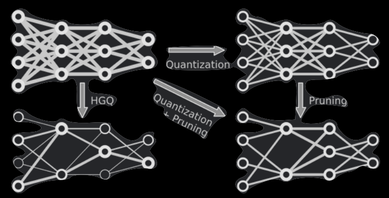


*Frontend (QAT)*

*Backend (compiler)*

QKeras

PyTorch Brevitas




HQG

QONNX

FINN AMD

hls4ml

Catapult HLS SIEMENS

Vivado™ HLS

HLS COMPILER

XILINX VITIS™

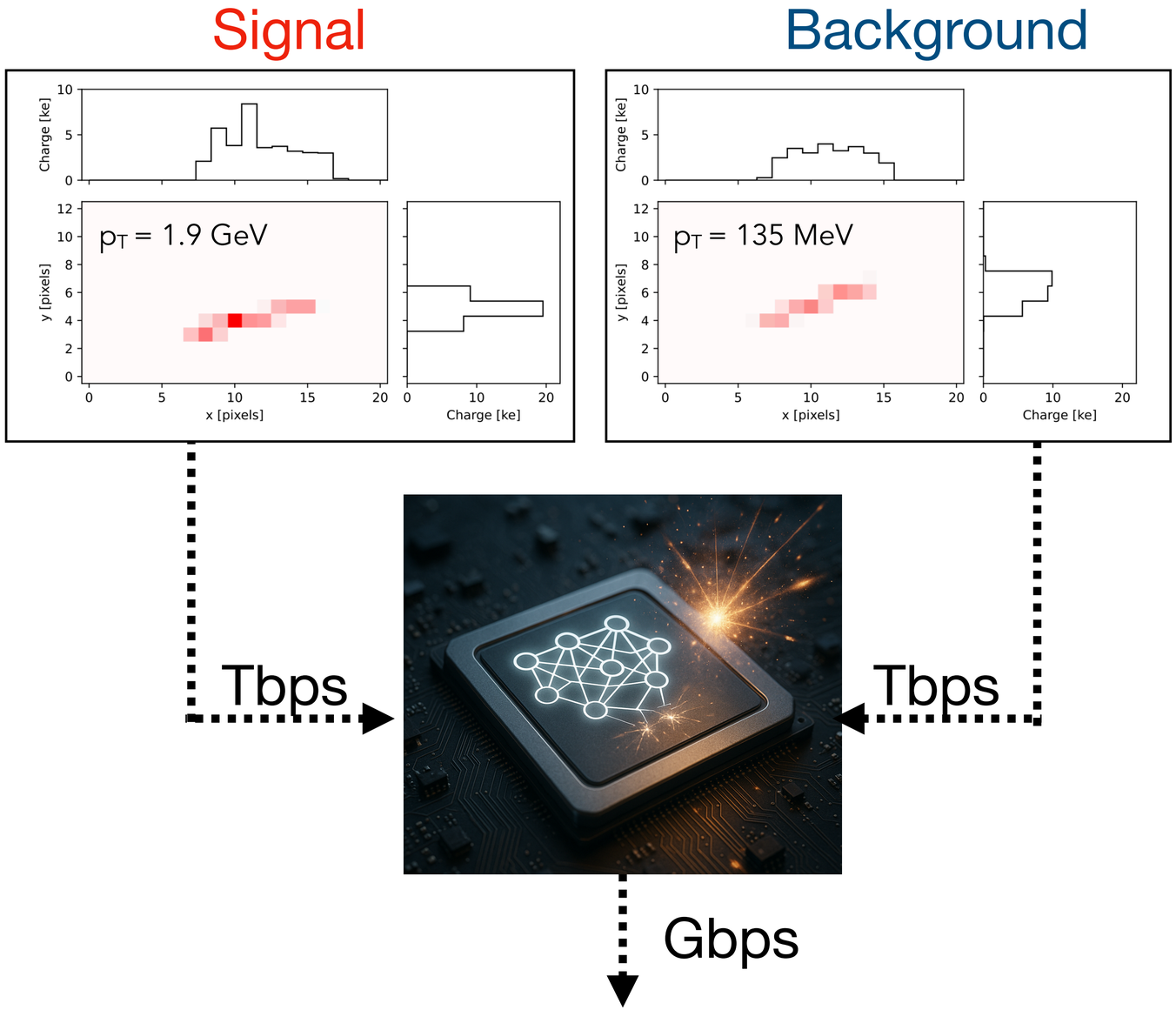
# Intelligent frontends

## Reduce on-detector

- **Subtle data patterns within single (pixel) layer! ML on ASIC** on-detector
  - filtering, and/or
  - featurizing
- **Challenges**

| Metric | Simulation | Target |
|---|---|---|
| Power | 48 mW | <100 mW |
| Energy / inference | 1.2 nJ | N/A |
| Area | 2.88 $mm^2$ | <4 $mm^2$ |
| Gates | 780k | N/A |
| Latency | 50 ns | <100 ns |

  - Radiation hardness



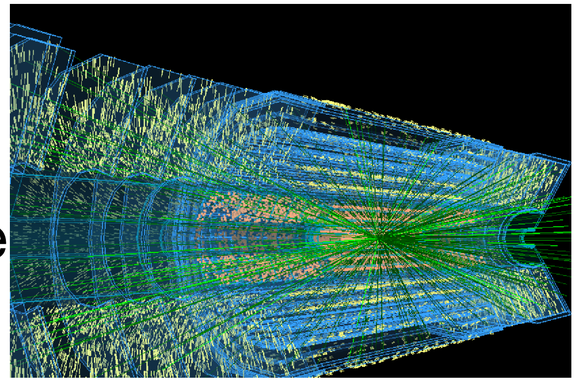
#11 #95 #247 #233 #93 #211 #272

# Smart Pixels

## Pixel readout ASIC with ML

- **frontend filtering**:
  discard low-$p_T$ tracks

- **feature extraction**:
  particle position+angle
  Mixture Density Model

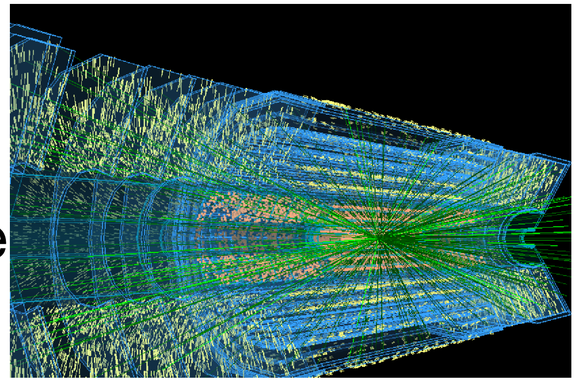


#11 #95 #247 #233 #93 #211 #272

# Smart Pixels

**Pixel readout ASIC with ML**

- **frontend filtering**: discard low-$p_T$ tracks
- **feature extraction**: particle position+angle Mixture Density Model

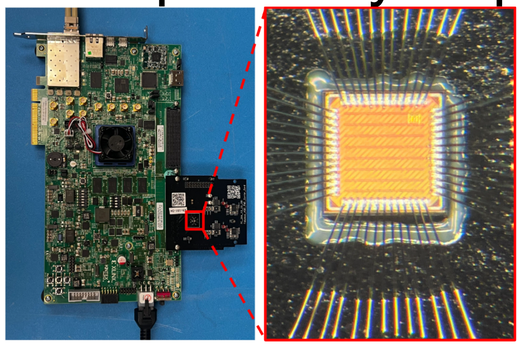


#95

# eFPGAs

Fully **reconfigurable logic** in ASIC design

- The pathway to put ML on-detector!



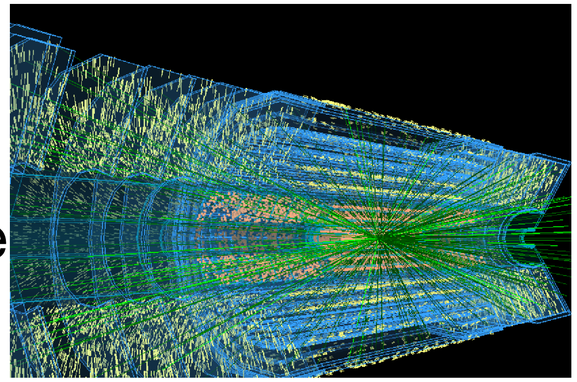
BDT classifier in 28nm CMOS ASIC

#11 #95 #247 #233 #93 #211 #272

# Smart Pixels

**Pixel readout ASIC with ML**

- **frontend filtering**: discard low-$p_T$ tracks
- **feature extraction**: particle position+angle Mixture Density Model

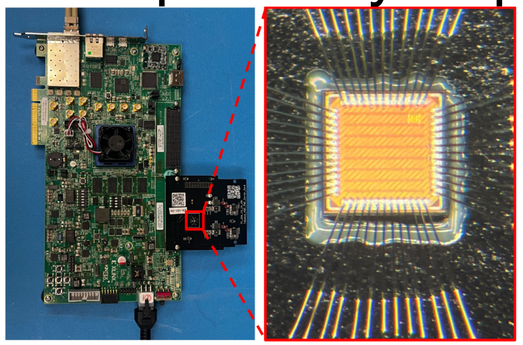


#95

# eFPGAs

Fully **reconfigurable logic** in ASIC design

- The pathway to put ML on-detector!

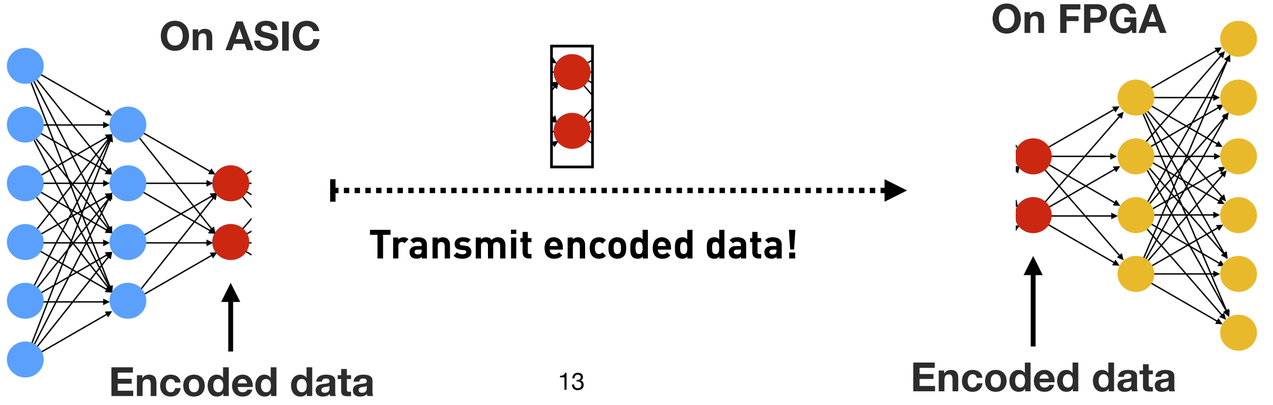


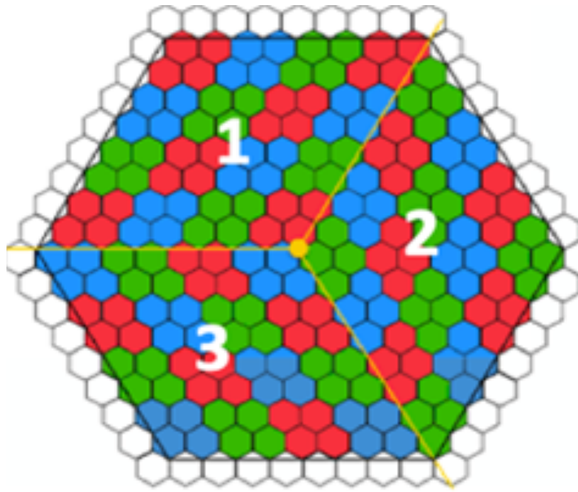
BDT classifier in 28nm CMOS ASIC

# Calorimeter data concentrator

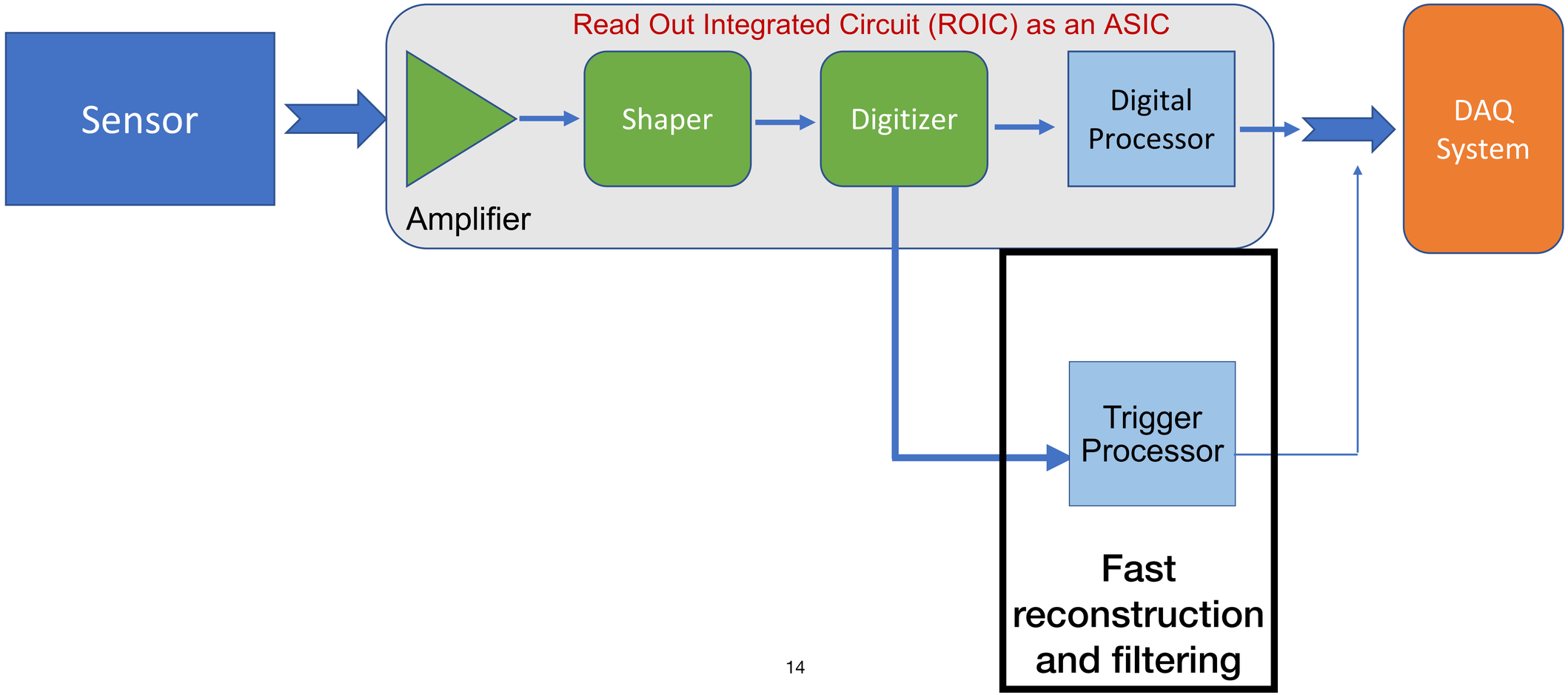
#11


10k ECONS with ML inside **in CMS HL-LHC**



# Intelligent back-ends



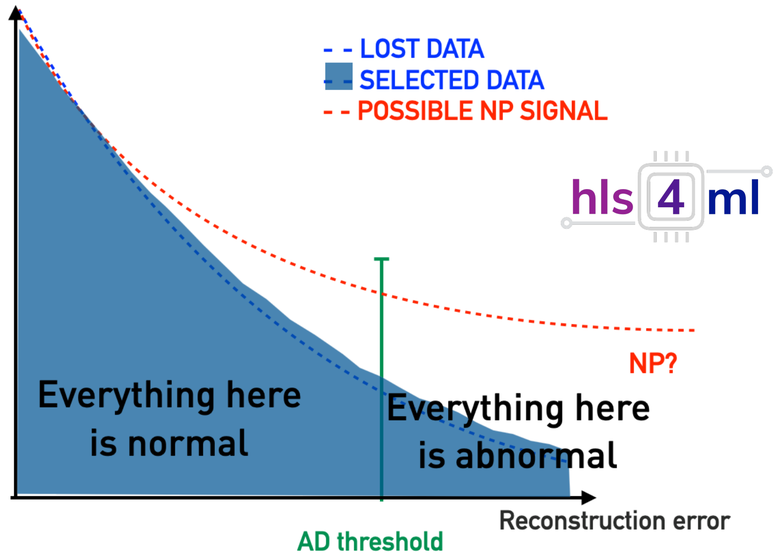
# Intelligent backends

## 2024: Neural hardware triggers making decisions in LHC experiments!

- Trigger-less for future $e^+e^-$ not guaranteed
- FPGA-based inference for improved triggering
  - in L1 trigger
  - as accelerators in HLT
- Possible through HEP experiment-agnostic tools!

**CMS:**

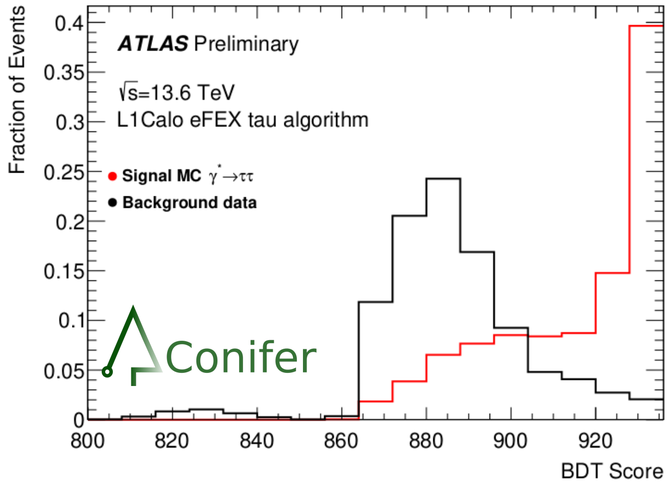
- Anomaly detection in 50 ns
- 300 events/second




CMS DP2023_079

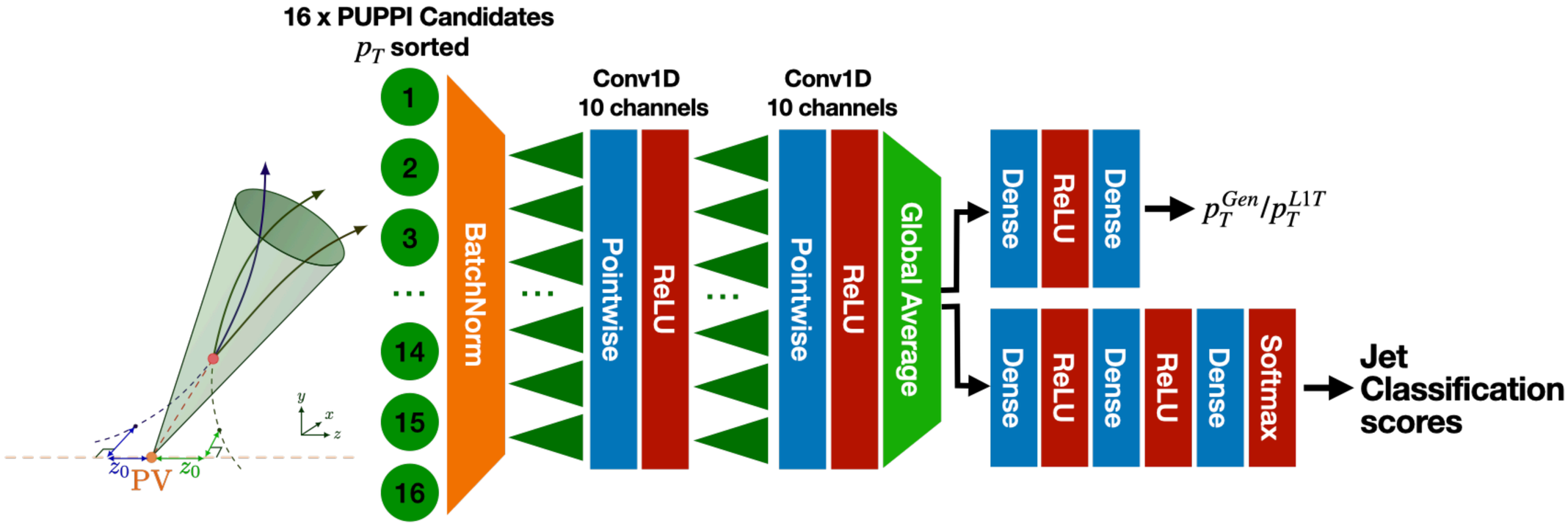
**ATLAS:**

- BDT selecting candidate $\tau$ lepton events in <100 ns




L1CaloTriggerPublicResults

# ML inference on FPGA

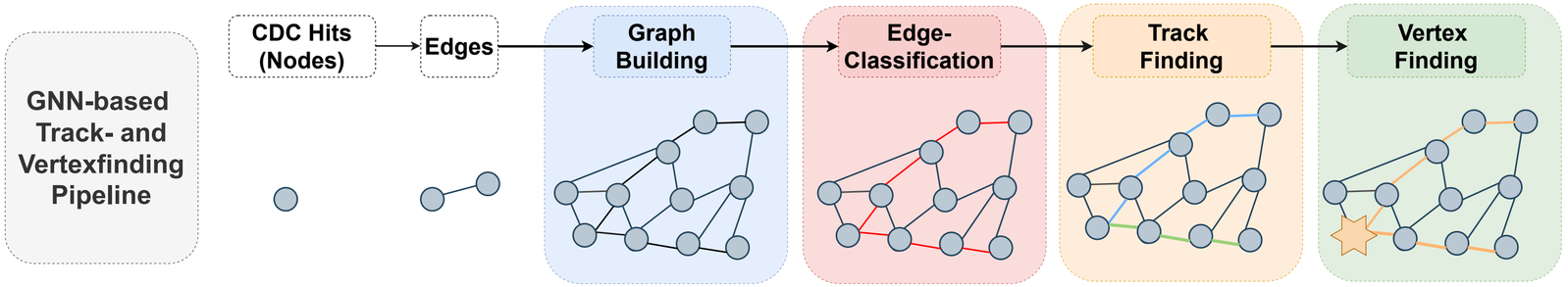


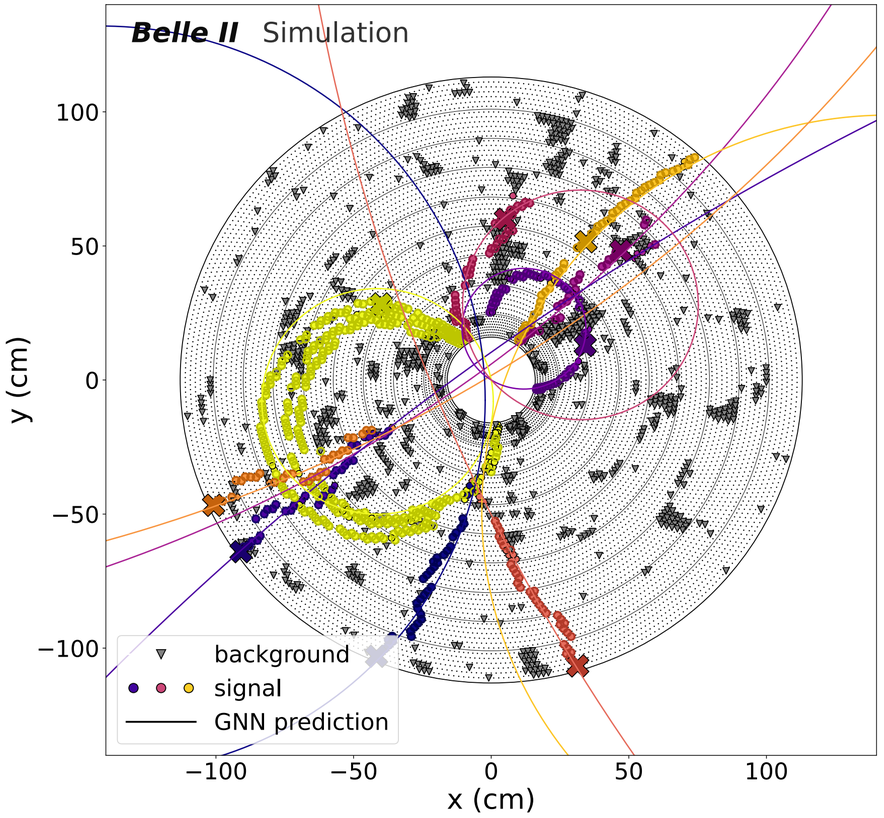

Lower-level information like tracks and particles in hardware triggers has lead to increased usage of set-based and graph-based architectures, like 250 ns DeepSet flavour tagging!

arxiv:411.13596

# Real-time tracking

## BELLE-II



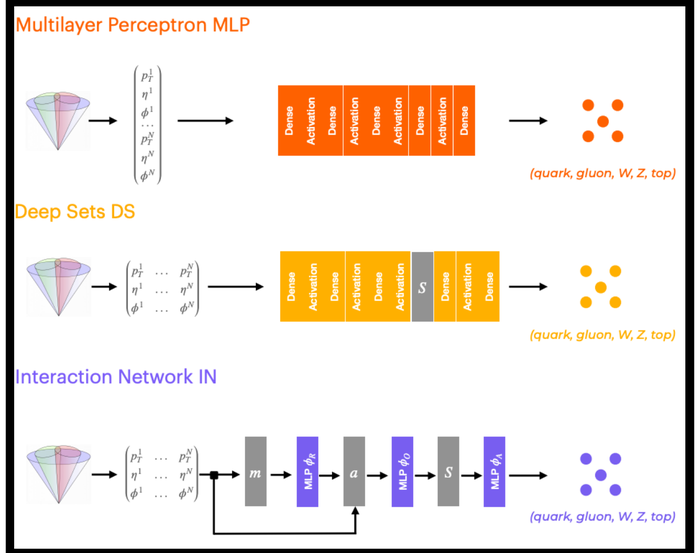




- Belle-2 uses ML for GNN-based offline tracking!
  - Work ongoing on bringing this to the hardware trigger
  - Graph building non-trivial (arxiv:2307.0728)
- Already use neural track hardware trigger for vertex-reconstruction, and azimuthal and polar angles of single particle track

**Well-established:**
MLPs, CNNs, DeepSets, GNNs, RNNs, symbolic regression, (variational) autoencoders, BDTs, IsolationTrees

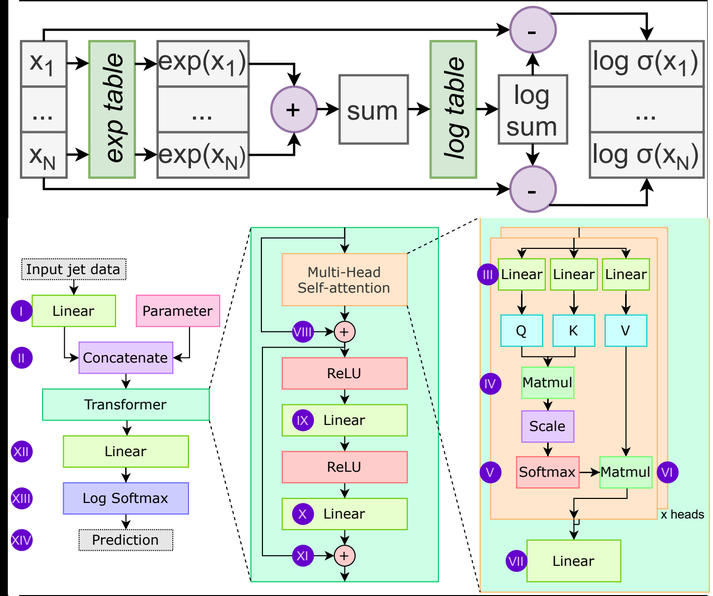
**Experimental:**
Transformers, large distributed CNNs (ResNet, VGG)

## Quantised Interaction Networks and Deep Sets in <160 ns




P. Odagiu et al. 2024

## Fully on-chip transformers in 90 ns (500k flops, 9k param)




IEEE ICFTP 2022
arxiv:409.05207

# ESPP proposal: AI R&D collaborations

**EuCAIF proposal for scalable, robust AI through cross-domain collaboration**

- **AI for Detector and Accelerator Control:** Accelerator performance, calibration, system monitoring.
- **AI for Detector Optimization:** Differentiable programming, reinforcement learning to maximize detector performance
- **AI for Event Reconstruction:** Tracking, calorimetry, end-to-end foundation models.
- **AI for Data Processing: Front-end electronics, trigger**

in close collaboration with the Fast Machine Learning Lab , CERN NGT, DRD-7 and EuCAIF AI-RDs

# Conclusion

- ML is essential to address unique data and processing challenges in HEP
  - Custom workflows and tools developed for extreme constraints
- AI/ML is ubiquitous in upcoming and future high-luminosity experiments
  - Intelligent processing near sensors needed to manage data from granular detectors
- Cross-experiment collaboration (e.g. DRDs, AI-RDs) is key to future success!