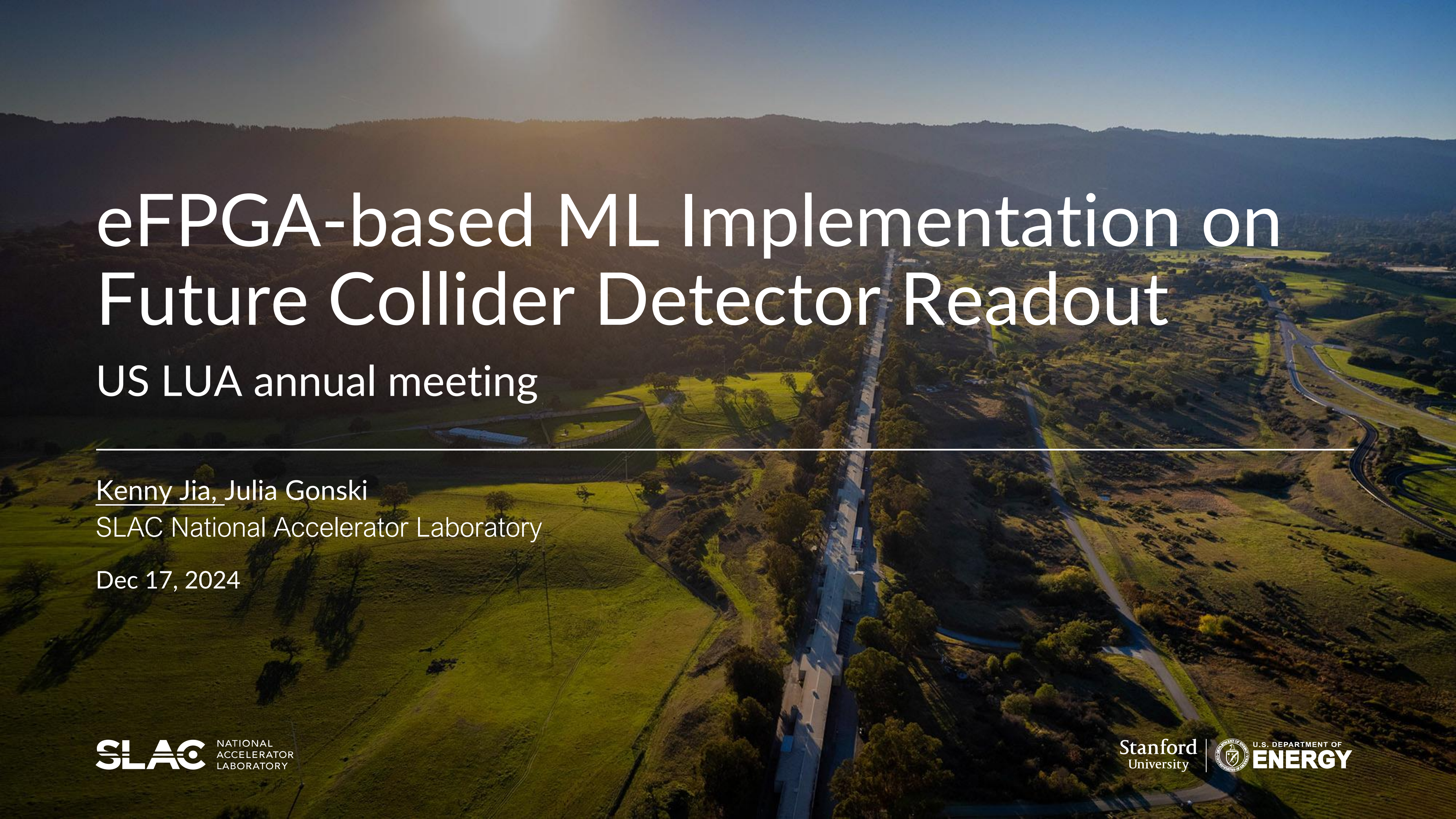


eFPGA-based ML Implementation on Future Collider Detector Readout

US LUA annual meeting

Kenny Jia, Julia Gonski

SLAC National Accelerator Laboratory

Dec 17, 2024

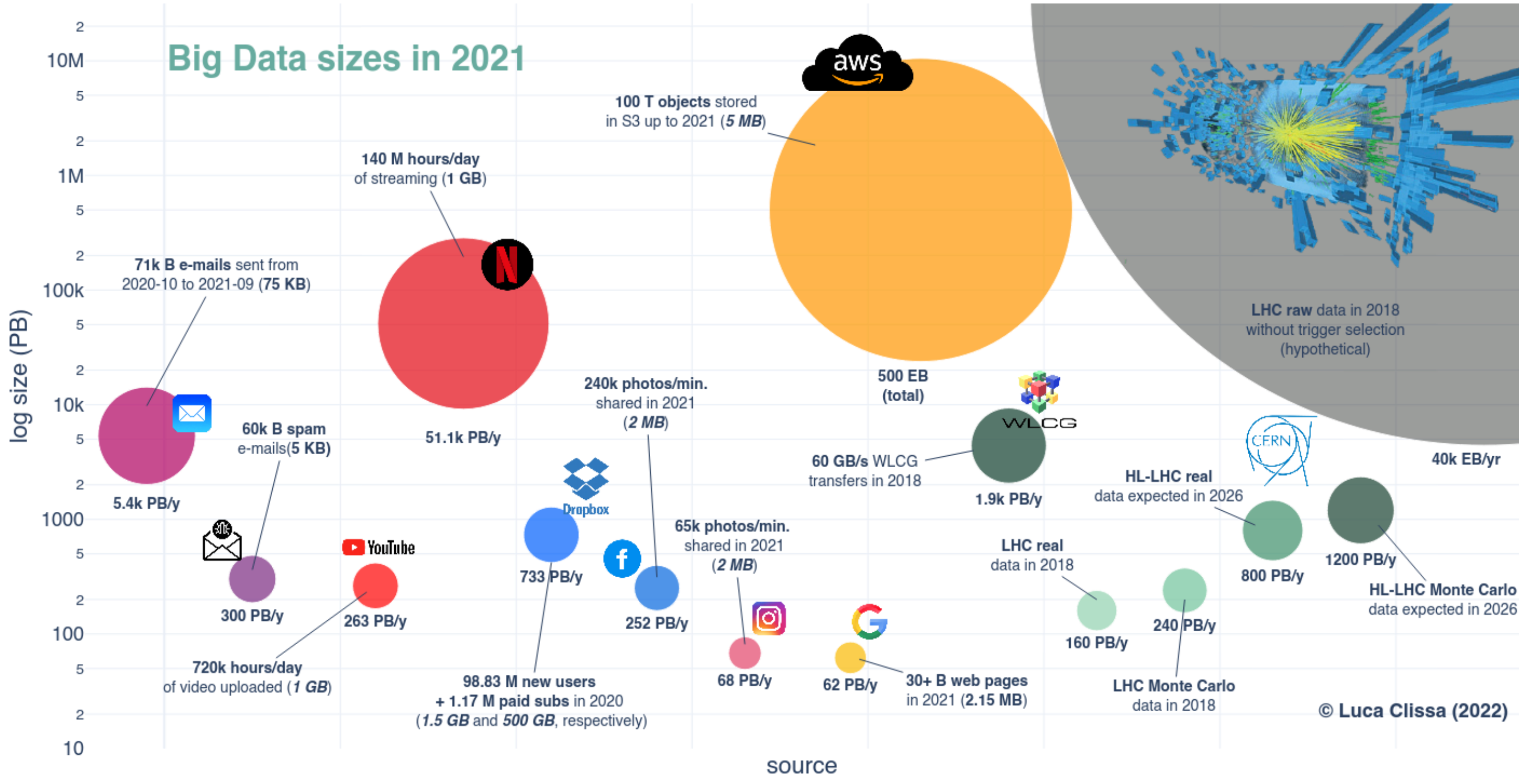
Outline

1. Motivation and Physics Context
2. eFPGA Technology
3. Machine Learning at the Front-End:
    1. Proof of concept BDT
    2. Autoencoder for lossy compression and anomaly detection
4. Conclusion

1

Motivation and Physics Context

Motivation and Physics Context



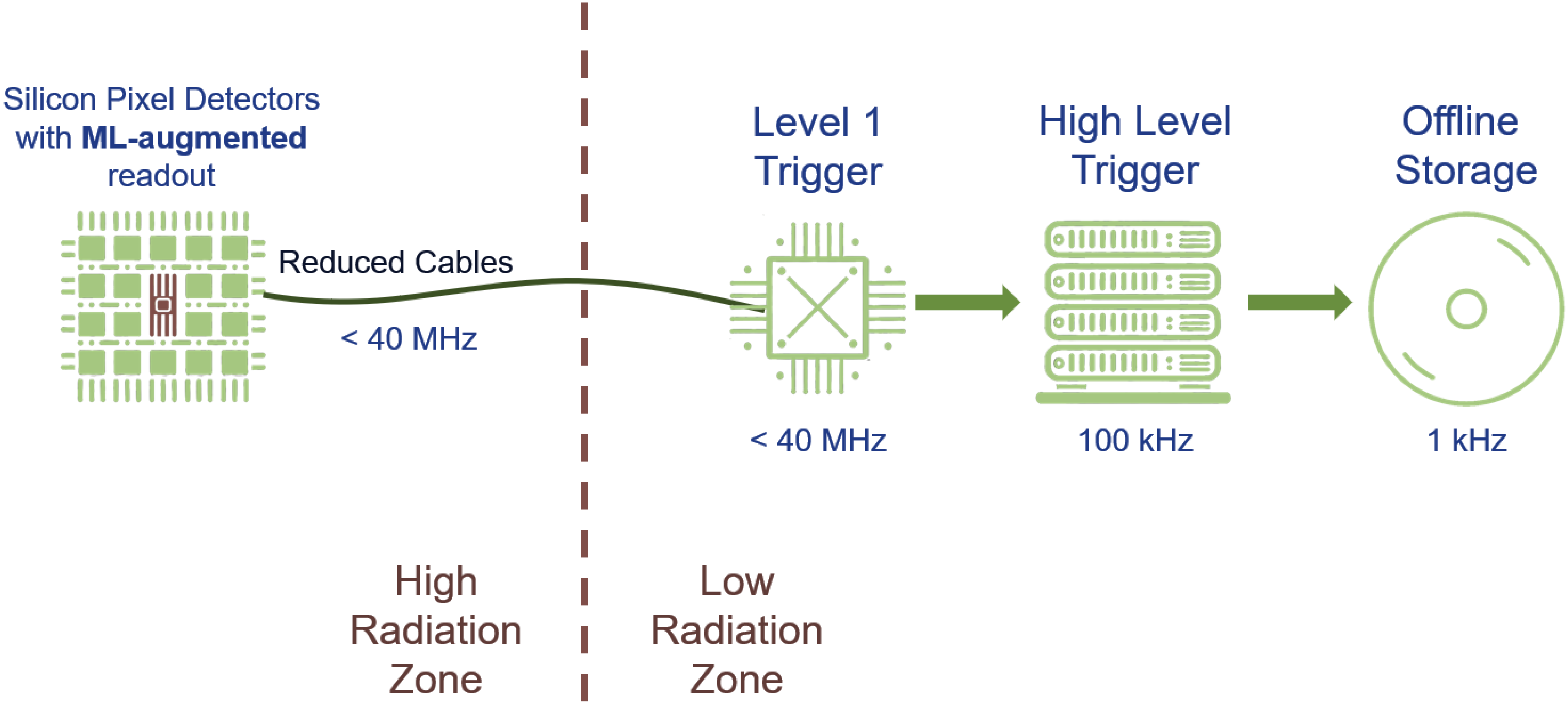
Pixel Detectors on collider:

- O(100) million pixels
- Petabyte per second data rate (more for future colliders!)

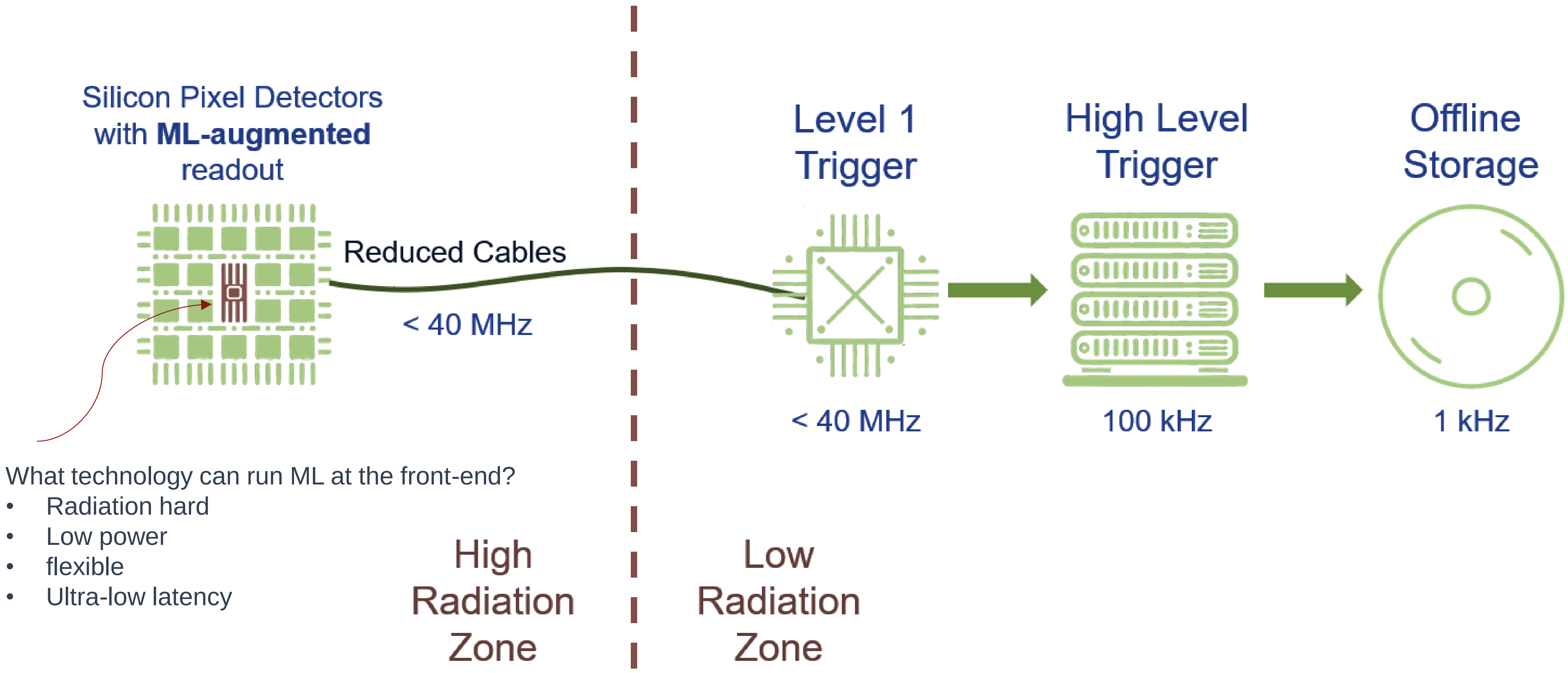
Can’t send everything off-detector

Challenge: how to effectively reduce the data volume transmitted off-detector while preserving useful physics information as much as possible?

ML-augmented readout system



ML-augmented readout system



Silicon Pixel Detectors with **ML-augmented** readout

Reduced Cables

< 40 MHz

Level 1 Trigger

< 40 MHz

High Level Trigger

100 kHz

Offline Storage

1 kHz

High Radiation Zone

Low Radiation Zone

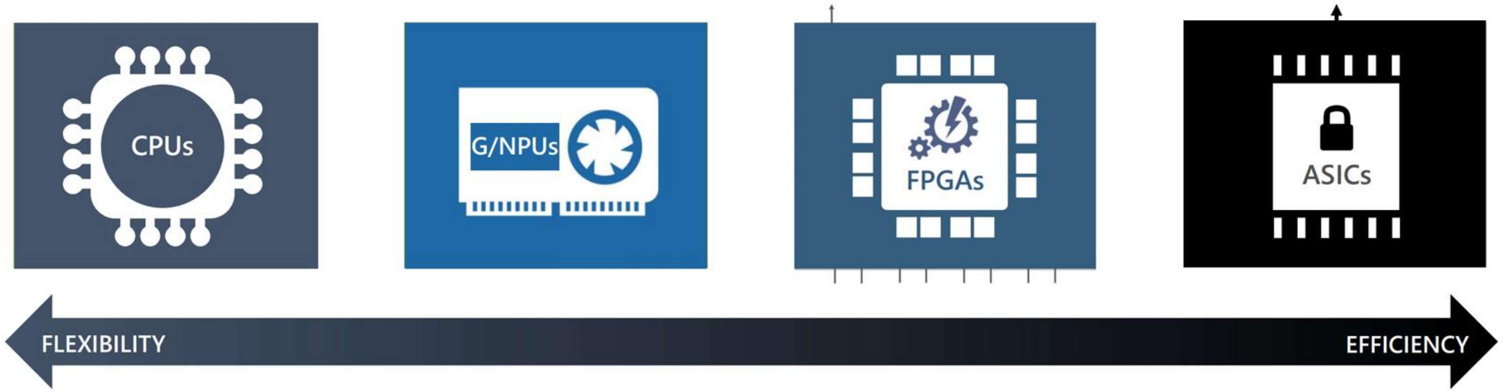
What technology can run ML at the front-end?

- Radiation hard
- Low power
- flexible
- Ultra-low latency

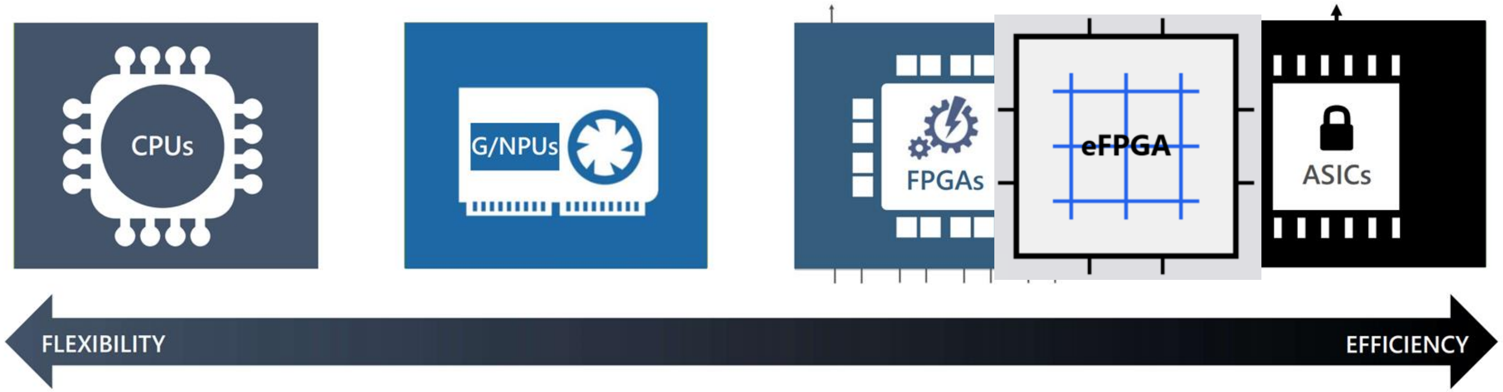
2

What’s Embedded FPGA?

Compute architectures



Compute architectures

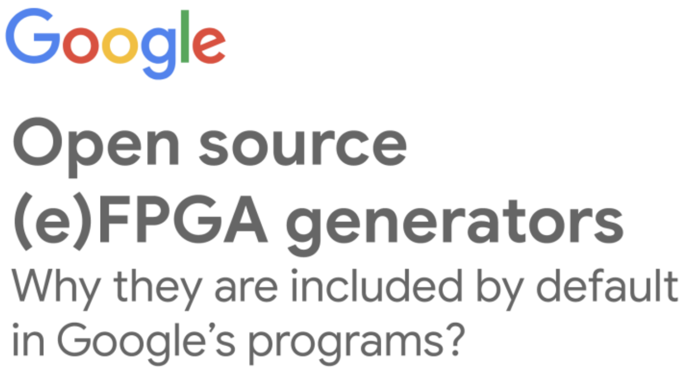


Embedded FPGAs (eFPGAs)

Basic idea is that you can put reconfigurable logic in your ASIC design.

- Full reconfigurability: can be re-configured just like a regular FPGA
- Power Efficiency: ASIC implementation means lower power than FPGA (“best of both worlds”)
- Development Time: “plug-and-play” FPGA fabric into ASIC
- Cost: no need for costly engineer hours or licenses to design an ML chip

Also, in use as hardware accelerators



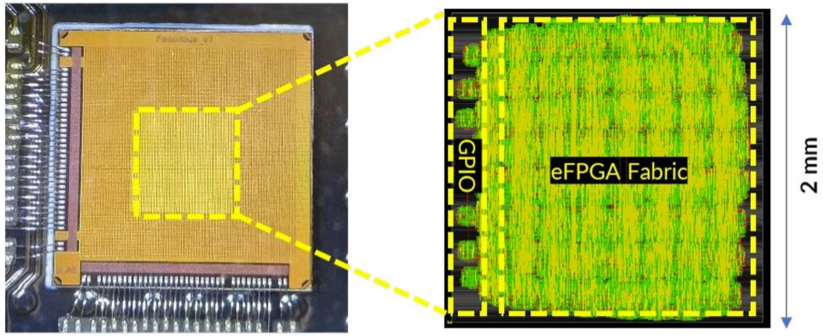
Google

Open source
(e)FPGA generators

Why they are included by default
in Google’s programs?

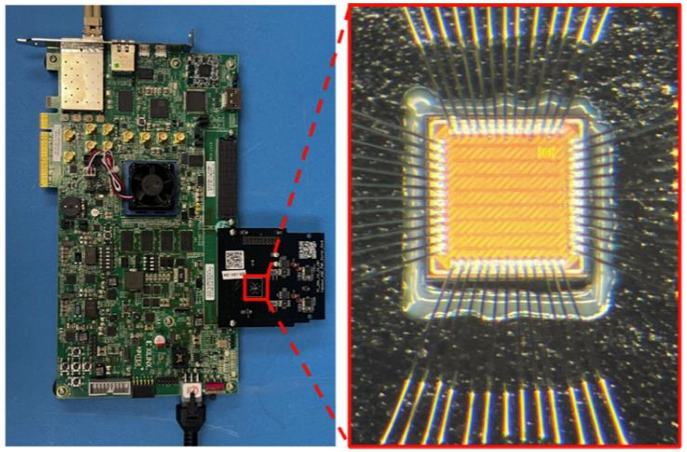
See Larry Ruckman's (CPAD 2024) talk for more

eFPGA Development at SLAC



Fabulous v1 ASIC

Floorplan of eFPGA fabric after Place&Route



- SLAC's Technology Innovation Directorate (TID) demonstrated an eFPGA design in a 130nm CMOS Multi-Process Wafer (v0)
- Subsequently designed a version 1 “proof-of-concept” eFPGA in 28nm CMOS in 2023 (v1), 1mm x 1mm

Results are published on 2024 *JINST* **19** P08023

Both are designed with **open-source** framework “FABulous” from University of Manchester. **Low cost and barrier to entry for institutions to participate in microelectronics design.**

FABulous
eFPGAs made easy

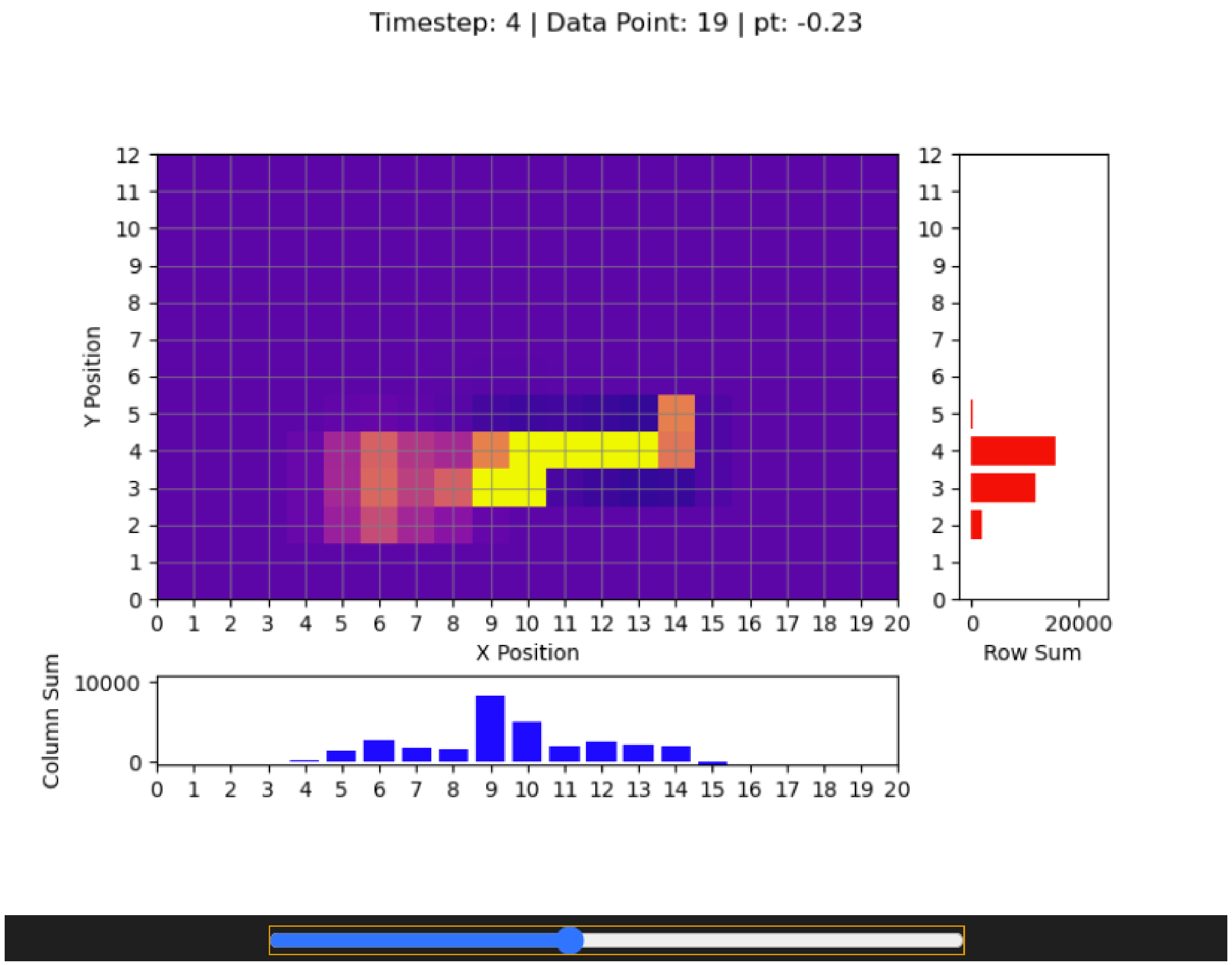
3

Machine Learning at the Front-End

Smart Pixel Dataset

We used the Smart Pixel Dataset*, which are pixel clusters produced by charged particles (pions) with real kinematics from CMS Run 2.

- 0.5 Millions of 20*13*21 (time × y position × X position) 2D "video" + y-local (y0)
- 13 truth info: positions, pT, angles…

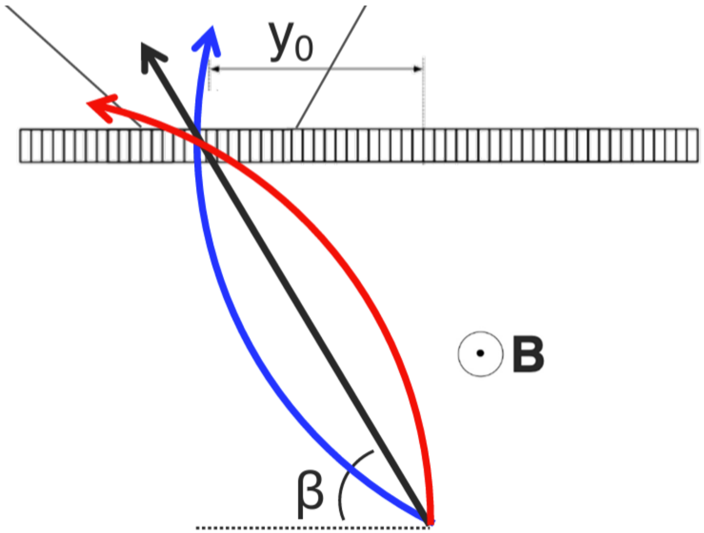


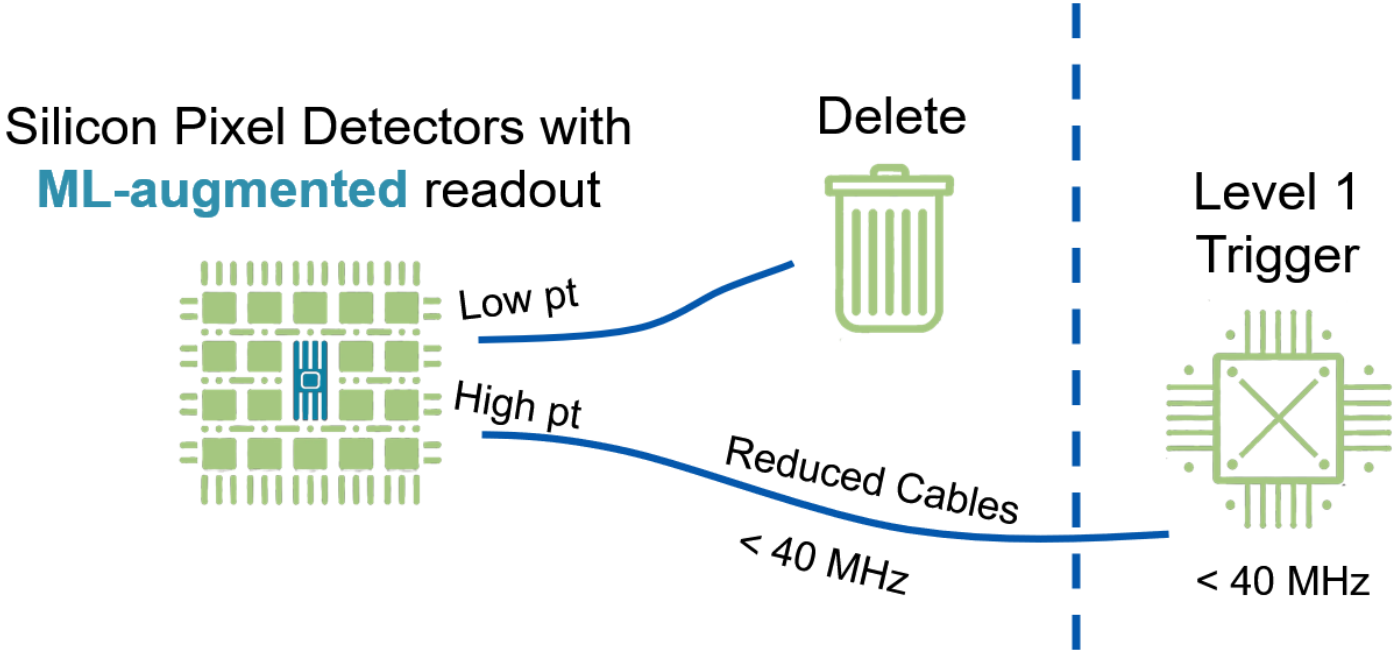
* https://zenodo.org/records/7331128

Proof-of-concept study

Reduce data rate by momentum classification using a **Boosted Decision Tree** with conifer.

Conifer





Silicon Pixel Detectors with **ML-augmented** readout

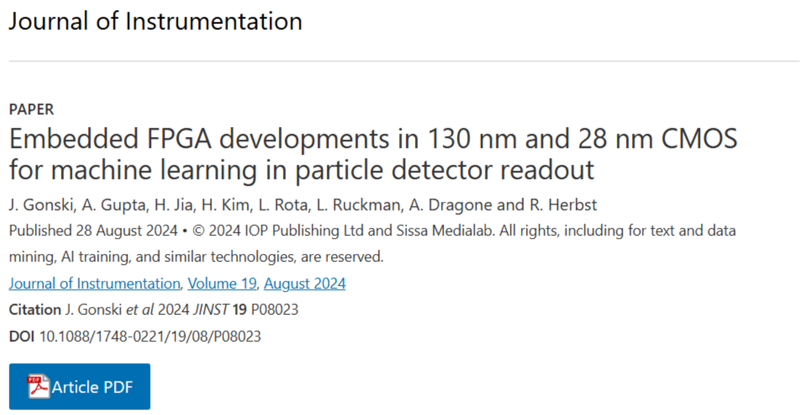
Delete

Level 1 Trigger

Proof-of-concept Results

We train a BDT to classify tracks with transverse momentum larger than 2 GeV, quantize and implement on the v1 eFPGA in 28nm CMOS. Use only 294 LUTs and nothing else (BRAM_18K, DSP, FF, URAM). Latency under 25ns. Hardware test achieves 100% accuracy compared to expected output!

Published 2024 *JINST* **19** P08023



Journal of Instrumentation

PAPER

Embedded FPGA developments in 130 nm and 28 nm CMOS for machine learning in particle detector readout

J. Gonski, A. Gupta, H. Jia, H. Kim, L. Rota, L. Ruckman, A. Dragone and R. Herbst

Published 28 August 2024 • © 2024 IOP Publishing Ltd and Sissa Medialab. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

Journal of Instrumentation, Volume 19, August 2024

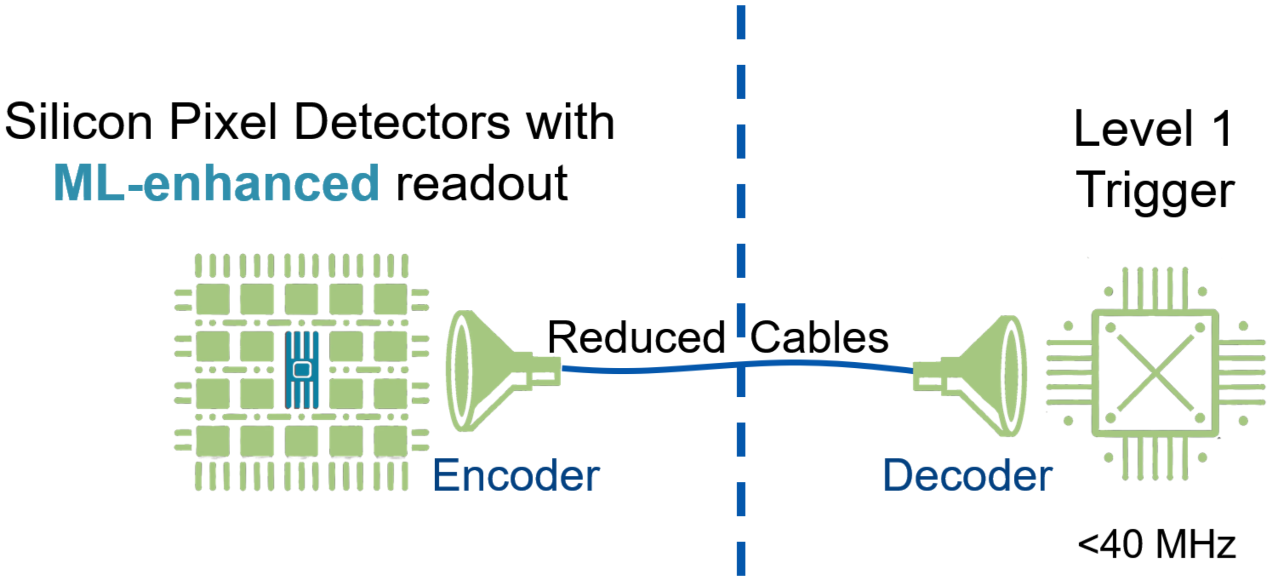
Citation J. Gonski *et al* 2024 *JINST* **19** P08023

DOI 10.1088/1748-0221/19/08/P08023

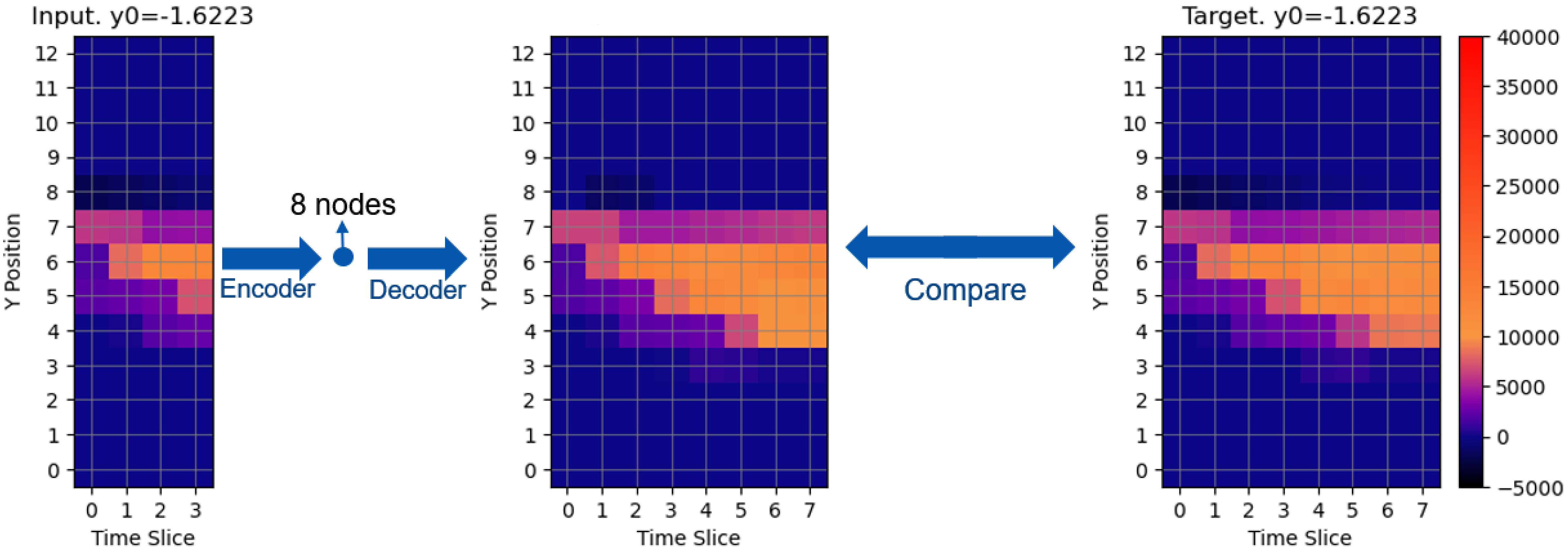
Variational Autoencoder for readout

Variational Autoencoder is a type of neural network which learns to compress and reconstruct input data. We propose to use them for **off-detector data compression** and **anomaly detection** for readout system.

- Latency constraint <25 ns
- Within eFPGA limited resources

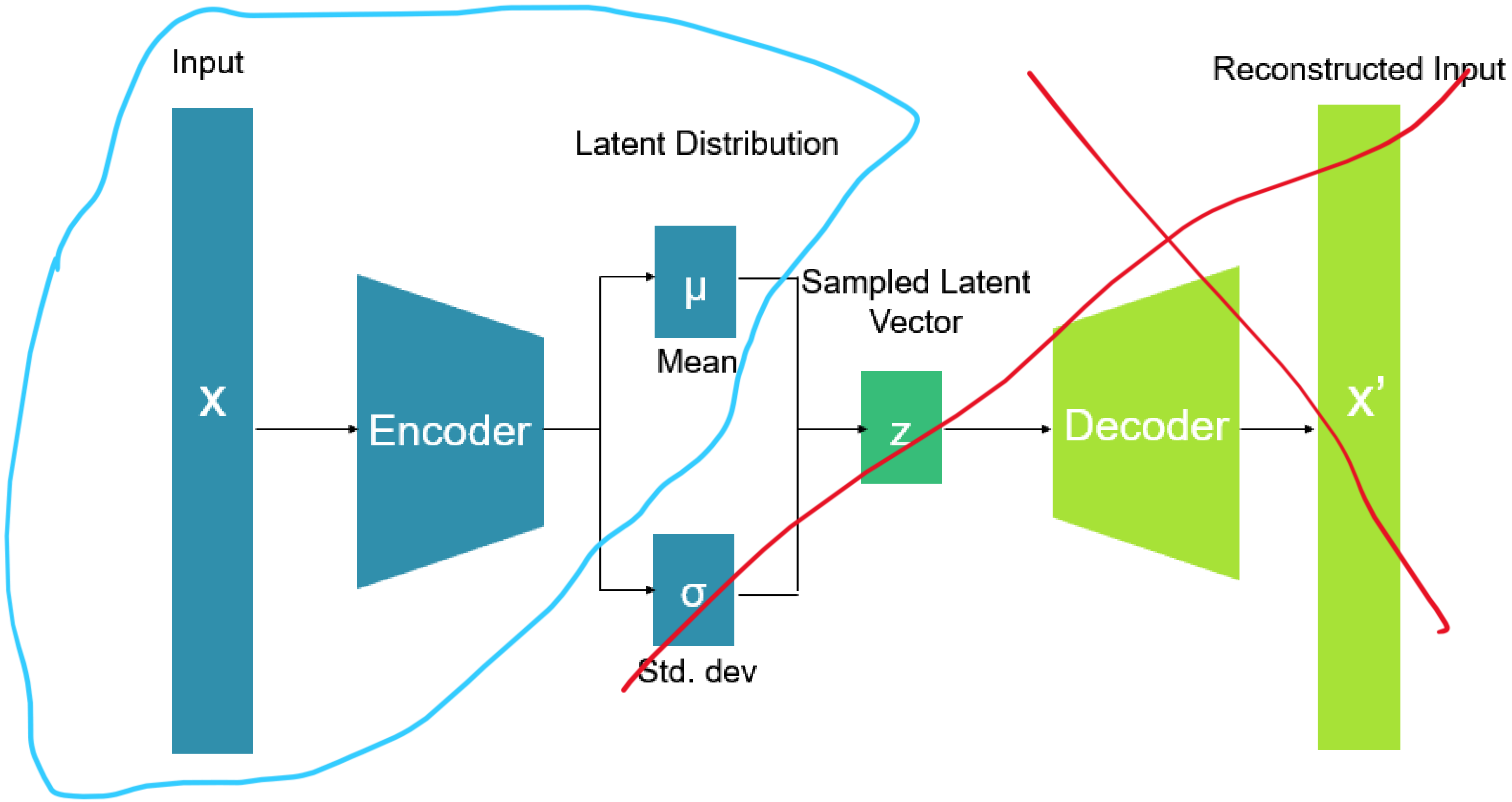


Example reconstruction



The model learns to reconstruct the whole piece with only 4 time slices!

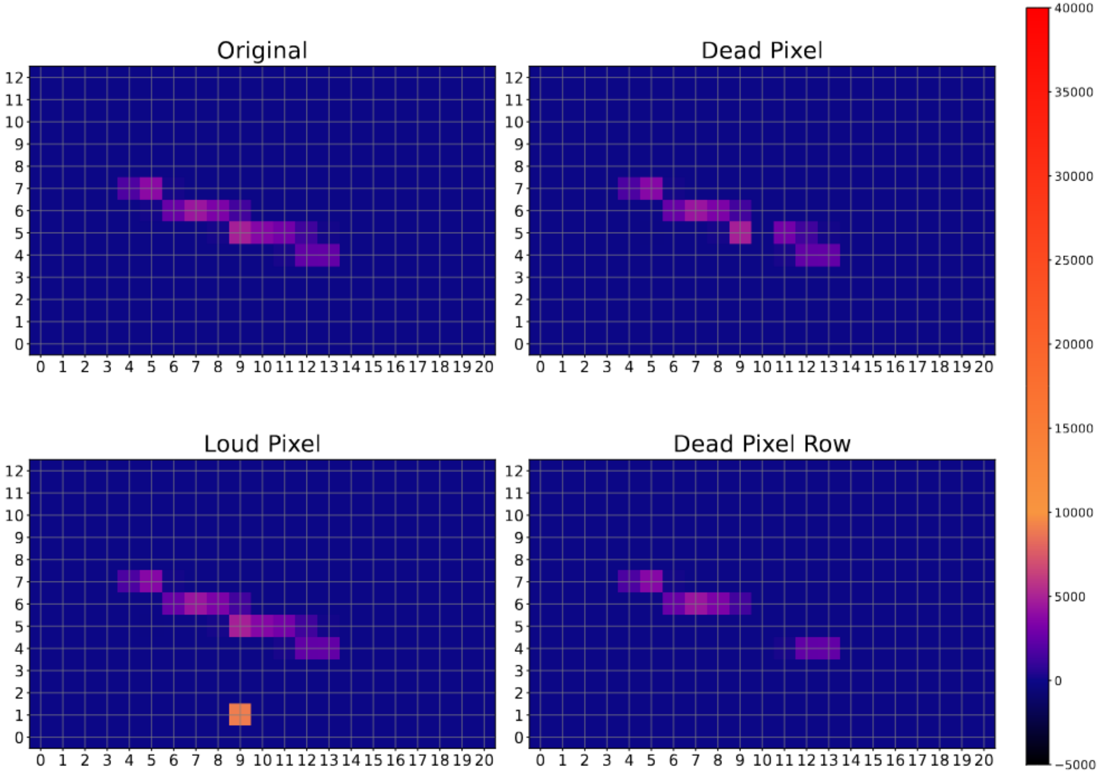
Variational Autoencoder (VAE)

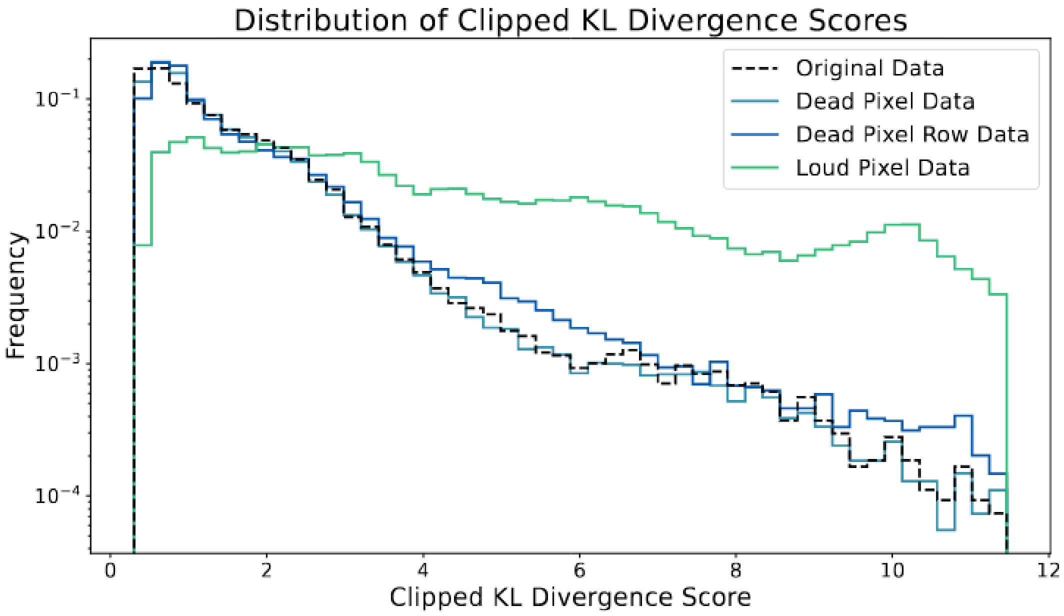


Kullback–Leibler divergence(KL) is a regularization term in training loss which helped shape the latent distribution into gaussian. Then the clipped KL divergence can be used as anomaly score with only the mean.

$$2{\cdot}\mathrm{KL} = \textstyle\sum_i \mu_i^2 + \xcancel{\sigma_i^2 - 1 - \log \sigma_i^2}$$

VAE for Defect monitoring





Clipped KLD can be used to capture detector defects.

Results on arXiv 2411.01118 and submitted to JHEP.

4

Conclusion

Conclusions

- Real-time machine learning embedded directly in particle detector hardware could revolutionize how these instruments operate at future colliders.
- eFPGAs is a promising hardware technology for deploying smart data readout at the edge in high energy collider experiments.
- Highly generalizable framework for different subsystems (silicon sensor charges, dual readout/LAr waveforms…)