Embedded FPGAs for Machine Learning in Future e+e- Detectors

Intelligence Across the Data Pipeline

- Detectors at a future Higgs factory can benefit from real-time machine learning in readout
	- Edge intelligence: feature extraction, classification, data compression at-source
	- Efficiency: lower computational power/storage needs for transmission & later DAQ stages (eg. trigger)
- Latency and radiation dosages require ML implementation in hardware/electronic (FPGAs, ASICs)

ML in Silicon Front-End Readout

- Future silicon pixel detectors will present exceptional challenges
	- \cdot Close to beamline = high occupancies/radiation
	- ‣ Very high granularity (25 μm) pixel pitch
	- ‣ Little room for services/cooling → minimize material budget & power density
- ML at the front-end to reduce off-detector data rate
	- \cdot HET factory: reduce cabling, increase granularity
	- ‣ Exascale (1015 bytes/sec) data rates anticipated at FCChh
- "Smart pixel" collaboration: study AI/ML to filter high p_T from pileup tracks (\lt 2 GeV) at source using pattern of deposited charge

HL-LHC Inner System

ML in Silicon Front-End Readout

• Future silicon pixel detectors will present exceptional abollenges \mathbb{R} is the basic to beam line \mathbb{R} in the high occupancies \mathbb{R} is the high occupancies of \mathbb{R} **Example 12 What hardware technology can implement** ↑ **ML** at the front-end? material budget & power density \cdot ML at algorithm: ASIC implementation rate • Ability to reconfigure \mathbb{R} - Z vs. WW vs. H vs. tt poles have different energies, backgrounds, occupancies: motivates readout algorithm optimization; reduces \cdot "Sm \cdot - Can also preserve option for "safe" non-ML operation mode high p $\overline{}$ from pileup tracks ($\overline{}$ GeV) at sources ($\overline{}$ usin<mark>g</mark> pattern of deposited charges charges and deposited charges of deposited charges of deposited charges of \sim "Smart Pileup" Pileup Track Filtering Track Filtering 4 HL-LHC Inner System • Lowest power, fastest latency (< 25 ns), and ability to radiation-harden upgrade need

ITk

eFPGAs

- "Embedded" FPGAs: reconfigurable logic in ASIC design for configurability ease of FPGA with low power/footprint of chip
- Patents of many commercial FPGAs recently expired
	- ‣ Open-source frameworks (eg. FABulous) allow for lowered barrier to entry for ASIC design

- Array of 8x8 tiles (scalable & configurable) \bullet
- Switch matrix \bullet

[[L. Ruckman\]](https://indico.slac.stanford.edu/event/8288/contributions/7652/)

10 October 2024 J. Gonski 6

Proof-of-Concept eFPGA Tapeouts [[2404.17701\]](https://arxiv.org/abs/2404.17701)

- SLAC designed prototype eFPGAs with FABulous **28nm eFPGA Test Setup** and taped out in 130nm & 28nm CMOS on TSMC MPW
	- ‣ Area: 1 mm2
	- ‣ Very small logical capacity (< 500 look-up tables)
- Physics performance: classify pileup from signal tracks
	- ‣ Model: boosted decision tree with depth 5, 440 LUTs and quantized to ap $fixed < 28,19>$
	- ‣ Configured to eFPGA and read back with 100% accuracy with respect to simulated expectation and quantized software result

➡Proof-of-concept for open-source design tools for eFPGAs ✅

28nm eFPGA Power [[2404.17701\]](https://arxiv.org/abs/2404.17701)

- Clock frequency scans (10-250 MHz) indicate no detected bit errors
- Extremely stringent power requirements for readout in Higgs factory vertexing/tracking detectors; $O(10)$ mW / cm² [[FCCW24\]](https://indico.cern.ch/event/1298458/timetable/#b-565605-physics-experiments-a)
	- ‣ Considerable power optimization expected from dedicated engineering design
	- ‣ R&D into new technologies, eg. silicon photonics/analog compute elements?

Front-End ML Architectures bandwidth. Figure 5 shows the VALE and the VAE architecture and the serve and the output score to serve as and potential anomalies in the tails of a distribution. A threshold on this score can be used to select the \blacksquare most anomalous events the value of the value of the \blacksquare bandwidth. Figure 5 shows the VAE architecture and the ability of the ability of the output score to serve as $$ **bandwidth. Figure 5 shows the VALE and the ability of the ability of the ability of the ability of the ability o**

• BDTs, neural nets: simple classification \sqrt \cdot BDTs, neural nets: simple classification $\sqrt{ }$

indicator of out-of-distribution events.

indicator of out-of-distribution events.

- Variational autoencoders can offer two front-end capabilities: t respecting to respect the typical constraints of a hardware trigger at a hardware trigger at an L Hallonal autochcodels can oner two hont-end capabilities. • Variational autoencoders can offer two front-end canabilities: variational adiocheodels can one two none end capabilities.
- **Data compression**: resource-constrained encoder on-chip followed by \blacksquare decoder off-detector only decoder off-defector a crucial proof-of-concept for performing and performing and performing and performing and performing and performing and performing anomaly detection with α and α and α and α and α $\overline{}$. Data compression: resource-constrained encoder on-chin followed by (AXOC1TL), and was run in the trigger test crate during 2023 with graphs 2023 with good rate stability in the trigger of the trigger test can be decoder of the total control with graphs and with graphs and with graphs and between the trianger test constraint teacher on the trigger test can be the trigger test can be started with α . Note that with good α only does this efort demonstrate a crucial proof-of-concept for performing anomaly detection within
- \rightarrow Anomaly detection: latent space variables can be used to flag inputs that appear anomalous and/or outliers opposes this effection. Internet and concept for concept for performance that \overline{P} Anumary detection. Fatern space variables can be used to flag imputs that \overline{P} appear anomalous any or outliers.

Autoencoders at the Front-End

- Model: low-latency (< 25ns) and resource-constrained (< 30,000 LUTs) VAE
- Achieve faithful reconstruction of 10-bit pixel values with just 8 latent dimensions
- to "clip" any Kalendary and simply use *u*se *y* and simply use *µ2 as the anomaly use* α • Outperforms on-chip classifier methodology in performance, resources, and latency, with just 8% of the original data with the sets of the sets of the original data $\overline{}$ and $\overline{}$ for events for each of the three anomalies, and its anomalies, and its anomaly score is anomalies, and its anomalies, $\overline{}$ $\overline{\textbf{10}}$. Figure 5 displays the anomaly score distributions of the standard track and $\overline{\textbf{10}}$ anomaly test sets. Each of the anomalies is reconstructed with a higher mean anomalies is reconstructed with a h transmitted off-detector
- score compared to the standard tracks, with a tail to the standard tracks, with a higher value of \mathcal{I}_n • On-chip latent space variable can separate several classes $\frac{1}{4}$. The state \mathbb{E} of \mathbb{E} metric is defined by choosing a threshold such that the false positive rate for anomalies \mathbf{r}_i \sim 1 and report the false discovery rate (ratio of non-anomalous tracks over the false over the false to the total tracks over the interesting over the interesting of \mathcal{A} . This metric is 48.8% for deal pixel and \mathcal{A} of anomalous pixel events from background

 $\sqrt{12}$ and $\sqrt{01}$ and $\sqrt{0}$ and $\sqrt{0}$

Figure 4. Display of the smart pixel simulated tracks and their pattern of charge

Track Reconstruction & Extrapolation

Higgs Factory Applications

- Reconfigurability of eFPGAs enables generic ML methodologies: applicable to wide variety of datasets & subsystems
	- ‣ Dual readout calorimetry: ML to extract Cherenkov C and scintillation S photon yields from single waveform
	- ‣ High granularity calorimetry: ML for pattern recognition of hits \rightarrow showers & energy regression
	- ‣ Liquid argon: ML to extract energy and timing from time-domain waveform

➡**Get in touch if interested!**

Dual Readout Waveform Analysis

LAr Waveform Analysis

Conclusions

- Detectors at a future Higgs factory can benefit from real-time ML for advanced DAQ systems
- Embedded FPGAs provide a low-power ASIC option for generic and reconfigurable ML at the front-end
- SLAC proof-of-concept FABulous eFPGA in 28nm implements small ML and verifies open-source design frameworks for future work
- Looking forward:
	- Tape out larger eFPGA for more complex algorithms, hardware verification, and power studies
	- Implement radiation-hardness and/or cryogenic tolerance
	- Hope to deliver eFPGAs as a viable readout technology for future Higgs factory detector designs!

28nm eFPGA Design

Figure 7. Block diagram of the 28nm CMOS ASIC design.

SLAC

FABulous Design Workflow

SLAC

Smart Pixel Dataset momentum particles arising from pileup. High-momentum particles are less curved in the magnetic \mathbf{R} mort Divol Dotocot \blacksquare $m_{\rm H}$ to model delta rays), a realistic field map resulting from the simulation of simulation \sim

field, therefore traversing fewer pixels compared to pileup particles. It is important to note that The issues composed of z ix id pixel andy with dox iz.d pin pitch, dominimonic sensitive magnetic field and is the ma • Sensors composed of 21x13 pixel array with 50x12.5 µm pitch, 30mm from beam line with B = 3.8 T

of Poisson's Equation, carrier continuity equation, and various charge transport models, and various S

- \cdot Track 8 denosited (x y) charge arrays with timestens of 200 ps $\frac{1}{2}$ is study is the drift and induced currents in the drift and induced currents in the pixel sensor. • Track = 8 deposited (x,y) charge arrays with timesteps of 200 ps
- ~550,000 tracks in dataset

(a) (b)

Figure 12. A diagram of the single tree BDT model used for proof-of-concept synthesis to the 28 nm eFPGA.

pileup based on the pattern of charge distribution across the sensor over the sensor over the sensor over the s data rate reduction by rejecting pileup tracks at the sensor level. This pileup classification model

Front-End VAE Performance classification task. The order of magnitude reduction in o↵-detector data rate, along

with the generality of the VAE compression method which can be easily adapted to adapted to adapted to adapted to a

Table 1. Summary of model performance metrics, namely latency, on-detector resources (LUTs, DSPs, FFs), background rejection (BR) for two fixed signal efficiencies (SE) on the pileup classification task, and the percent of the original data volume that is transmitted off the detector. Two models are shown: the VAE scheme which includes an on-detector encoder followed by off-detector decoder and classifier stages, and a classifier that can fit on-detector requirements.

Smart Pixel Anomalies for VAE these anomalies as they appear in the pixel sensor, alongside a standard smart pixel

track for reference.

Figure 4. Display of the smart pixel simulated tracks and their pattern of charge deposition across the simulated sensor of the smart pixel dataset, including a typical background track (top left), along with the three types of anomalies, namely a dead pixel (top right), loud pixel (bottom left), and a dead pixel row (bottom right).