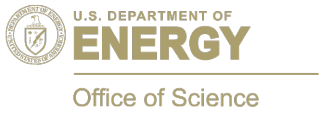
Embedded FPGAs for Machine Learning in IDEA Subsystems

Julia Gonski

19 November 2024

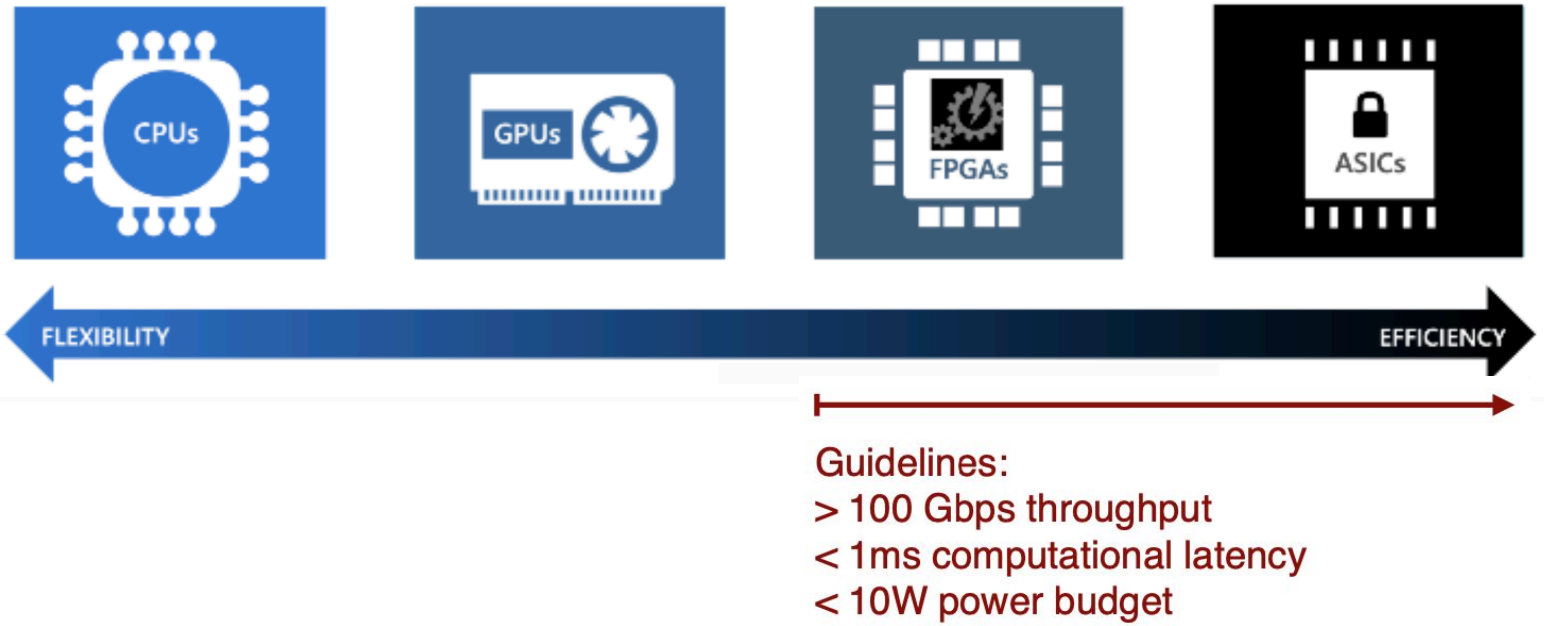
IDEA Study Group Meeting



Intelligence Across the Data Pipeline

SLAC

- 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/electronics** (FPGAs, ASICs)



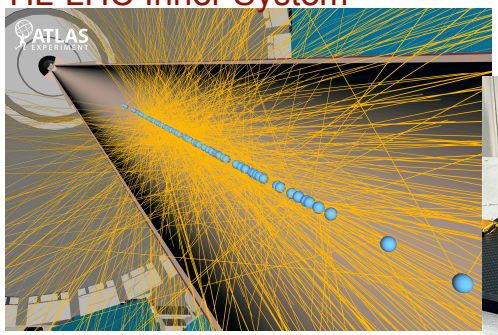
ML in Silicon Front-End Readout

- Future silicon pixel detectors will present exceptional challenges
  - Close to beamline = high occupancies/radiation
  - Very high granularity (25 μm) pixel pitch
  - Little room for services/cooling → minimize material budget & power density

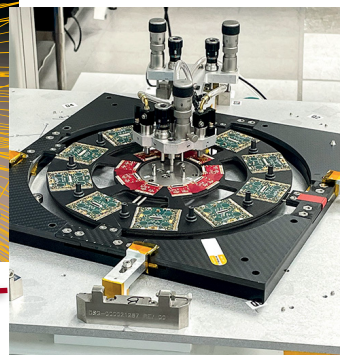
- ML at the front-end to **reduce off-detector data rate**
  - **HET factory: reduce cabling, increase granularity**
  - Exascale (10^{15} bytes/sec) data rates anticipated at FCChh

- "Smart pixel" collaboration: study AI/ML to filter high p_T from pileup tracks (< 2 GeV) at source using pattern of deposited charge

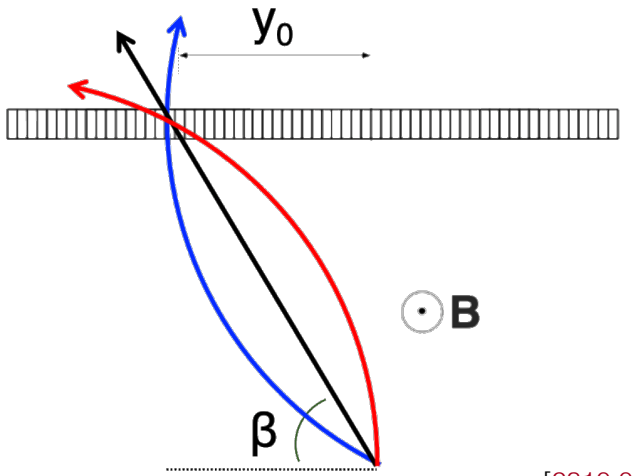
HL-LHC Inner System



ITk



"Smart Pixel" Pileup Track Filtering



[2310.02474]

ML in Silicon Front-End Readout

SLAC

- Future silicon pixel detectors will present exceptional challenges
  - ...
  - ...
  - ...
- ML a... rate
  - ...
  - ...
- "Sm... high... usin...

HL-LHC Inner System

ITk

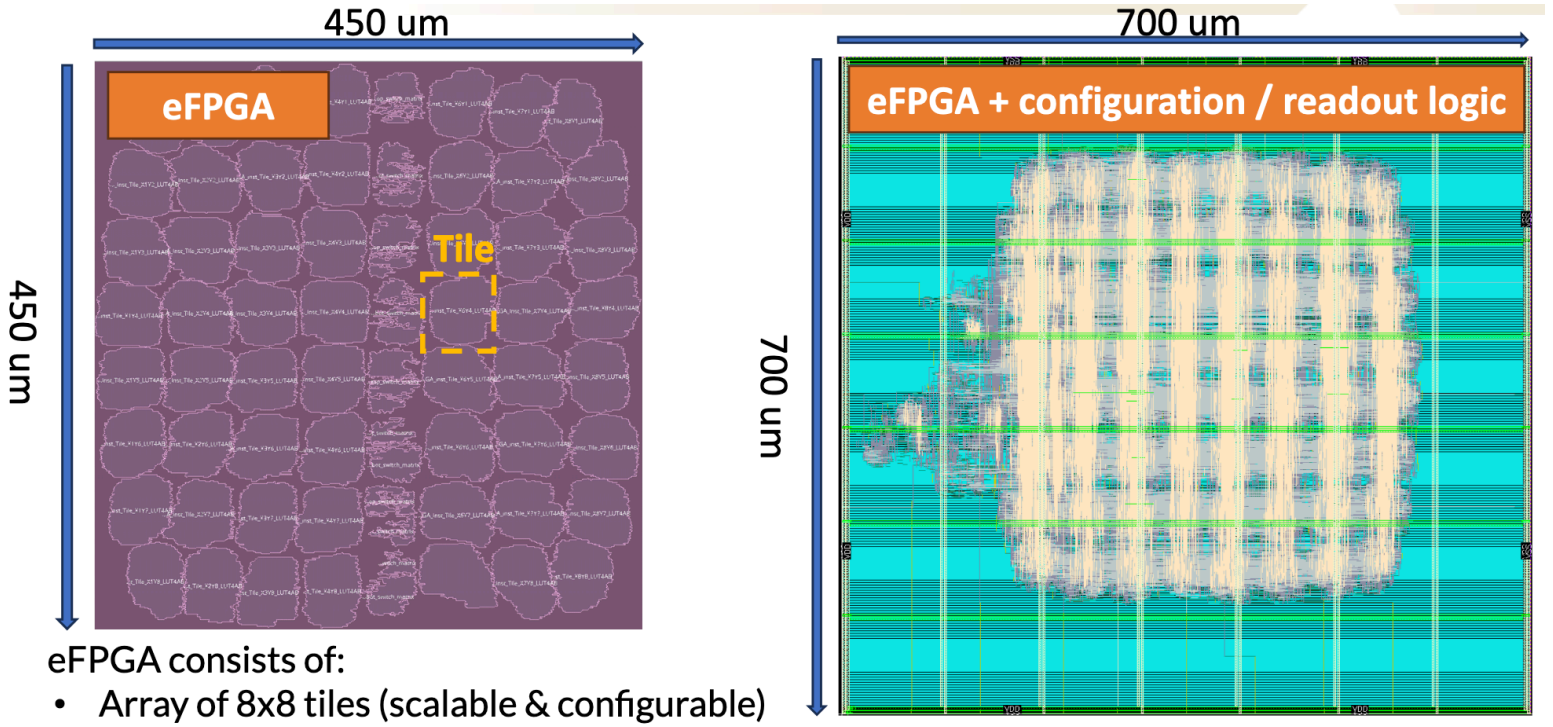
What hardware technology can implement ML at the front-end?

- Lowest power, fastest latency (< 25 ns), and ability to radiation-harden algorithm: **ASIC implementation**
- Ability to **reconfigure**
  - Z vs. WW vs. H vs. tt poles have different energies, backgrounds, occupancies: motivates readout algorithm optimization; reduces upgrade need
  - Can also preserve option for "safe" non-ML operation mode

[2310.02474]

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



eFPGA consists of:
- Array of 8x8 tiles (scalable & configurable)
- Switch matrix

[L. Ruckman]

Proof-of-Concept eFPGA Tapeouts

[2404.17701]

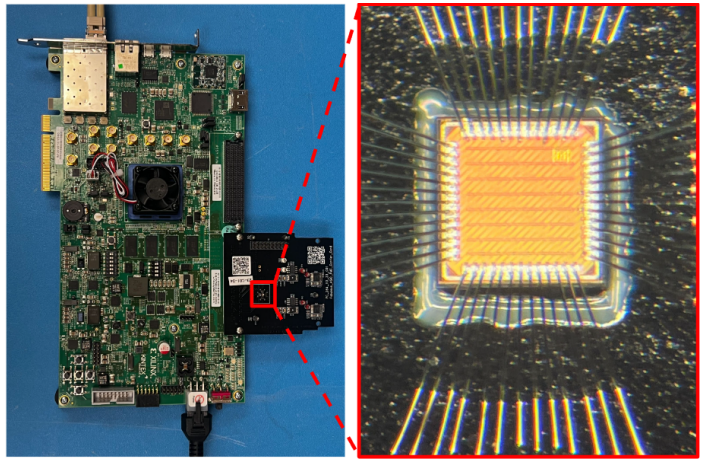
SLAC

- SLAC designed prototype eFPGAs with FABulous and taped out in 130nm & **28nm** CMOS on TSMC MPW
  - Area: 1 mm^2
  - 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 Test Setup

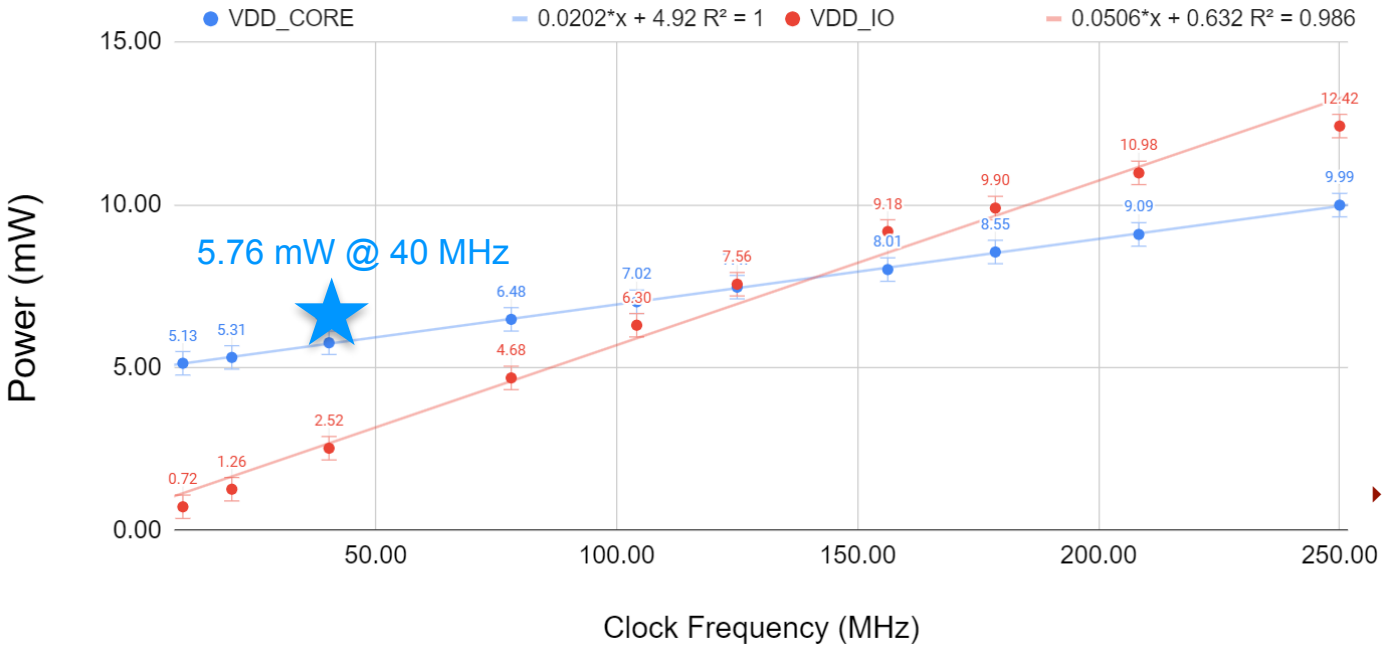


28nm eFPGA Power

[2404.17701]

SLAC

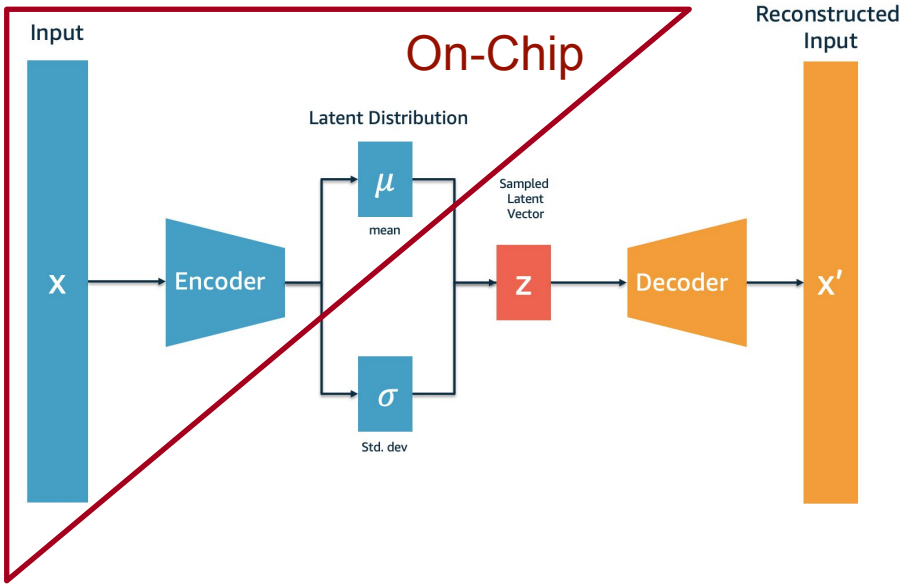
- Clock frequency scans (10-250 MHz) indicate no detected bit errors
- Extremely stringent power requirements for readout in Higgs factory vertexing/tracking detectors; O(10) mW / cm^2 [FCCW24]
  - Considerable power optimization expected from dedicated engineering design
  - R&D into new technologies, eg. silicon photonics/analog compute elements?

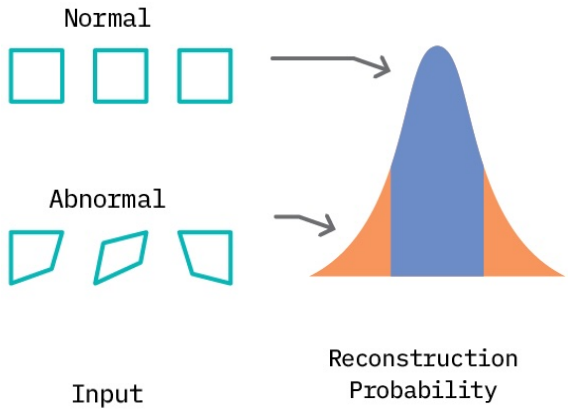


- Area: 1 mm^2

Front-End ML Architectures

- BDTs, neural nets: simple classification ✓
- **Variational autoencoders** can offer two front-end capabilities:
  - **Data compression**: resource-constrained encoder on-chip followed by decoder off-detector
  - **Anomaly detection**: latent space variables can be used to flag inputs that appear anomalous and/or outliers



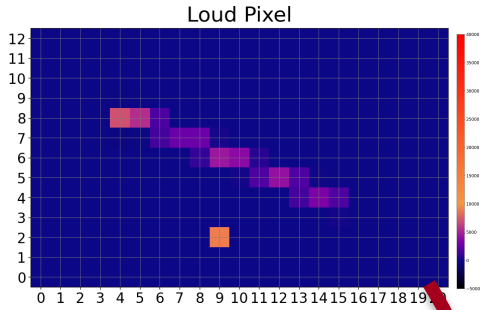


Autoencoders at the Front-End

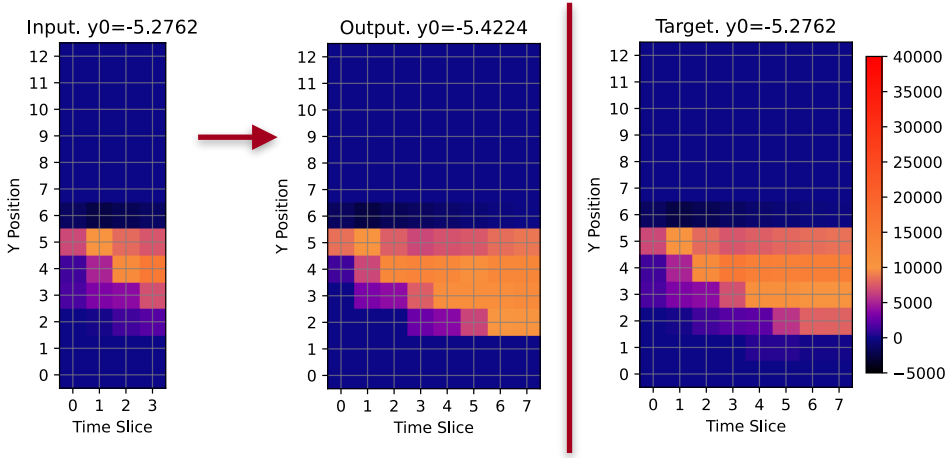
[2411.01118]

SLAC

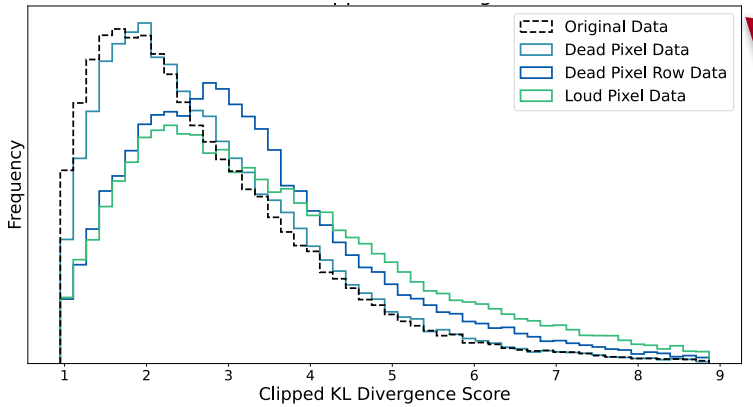
- Model: low-latency (< 25ns) and resource-constrained (< 30,000 LUTs) VAE
- **Achieve faithful reconstruction of 10-bit pixel values** with just 8 latent dimensions
- **Outperforms on-chip classifier methodology** in performance, resources, and latency, with just 8% **of the original data transmitted off-detector**
- On-chip latent space variable can separate several classes of **anomalous pixel events** from background



Track Reconstruction & Extrapolation



Anomaly Detection

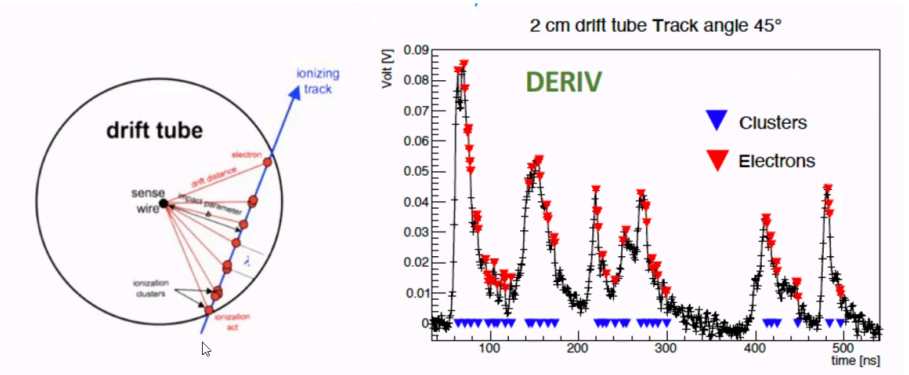


Applying to IDEA

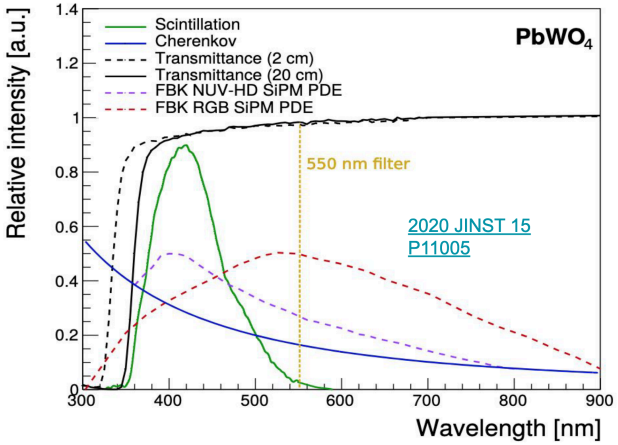
- "Edge" (at-source) ML can reduce off-detector data rate, minimize cabling, and enhance efficiency of subsequent DAQ steps
  - Silicon vertexing: on-pixel pileup track filtering
  - Drift chamber: cluster counting
  - Dual readout calorimetry: real-time waveform analysis for C/S extraction

➡ **Please get in touch with ideas/ datasets!**

Drift chamber cluster counting



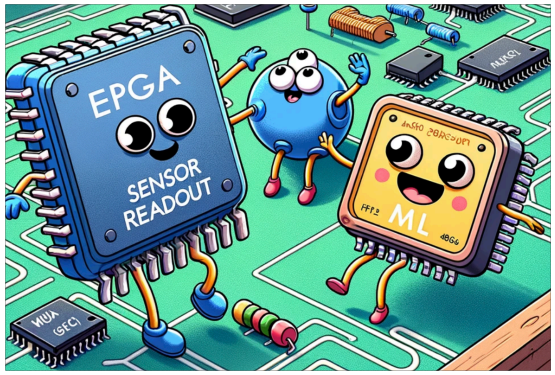
Dual Readout Waveform Analysis



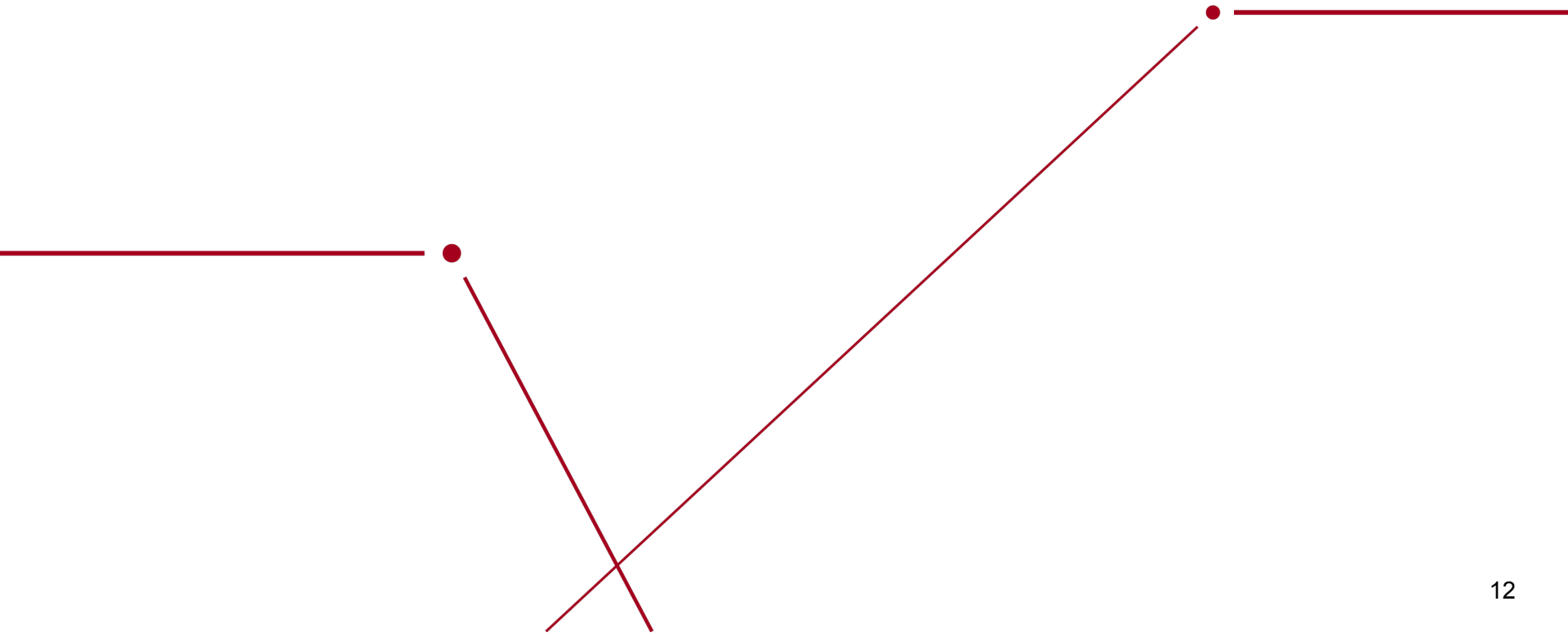
G. Cummings

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!



Backup



28nm eFPGA Design

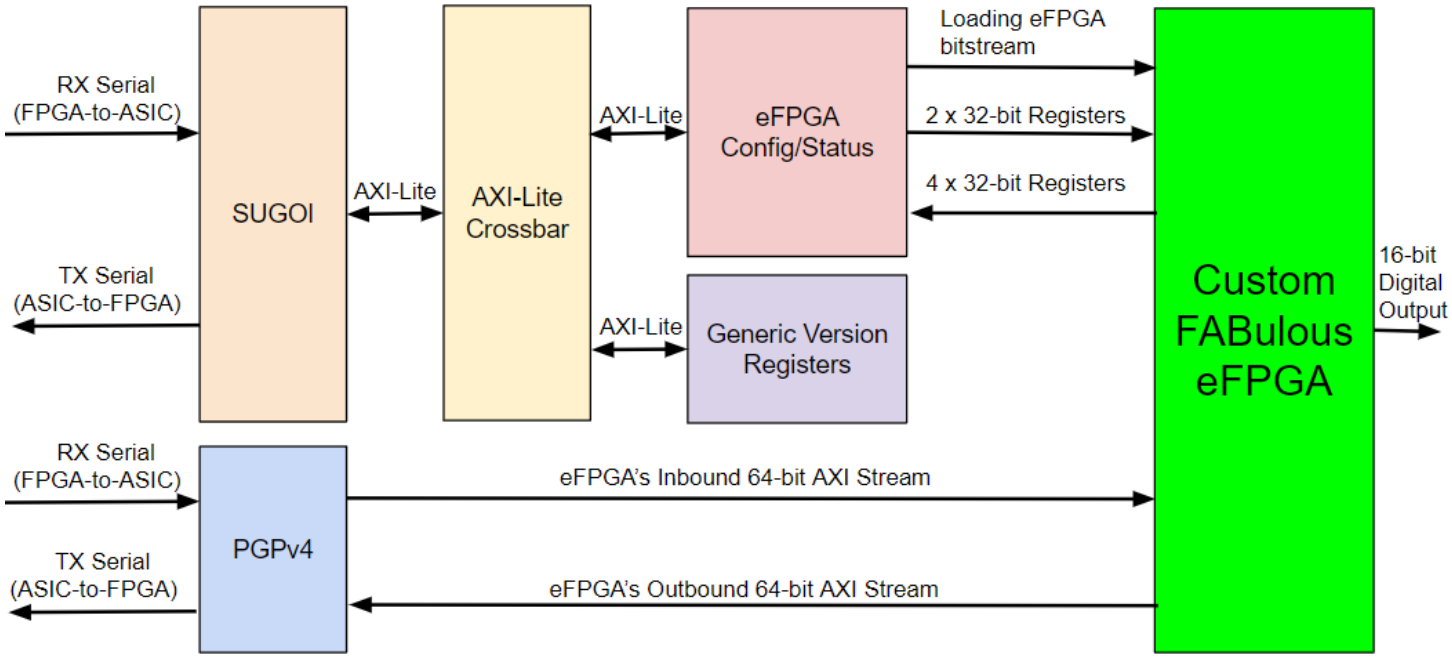
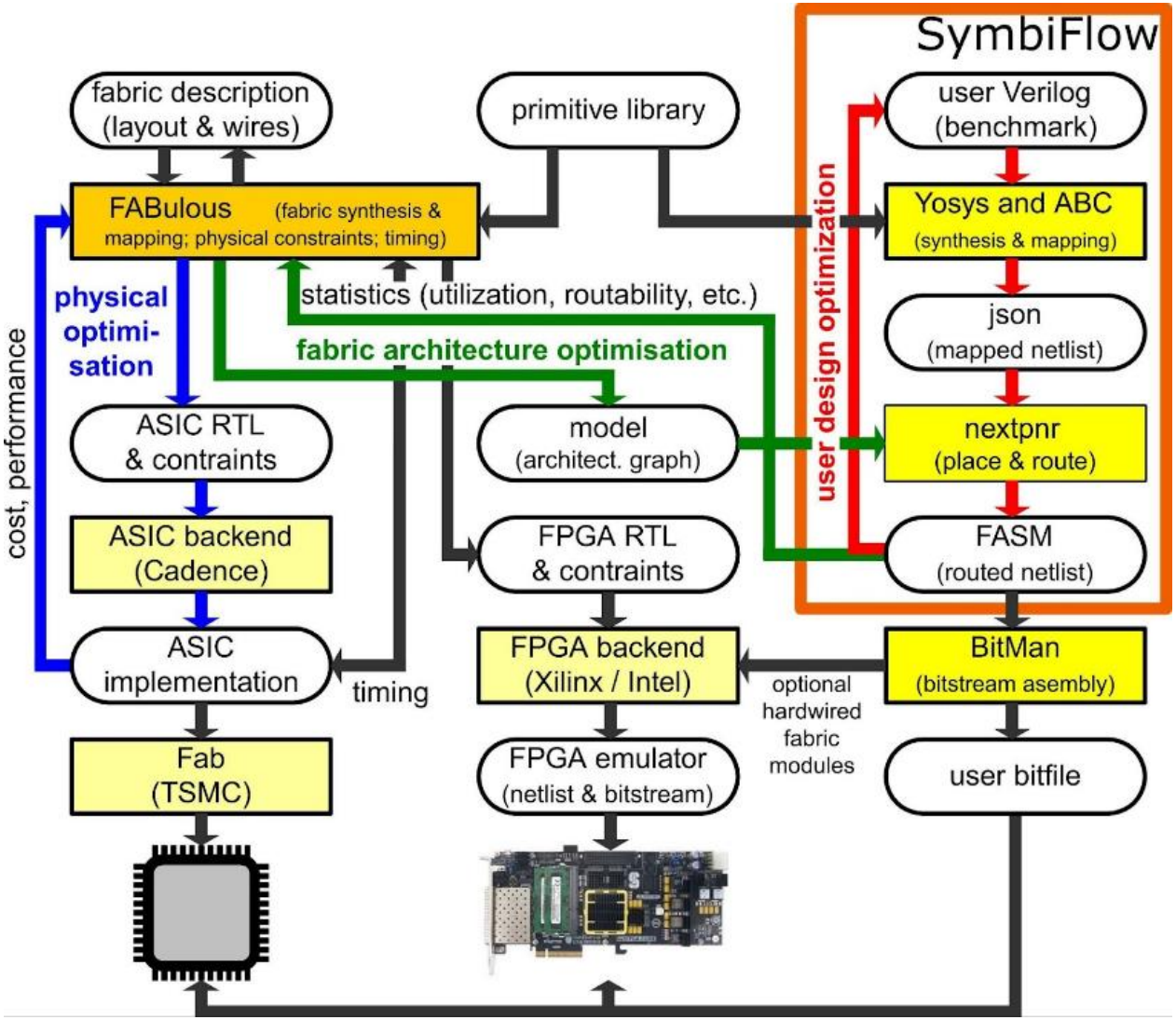


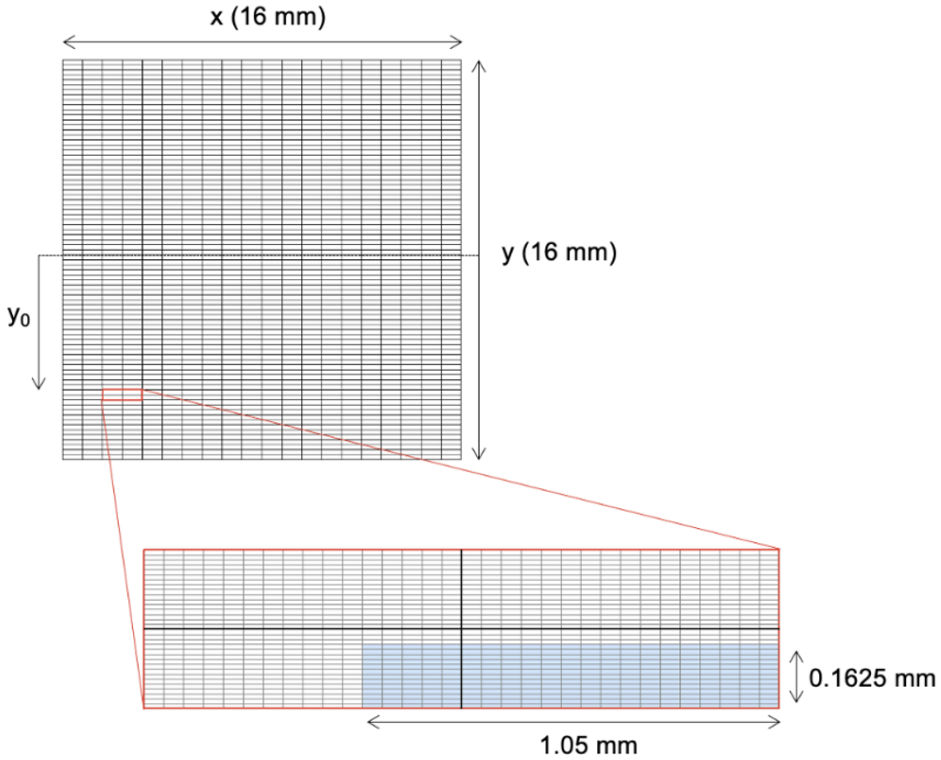
Figure 7. Block diagram of the 28nm CMOS ASIC design.

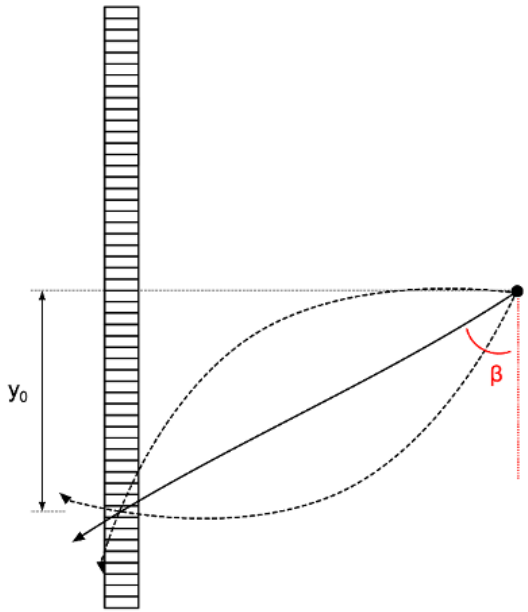
FABulous Design Workflow



Smart Pixel Dataset

- Sensors composed of 21x13 pixel array with 50x12.5 μm pitch, 30mm from beam line with B = 3.8 T
- Track = 8 deposited (x,y) charge arrays with timesteps of 200 ps
- ~550,000 tracks in dataset

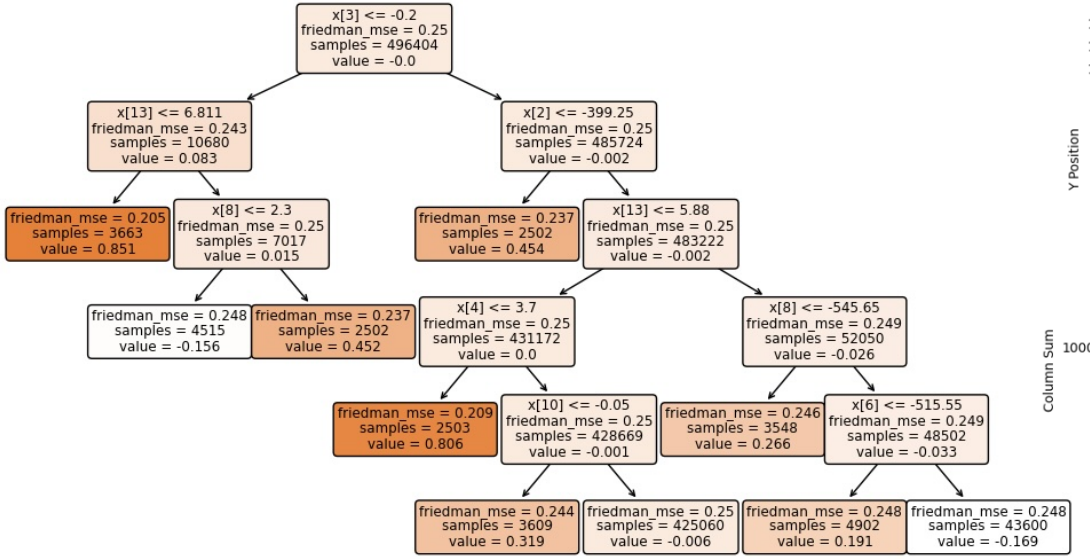




https://doi.org/10.5281/zenodo.10783560

28nm Boosted Decision Tree

Architecture



Example Input Track

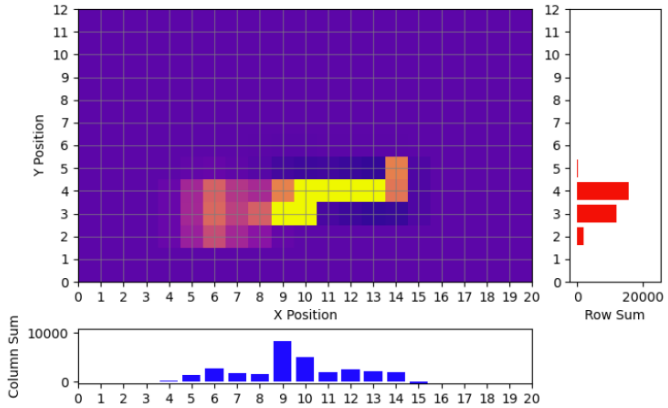


Figure 12. A diagram of the single tree BDT model used for proof-of-concept synthesis to the 28 nm eFPGA.

Performance

Signal Efficiency	Background rejection
96.4%	5.8%
97.8%	3.9%
99.6%	1.1%

Front-End VAE Performance

SLAC

	VAE + Off-Detector Classifier	On-Chip Classifier
Latency [ns]	15	25
LUTs	27,629	38,394
DSPs	680	723
FFs	850	931
BR @ SE=0.93	0.36	0.32
BR @ SE=0.98	0.23	0.18
Data Compression (%)	7.6	82

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

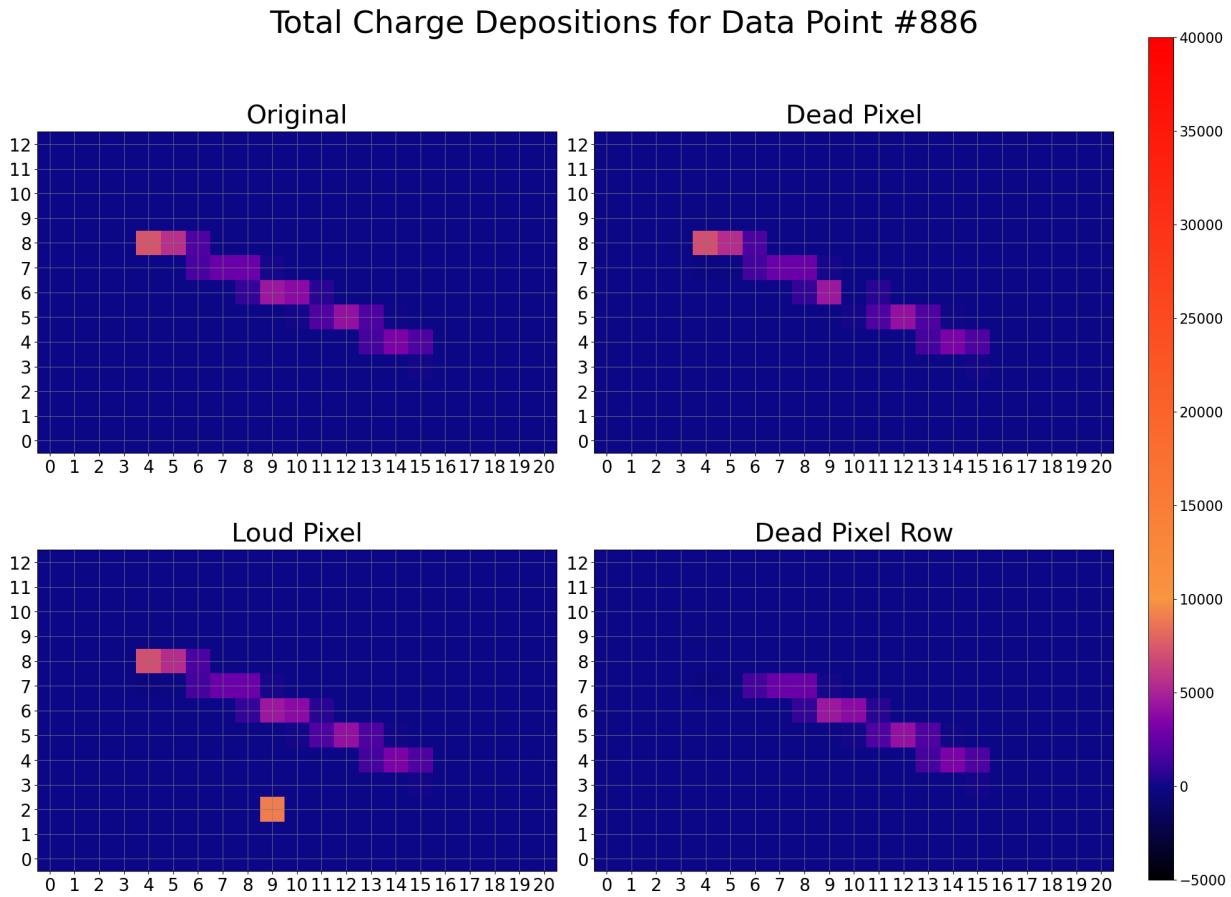


Figure 4. Display of the smart pixel simulated tracks and their pattern of charge deposition across the simulated sensor of the smart pixel dataset, including a typical background track (top left), along with the three types of anomalies, namely a dead pixel (top right), loud pixel (bottom left), and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