Tools (e.g. for streaming DAQ, fast ML, automation/self running DAQ,…)

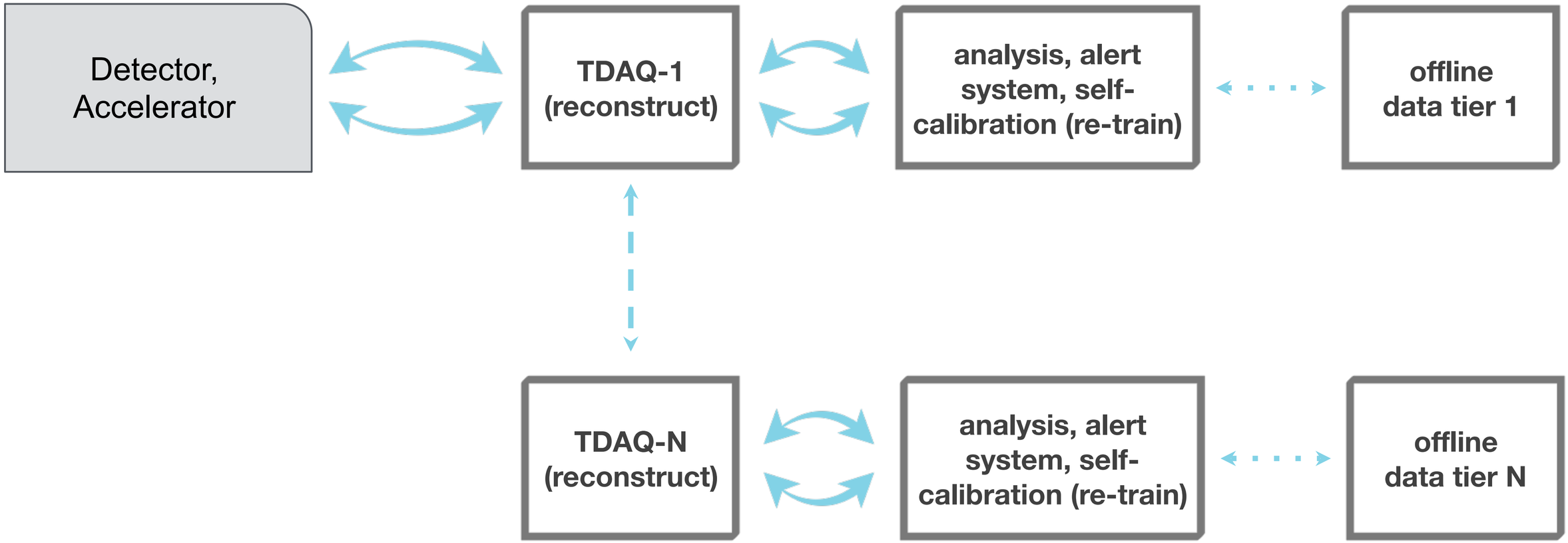
Mia Liu, Nhan Tran, Fermilab + input from many in Fast ML and broader community!

DOE Basic Research Needs Study (Community meeting for TDAQ)
December 3rd, 2019

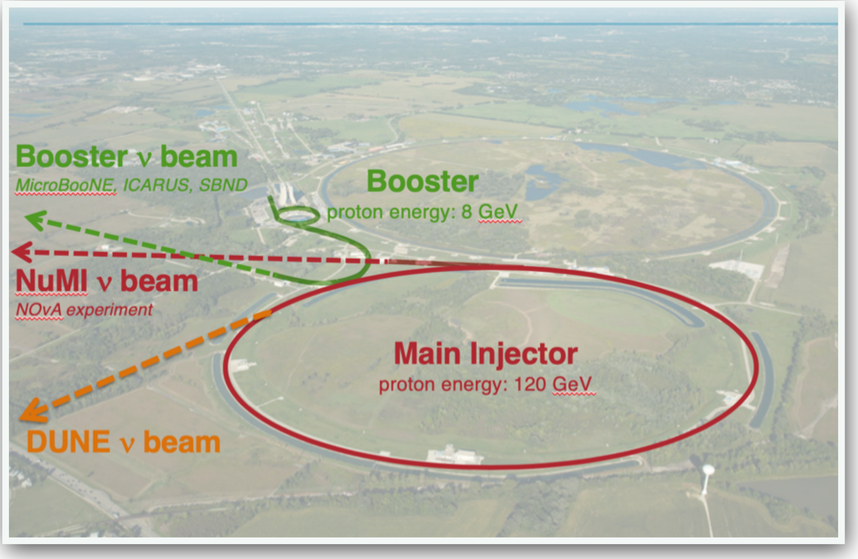
The dream TDAQ

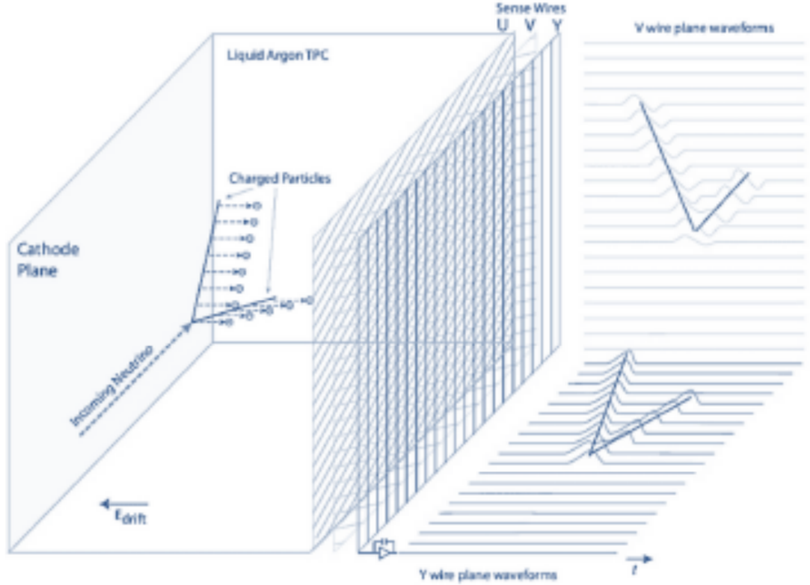
- **Powerful intelligent algorithms**
  - Sophisticated algorithms
  - Training/updating on the fly
- **Autonomous, self-calibrating**
  - Safe with minimal down-time
- **Analyze everything, no data loss**
  - Modular, multiple processing layers

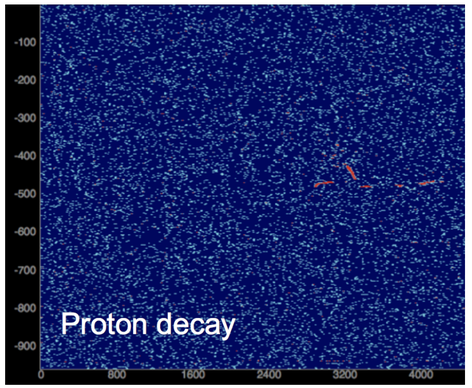
Generic system

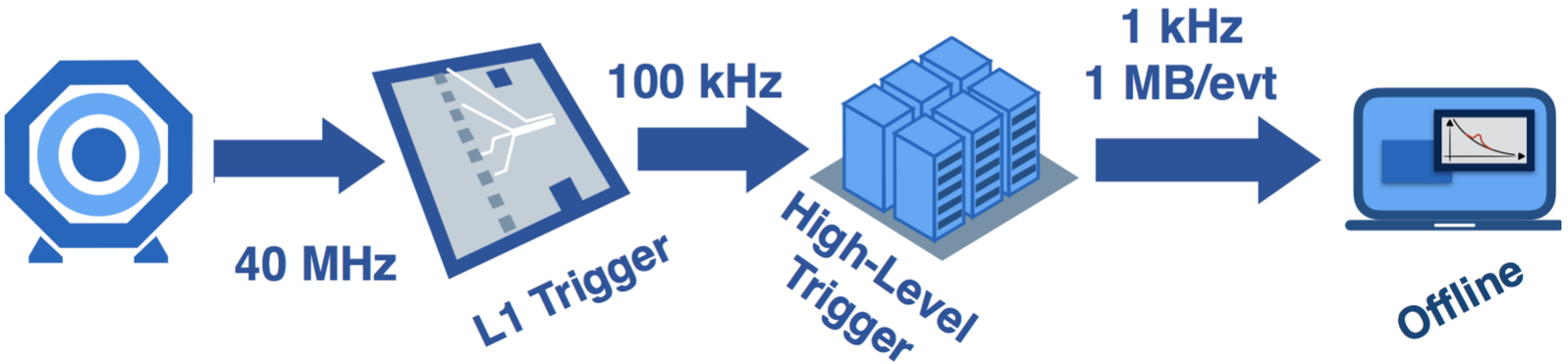


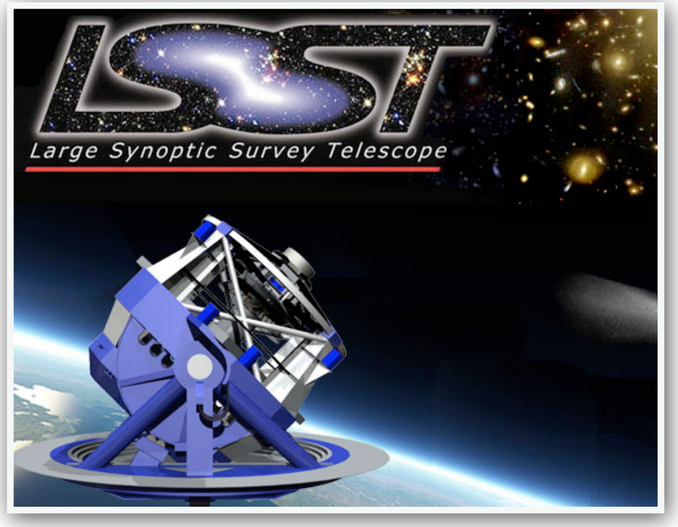
Specific systems



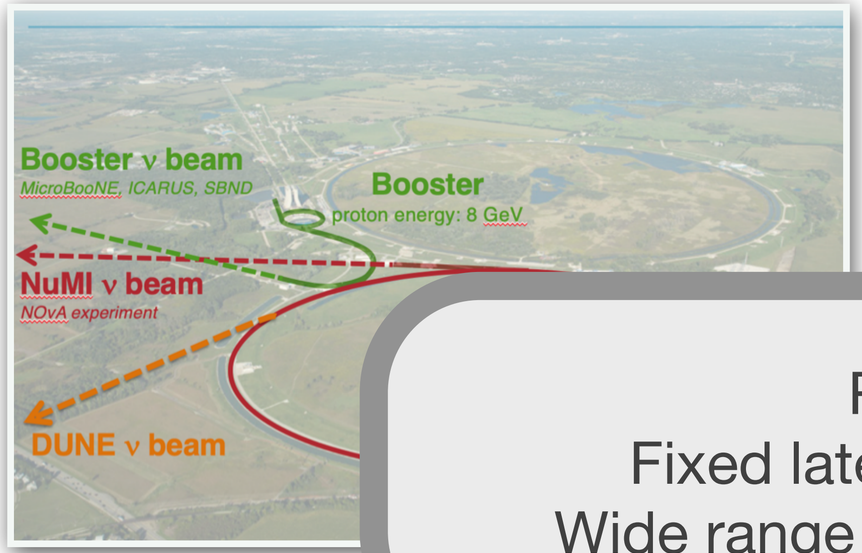


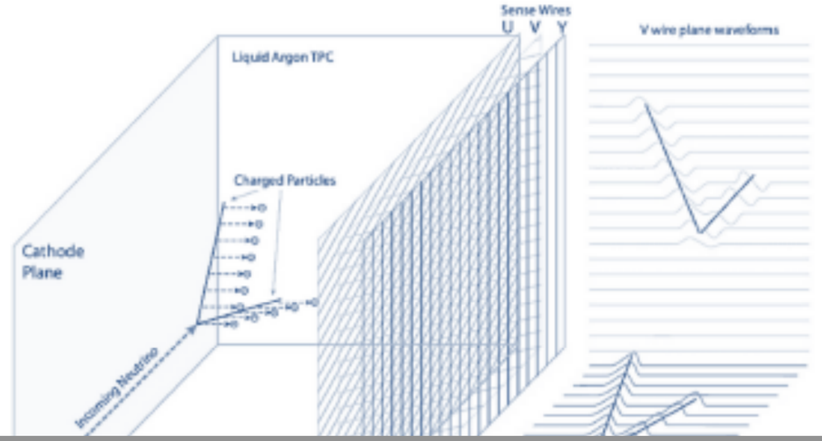


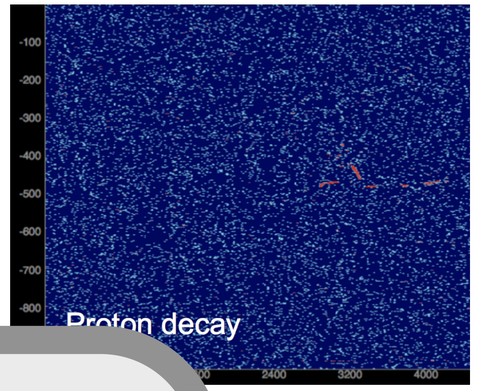




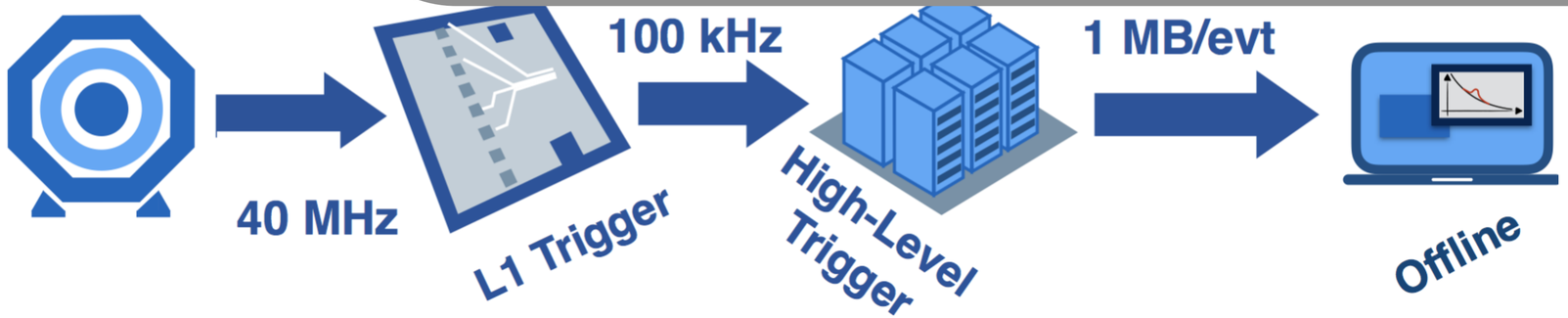
Specific systems

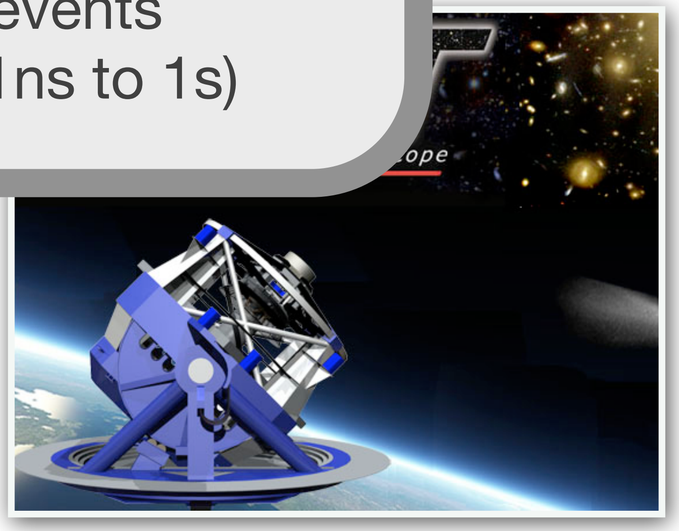




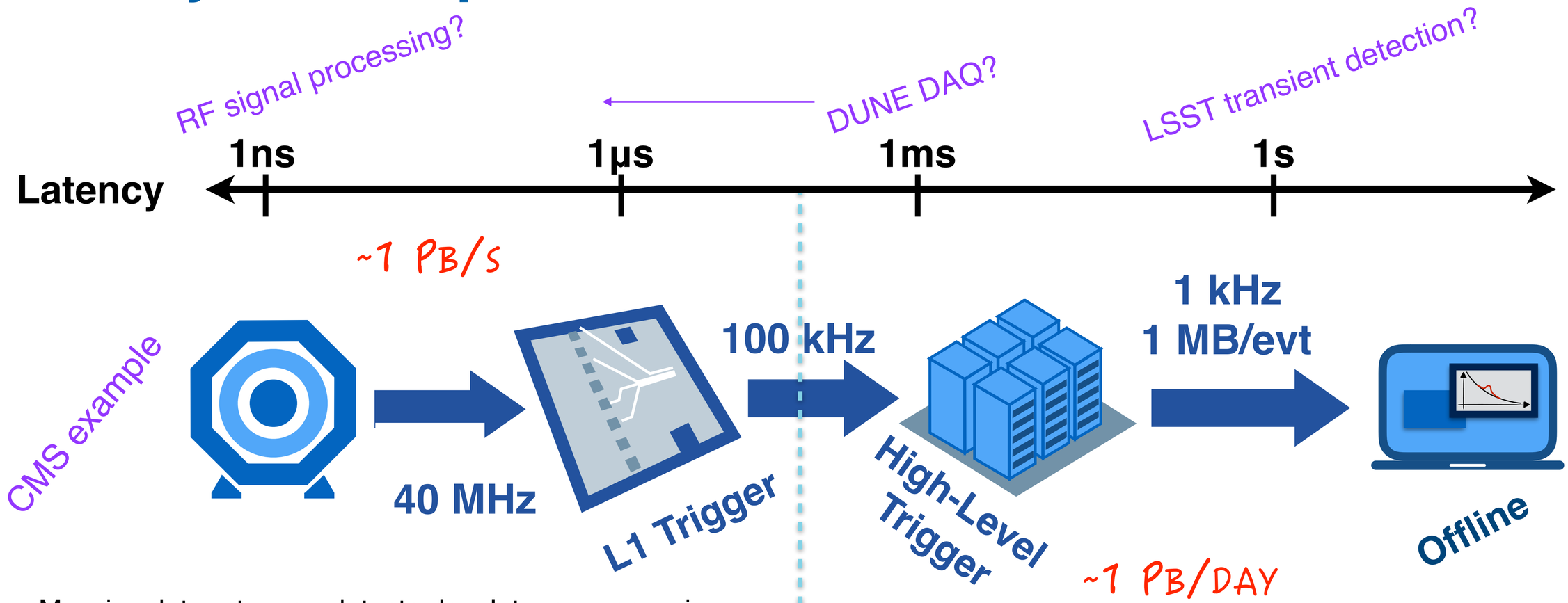


Real-time controls, trigger, alerts
Fixed latency/clock to transient/streaming events
Wide range of detector scales and timelines (1ns to 1s)





Latency landscape



Massive data rates, on-detector low-latency processing

Extreme environments: low-power, cryogenic, high-radiation

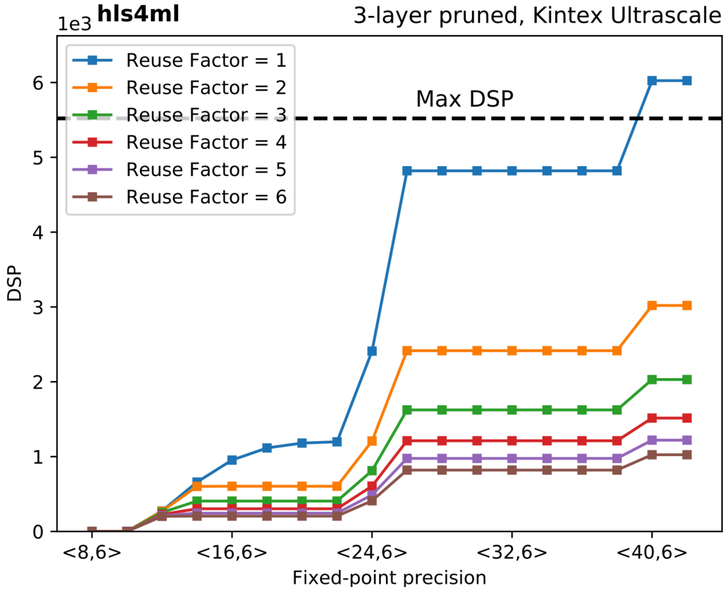
Computing challenges: Need to investigate in how to integrate heterogeneous computing platforms

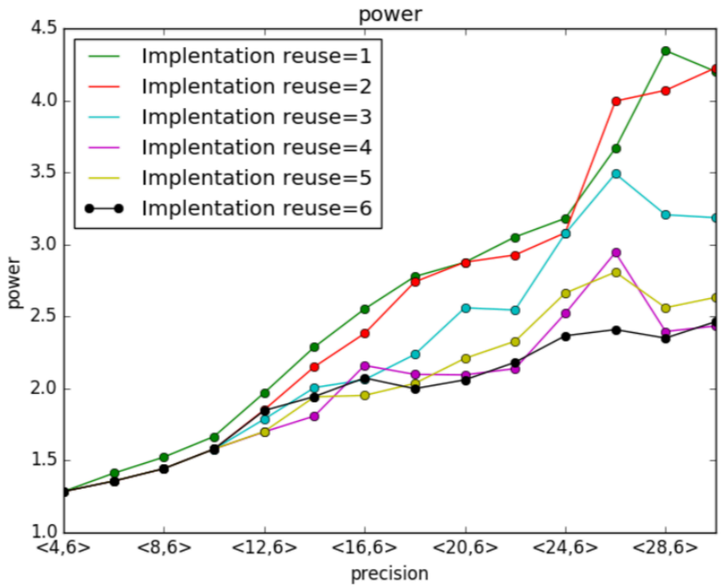
On-detector sophisticated algorithms

[https://arxiv.org/abs/1804.06913], [fastmachinelearning.org/hls4ml]

ML in the hardware trigger

- All FPGA design
  - Flexible: many algorithm kernels for processing different architectures
- Application and adoption growing across the LHC and beyond!
- Growing interest with many on-going developments
  - CNNs, Graphs, RNNs, auto-encoders, binary/ternary
  - Alternate HLS (Intel, Mentor, Cadence)
  - Co-processors, multi-FPGA
  - Intelligent ASICs
- See Phil's talk



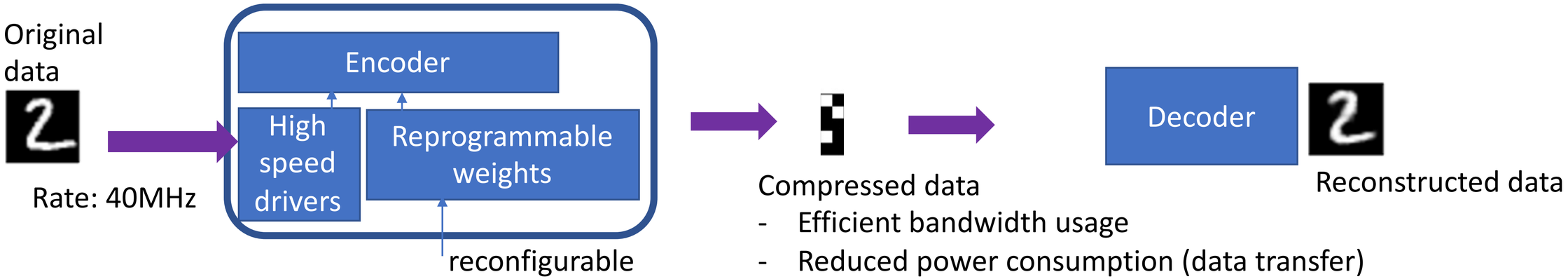


> 5000 parameter fully connected network in 100 ns

hls4…ml…4asic?

Hardware acceleration with an emphasis on co-design and fast turnaround time

First project: Autoencoder with MNIST benchmark (28 x 28 x 8-bits @ 40 MHz)

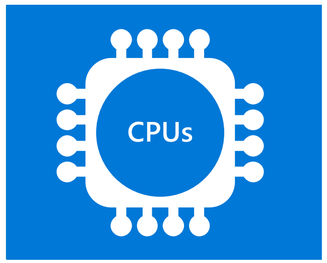


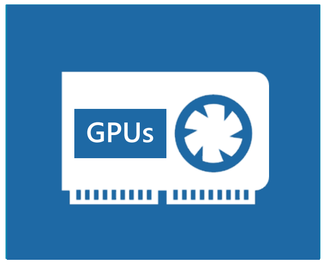
Enable edge compute : e.g. data compression
Programmable and Reconfigurable: reprogrammable weights
Hardware – Software codesign: algorithm-driven architectural approach
Optimized Mixed signal / Analog techniques: Low power and low latency for extreme environment (ionizing radiation, deep cryogenic)

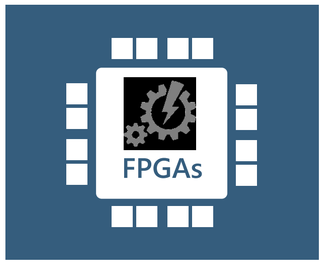
First tests of 1-layer design
Latency: 9ns
Power (FPGA, 28nm) ~ 2.5 W
Power (ASIC, 65nm) ~ 40 mW
Area = 0.5mm x 0.5mm

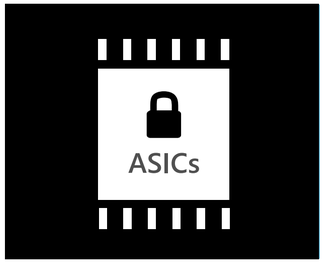
FNAL, NW, Columbia, work-in-progress

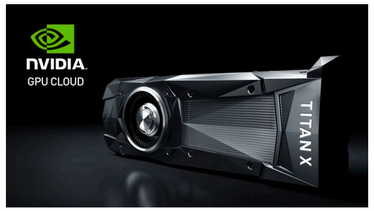
Off detector: heterogeneous computing

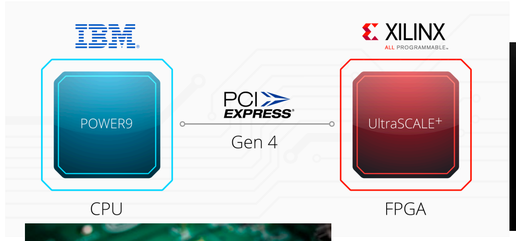


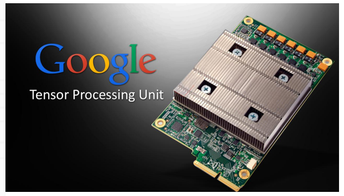


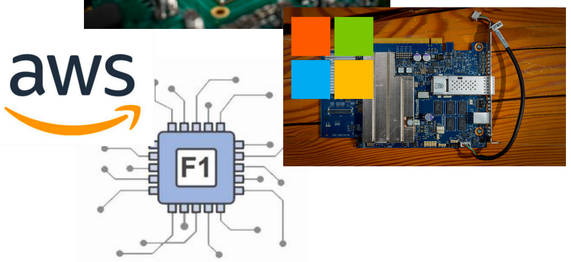








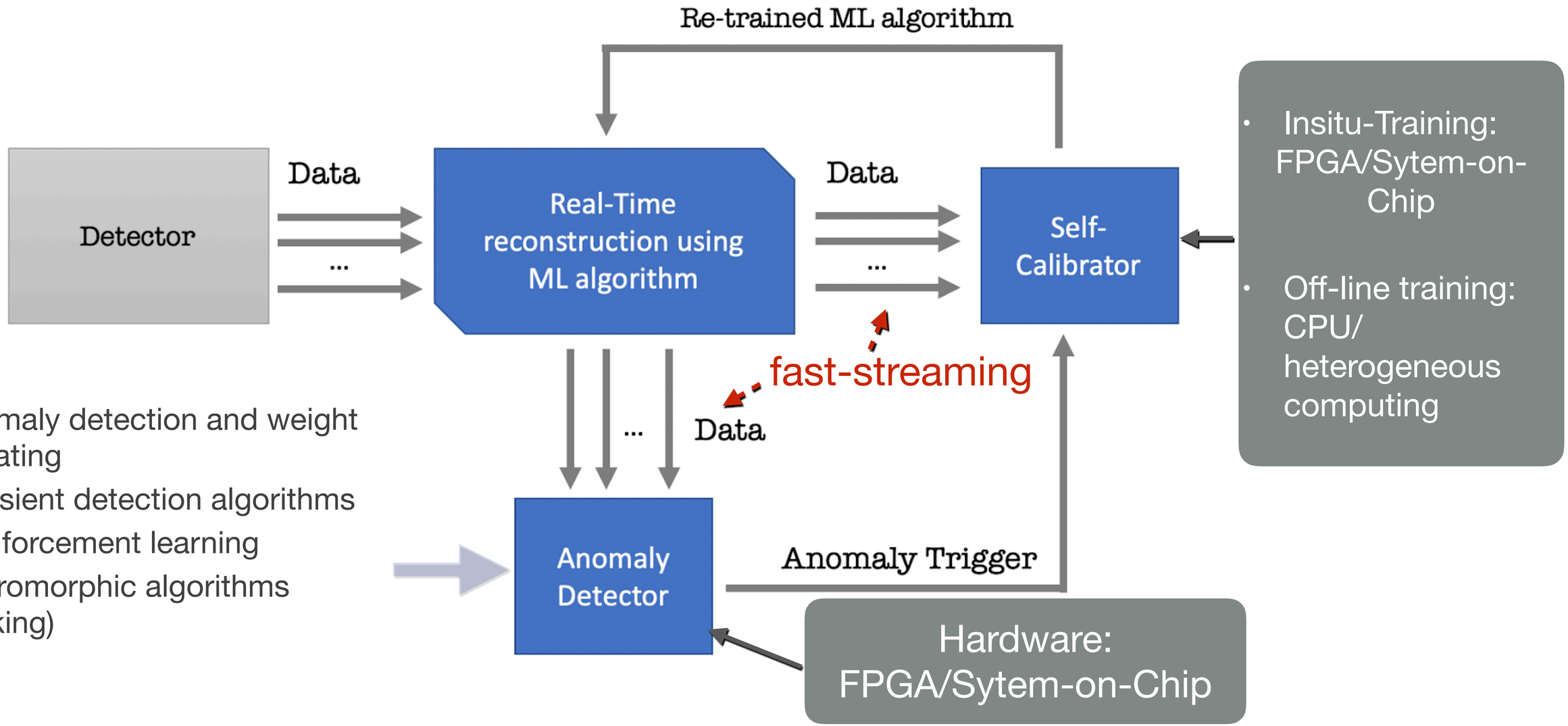




Advances in heterogeneous computing driven by **machine learning**

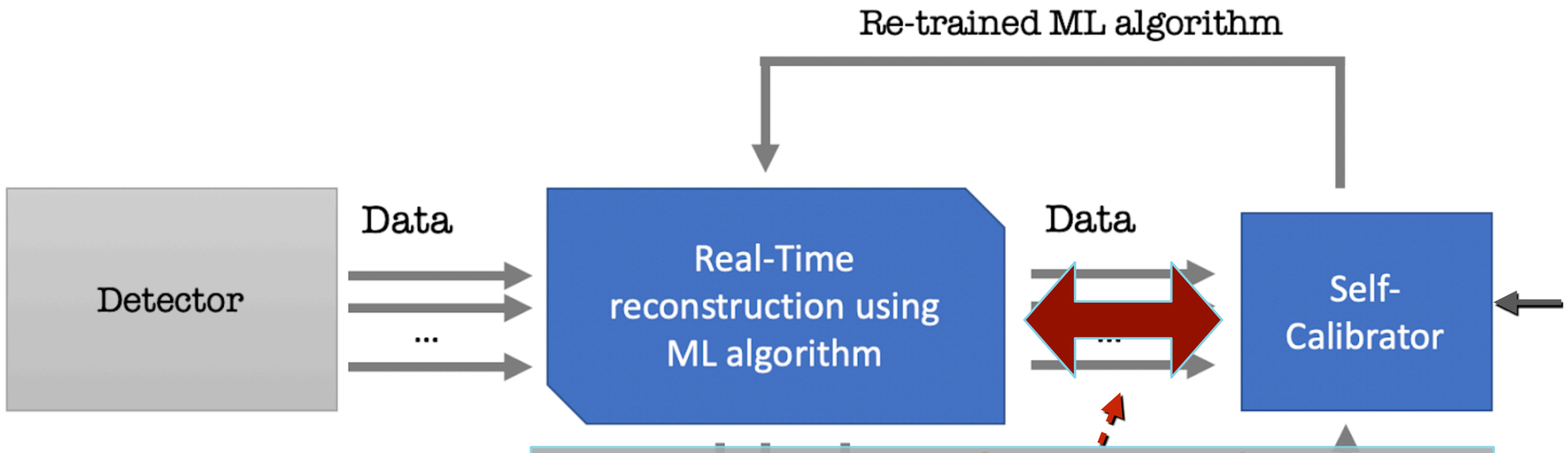
- Opportunities for deploying **accelerated heterogeneous compute** for real-time analysis
- How best to integrate into a given TDAQ workflow
  - ML/not ML
  - Service or direct connect
  - GPU, FPGA, ASIC
- Proof-of-concept for ML with FPGAs as a service, https://arxiv.org/abs/1904.08986

Autonomous, self-calibrating detector



- Anomaly detection and weight updating
- Transient detection algorithms
- Reinforcement learning
- Neuromorphic algorithms (spiking)

Autonomous, self-tuning accelerator



- Insitu-Training: FPGA/Sytem-on-Chip
- Off-line training: CPU/ heterogeneous computing

- Anomaly detection and weight updating
- Transient detection algorithms
- Reinforcement learning
- Neuromorphic algorithms (spiking)

For accelerator applications, constant tuning/feedback loop required

Tools for dream

New algorithms

Electronics hardware and infrastructure

Systems designed for operations and control

- **Powerful intelligent algorithms**
  - FPGAs designed for ML and vice versa
  - Opportunities for heterogeneous hardware (e.g. Versal)
  - Push up to the frontest end (ML in ASIC, reconfigurable weights)
  - New types of algorithms beyond classification & regression
- **Autonomous, self-calibrating**
  - Automation for (a) when conditions have changed (b) what actions to take
  - Fast DAQ paths with deep buffers for monitoring individual channels, how to deal with different time scales?
  - Training and recalibration “offline-system” (GPU…) or small-scale in situ (ARM processor, in FPGA)
- **Analyze everything, no data loss**
  - Modular, portable, multiple processing layers
  - Streaming fast analysis - accessible programming paradigms; SoC R&D
  - Data storage - Affordable, new/different storage technologies for persistent (parked) datasets

Extra