Dec 10, 2019

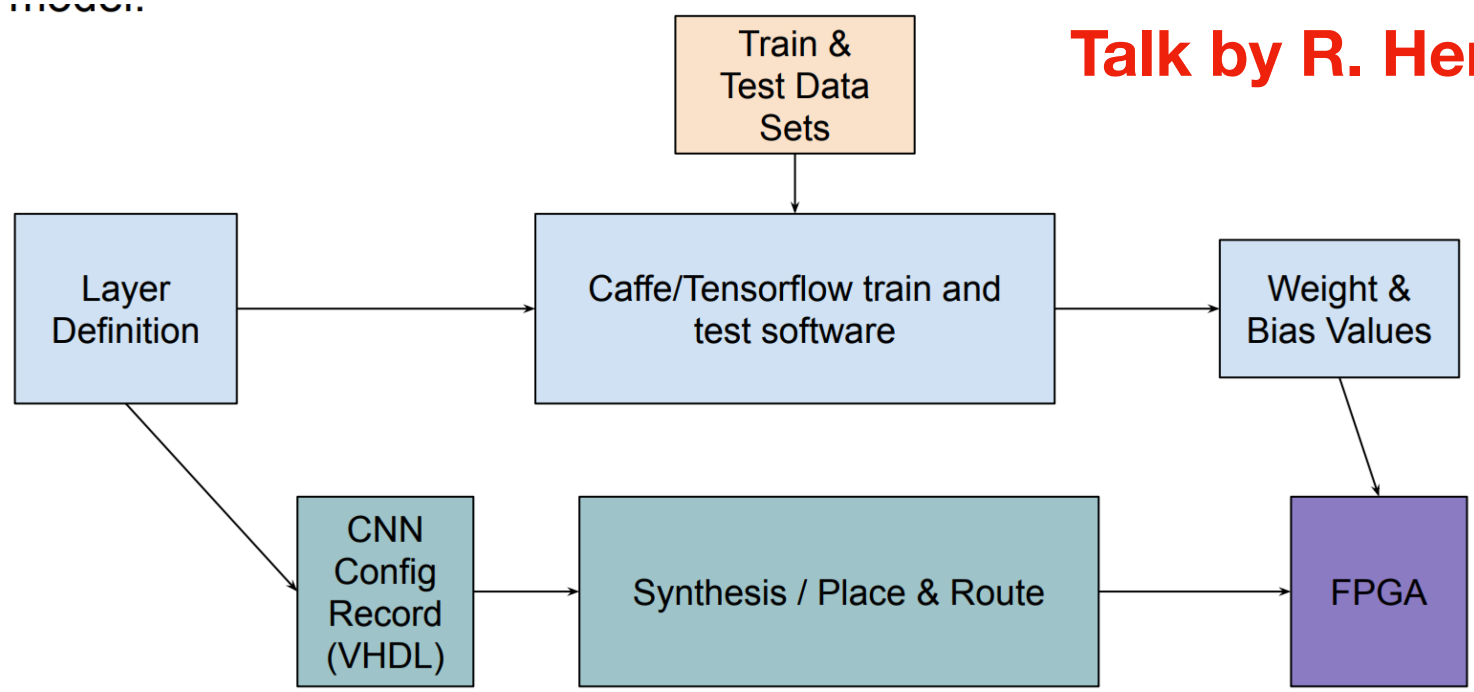
Trigger, DAQ and Machine Learning

CPAD 2019

Verena Martinez Outschoorn and Isobel Ojalvo

DAQ Concepts Edge ML

Talk by R. Herbst

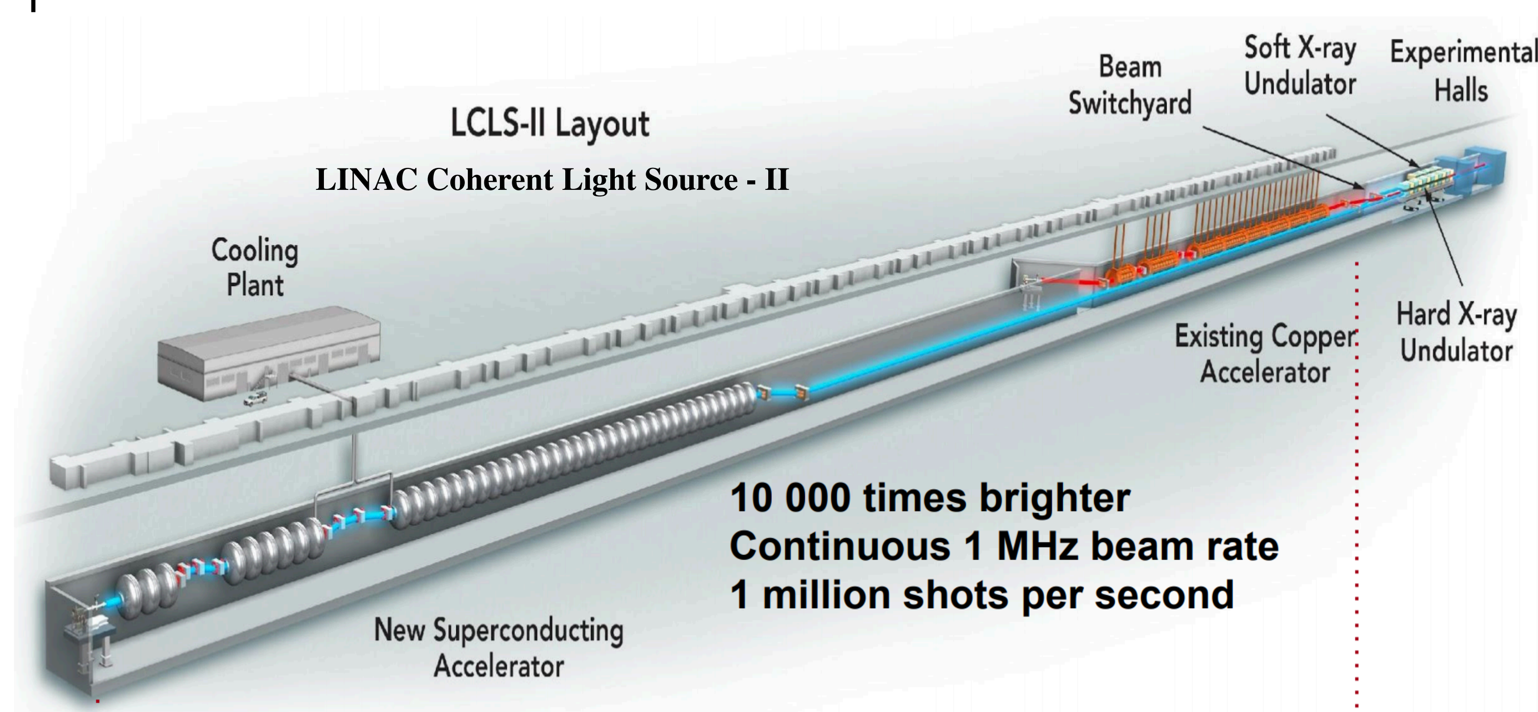


Framework to provide a configurable VHDL based **inference engine**

- Layer Types supported: Convolution, Pool & Full

Developed as a proof of concept but applicable for many HEP experiments

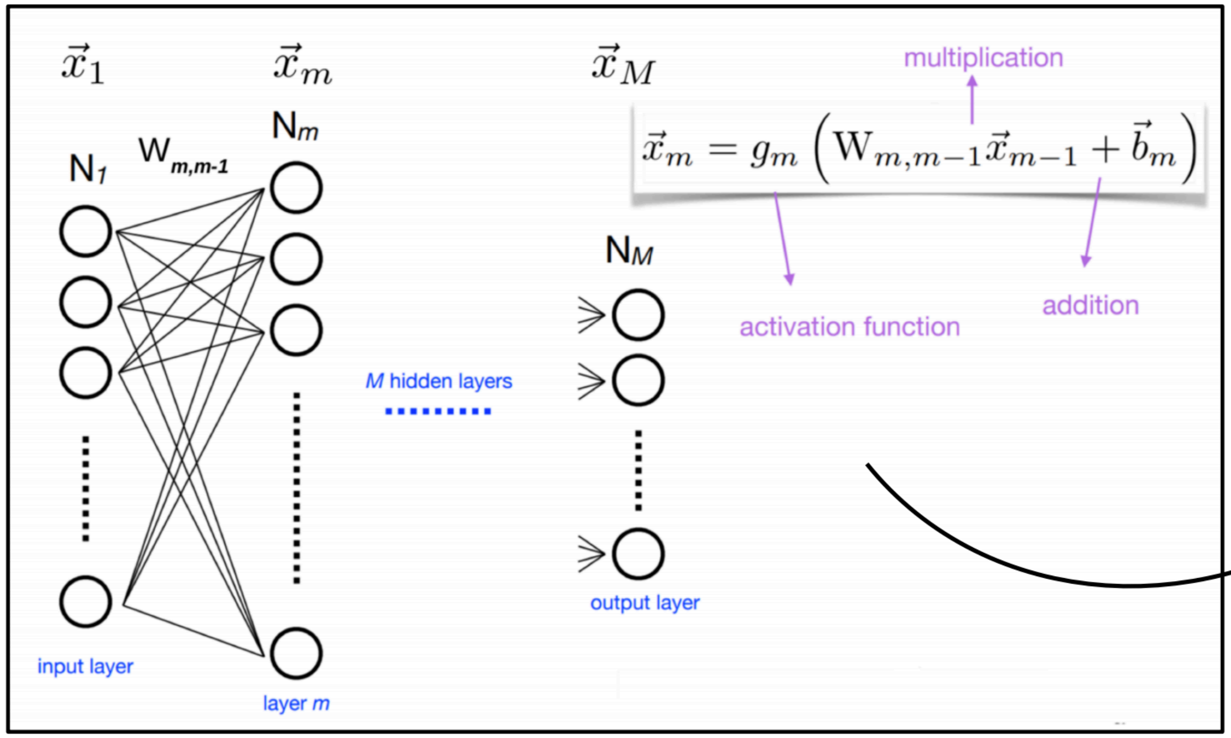
developed for
Linac Coherent Light Source II

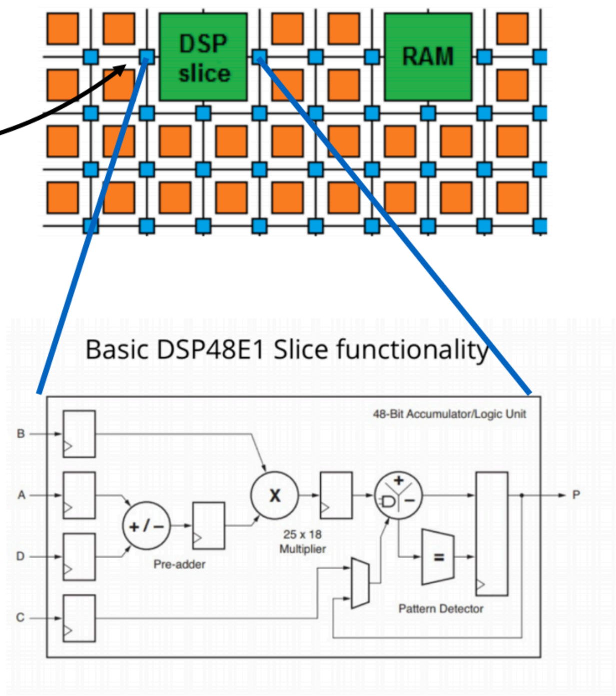


DAQ Concepts Edge ML

Talk by S. Jindariani

Typical NN operations:





Maps naturally into the functionality of DSP slices available in modern FPGAs

Device	# of DSPs
Kintex-7 325T	840
Virtex-7 690T	3600
Kintex UltraScale KU115	5500
Virtex UltraScale+ VU9P	6800

- DNN
  - Including support for large layers
- Binary and Ternary DNN
  - Low precision (1 or 2 bit) weights performance
  - Implemented in LUTs
- Conv1D and Conv2D (small)
  - Large Convs and Binary/Ternary coming soon
- Other features
  - Batch normalization
  - Various activation functions
  - Tools for comparing C and RTL simulation results

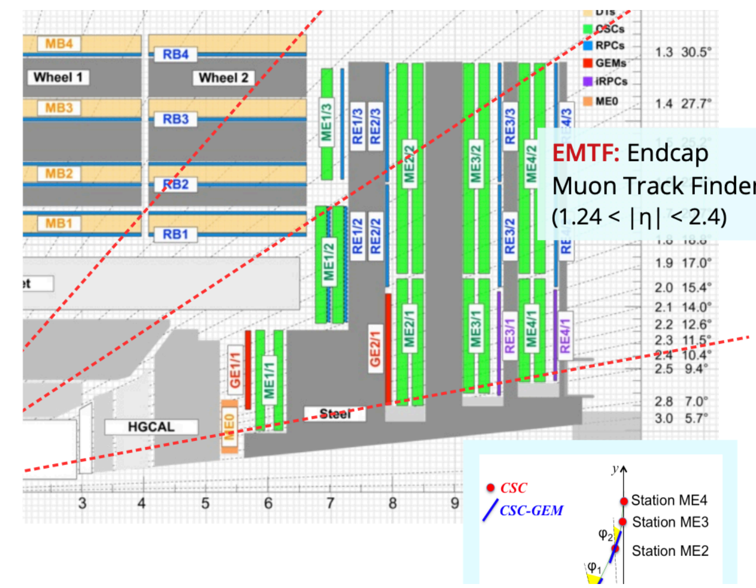
Machine Learning based algorithm for reconstructing prompt and displaced muons at Level-1 in CMS detector

Talk by J. F. Low

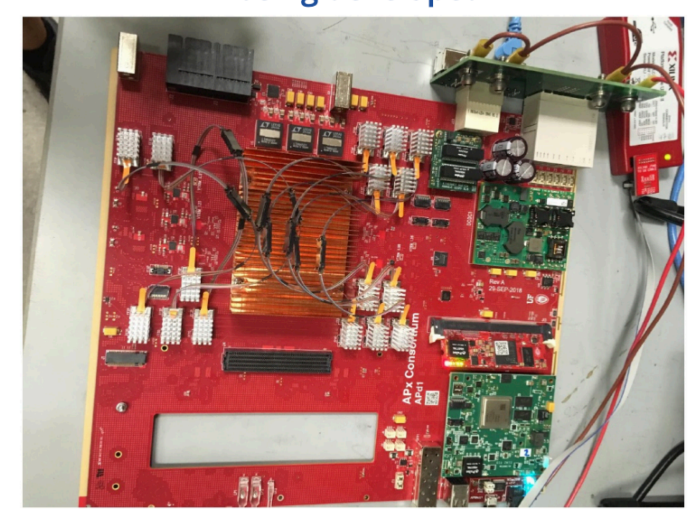
At CMS L1 muon transverse momentum assignment has used ML for inference since LHC Run-1

- Traditionally Used LUTs
- NN Inference also should be possible!
  - Trade off in FPGA resource usage

For **Phase-2 the EMTF algorithms will evolve** to incorporate **new detectors**, **pile up**, **maintain efficiency**, also **incorporate displaced Muon ID**



APd board being developed



Displaced EMTF++: NN performance

HLS estimates

== Utilization Estimates

* Summary:

Name	BRAM_18K	DSP48E	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	6	-
FIFO	-	-	-	-	-
Instance	39	2420	69109	90580	-
Memory	-	-	-	-	-
Multiplexer	-	-	-	1404	-
Register	0	-	4280	32	-
Total	39	2420	73389	92022	0
Available	4320	6840	2364480	1182240	960
Utilization (%)	~0	35	3	7	0

Looking into the Phase-2 APd board [3] with Virtex US+ VU9P FPGA, which has 3X more LUT & FF, and 2X more DSP.

NN should comfortably fit in the VU9P (DSP usage is 35%)

32 clk @ 333 MHz ≈ 100 ns latency

hls4ml

Detect New Physics with Deep Learning

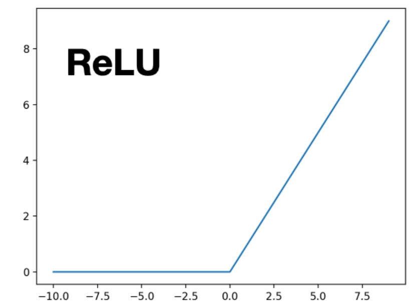
Talk by Z. Wu

Example AE Model

- Train with simulated ZeroBias event at 200 pileup
- Use simulated Puppi Jet/MET/MHT inputs (18 inputs) with preprocessing

- Activation function: ReLU
- Loss function: L1Loss
- Training - validation ratio : 0.8
- Number of epochs: 100-200 epochs
- Number of layers: 8 layers

$$\ell(x,y) = L = \{l_1,\dots,l_N\}^\top, \quad l_n = |x_n - y_n|,$$

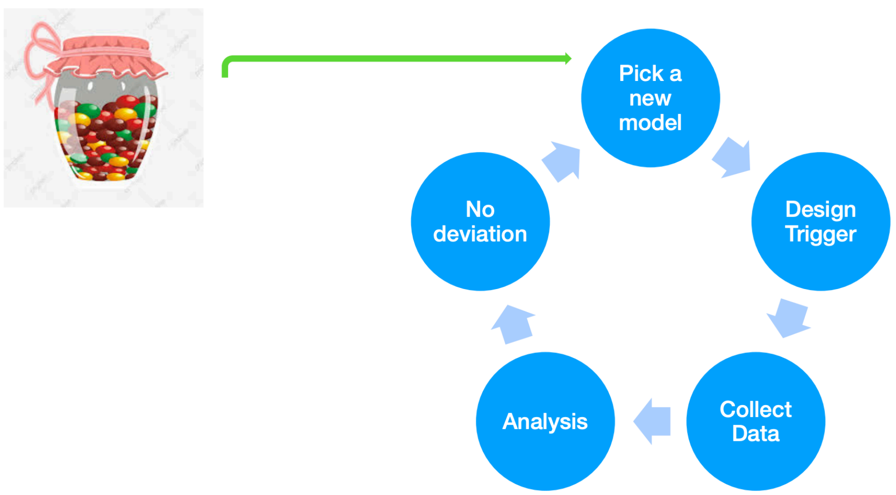


- Model is designed with simplicity for firmware implementation and resource/latency requirement

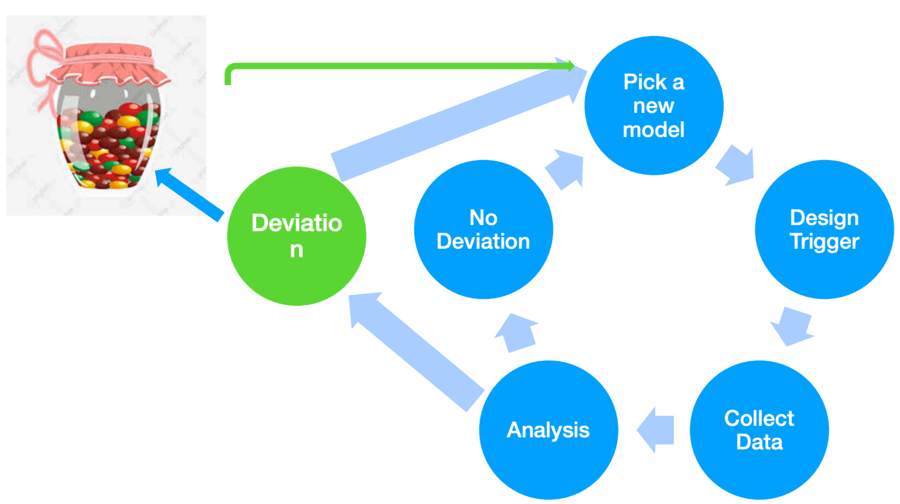
Illustration by Jeff Lewonczyk

Not to claim a discovery!
But to give an idea of what Exotic Signals to integrate into our trigger menus

Traditional Workflow of Searches



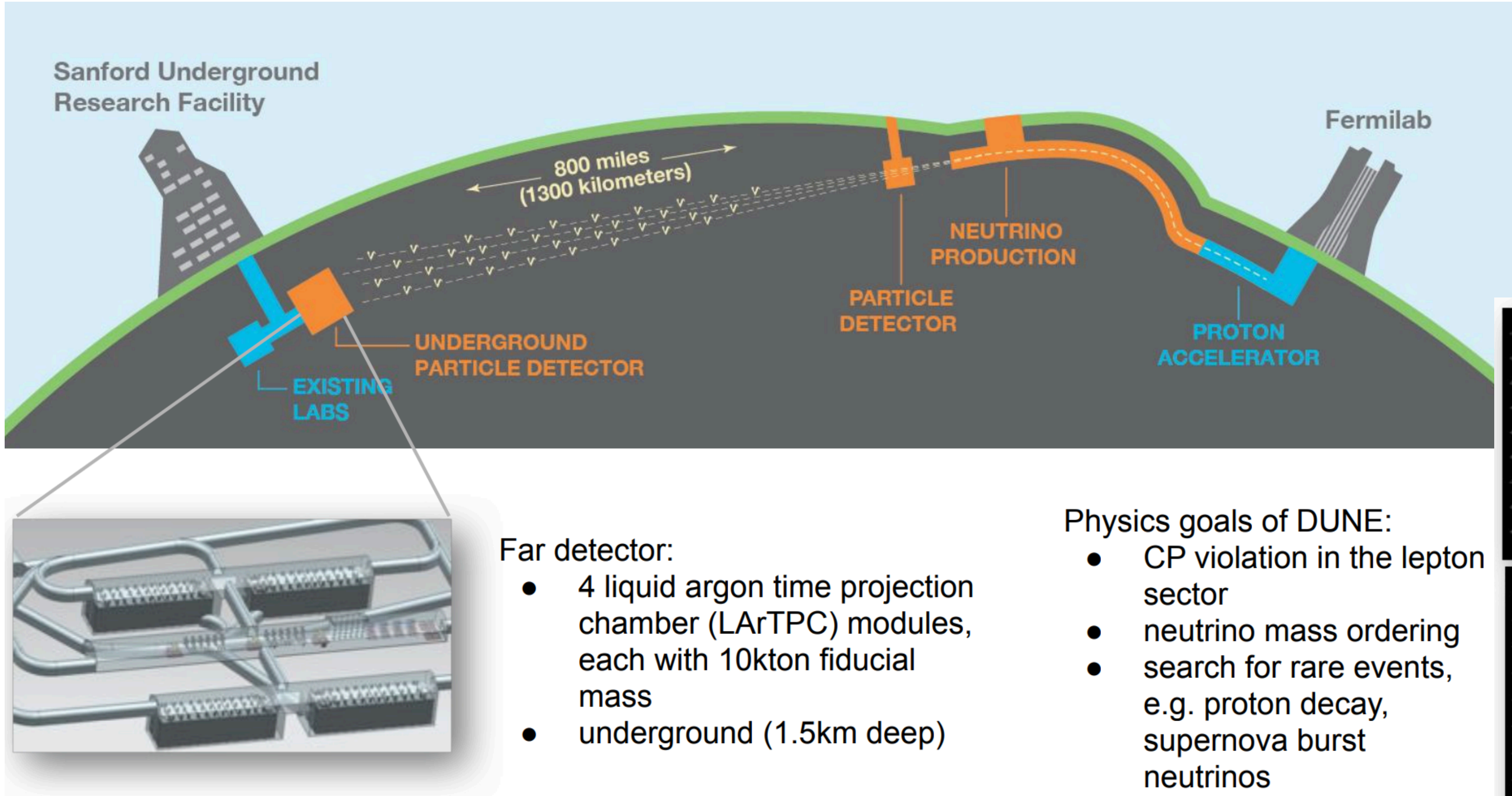
Auto Encoder Workflow of Searches



CPAD Trigger/DAQ/ML Dec 10, 2019

Machine Learning-based Trigger for DUNE

Talk by G. Ge

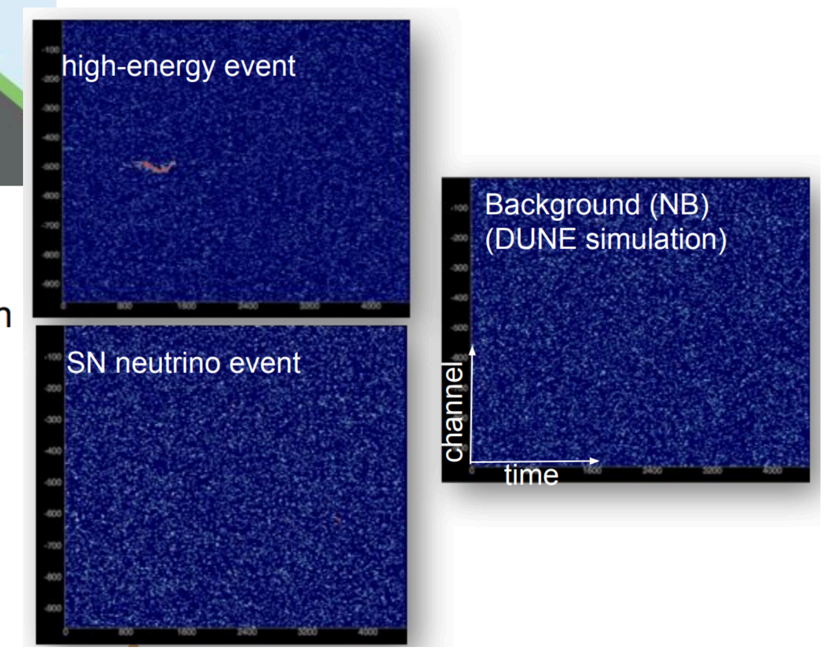


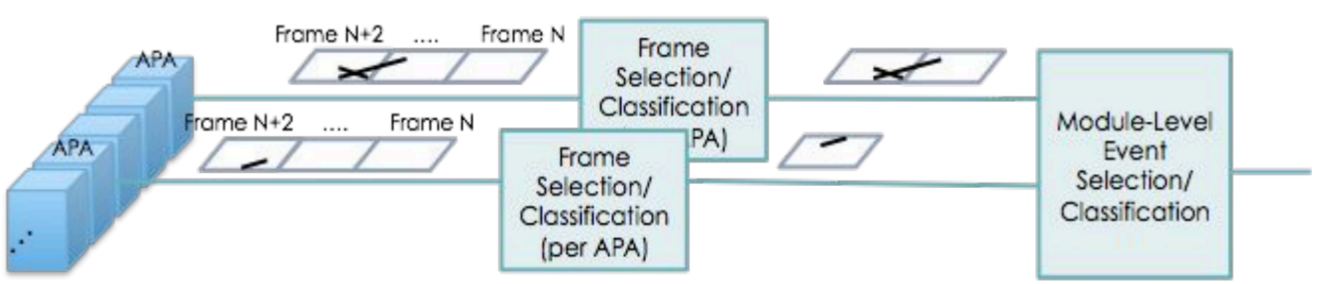
Far detector:

- 4 liquid argon time projection chamber (LArTPC) modules, each with 10kton fiducial mass
- underground (1.5km deep)

Physics goals of DUNE:

- CP violation in the lepton sector
- neutrino mass ordering
- search for rare events, e.g. proton decay, supernova burst neutrinos





1. Low-level:
CNN-based APA-frame selection and reweighting

2. Module-level:
APA-frame coincidence across module and over 10 seconds

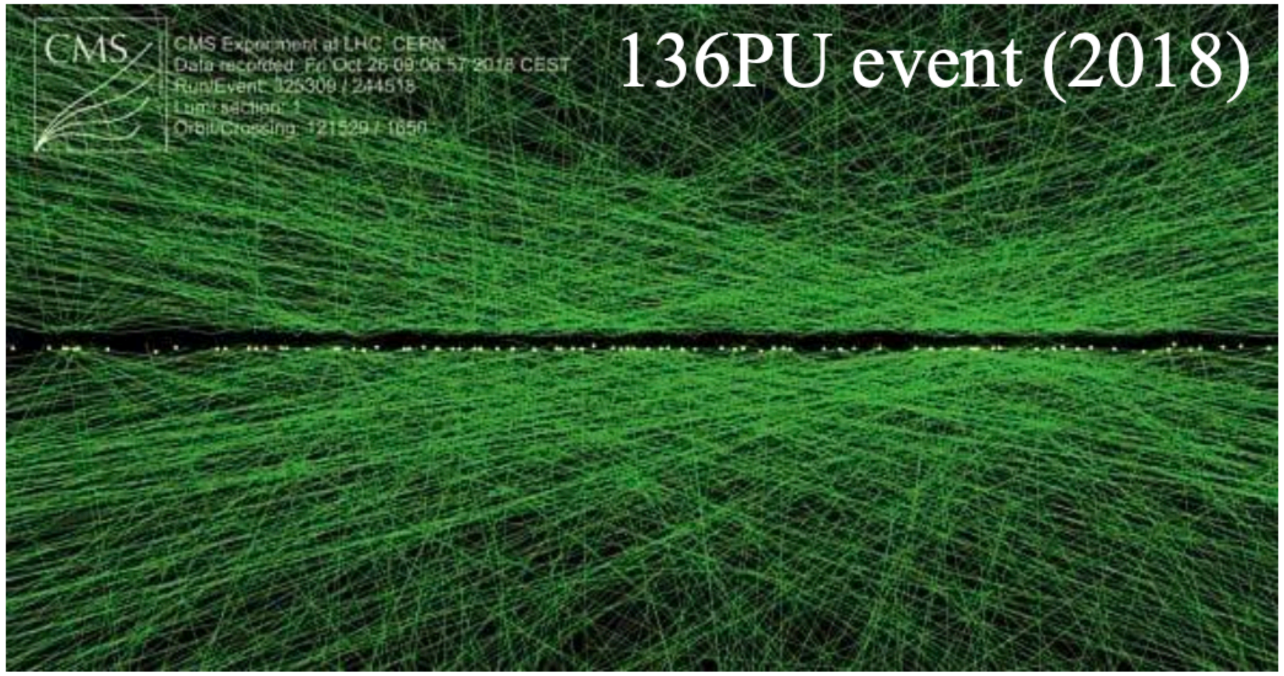
Performance and power analysis of CNN_s:

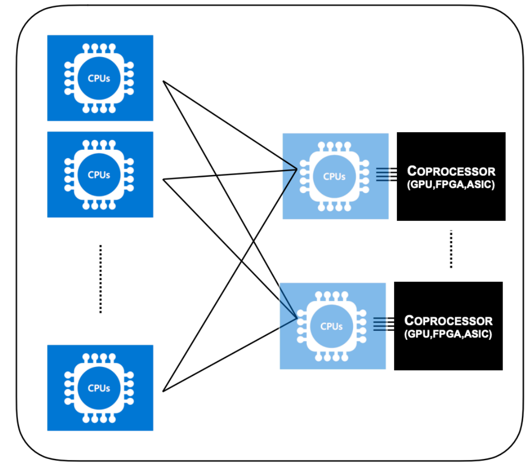
Platform	Model	Time (s)	Power (W)	Energy Efficiency (img/s/W)
ARM C-A53	CNN_s	0.0855	2.871	4.074
FPGA	CNN_s	0.0511	1.110	17.630

**G. Karagiorgi, Y. Jwa, G. di Guglielmo, L. Carloni; DOI: 10.1109/NYSDS.2019.8909784*

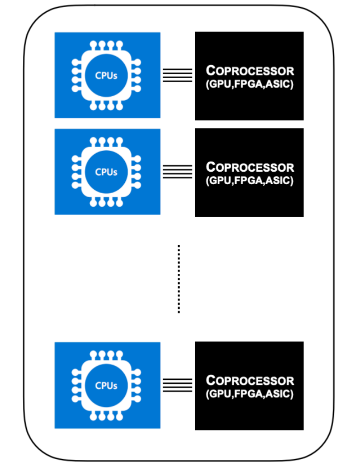
Accelerated Machine Learning Inference as a Service

Talk by N. Tran



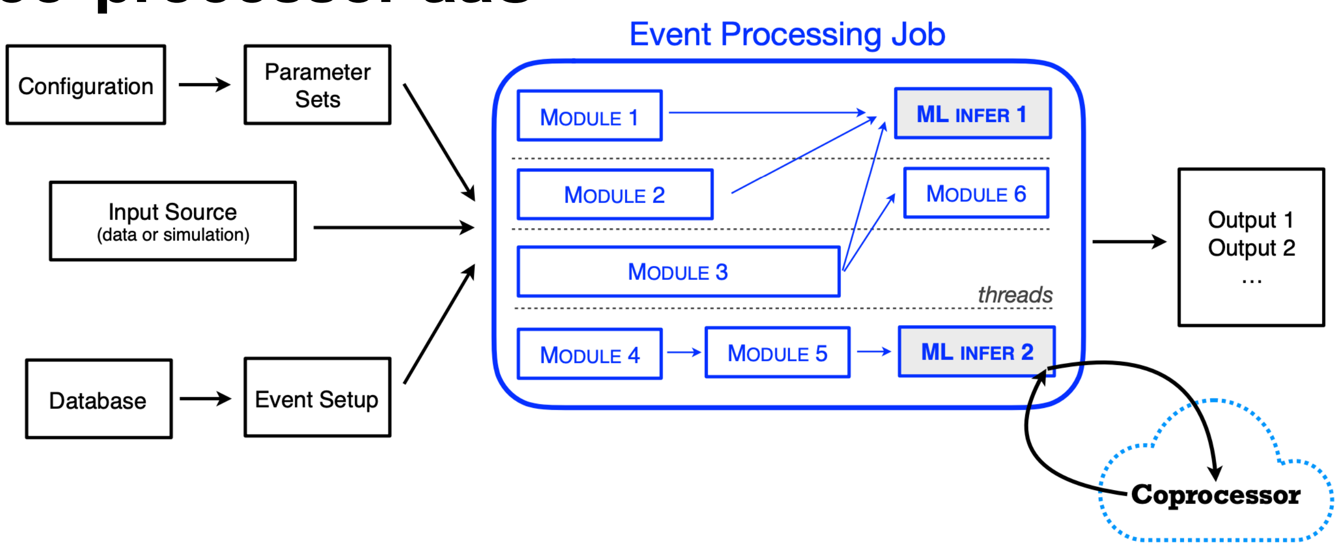


Pros:
scalable algorithms
scalable to the grid/cloud
heterogeneity (mixed hardwares)



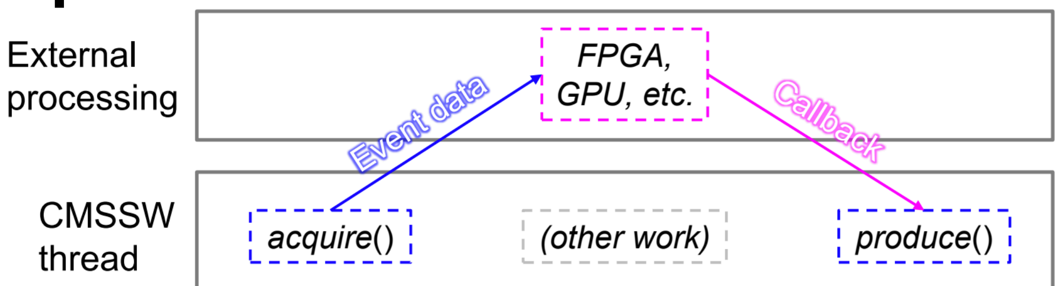
Pros:
less system complexity
no network latency

co-processor aaS



	HCal Reco Network	Resnet-50 (Top tag) Network
CPU (single-thread)	67 inf/s	0.6 - 2 img/s (depends on CPU)
GPUaaS w/TensorRT	333 inf/s (batch 16000)	140 img/s (batch 1) 667 img/s (batch 32)
FPGA (batch 1)	500 inf/s (batch 1)	660 img/s (Brainwave, aaS)

co-processor aaS



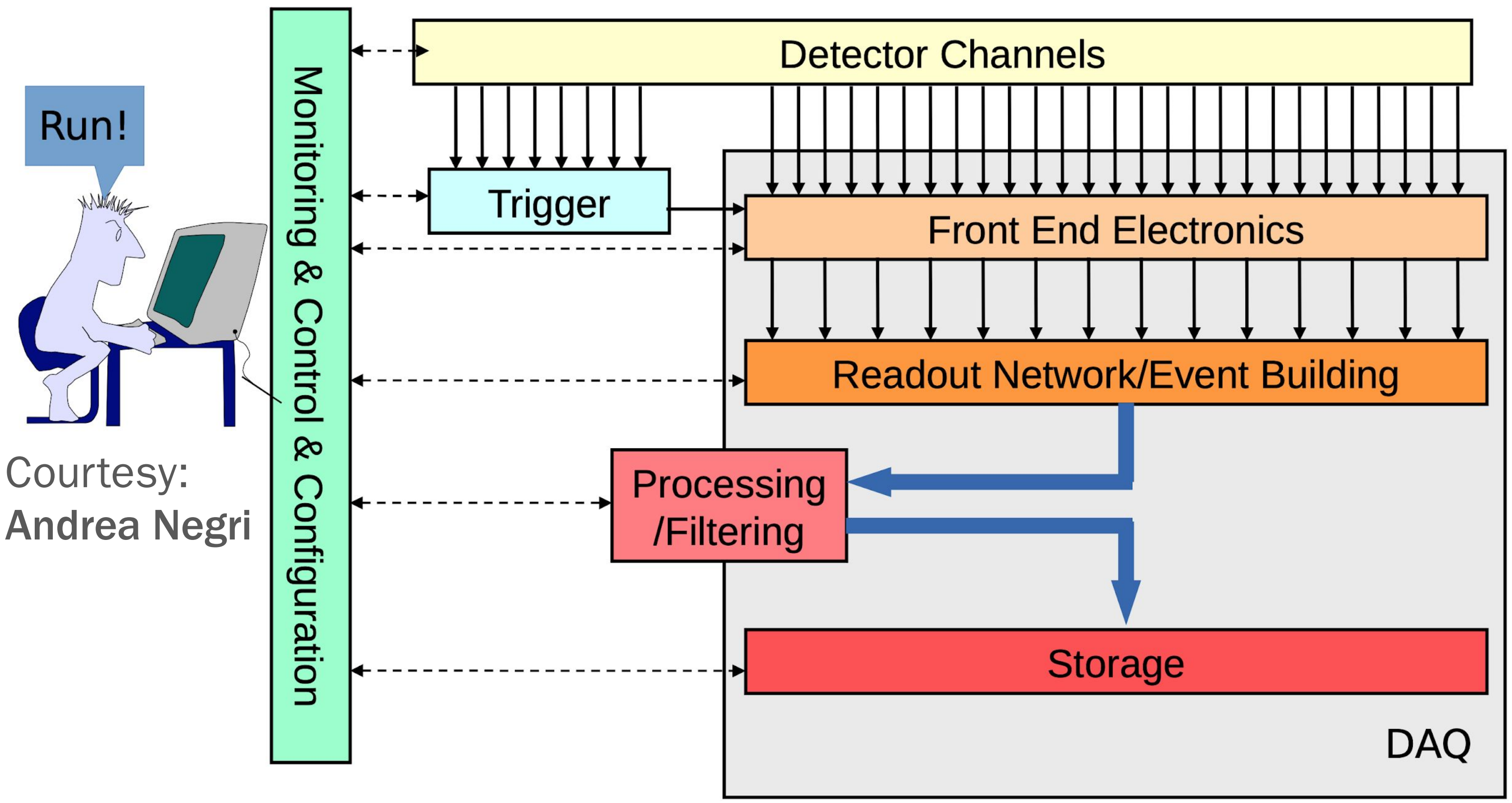
SONIC

Services for **O**ptimized **N**etwork **I**nference on **C**oprocessors

CPAD Trigger/DAQ/ML Dec 10, 2019

Overview of Trigger & DAQ Systems

Talk by K. Chen



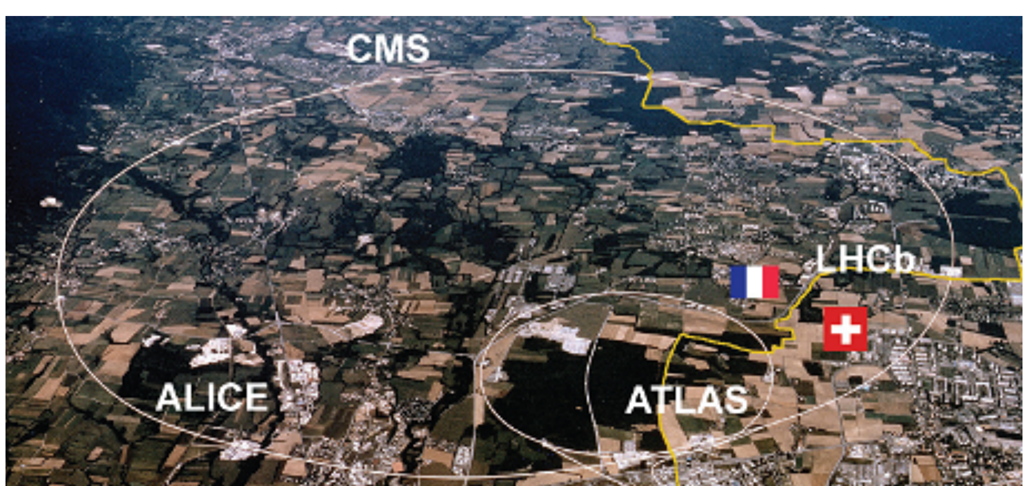
Courtesy: **Andrea Negri**

- Triggered readout
  - ATLAS
  - CMS
  - ALICE
- Streaming readout
  - LHCb (Run-3)
  - EIC (in R&D)
- Hybrid readout
  - sPHENIX
  - ProtoDUNE-SP
  - SBND
  - DUNE

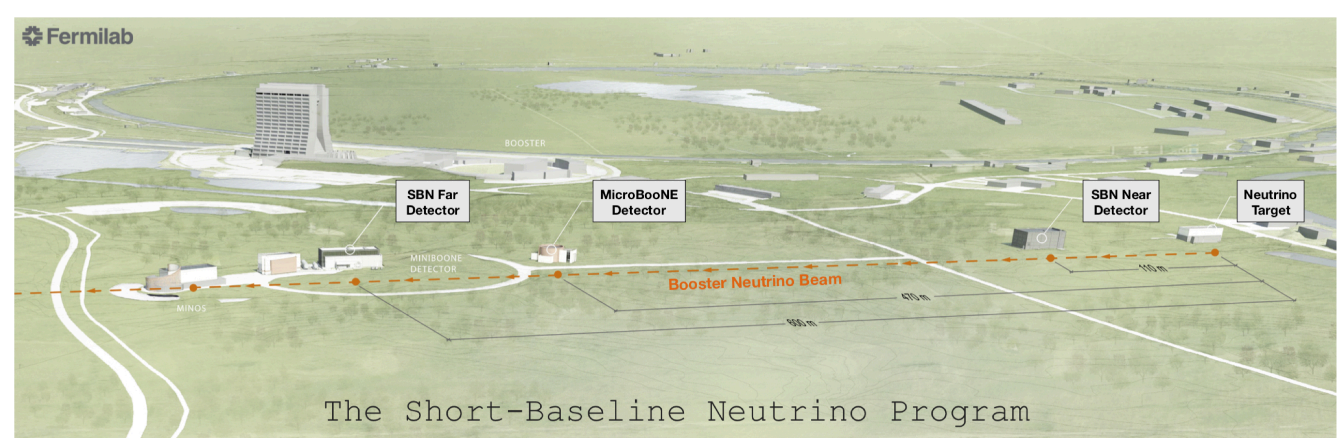
Intensity Frontier

Energy Frontier

Energy Frontier



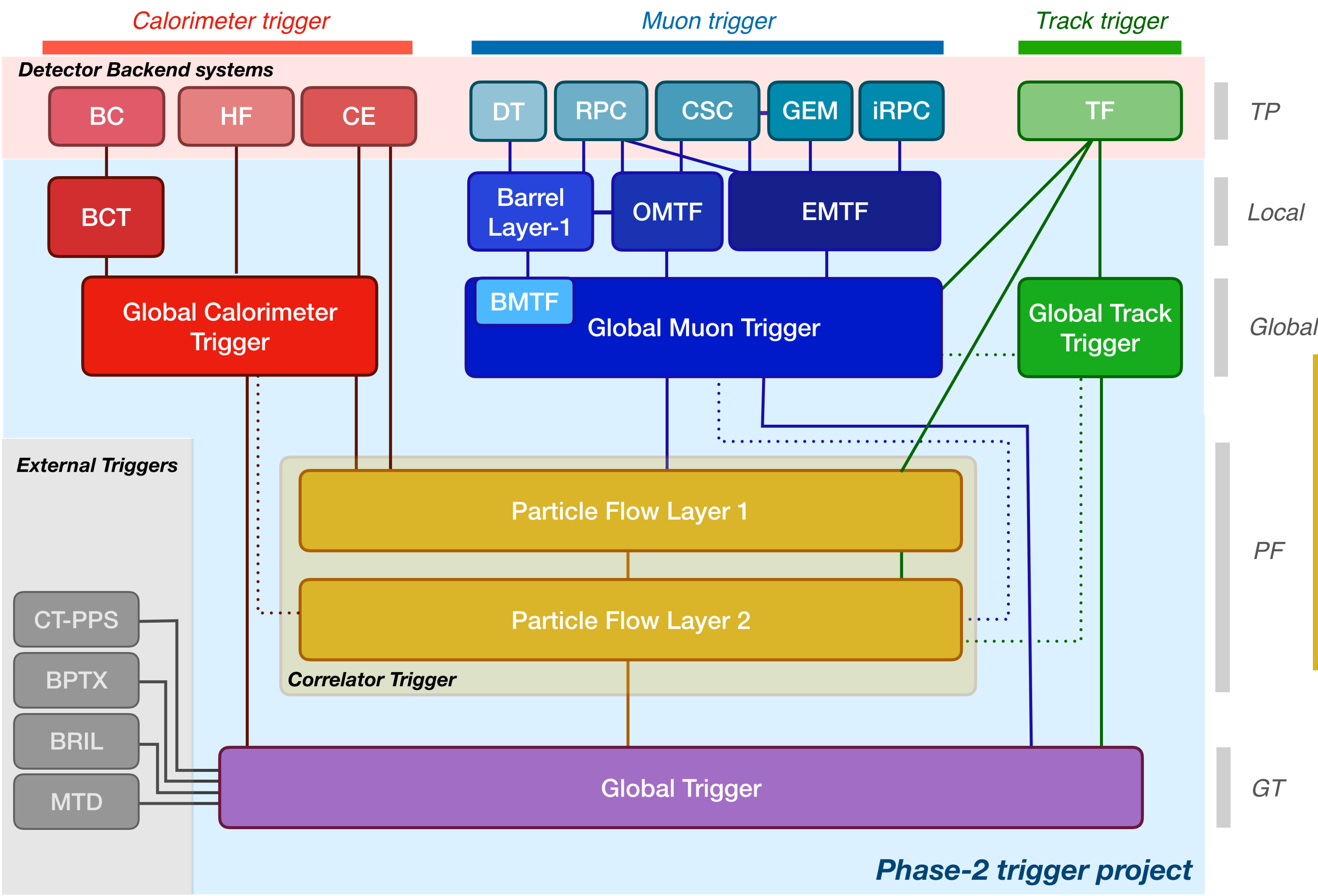
Intensity Frontier



Triggered Readout - CMS

Talk by C. Herwig

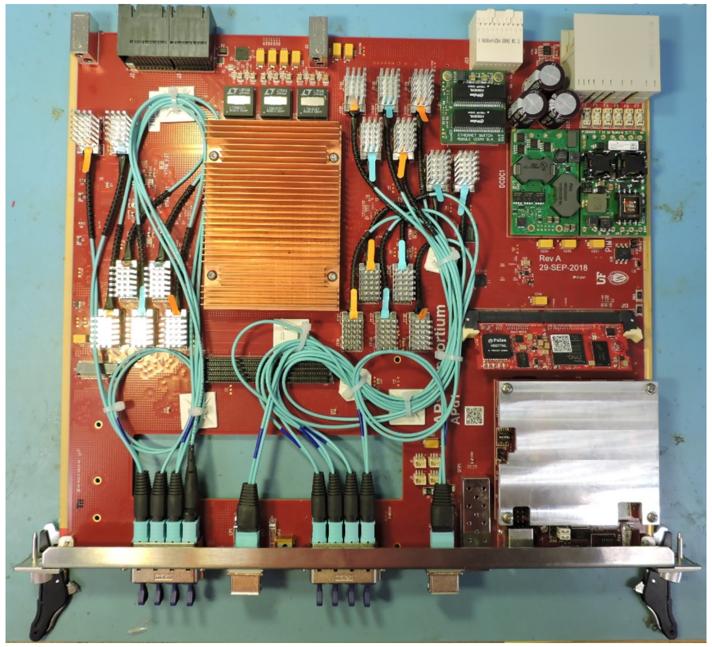
Example of CMS Phase 2 Architecture (for HL-LHC)



Combine detailed **Calorimeter** & **Muon** Information with

track trigger at L1, p_T>3-4 GeV, Vertices

Implement offline-like algorithms
Layer 1 → run "particle-flow candidate" reconstruction & pileup
Layer 2 → run algorithms on candidates



Sophisticated algorithms to combine information from all sub detectors at 40MHz
Algorithms with latency of O(100ns) implemented in FPGAs using ATCA hardware

Similar strategy pursued by ATLAS

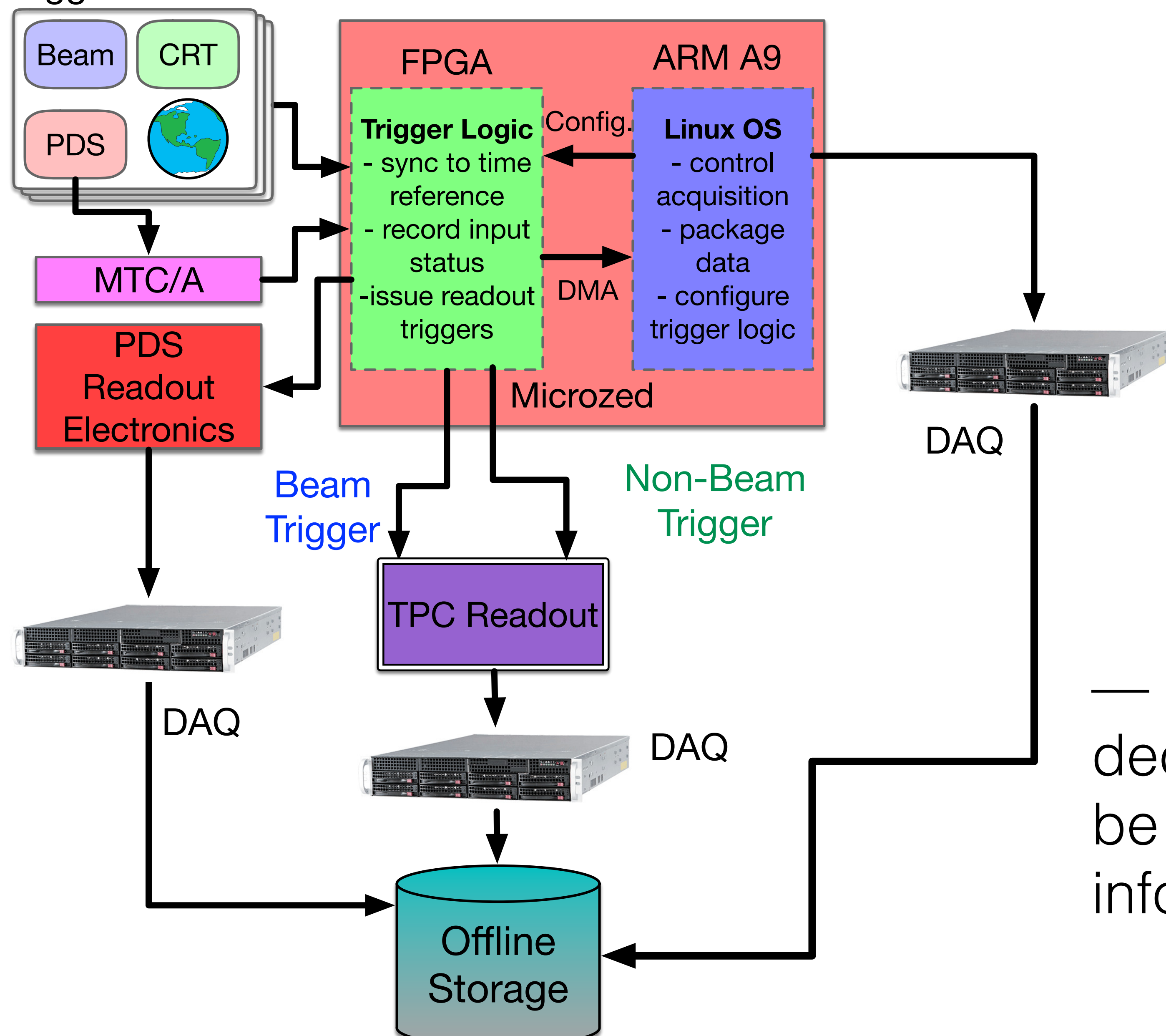
Triggered Readout - SBND

Talk by D. Rivera

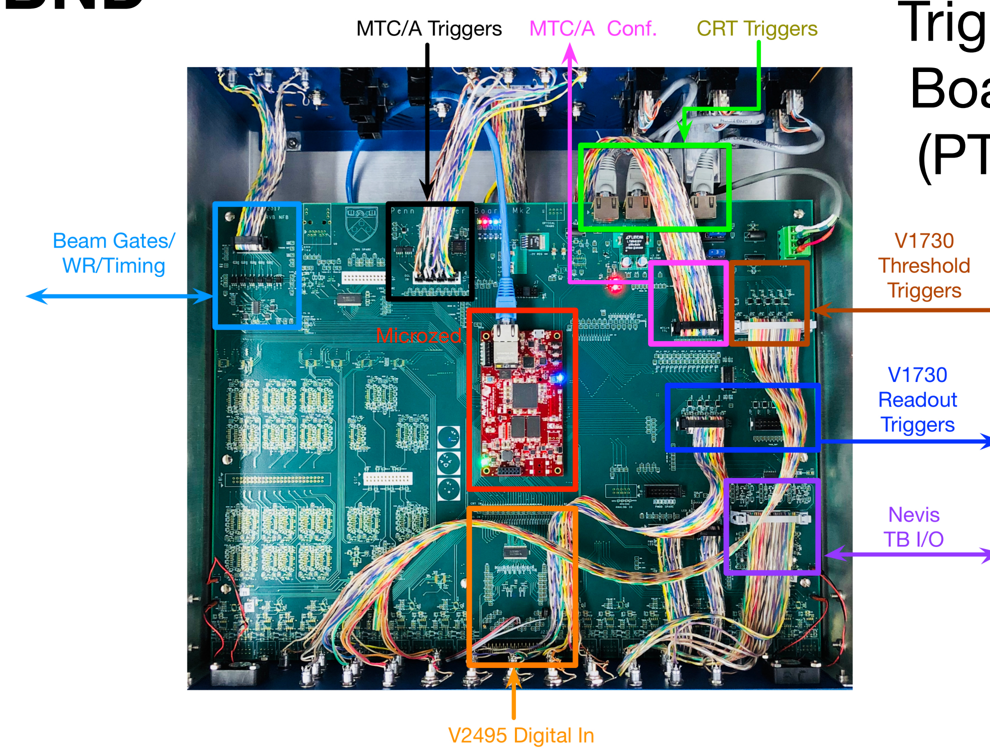
Trigger decision is critical for LArTPC due to slow drift and high granularity of detectors

— Data rates and storage increasingly become an issue

Example from SBND



Penn Trigger Board (PTB)



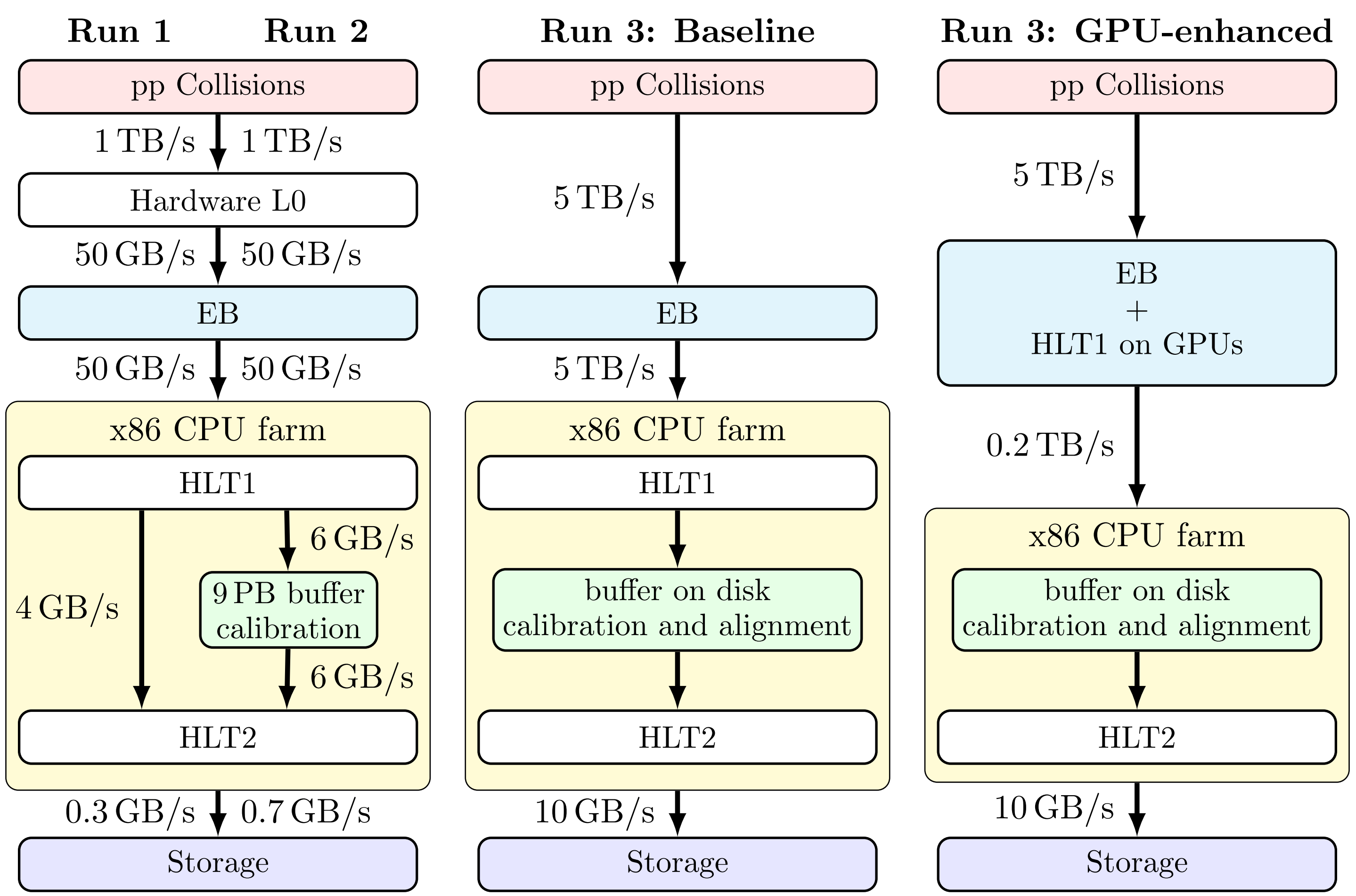
— Hardware trigger implemented to decide whether or not the TPC should be read out based on combination of information from several key sources

Real-Time Reconstruction - LHCb

Talk by D. Craik

Several interesting physics signals are high rate processes at LHCb
— improved sensitivity by accessing event information early on
LHCb performs analysis in real time
— Data is buffered before final stage of trigger to derive calibrations & alignment
— Perform reconstruction at bunch-crossing rate with same quality as offline for most objects
— Full raw event is no longer stored, reduce load on offline reconstruction

Already successfully used for several results & plans for extension for next run

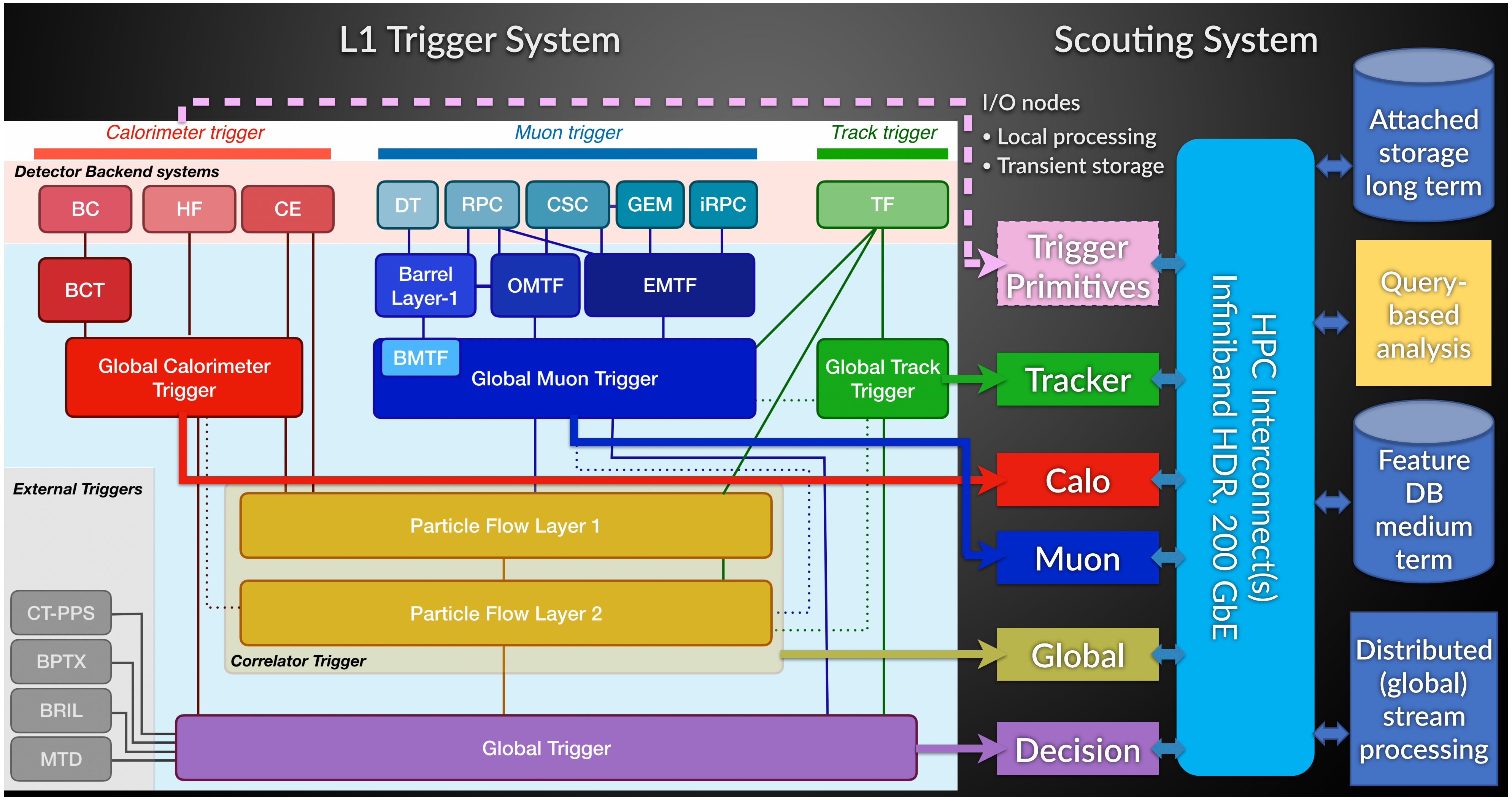


Several options explored e.g. use alternative processors (GPUs)

Real-Time Analysis - CMS

Talk by R. Mommsen

CMS is planning a 40 MHz real-time analysis stream for HL-LHC
— Interesting for physics and as diagnostic & monitoring tool



Gained experience in Run 2, plans for expansion in Run 3

Successful implementation requires R&D activities on several fronts

— HW inference engines
— Stream processing
— Distributed algorithms (MPI)
— NVRAM latency
— Searchable Feature DB
— Key-value store to assemble and buffer event fragments

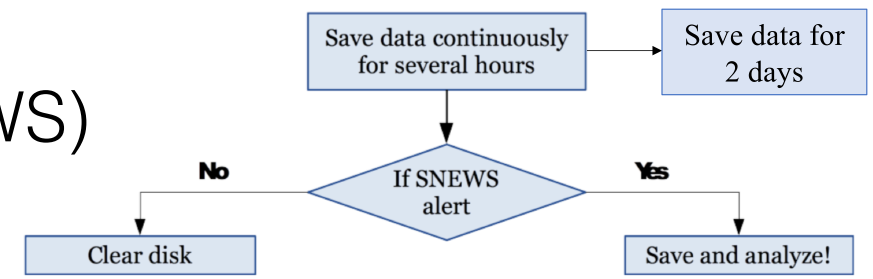
Continuous Readout - MicroBooNE

Talk by I. Ponce

MicroBooNE's Continuous Readout Stream targets seeks to observe supernova signal

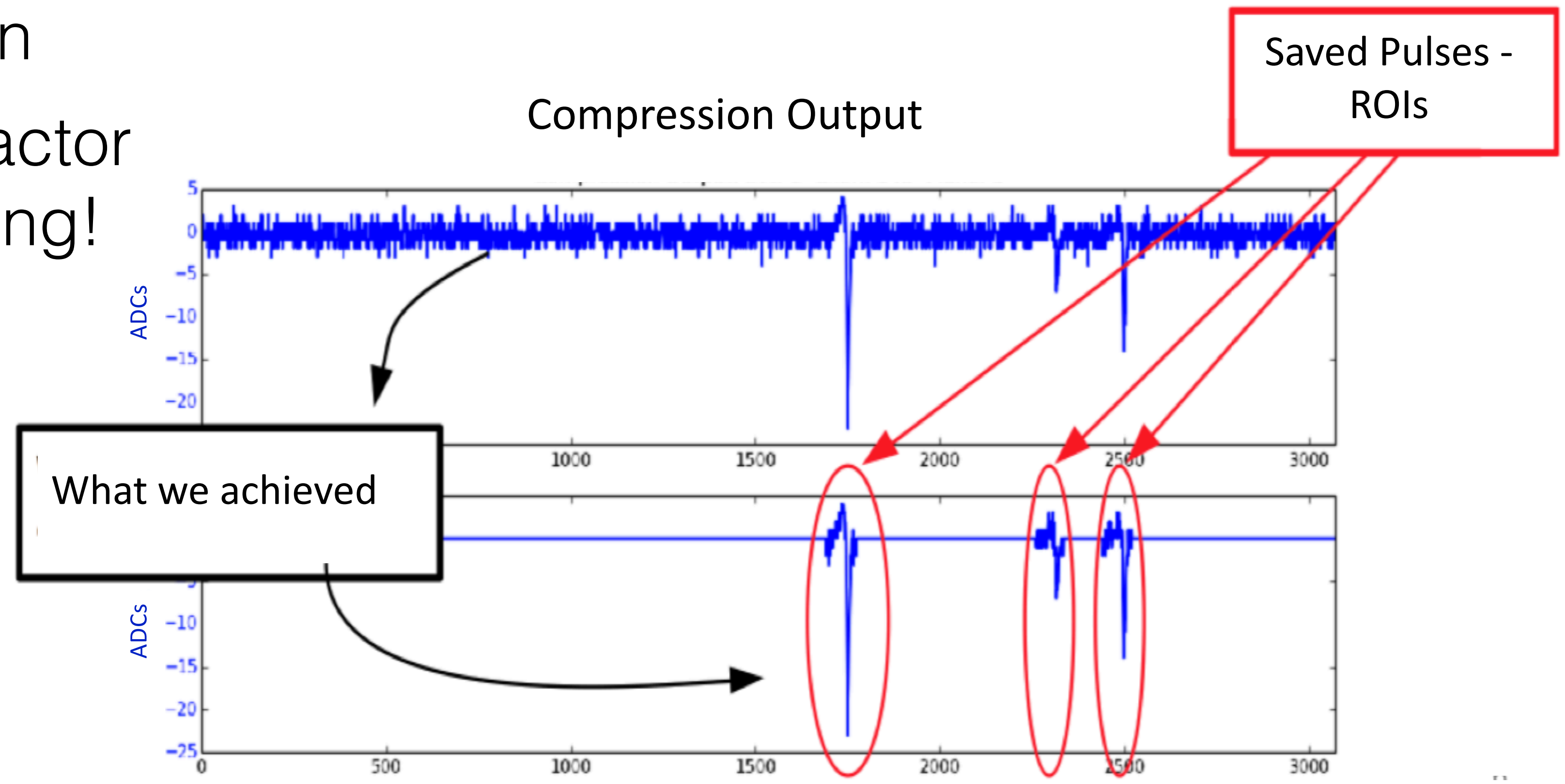
— Reads out data continuously and stores it until external trigger is issued

- Supernova Early Warning System (SNEWS)



— Requires data compression

- Achieved data reduction factor of 80 & successfully operating!

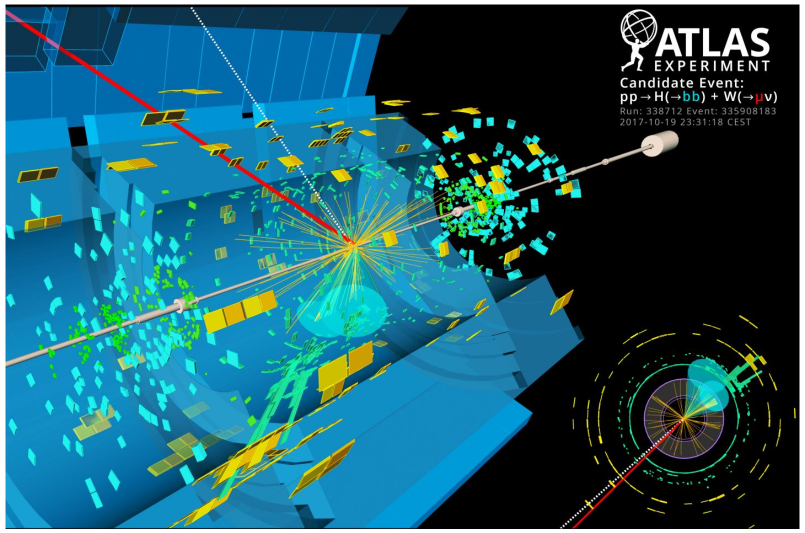


Mechanisms for reducing the transmitted data volume is another key area of future R&D

Common Challenges and R&D

Talk by K. Chen

Experiments with large data rates over many links require R&D



ATLAS raw data from detector: **1Pb/s**

FCC-hh: ~10Pb/s

Crucial for future colliders!

Exploring high-bandwidth COTS solutions
— Terabit Ethernet: 800 Gb/s and 1.6 Tb/s may become IEEE standard in 2025

Data transmission: higher bandwidth, radiation hard, lower mass, lower power consumption

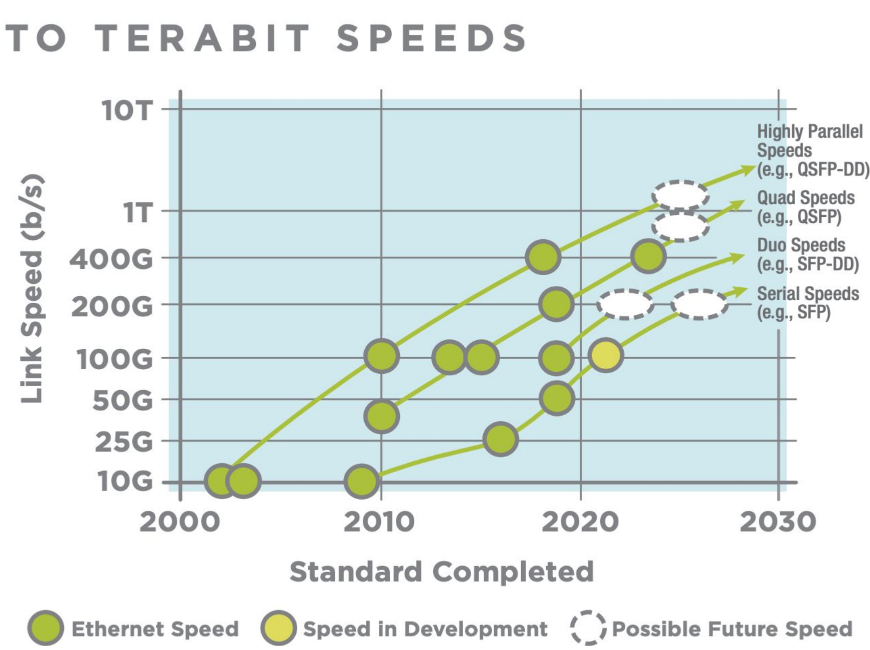
Electrical links between front-end ASIC and high-speed transmitter
- For RD53, ATLAS: up to 6 m @ 1.28 Gbps;

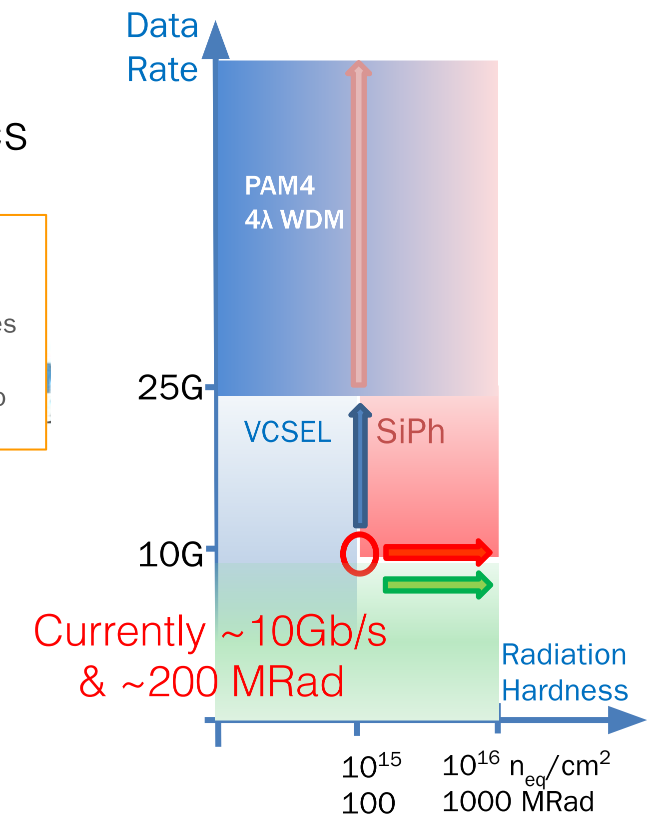
High-speed fiber optical links:
- R&D towards 28G/56G

Wireless transmission:
- R&D by groups like WADAPT for tracking detector: 60G band and 240G carrier have been demonstrated.
- Data rate 1/10 carrier frequency (OOK, BPSK)

Radiation hard high-speed serializer and optoelectronics

28Gbps NRZ / 56Gbps PAM4 Transmitter with **28nm** CMOS
Si-Photonics: integration of optoelectronic devices in a "Photonic Si chip", by using WDM: 40Gbps NRZ is possible. **Mach-Zehnder Modulator** is also insensitive to NIEL.





Exciting developments in Trigger DAQ and ML!

Thank You!