Summary of TDAQ session

Stewart Martin-Haugh (RAL), Federico Alessio (CERN), Zhen-An Liu (IHEP)

TDAQ session

- Can’t do justice to all the talks in time available
- All talks available here: https://indico.ihep.ac.cn/event/9960/session/16/#20191119

Experience from other experiments

- Buffers: DUNE, ATLAS, LHCb
- Life without a hardware trigger: LHCb, DUNE
- Life with a hardware trigger: ATLAS
- Networking: ATLAS
- Common hardware: share electronics between detectors and subdetectors (e.g. FELIX, PCIe40)
- Technology from other experiments that could fit CEPC use case:
  - artdaq framework, GBTX, VeloPix

DUNE Requirements

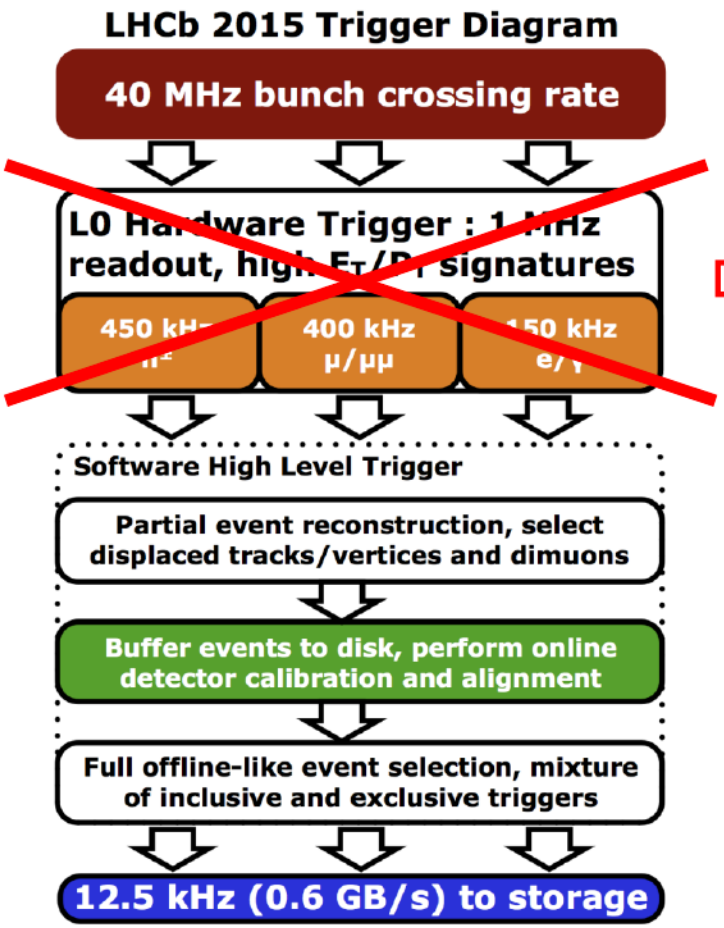
- (some) Requirements
  - Time resolution < 1μs, goal is <100 ns;
  - 1.5 TB/s DAQ readout throughput of TPC and photon detector per single phase module;
  - 10 Gb/s average storage throughput per single phase detector; 100 Gb/s peak;
  - Readout window: 10 μs (calibration) to 100 s (SNB), typically 5.4 ms for triggered interactions;
  - Ability to record 30-100s worth of raw data for a Supernova trigger;
  - High trigger efficiency (99% for particles > 100 MeV; 95% in the first 10s of SNBs)
  - Goals is to get zero dead time (small deadtime would not compromise physics sensitivity);

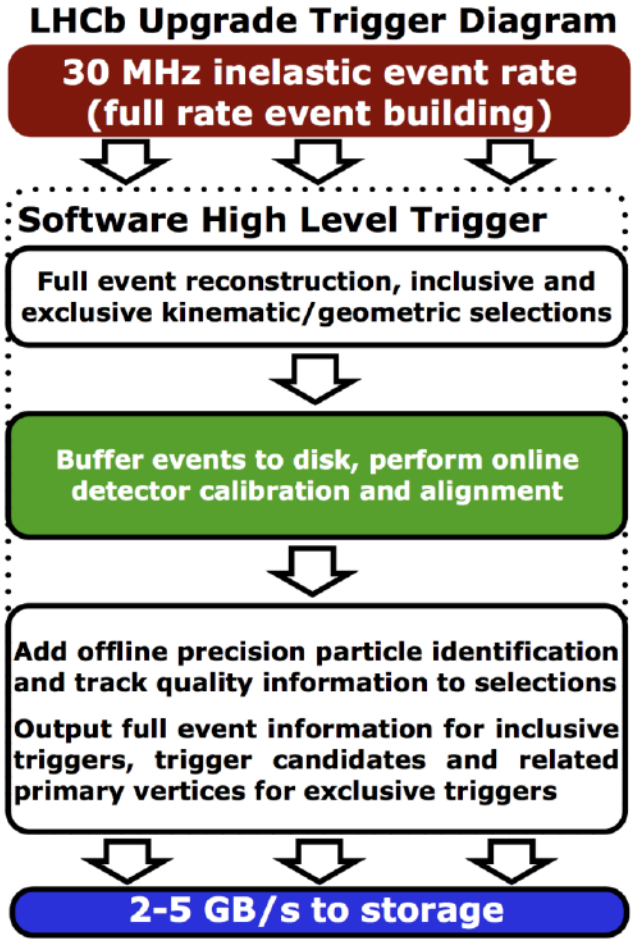
Back-End DAQ (I)

- Back-end DAQ is all about moving data across the network…
- Based on Fermilab's artdaq framework;
- artdaq has been used by many experiments:
  - NOvA Test Beam, Mu2e (tracker, calorimeter, CRV), many LArTPC experiments (SBND, ICARUS etc).
- artdaq features:
  - Provide a common framework for back-end DAQ;
  - Highly configurable, can deal with wide range of event rate and size;
  - Experiments put most efforts on implementing hardware specific code in the form of BoradReader;
  - artdaq provides subsequent data transfer, event building, file I/O etc;
  - also provides system configuration, process management, monitoring etc.

Upgrade strategy: triggerless readout

→ remove the first-level hardware trigger!



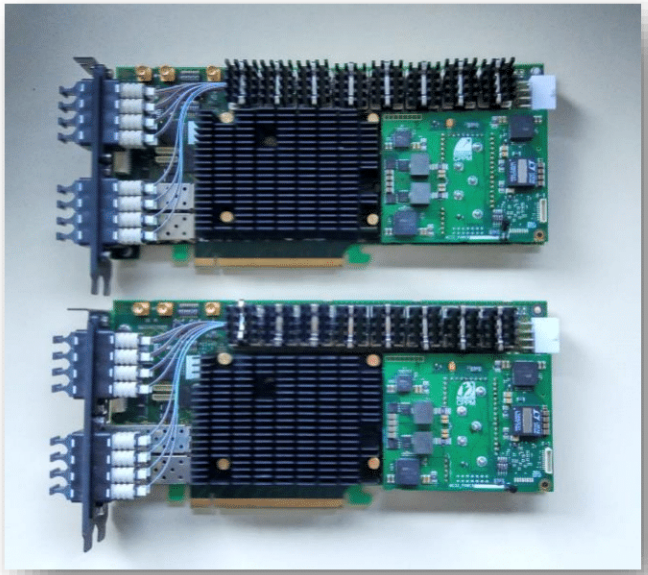


A common and generic hardware

LHCb developed a custom-made hardware readout card: PCIe40

- PCIe Gen 3 x8x8 pluggable card in commercial server
  - Validated up to ~90 Gb/s sustained
- 48 bidirectional or unidirectional high-speed links
  - Used at 5 Gb/s with custom protocol
  - MiniPod optics onboard
  - 12 links MPO connectors front panel

  + 2 dedicated SFP+/PON links for timing distribution
- Altera Arria X FPGA w/ embedded transceivers
  - Custom made protocol (CERN GBT)
- ~500 cards being produced
- Designed at CPPM, produced at FEDD, tested and validated at CERN

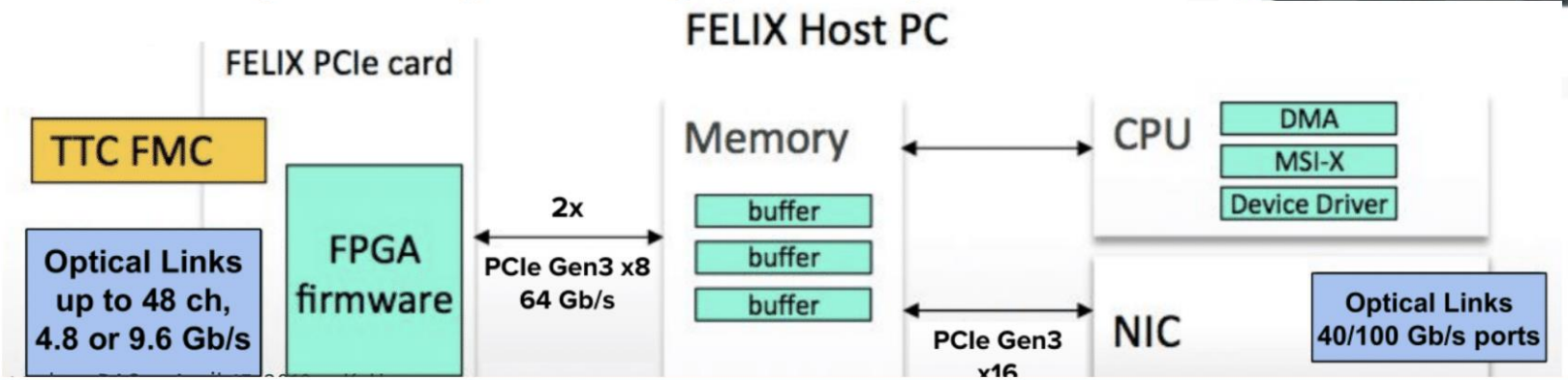


same hardware used for readout, supervision, controls

→ firmware defines the flavor of the card

Upstream DAQ (III)

- FELIX @ ProtoDUNE-SP
  - PCIe card on host computer;
  - P-to-p link throughput;
  - 10 links to host memory over 1 FELIX card;
  - HW aided data compression using Intel's QAT;
  - Full I/O over InfiniBand for ProtoDUNE;
  - Need much less network I/O for DUNE;
  - Need longer and higher throughput storage for DUNE.

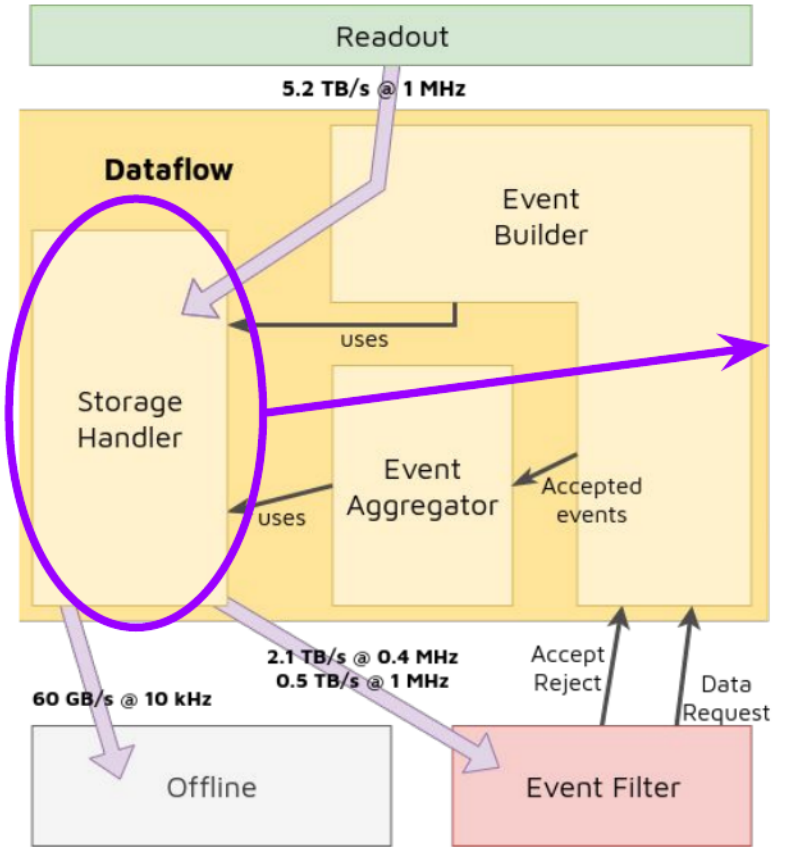


ATLAS Storage and Networking

A New Dataflow System

Novel design persistent storage to completely decouple detector readout from event selection

Support:
- recording of all read-out data at **5+ TB/s**
- transfer of read-out data to event filter: **2.5+ TB/s**
- buffering of read-out data for O(10) minutes: **3+ PB**
- recording of selected events at 60 GB/s
- buffering of selected events for 48 hours: **10+ PB**
- Elementary block size: **10 kB**



Challenge

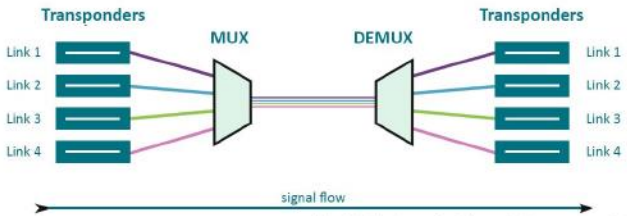
Large-volume
High-throughput
Distributed storage system

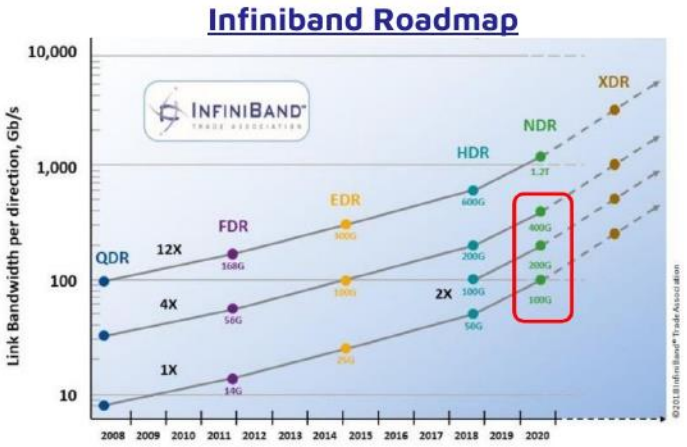
Single "namespace" for all data

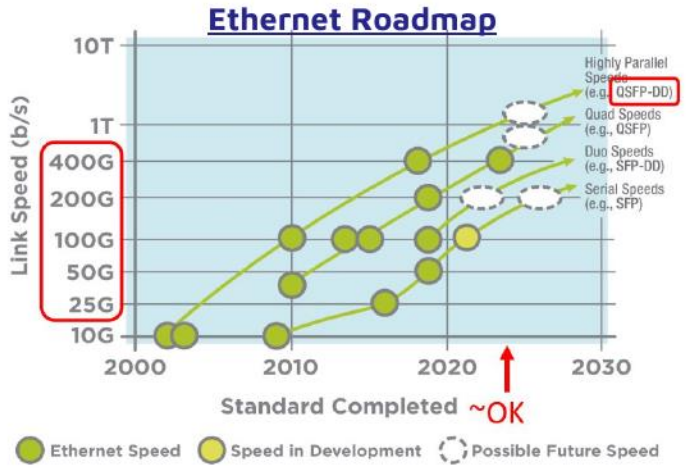
ATLAS Storage and Networking

Technologies under Evaluation

- Infiniband
  - Forwarding ("routing") policy is simple
  - Single-speed network
  - Slightly less expensive per port (at given speed)
    - But operations more complex
    - Smaller community
  - ⇒ Considered as a fallback solutions if Ethernet do not reach what we expect
    - For the readout network only, as the Event Filter farm is already equipped with Ethernet
- Ethernet technologies already available (100 and 200 GbE)
  - Now we want: high density, low power, lower price
- WDM (wavelength division multiplexing) to reduce the number of physical long-range fibers
  - Trade-offs like fibers vs. transceivers cost



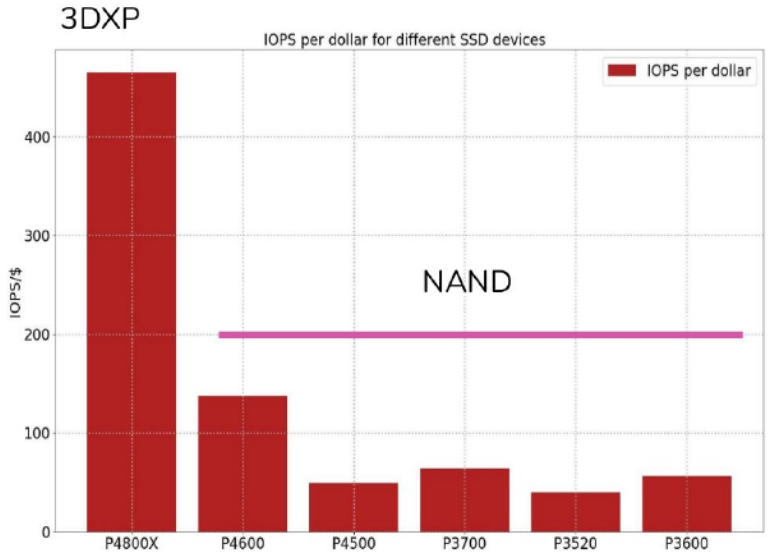




Hardware Storage Technology Evaluation

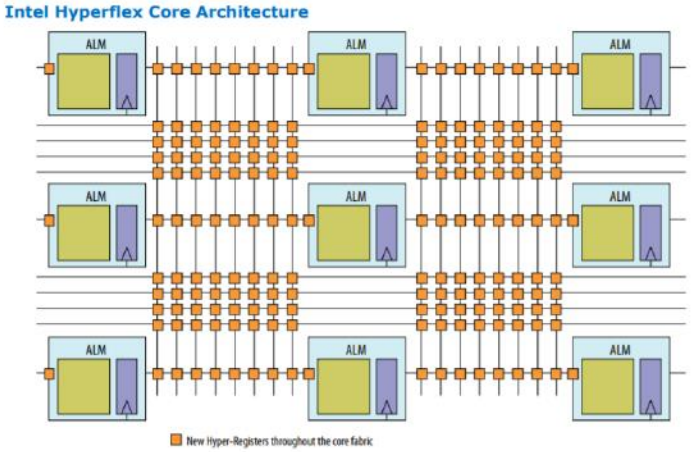
- Problem: SSDs wear out when written
- With current NAND SSDs, all SSDs in our system would have to be replaced every year

- 3D-XPoint technologies (Intel, Micron) offers much better endurance: 40 x higher
- IOPS/$ much higher



Future FPGA

- Intel Agilex
  - Architecture closer to traditional FPGA



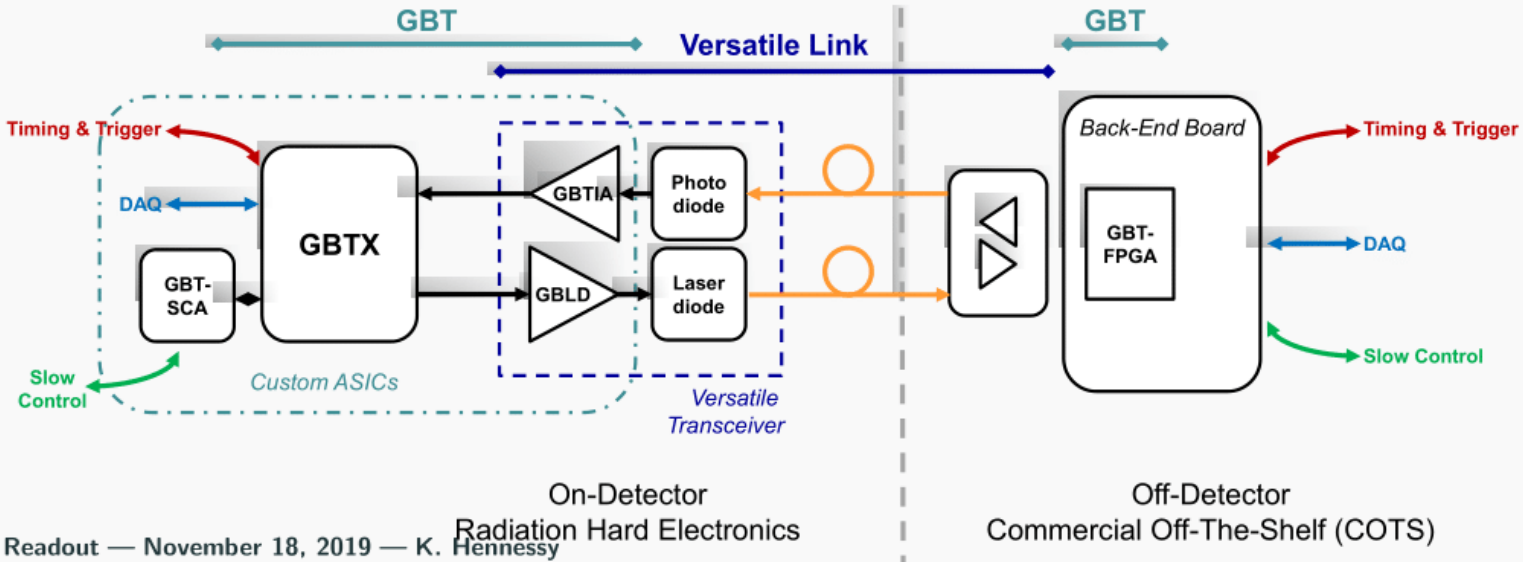
- Xilinx Versal
  - Prime Series
    - Upgrade to Zynq SoC
  - AI Core Series
    - Automatic search for new physics?
      - Needs R&D

Summary

- FPGAs was originally used as the prototyping for ASIC in ATLAS TDAQ.
- At Run1, FPGAs displaced most of ASICs in the original technical design.
- At Run3, FPGAs with high density high speed links dominated trigger upgrade design.
- At Run4, increased L1 trigger latency (due to detector frontend upgrade) allows a step change in trigger architecture.
  - Next generation FPGAs will run iterative algorithms in real-time system.
  - Even higher speed links (25G or 50G) allows data aggregation for full event process.
- ATLAS time scale much longer than industrial norm
  - Typically running electronics for some 20 years without large-scale upgrade
    - Original L1Calo was base on 9U VME designed in early 2000, and part of it will continue until the end of RUN3 in 2024.
- FPGA technology has been advancing very fast, and designing with modern large FPGAs is challenging.

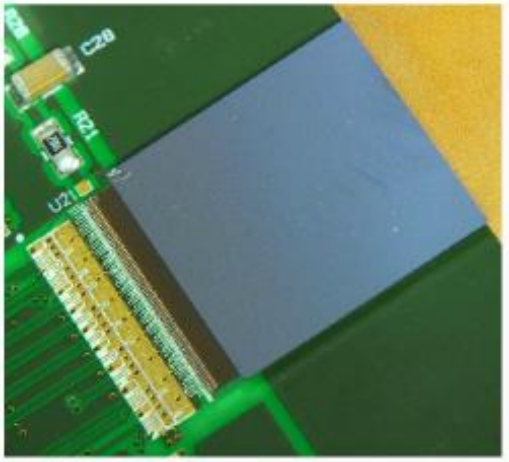
GBT - ECS interface

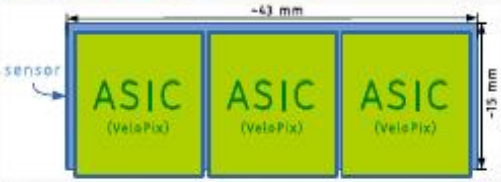
- The **GBTX** chip is a radiation tolerant chip for LHC experiments
- GBT Protocol can utilise three logical data paths
  - Trigger and Timing Control (TTC)
  - Slow Control (SC) - via companion SCA chip
  - Data Acquisition (DAQ) - *(NOT used for VELO)*
- All three logical paths can be encapsulated on a single physical interface



VeloPix ASIC

- Front-end ASIC driving the design of the VELO data acquisition system
- Operates at LHC clock rate $\sim$ 40MHz
- Designed for high radiation tolerance and low power consumption
- Custom output serialiser - Gigabit Wireline Transmitter (GWT)
- Slow control via SLVS protocol
- 12 VeloPix chips per module
- 20 readout links (more links for hotter chips)



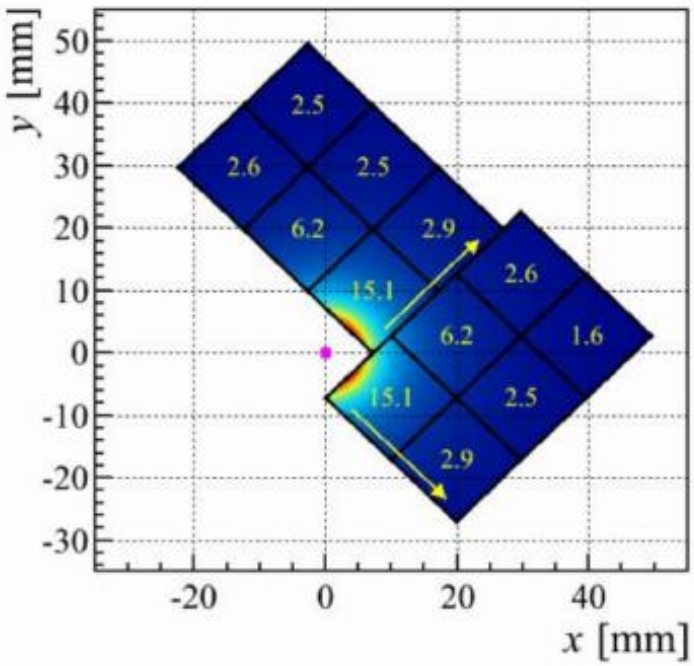


VeloPix ASIC

- Readout is data driven - *only* read out when they have "hits above threshold" (a.k.a. zero-suppression)
- **Binary readout** @ 40 MHz
- VeloPix is optimised for high speed readout

Peak hit rate	900 Mhits/s/ASIC
Max data rate	19.2 Gb/s
Total VELO	2.85 Tb/s

- Power consumption < 1.5 W$\cdot cm^{-2}$
- Radiation hard 400 Mrad, and SEU tolerant
- Non-uniform radiation dose



Data rate [Gbit/s] for hottest module.

Average data rate

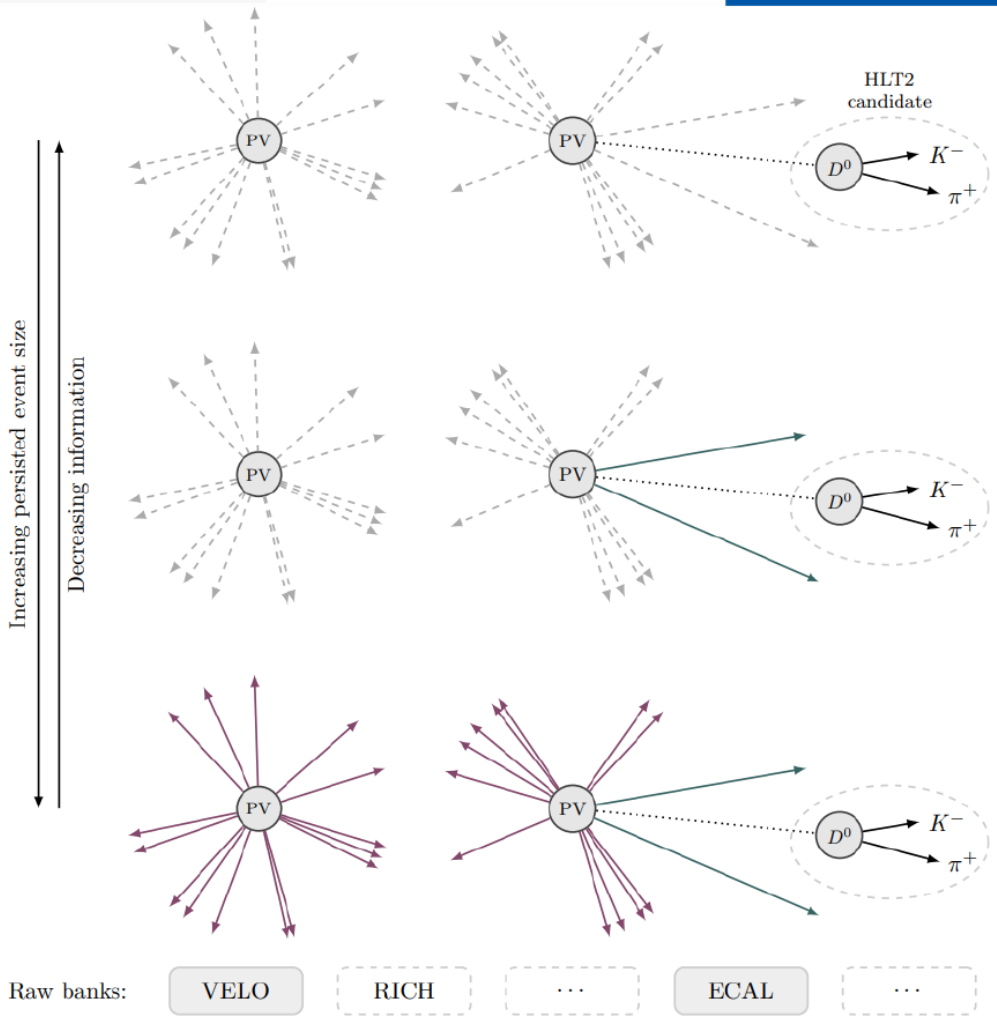
Turbo stream

- Given the bandwidth hard limits, do we need to save all information about all events?
- Select what we want to save

- Turbo (2015)
  - Keep only objects used for trigger

- Turbo SP (2017)
  - Objects used for trigger + special selection

- Turbo++ (2016)
  - All reconstructed events
  - Raw event is dropped



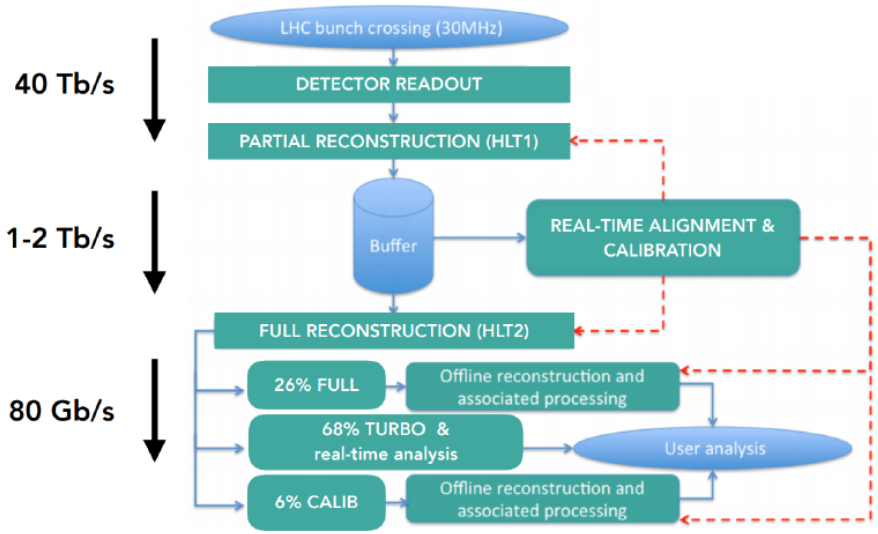
Requirements/Need-to-know(s)

- Backgrounds in tracker
  - Need an estimate for the total #hits per layer, probably only vertexer
- nTracks in tracker
  - CDR - All in all: Somewhere around 30 hits/25ns/Layer need to leave the detector, call it 100 for good measure
- Per track timing?
  - Do we need per hit? Depends on trigger setup/other detectors and computing model
- Trigger scenario
  - Do you want to trigger or go full software
  - Doesn't necessarily blow the data transmission budget: e.g. 30 hits at an excessive 40 bits would be 1200 bits per event, i.e. 50 Gbit/s for a full layer of the tracker - on the edge with rad-hard technology 20 years ago, easy to imagine now (even in rad-hard) - inner layer is probably about O(1000) front-end chips, i.e. 50Mbit/s/chip.
  - Need a concept for computing, more than for tracker readout?
- Other subsystems: Can most detectors be fully read out, or would one subsystem imply an early trigger for raw data - not worth it to consider untriggered readout for tracker if other detectors imply reduced rate very early

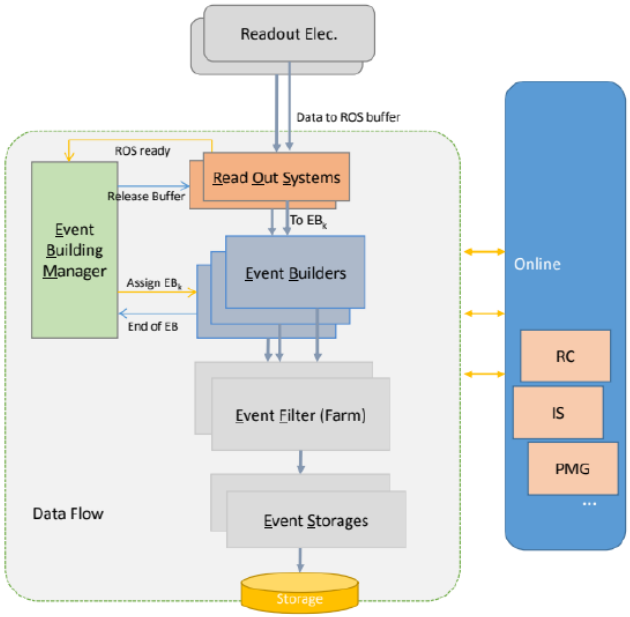
RTA and CepC

- Possibility of using Turbo-like and RTA approach in CepC experiments?
- CepC: cleaner environment and smaller event size
- Expected total raw date rate ~ 2 TB/s on 100 kHz L1 trigger (CepC TDR)
- RTA and Turbo-like system could save computing resources and improve versatility of detectors

LHCb (RUN III)



CepC (TDR)



IHEP-CEPC-DR-2018-02

Open questions for CEPC

- Probably don't need a trigger for ZH or WW running
- Need trigger for Z-pole running (x10 higher rate)?
  - All-software trigger?
- Which subdetectors should be triggered on?
  - e.g. vertex detector rate may be prohibitive
- Consider front-end specifications early, driven by TDAQ requirements