

# Trigger and DAQ at large HEP experiments

Srećko Morović – UC San Diego

Sarajevo School of High Energy Physics 2022

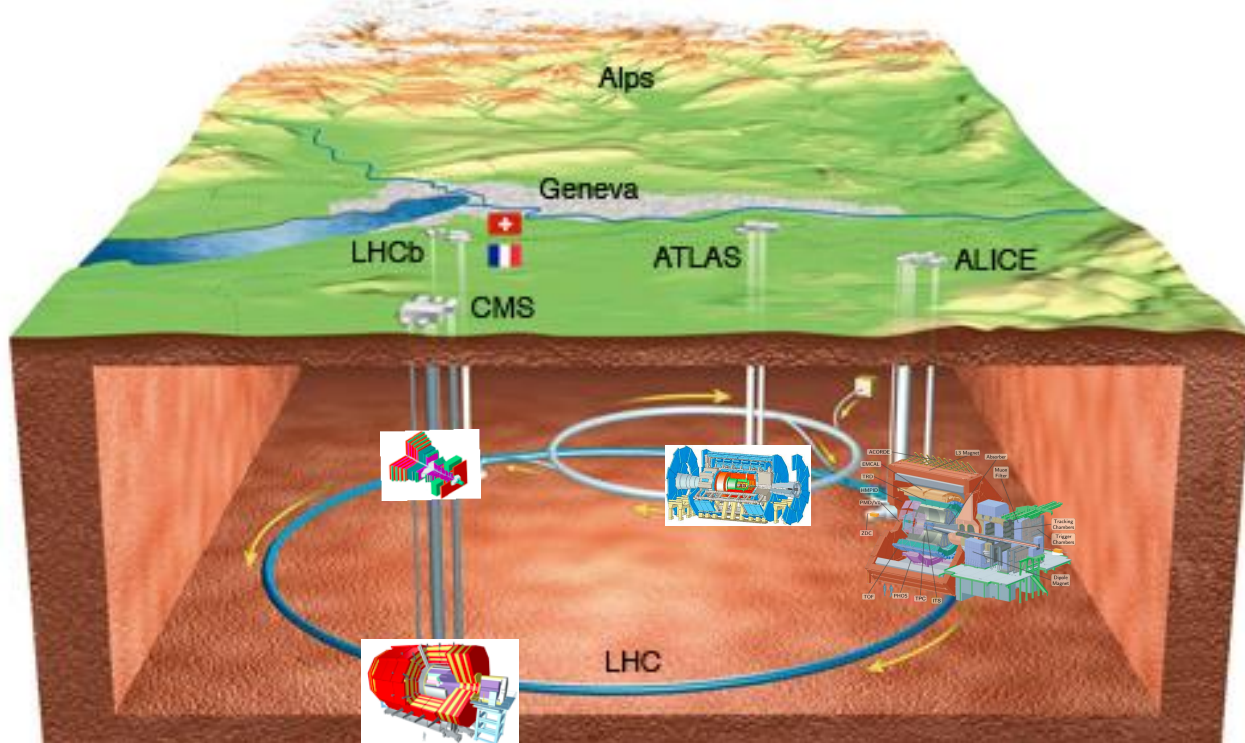
14.10.2022

# Introduction

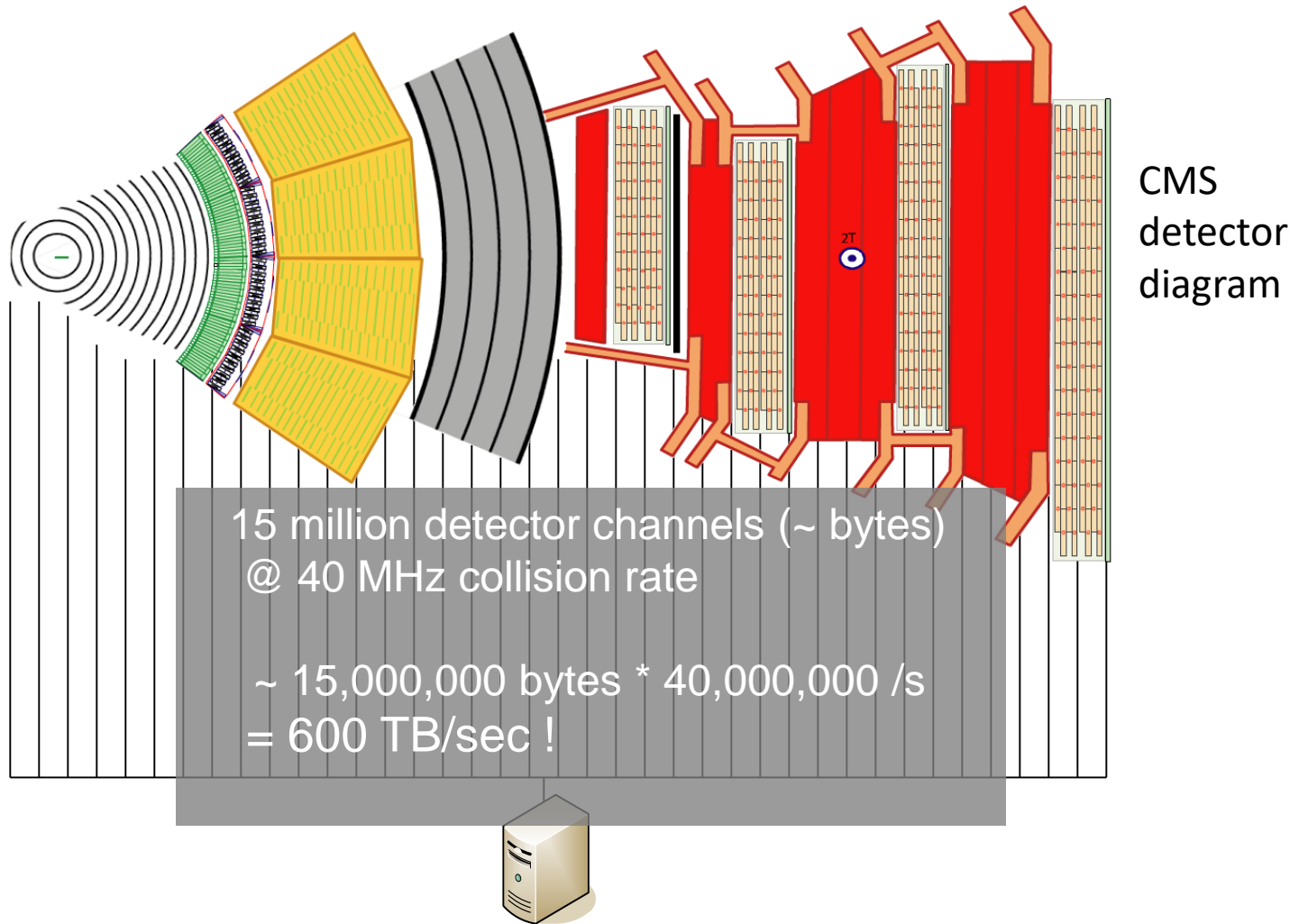
- Aim of this talk
  - Introduce basic concepts of Trigger and Data Acquisition (DAQ)
    - Mostly in the context of large HEP (LHC) experiments
    - Focus on technical more than on algorithmic aspects of Trigger
  - Followed up by
    - Description of DAQ/Trigger designs used by the LHC experiments
      - Focus on new upgrades for Run 3 (2022 +)
    - Discussing future/evolution of Trigger and DAQ systems
      - Through High-Luminosity LHC upgrades (~ 2029 +)

# Largest HEP experiments

- LHC – short introduction
  - superconducting dipole ring: 27 km circumference
  - proton-proton, proton-Pb and Pb-Pb collisions
  - collision energy:  $\sqrt{s}$  14 TeV (design), achieved: 13 TeV
  - 4 large detector experiments
    - State of art detector, data acquisition and trigger hardware



# Data collection problem



- Filtering, reading and storing of interesting collision data  
→ Role of the Trigger and Data Acquisition System and Trigger

# Trigger

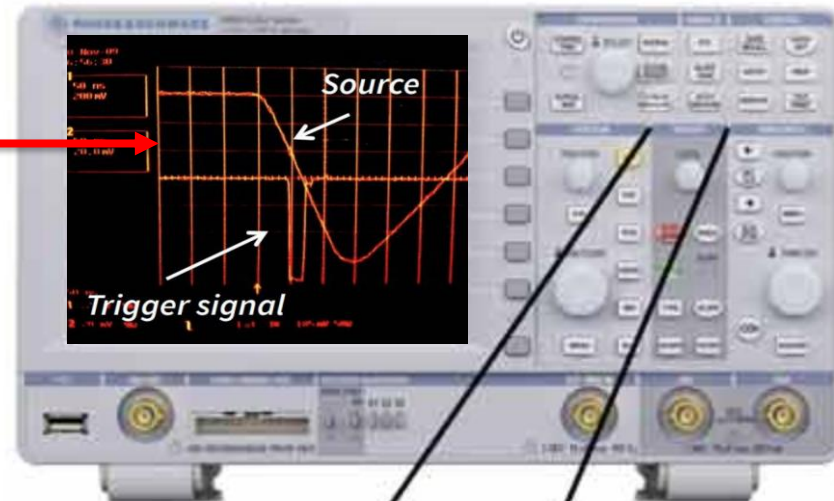


- Detector
  - Device capturing events
  - Event: e.g. proton-proton collision that we **want** to record
- Triggering
  - a process to rapidly decide if you want to keep data recorded by a detector
  - Separates interesting events from background
- Reason to use trigger
  - Not all data is interesting
    - Too high rate to keep all of it
  - Filtering beneficial in reducing event rate for
    - read-out from detector
    - Storage
    - Less data volume to process and analyze later (“offline”)



# Trigger example - oscilloscope

Pulse  
threshold  
for  
triggering  
readout



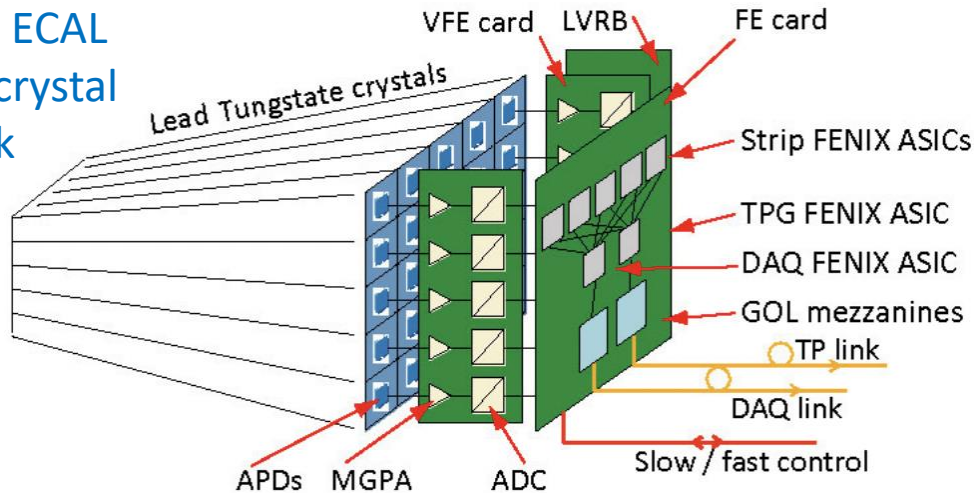
## Two data paths

- **Trigger path** – simple signal (spike) fed into the trigger decision unit
- **Data path** – full resolution pulse shape
  - read AFTER trigger accepts this *event*
  - Non-triggered events are not read



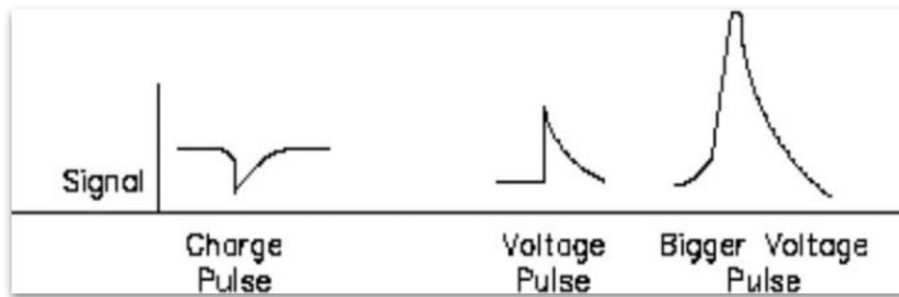
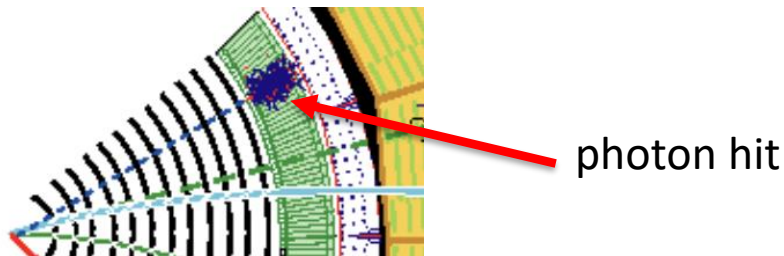
# Trigger example – CMS ECAL

CMS ECAL  
5x5 crystal  
block



FED (front-end)  
on-detector  
electronic  
modules

DAQ, Trigger &  
control digital  
links

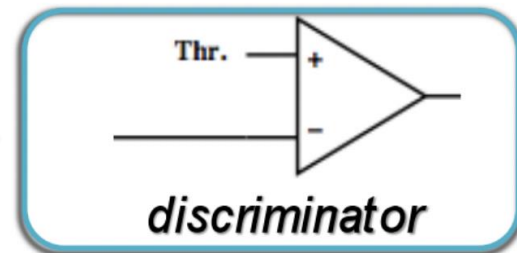


Frontend – amplification, digitization

Trigger  
path  
readout

40 MHz

Discriminator



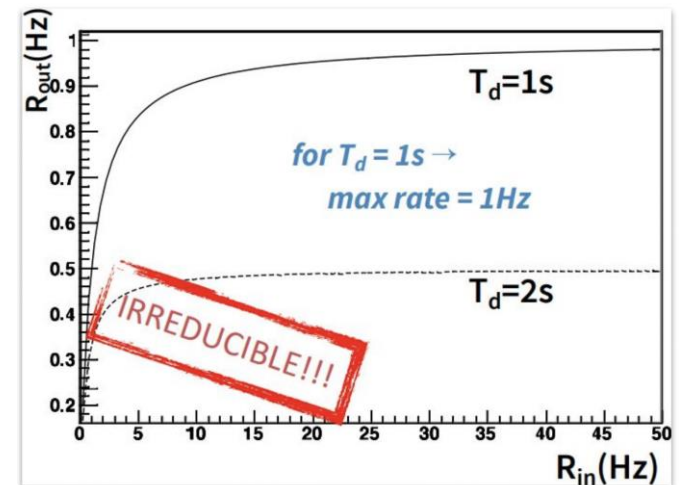
Data path readout  
100 kHz → DAQ

# Trigger efficiency

- High efficiency - capture high fraction of signal 'events'
  - Achieve low **deadtime** - time when trigger is busy
  - potentially lost for data acquisition due to inability to trigger
- $R_{in}$  = average trigger rate (target)
- $R_{out}$  = readout rate
- $T_d$  = processing time of one event
- Fraction of lost events:  $R_{out} * T_d$
- $R_{out} = (1 - R_{out} * T_d) * R_{in}$
- High efficiency:  $R_{in} * T_d \ll 1$
- Important: **fast decision**

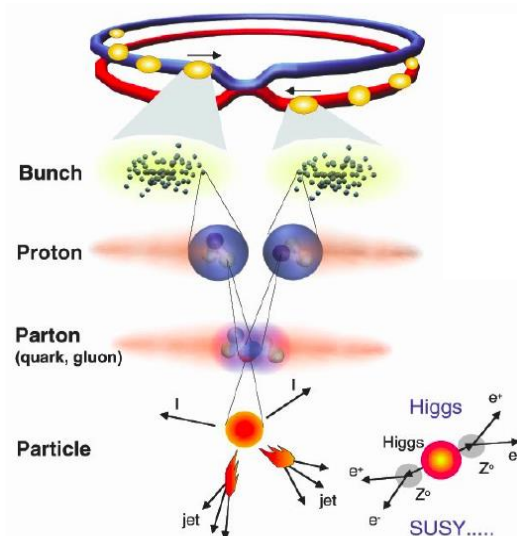
Fraction of captured events:

$$\frac{R_{out}}{R_{in}} = \frac{1}{1 + R_{in} * T_d}$$

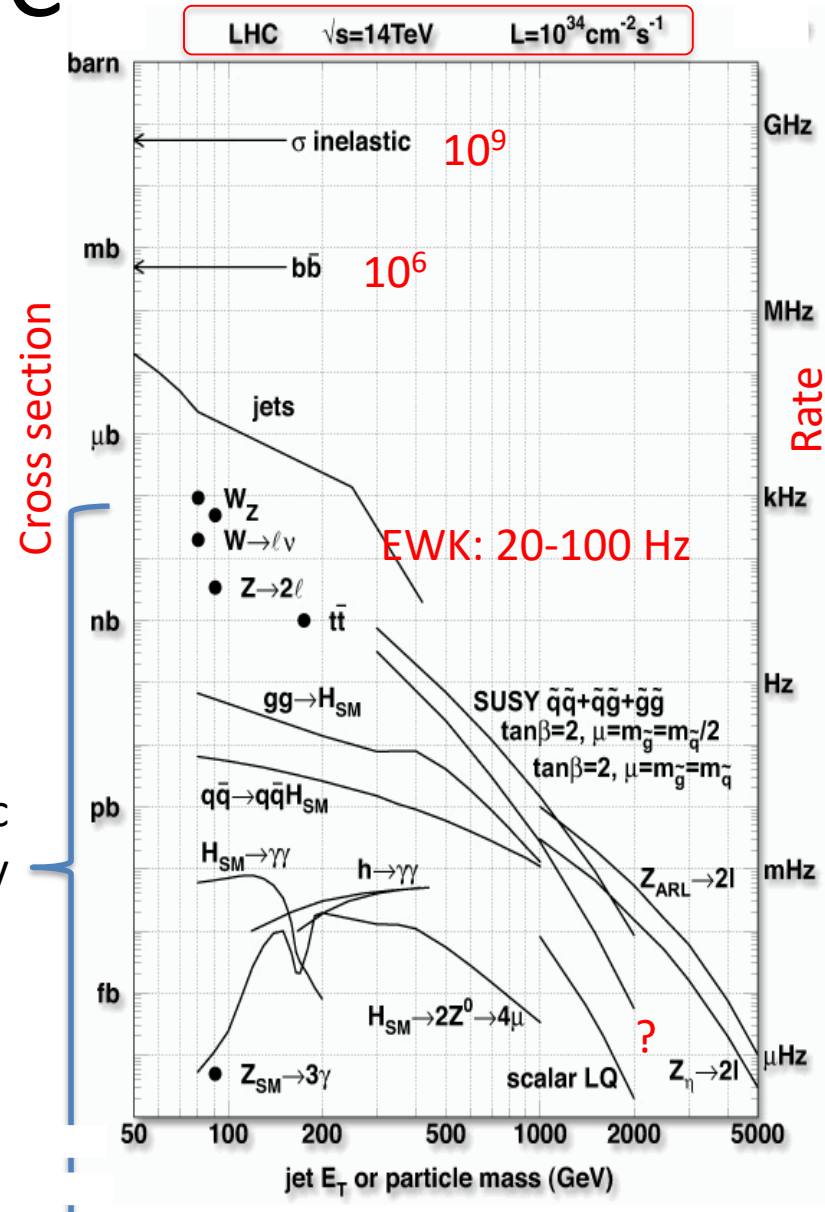


# Physics production at LHC

- Cross sections of processes at LHC (pp) span many orders of magnitude
- Huge rate of mainly “uninteresting” collisions
  - Low  $p_T/E_T$
  - Dominated by inelastic pp scattering
  - Only interesting to keep in smaller quantities (low- $E_T$  physics, cross-checks, calibrations etc.)

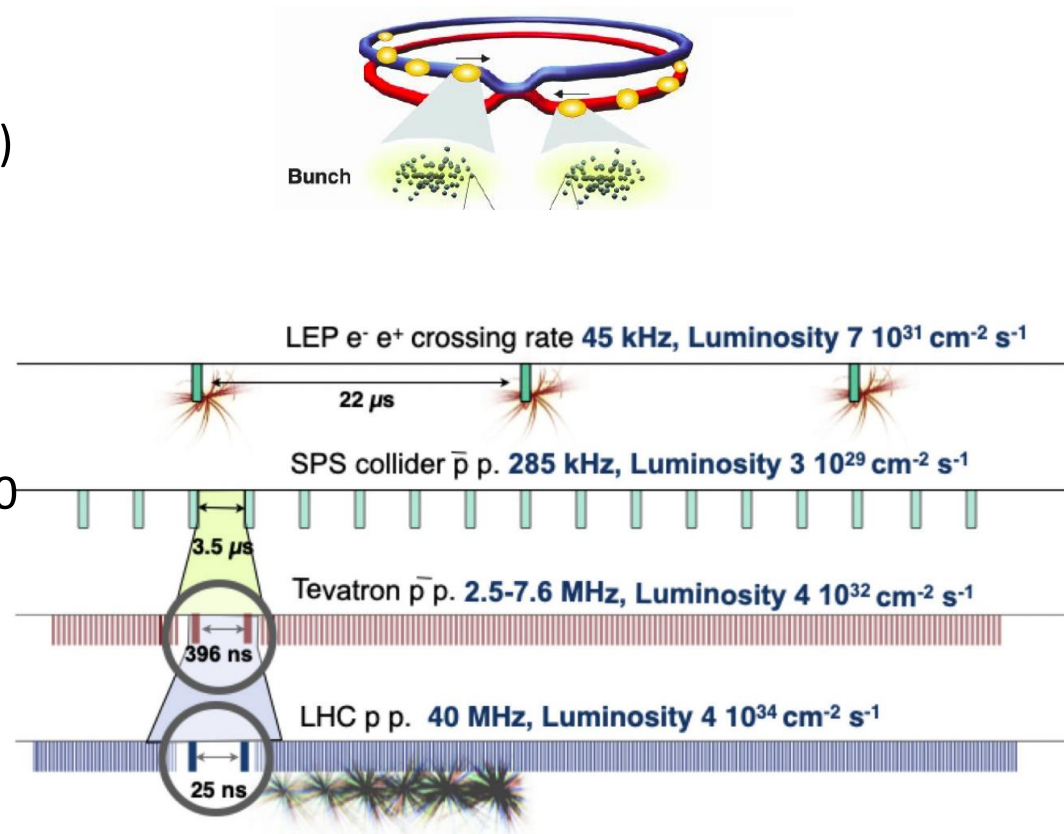


Filtered(accepted) by Trigger

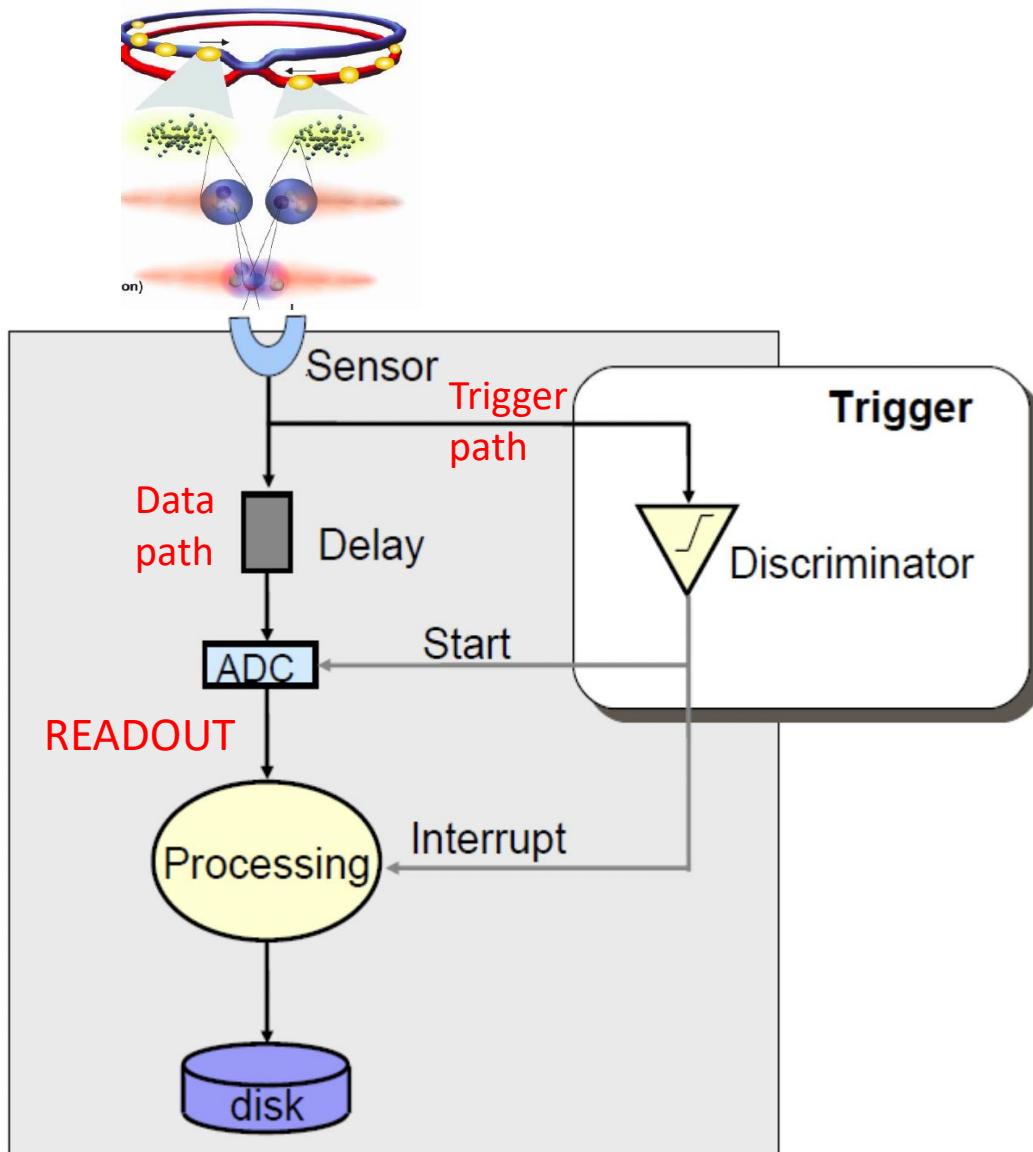


# Trigger – modern particle experiments

- Small intervals between collisions event
  - Appear when particle bunches cross each other (bunch-crossing)
- LHC:
  - 40 MHz pp crossing rate
  - 25 ns spacing between potentially triggered events (bunch collisions)
  - In special periods, lower rate (50 kHz) of lead-ion collisions
- Trigger requirement
  - decision needed every 25 ns
    - Commonly using dedicated electronics (e.g. using FPGA chips)
    - Typical latency ( $T^d$ ) ~ in microseconds
      - electronics + cabling

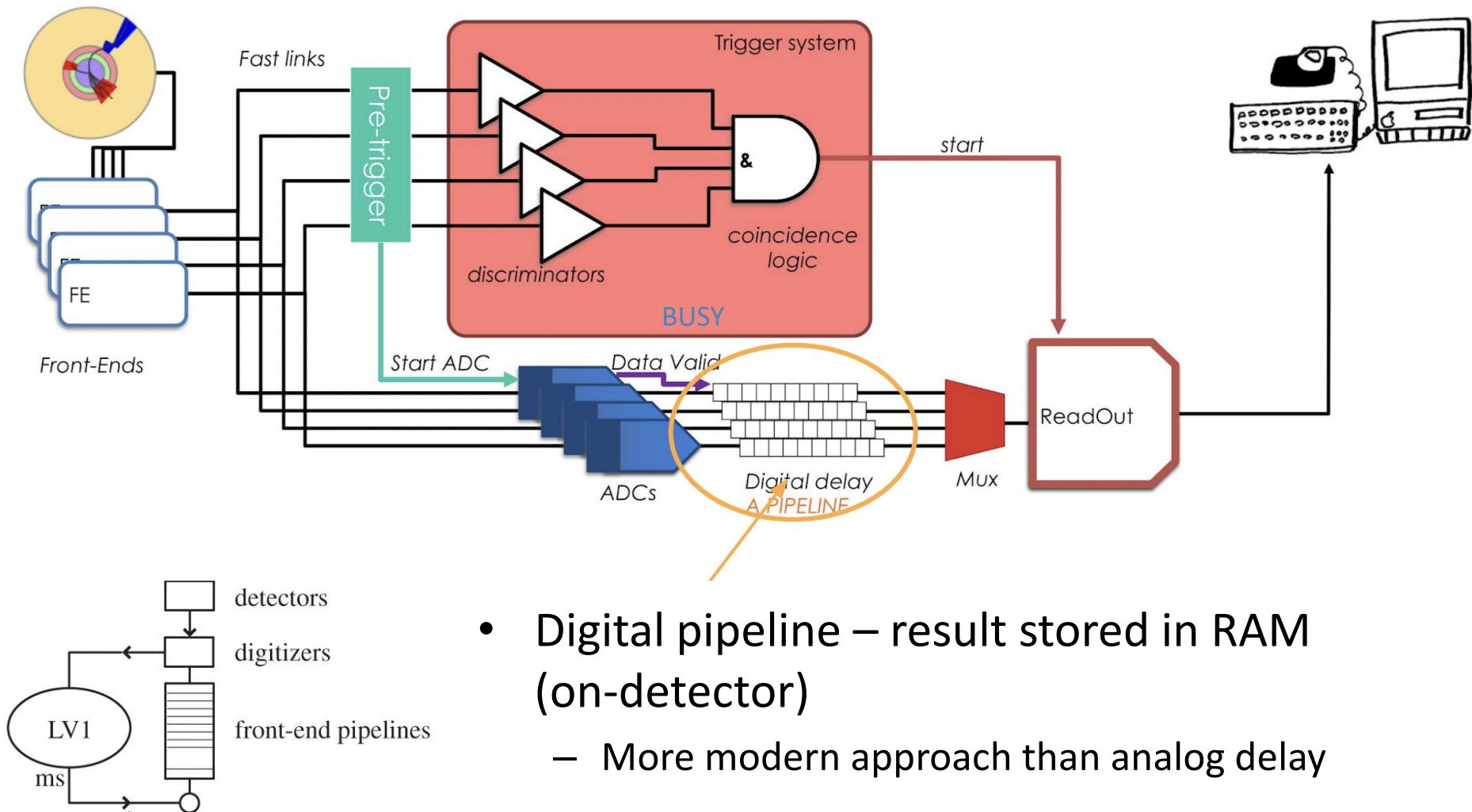


# Trigger delay



- Delay
  - Stores previously recorded signal
  - Analog or digital (buffer)
- Solves deadtime issue with event rate  $\gg 1 / T_d$
- Compensates for trigger latency
  - Multiple sequential events 'stored' (queued) while their trigger decision is being evaluated
- Implemented in hardware installed on-detector
  - radiation-resistant electronics (LHC)

# Trigger – real life



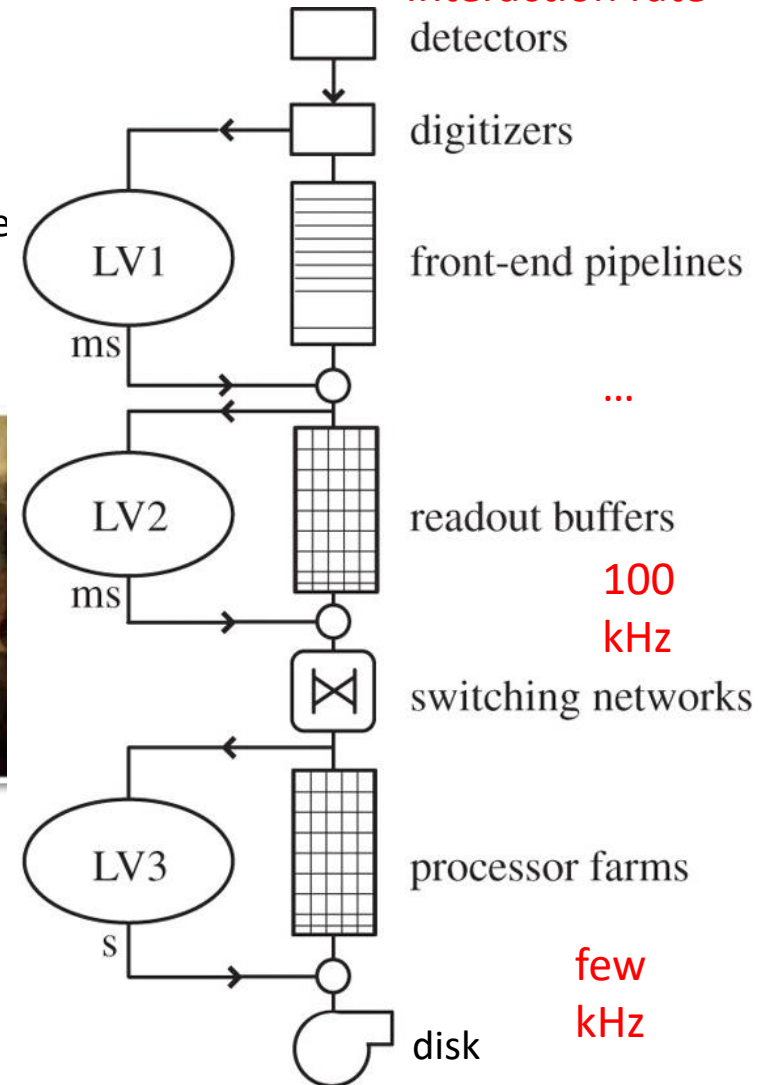
- Digital pipeline – result stored in RAM (on-detector)
  - More modern approach than analog delay

# Multi-layer triggers

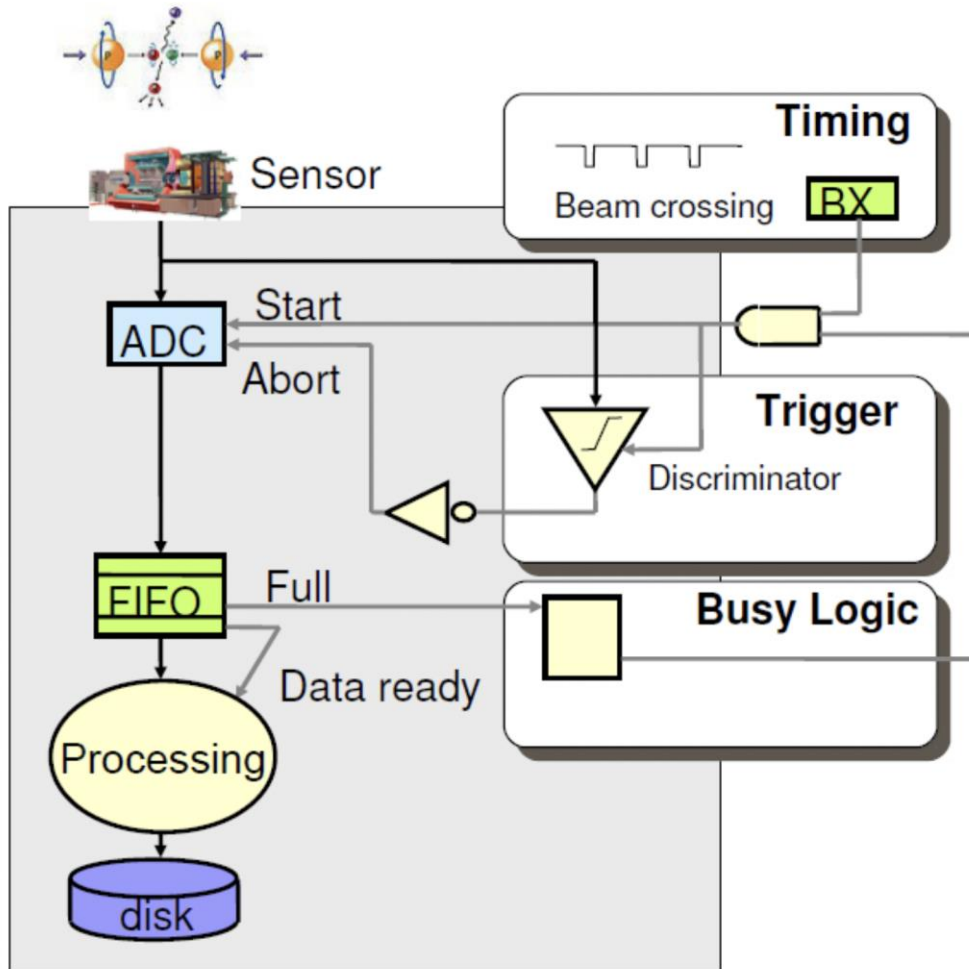
- Adopted in large experiments
  - Allows faster decision early on
    - faster reduction in accept rate and data bandwidth
    - more complex (slower) analysis and filtering late on preselected events



Experiment	Levels
ATLAS	2
CMS	2
ALICE	0 or 1
LHCb	2



# Synchronous trigger and readout



- System phase-locked to a **clock**
  - Such as coincidence with bunch crossing
  - all data moves at clock steps
  - Data in pipelines either discarded or sent forward
  - No deadtime
    - Fixed / deterministic latency behavior (decision guaranteed in time)
  - Disadvantages
    - Expensive hardware (customized, high-frequency, phase-locked)
    - Complex synchronization and alignment across the whole detector

DAQ - decoupled via buffering (FIFOs)

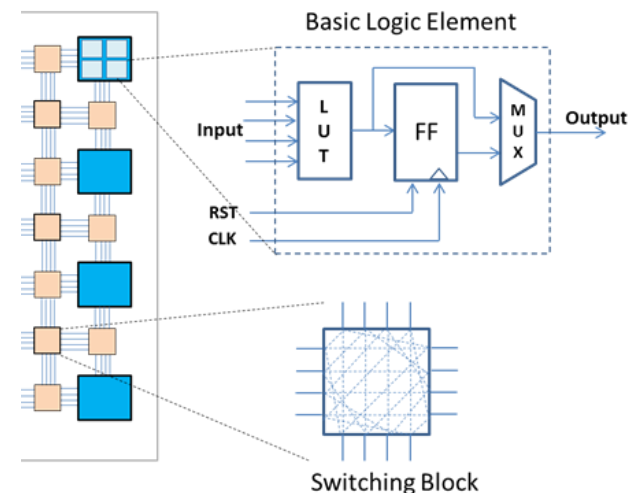
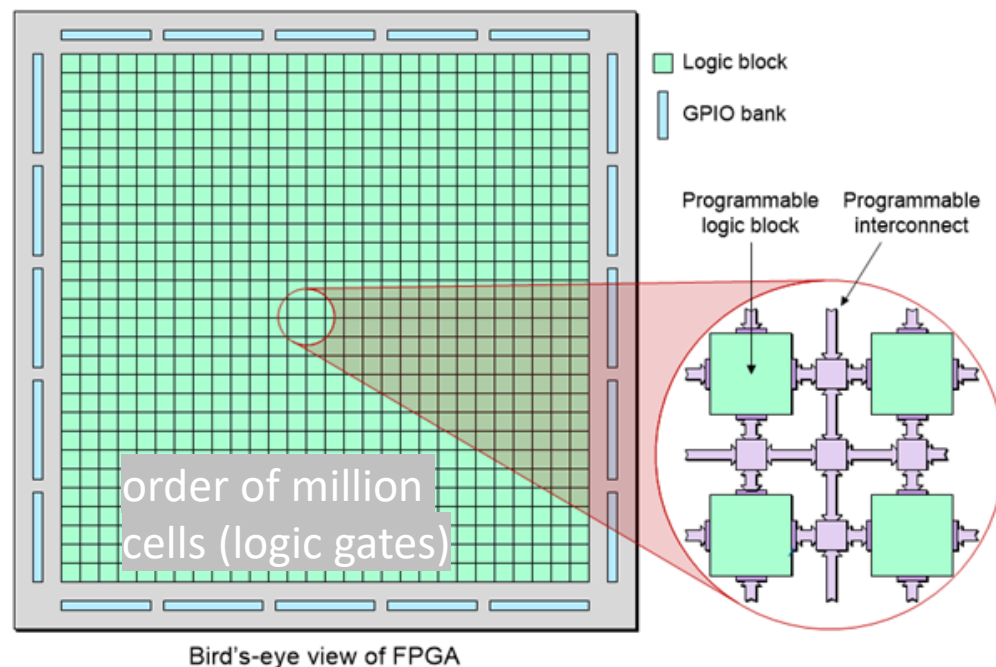
- Asynchronous part of the system
- Buffer accommodates fluctuations of accepted L1 rate
  - triggering is of random nature (follows **Poisson** distribution)
- **BUSY** logic - Used to stop trigger/readout when buffer is full (“**backpressure**”)

# Implementation - L1 Trigger

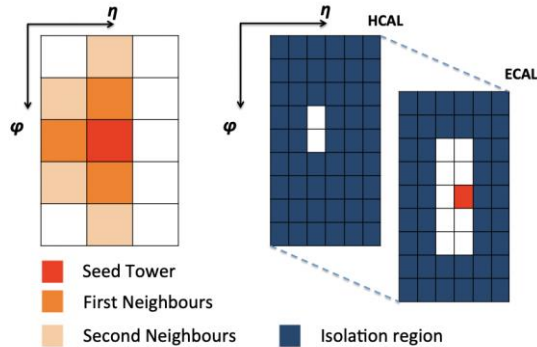
- Needs to decide within  $O(\mu\text{s})$  latency boundary
- Computers (PCs)
  - CPUs
    - Designed to execute general purpose code
      - Operations (instruction set) performed on a set of registers, advanced branch prediction, caching, memory (pre)fetching...
      - Run an OS to schedule execution (often on multi-core) and I/O
    - 4 GHz processors (typically):
      - In 25 ns interval can run only  $\sim 100$  operations / bx. (per core) !
    - Non-deterministic latency (hardware and operating system)  $\sim ms$  range)
      - not suitable for trigger operation (in HEP experiments)
- Hardware
  - ASICs – logic specialized for a specific task imprinted into silicon
    - Fast (parallelism) but limited programmability
    - High costly/unit (large R&D)
  - FPGAs
    - Chips with low-level logic programmable by users

# FPGAs

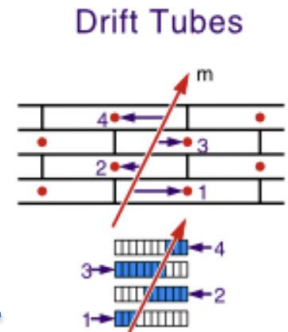
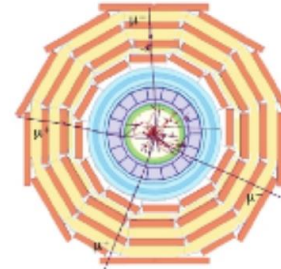
- Field-programmable gate arrays
  - Programmable logic and routing – **defined via firmware**
  - Advantages
    - Massively parallel logic operations
    - Clock synchronization, low latency
    - Updates can be deployed by firmware reload
  - Can come with special elements (DSPs, I/O etc.)
  - Con's:
    - Low-level logic, very difficult to program efficiently
    - Often requires designing custom boards (R&D effort)
    - Cost of chip + components
  - **Popular choice for L1-trigger implementation**



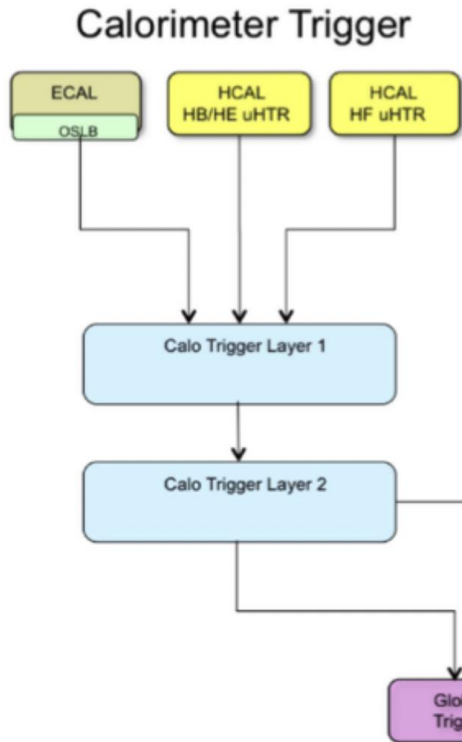
# Example – CMS L1 Trigger



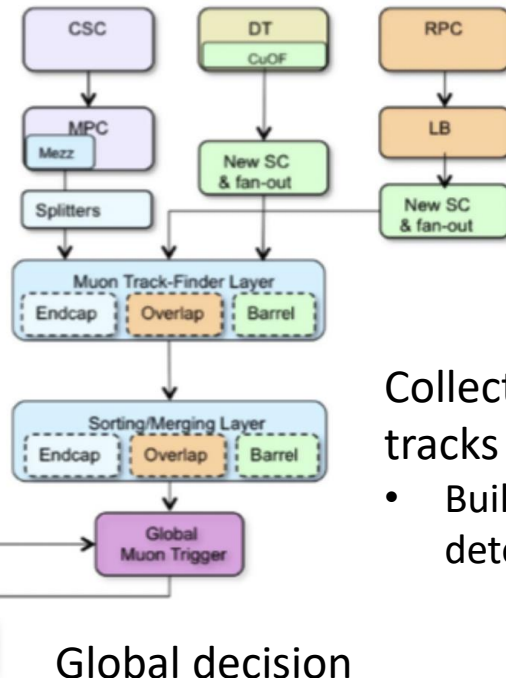
- Input: coarse data (**trigger primitives**) from on-detector front-ends (FEDs)



Calorimeter cells (ECAL and HCAL)



## Muon Trigger

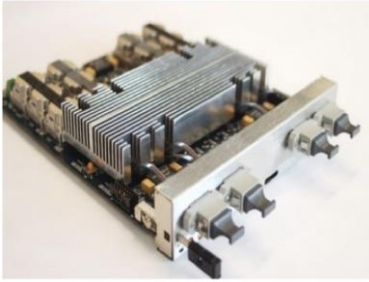


Pipelined logic  
→ from simpler to more complex objects

Collection of muon tracks

- Built from several detector types

# Example – CMS L1 Trigger hardware



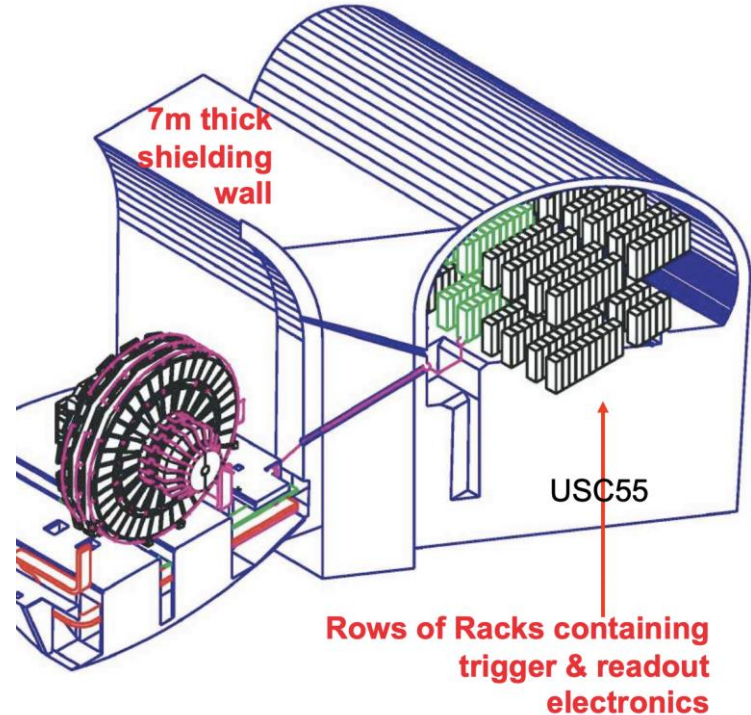
MP7 Board with large FPGA and optical input links ( $\mu$ TCA bus)

- Xilinx Virtex 7 FPGA
- Optical input and output



$\mu$ TCA Rack with installed Calorimeter  
L1 trigger boards

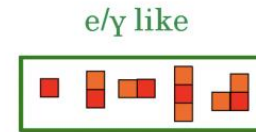
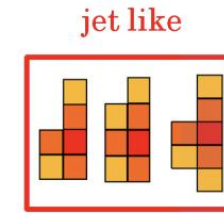
6 Gb/s (input) per card



- Location:
  - Underground counting room
  - away from the high radiation environment (cavern)

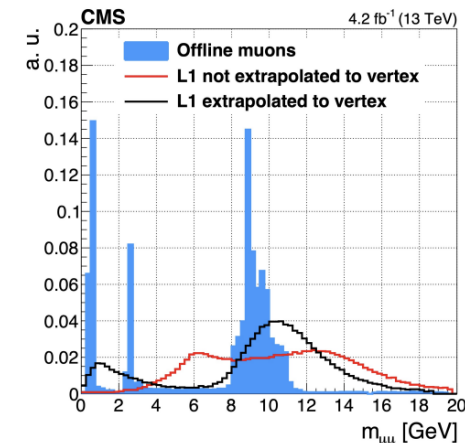
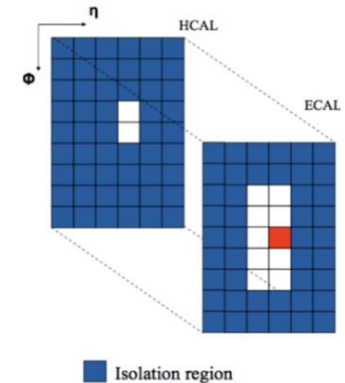
# Trigger reconstruction and selection

- Identifies range of objects: muons,  $e/\gamma$  muon, jets, MET, combined (mass)...
- For example, identification  $e/\gamma$  candidates in calorimeter using methods such as:
  - Clustering and cluster (shape) identification
  - low hadronic deposit versus EM:  $ET_H \ll ET_{EM}$  (hadronic isolation)
- Important requirement: Good rejection of backgrounds
  - Ability to discern between “fakes” (like jets identified by leptons)
  - Improves generally with better resolution, calibration



CMS

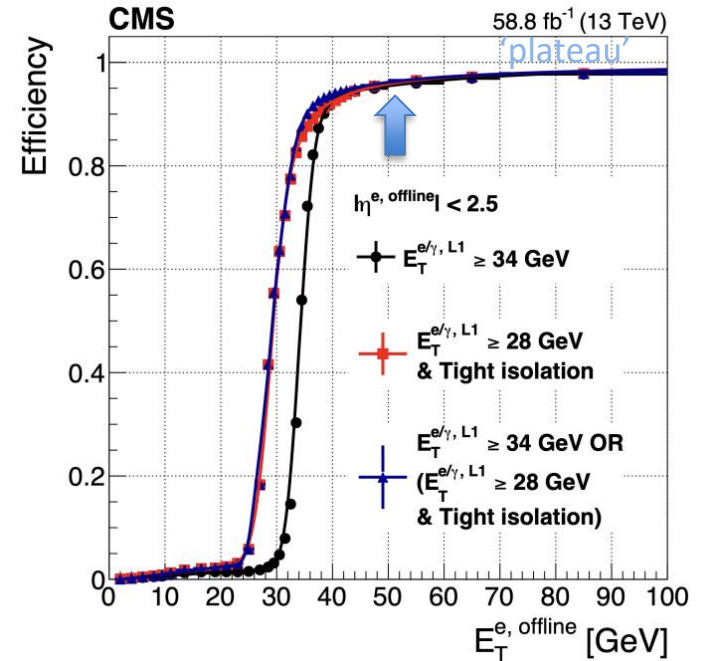
Examples of cluster shapes. Shape veto is used to exclude jet-like candidates.



Di-muon trigger

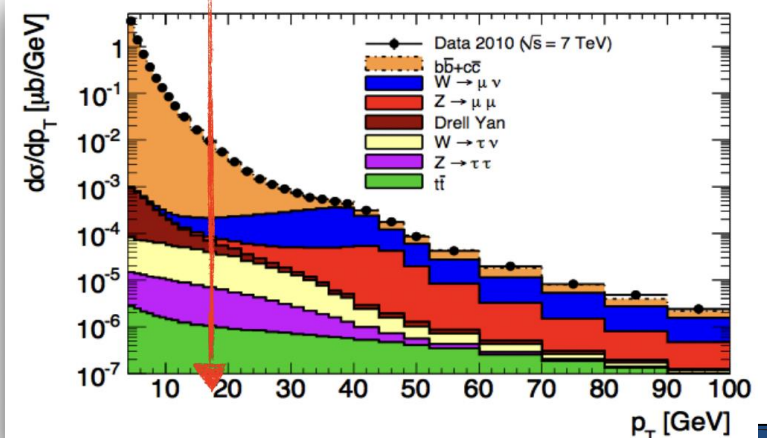
# Trigger selection

- Turn-on curve
  - Signal detection efficiency vs.  $P_T$  or energy function
    - “Turn-on”- kinematic rang where efficiency improves
  - Trigger **threshold** usually chosen in the good signal efficiency area (**plateau**)
    - Such that L1 rate is under control (background rejection)
      - Needs good rejection of “fakes”
      - Even physical processes producing real objects can produce high rate (irreducible background)



CMS – L1 electron efficiency

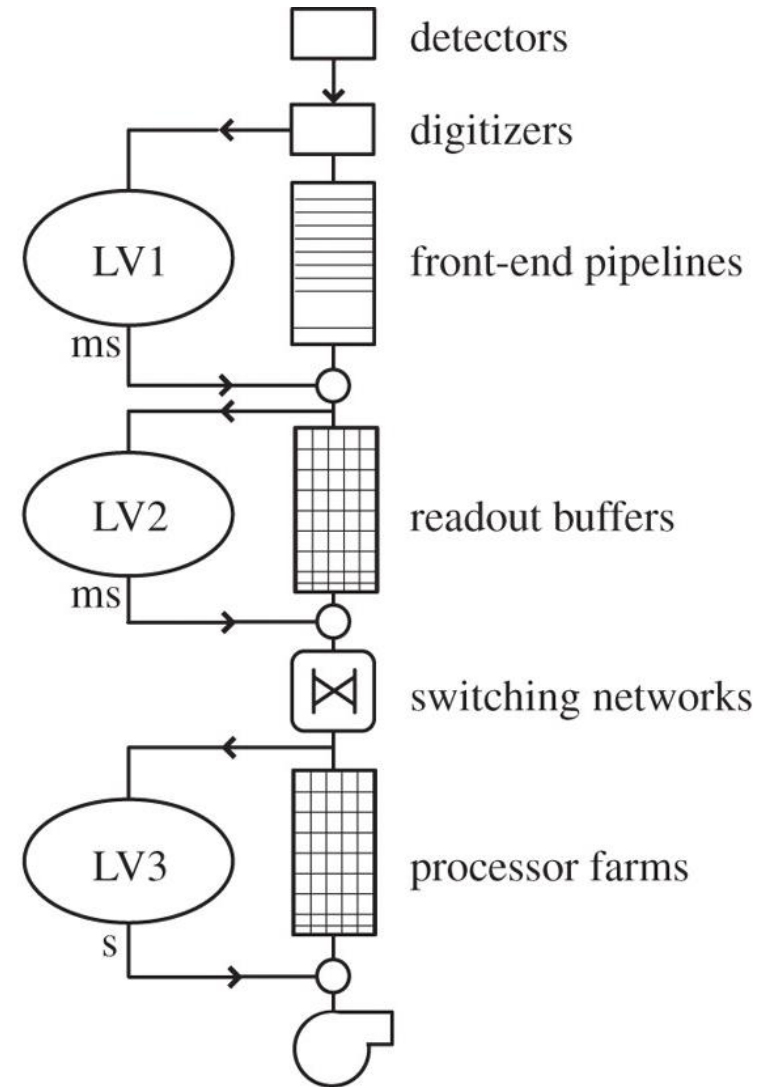
Inclusive single muon  $p_T$  spectrum



# Data acquisition system - DAQ

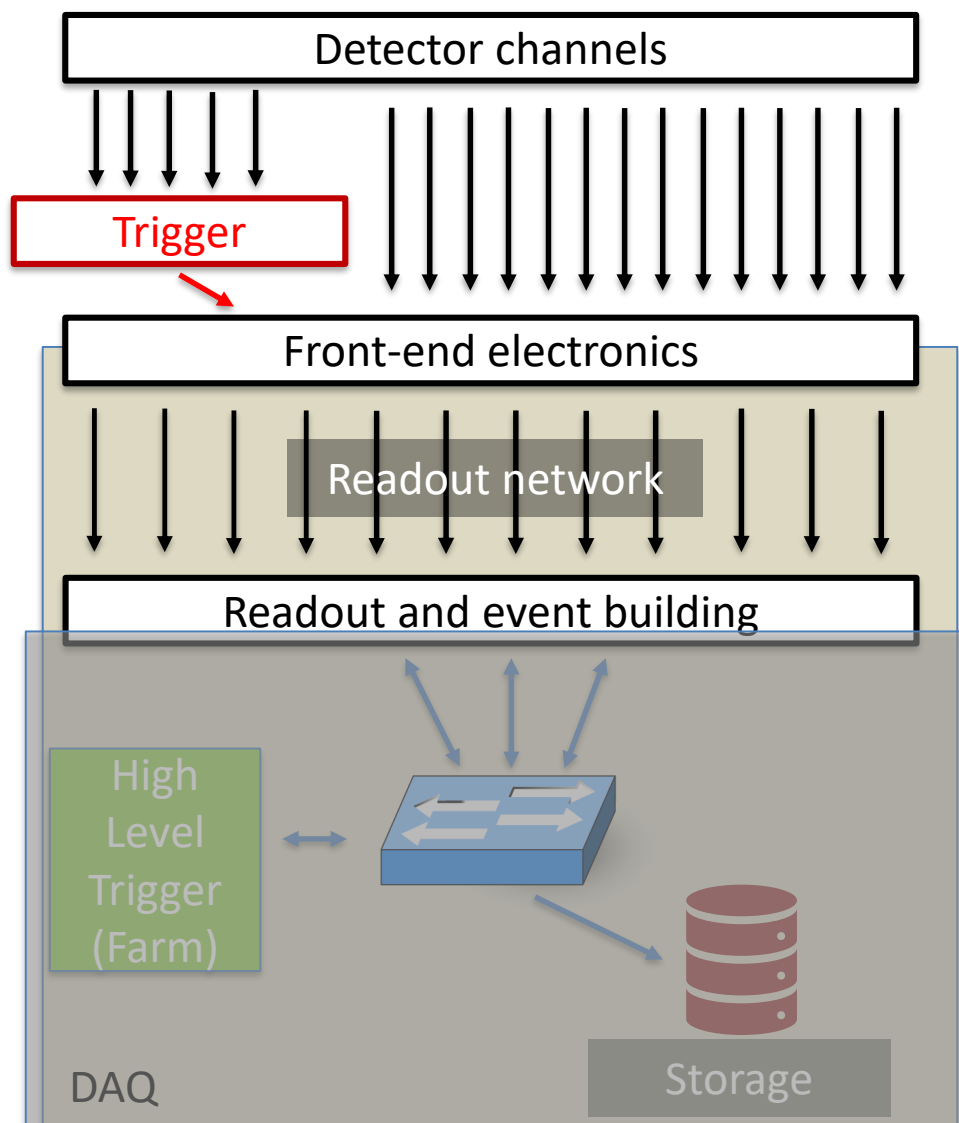
- Tasks

- Gather data produced by detectors (**readout**)
- Often coupled with several levels of data filtering (triggering) - **TDAQ**
  - Data feeded other trigger levels (High Level Trigger)
  - Can also be triggerless (**streaming**)
- Combines readout from multiple sensors into a single object per event (event building)
- Storage of events accepted by the trigger
- Control, configuration, monitoring...



# DAQ architecture - Readout

- Readout
  - L1-accepted event is read out by DAQ (synchronously)
  - Buffering layer – decoupling from timing constraints and trigger fluctuations
  - Usually custom electronics (detector interfaces)
  - transition to commercial equipment
    - More cost effective, can leverage High-Performance-Computing (computer cluster) technologies
  - Includes (typically) a networking layer

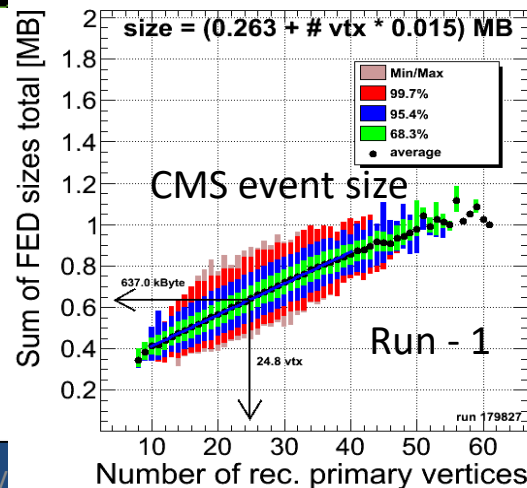
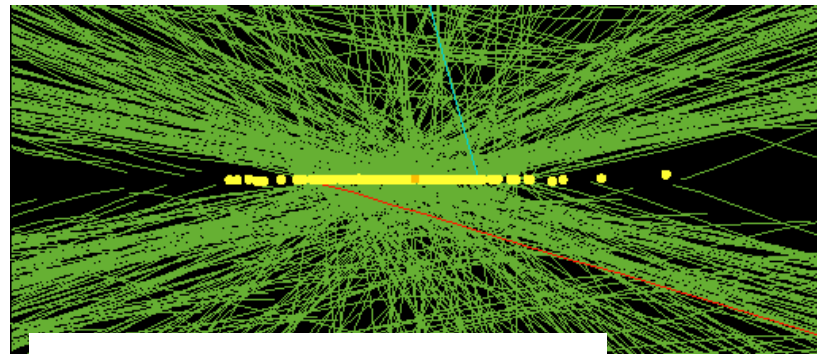


# Readout bandwidth

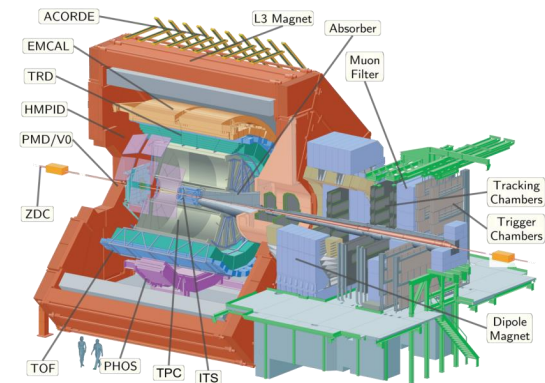
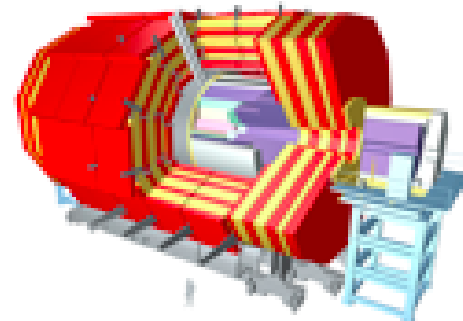
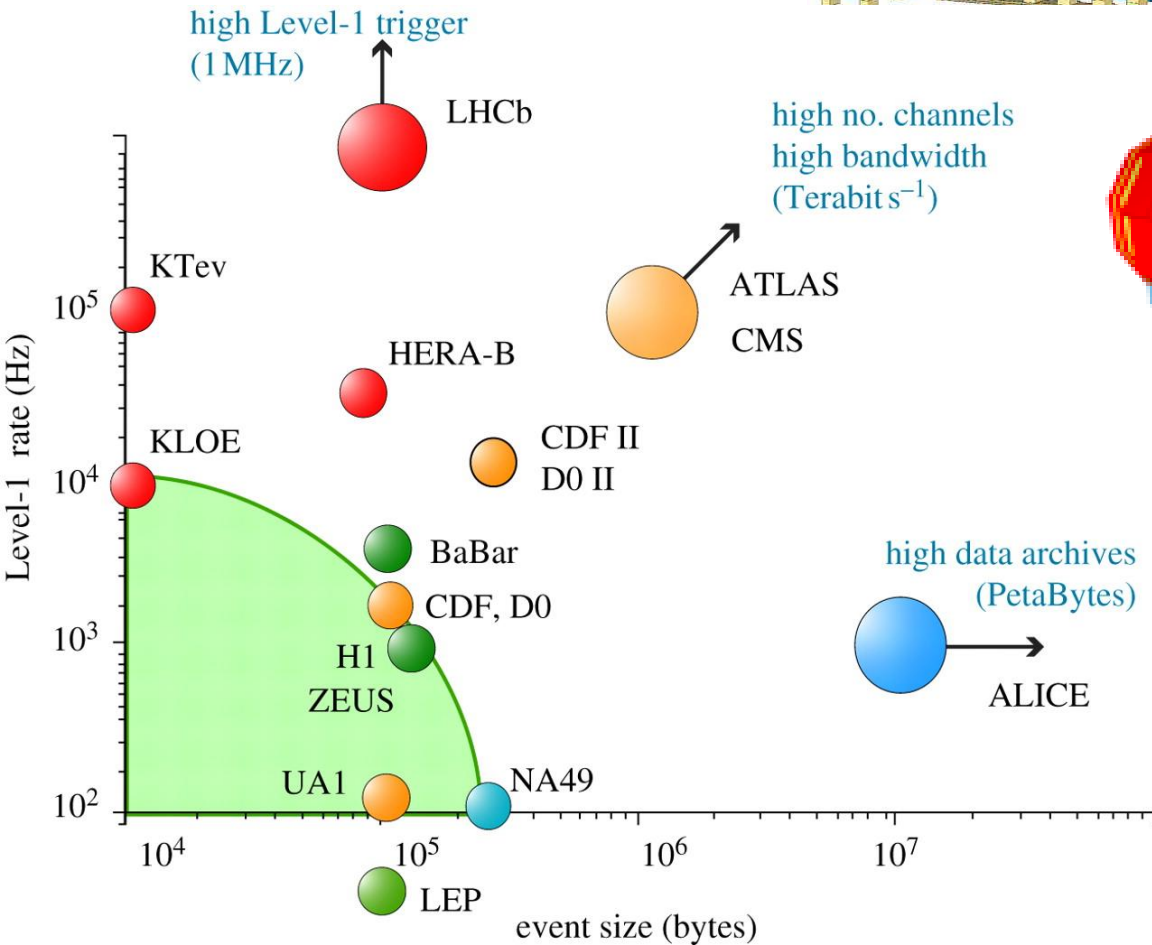
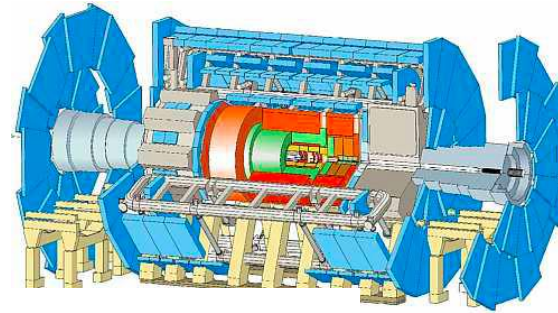
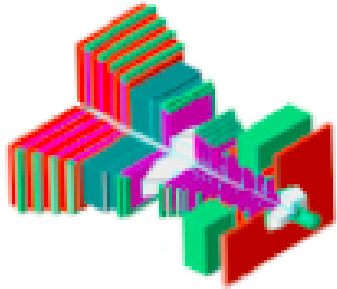
- CMS tracker: 75 million channels
- But typical event (CMS) is < 2 MB (?)
- Not all channels are read out - **zero suppression**
  - Skipping readout of empty channels (no signal detected)
  - A very simple form of compression
  - large size reduction often possible (x10)
- Read-out event size is pileup-dependent!
  - detector occupancy (% active channels) increases with pileup

PILEUP - multiplicity of inelastic particle (pp, ion-ion etc.) collisions in the same bunch-crossing – **scales with luminosity**

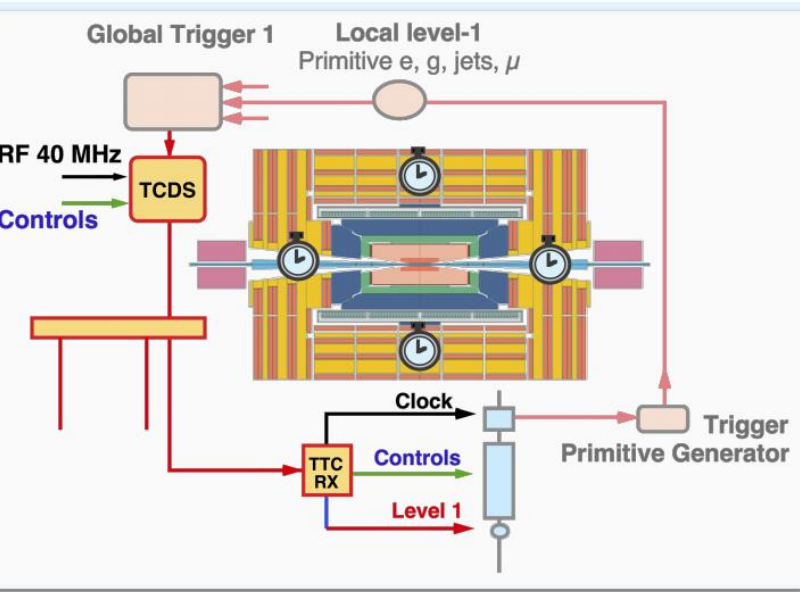
High x-section → probability of collision per bunch-crossing > 1 !  
dominantly soft (low –  $p_T$ )  
→ average: 50 - 60 in LHC run 3



# DAQ design requirements for LHC experiments

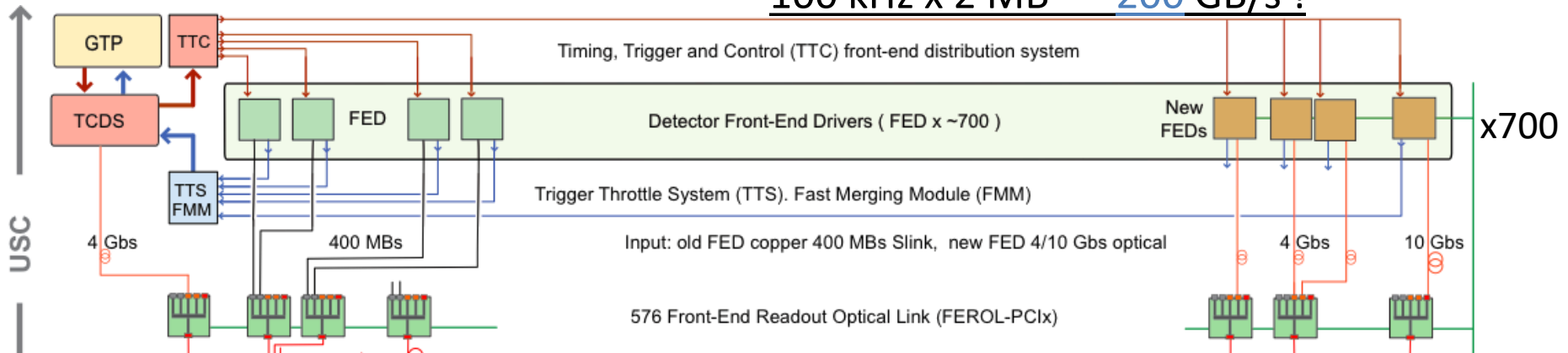


## Example - CMS L1-trigger and readout



- Detector and trigger synchronized to 40 MHz LHC clock
  - Clock propagated by **timing and control distribution system** (TCDS)
  - L1-Trigger decision distributed to FEDs (detector front-end electronics)
    - ~ 700 FEDs send event fragments to readout cards

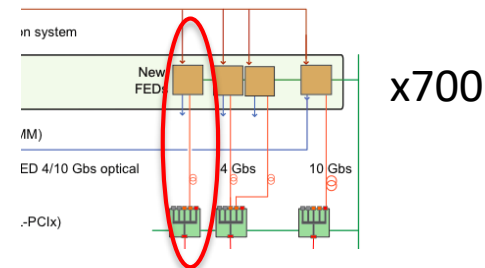
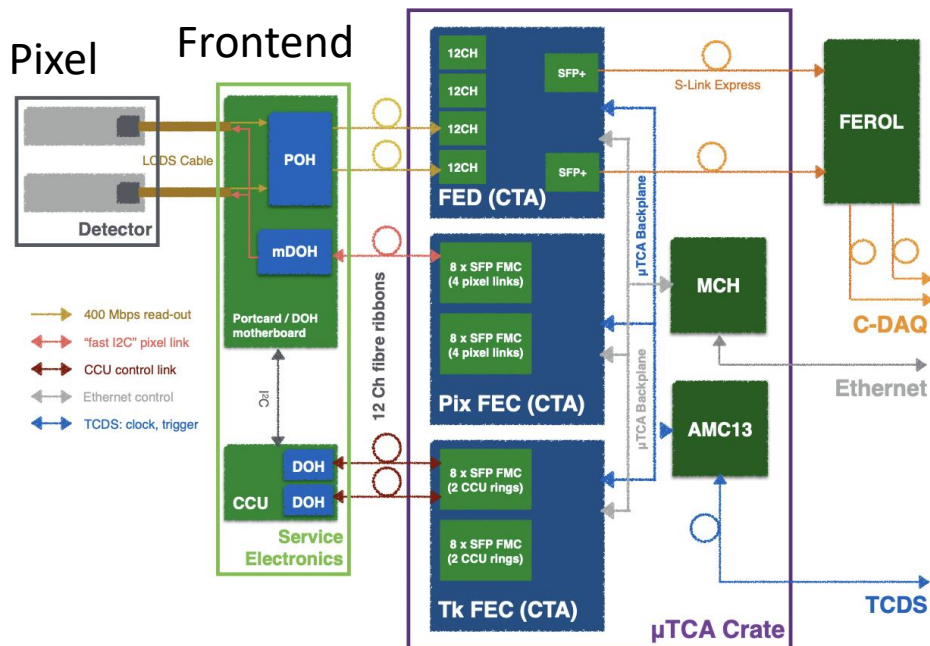
100 kHz x 2 MB - ~ 200 GB/s !



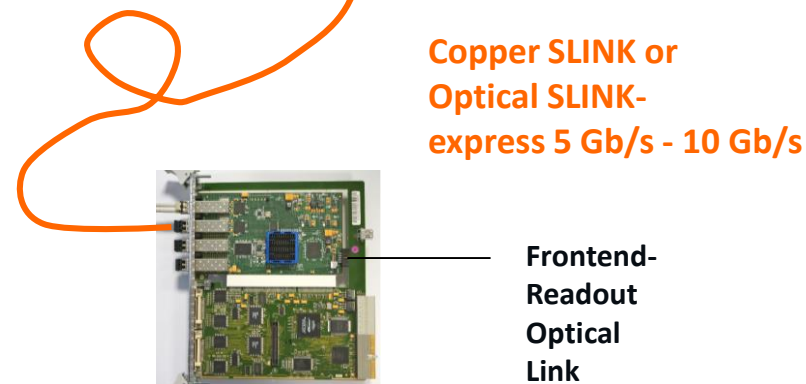
# CMS DAQ Readout hardware

- ➔ Readout electronics based on VME bus (legacy) or  $\mu$ TCA bus (telecommunication equipment standard)

Backend  
(service cavern)

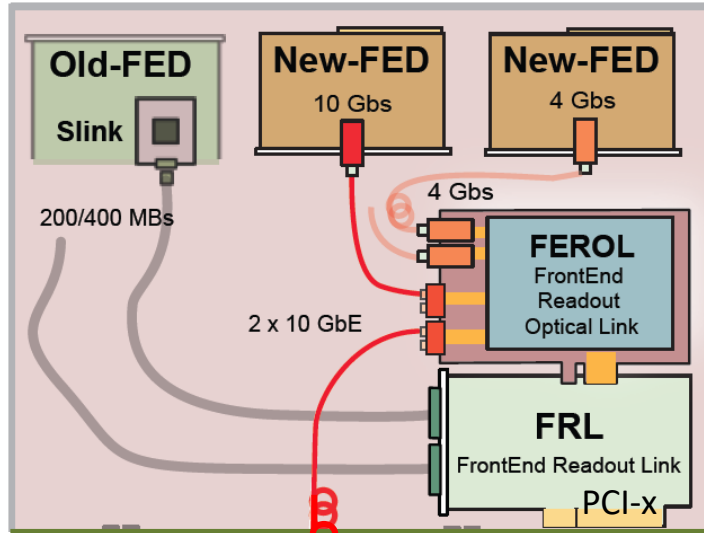


Fragments 2..8 kB



# Frontend-Optical Link & Data Concentrator (CMS)

FEROL

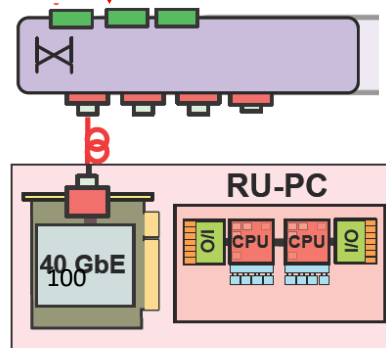


underground  
(CMS service cavern)

10 Gb/s simplified TCP/IP  
from an FPGA

surface

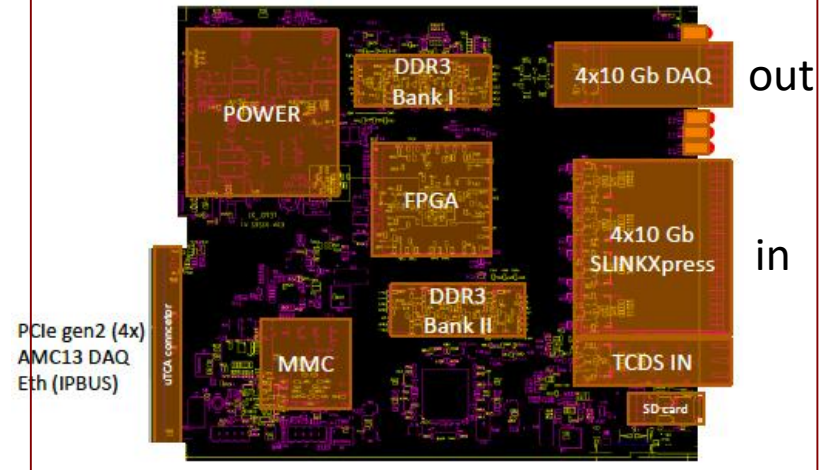
Readout Unit PC



Data concentration:  
10/100 Gb/s Ethernet  
switch

FEROL-40

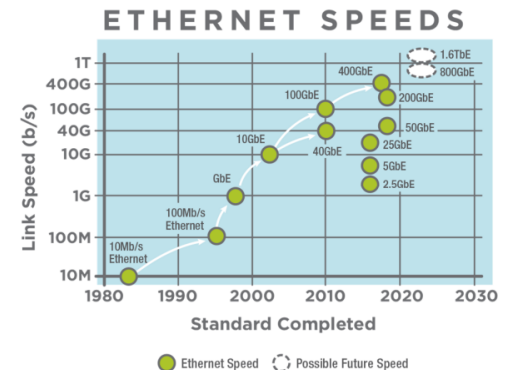
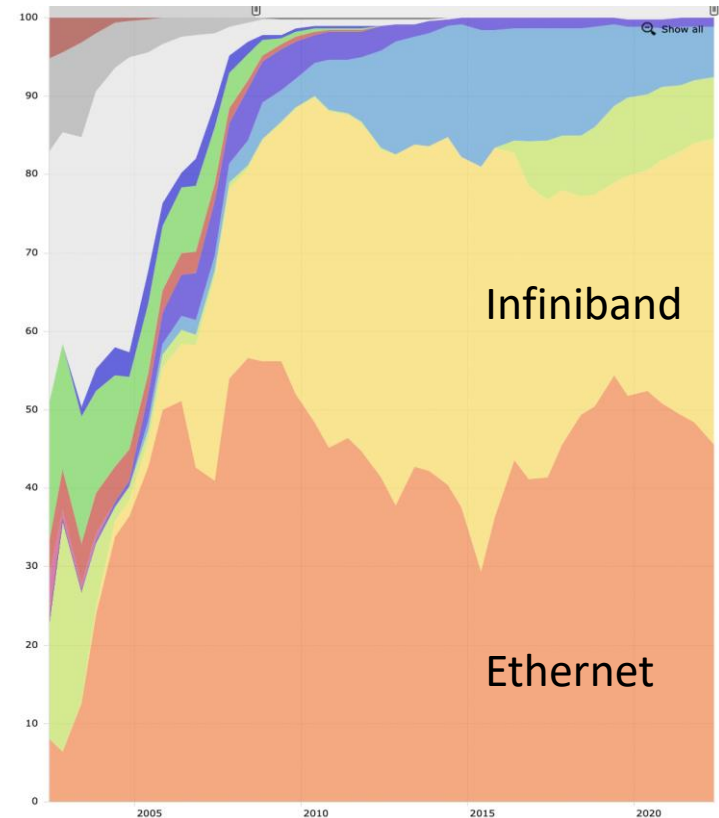
- 4x 10 Gb/s inputs and outputs



Another use-case for FPGAs  
Peformant for data buffering  
and I/O, output protocol  
(TCP/IP)

# Switched networks

- networks for data centers are main drivers for fast (wired) interconnect technologies
- Main contenders:
  - Infiniband (low latency, good congestion control, hardware acceleration/Remote-DMA)
  - Ethernet (10 ... 100 + Gbit/s)
    - Relies on standard TCP/IP
      - Client Implementation in software
      - scalability limited with high rate of packets, bandwidth
    - Recently retrofitted with infiniband-like features, like RDMA over Ethernet - RoCE
- Single-switch or a mesh (if the scale of the I/O requires it)
- Common drawback
  - Congestion – if lines where data is routed become oversubscribed
  - Two differing approaches to managing it:
    - Lossy network – drop packets that don't fit into switch buffers and retransmit
    - Lossless network – buffer everything and pause transmission when full



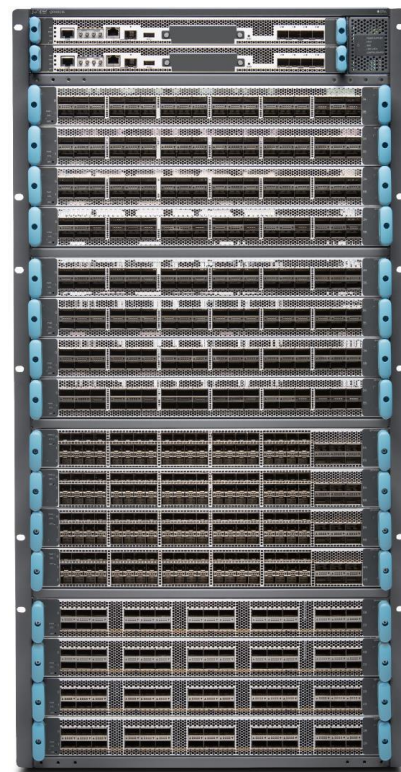
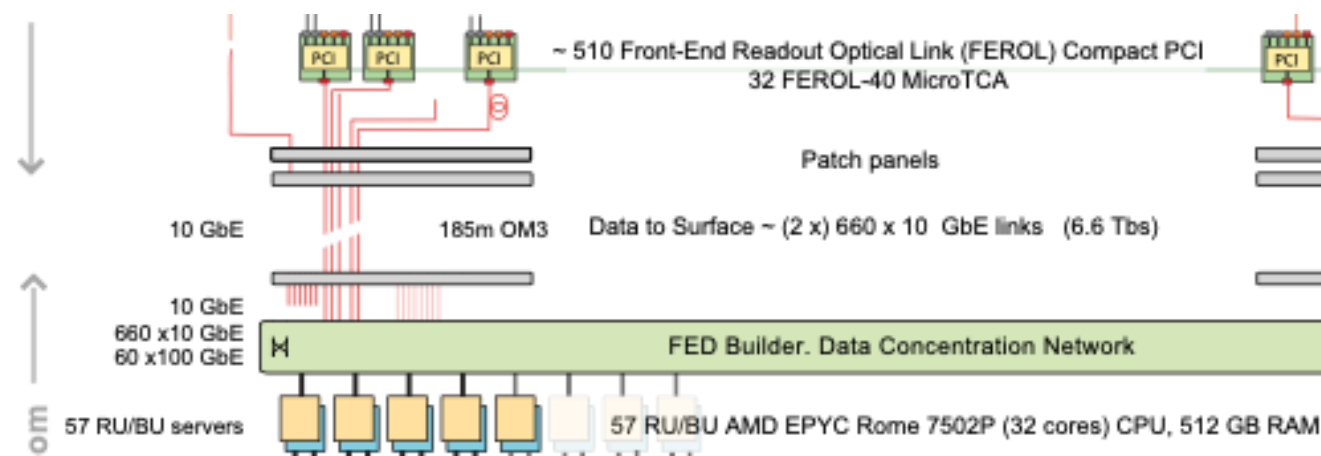
# Example - CMS readout network

- FEROLs → data concentrator network → Readout-Unit (RU) PCs
- 10 → 100 Gbit Ethernet network
- Also a pre-event building stage
  - 510 sources → ~ 50 'superfragments'

For run 3:

- fully handled by a single chassis-based switch

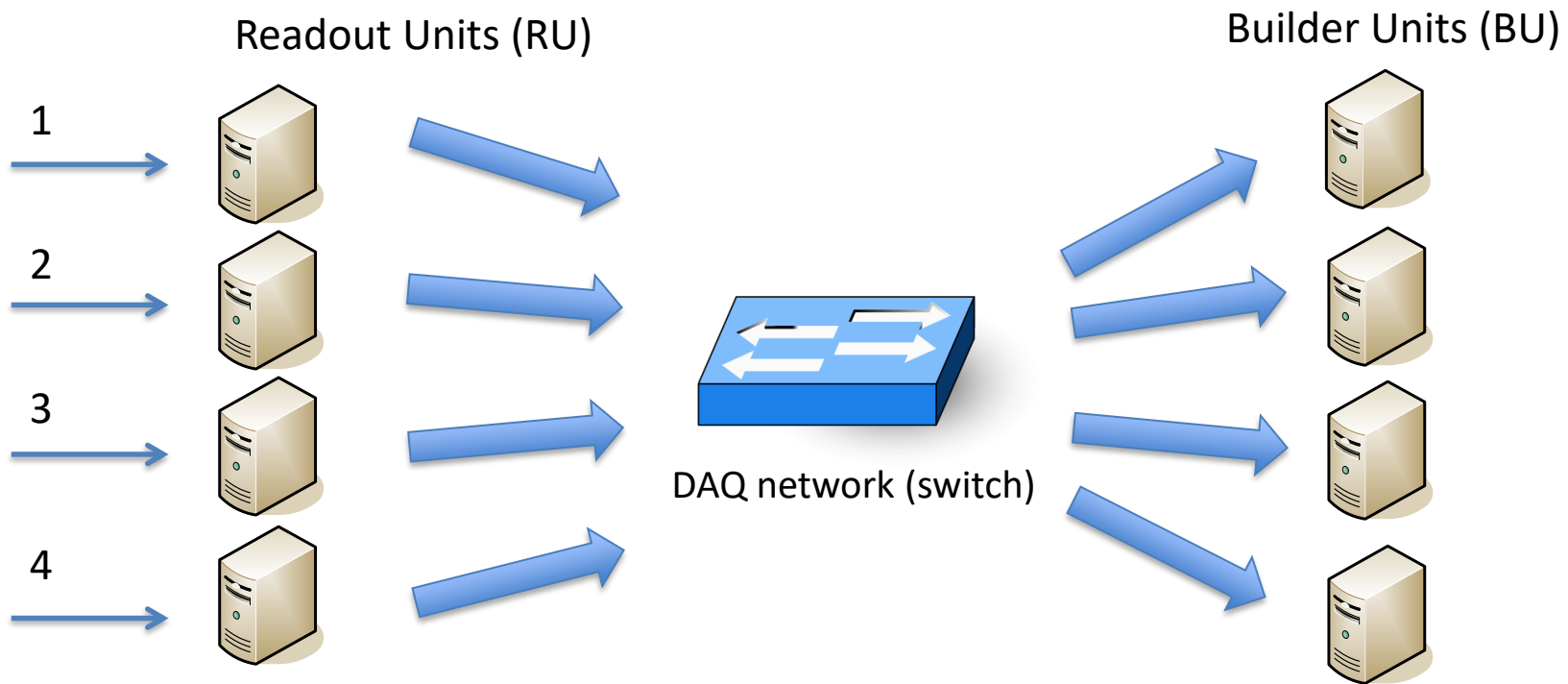
**Juniper QFX10008**



Juniper  
QFX10008

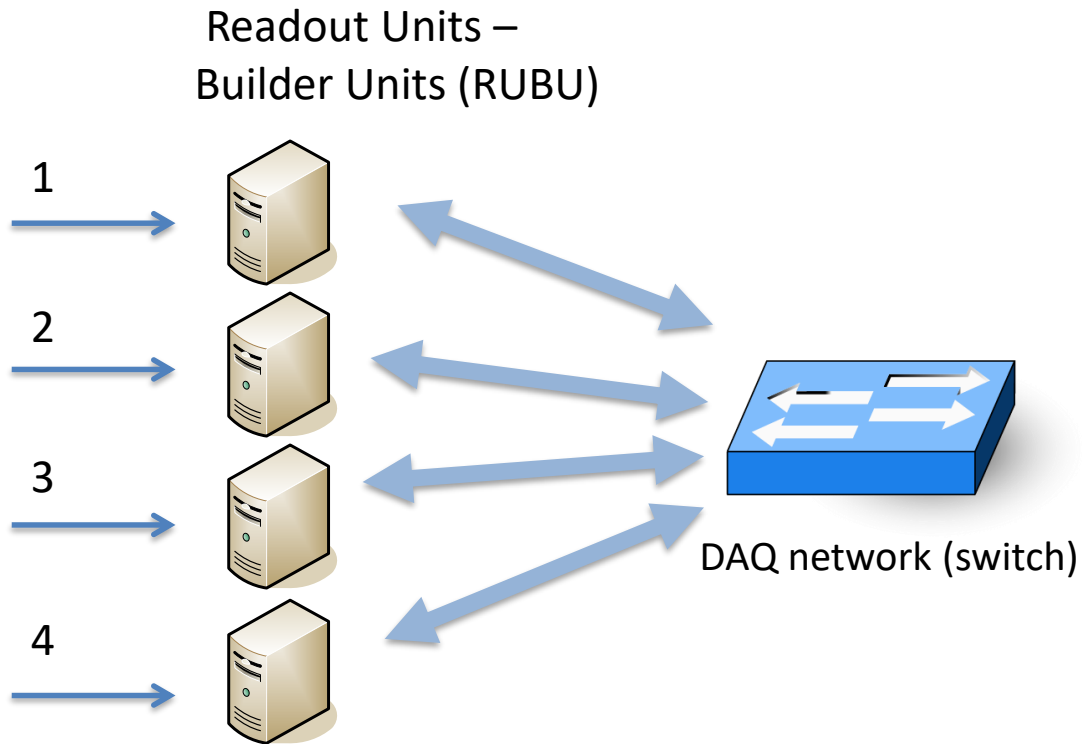
# Event Building

- Commonly implemented using switch network



# Event Building – folded architecture

- Merged functionality of readout and build unit

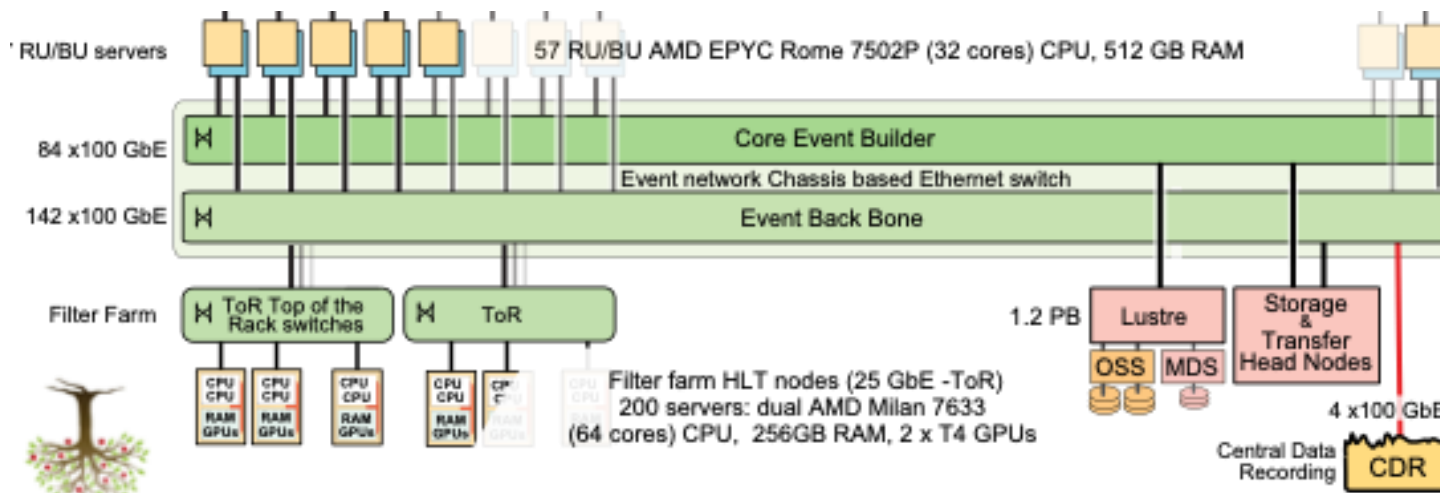
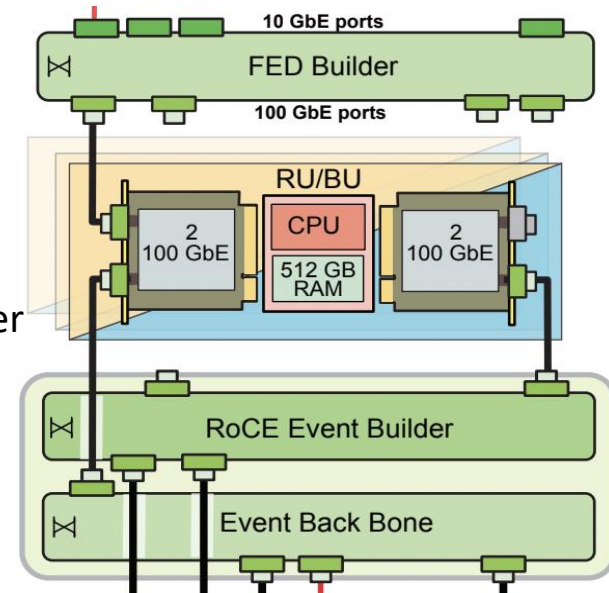


# Event building hardware - CMS

- On another chassis-based Ethernet switch (2<sup>nd</sup> layer Ethernet):
  - Event Building network
    - approx. 50 x 50 links
  - Event Backbone network
    - To ~ 200 HLT CPUs (via top-of-rack switches)
    - Serving High-Level-Trigger + storage I/O (input + output)

~ 50 I/O nodes - RU/BUs:

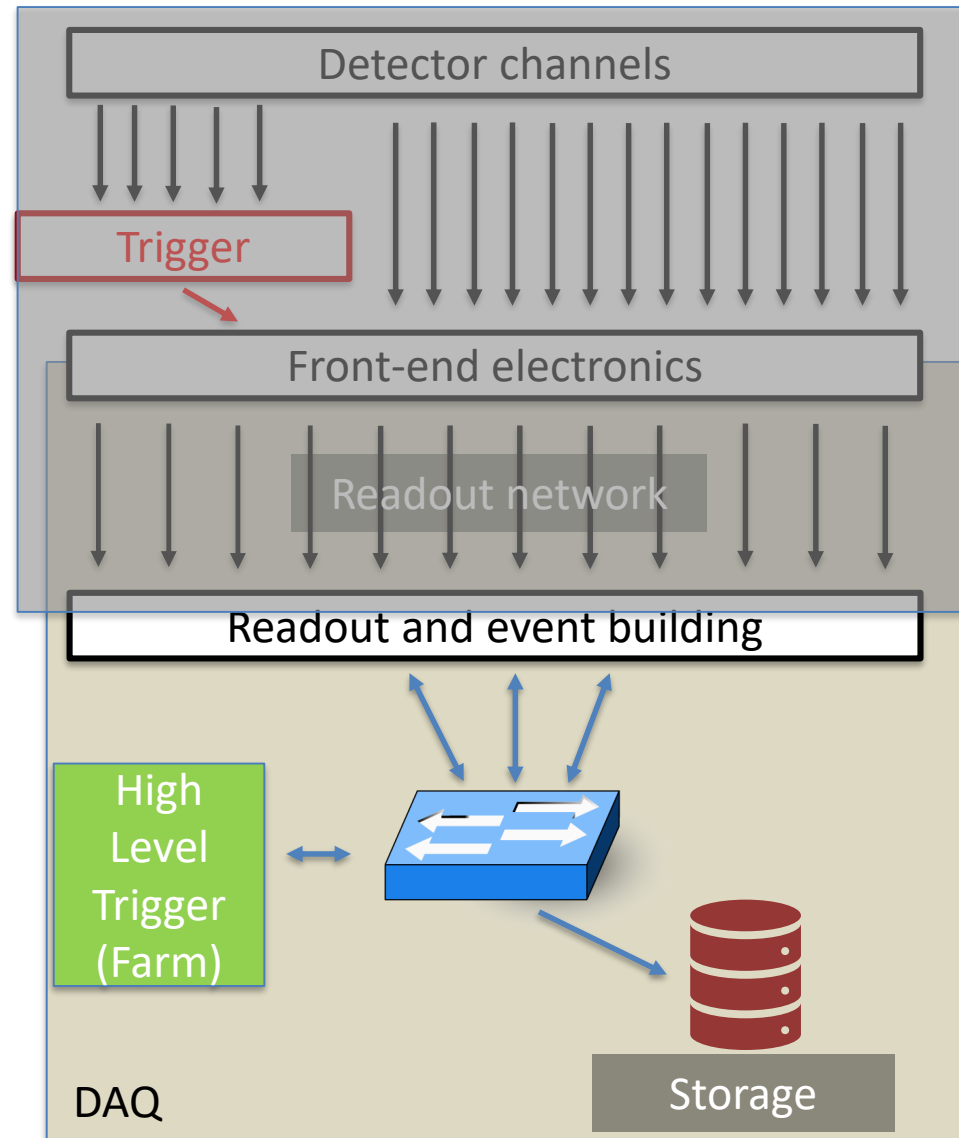
- Folded EvB
- 3 x 100 Gbit/s
- Use RDMA over converged Ethernet ([hardware acceleration](#))



Event-building switch And servers

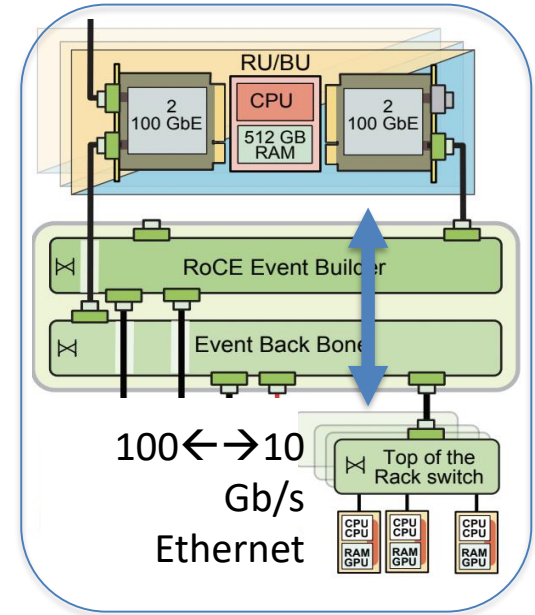
# High Level Trigger (HLT)

- Goal – further x100 rejection/filtering
- Commonly: general-purpose computer farm running HLT **software**
  - Fine-grained reconstruction/selection algorithms
  - Full-resolution event data (no “primitives”)
  - CPUs + increasingly GPUs (possibly other accelerators)
- More processing time (latency) allowed than for L1 (< second)
  - far more buffer space – often in computer RAM
    - 100+ GB per server
- Output (selected events) – final data selection → saved permanently to storage
  - In the range of *few kHz*



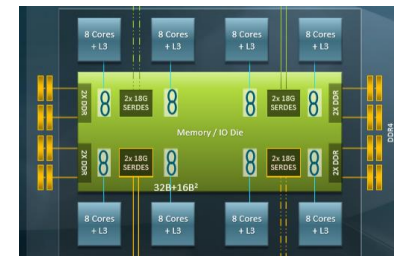
# HLT Filtering in CMS

- Filter Farm cluster - executing software processes
  - Significant CPU and GPU power – 25k CPU cores + 400 GPUs
    - [Heterogeneous architecture](#)
  - Timing budget:  $100 \text{ kHz} \times [\sim 250 \text{ ms / event}]$  (CPU)
  - Runs a common CMS Software (CMSSW) framework – code sharing between HLT and offline analysis
- Event data moved to/from nodes using 100/10 Gb/s Ethernet network
  - Accepted data copied back to I/O nodes and on-site storage system (and, finally, to CERN [Tier0](#) for permanent repacking & storage)
  - Standard data transfer protocols used: [remote filesystem \(NFS\)](#) over TCP/IP
    - “File based Filter Farm” - F3
  - Total: 20 TB of RAM buffer for I/O available



2022  
CMS HLT  
Farm

CPU (x2)	AMD Milan 7763 64 cores @2.45 GHz 256 GB RAM
GPU (x2)	Nvidia Tesla T4
# servers	200
# cores	25600
kHS06	645



AMD  
Epyc CPU  
architect-  
ure

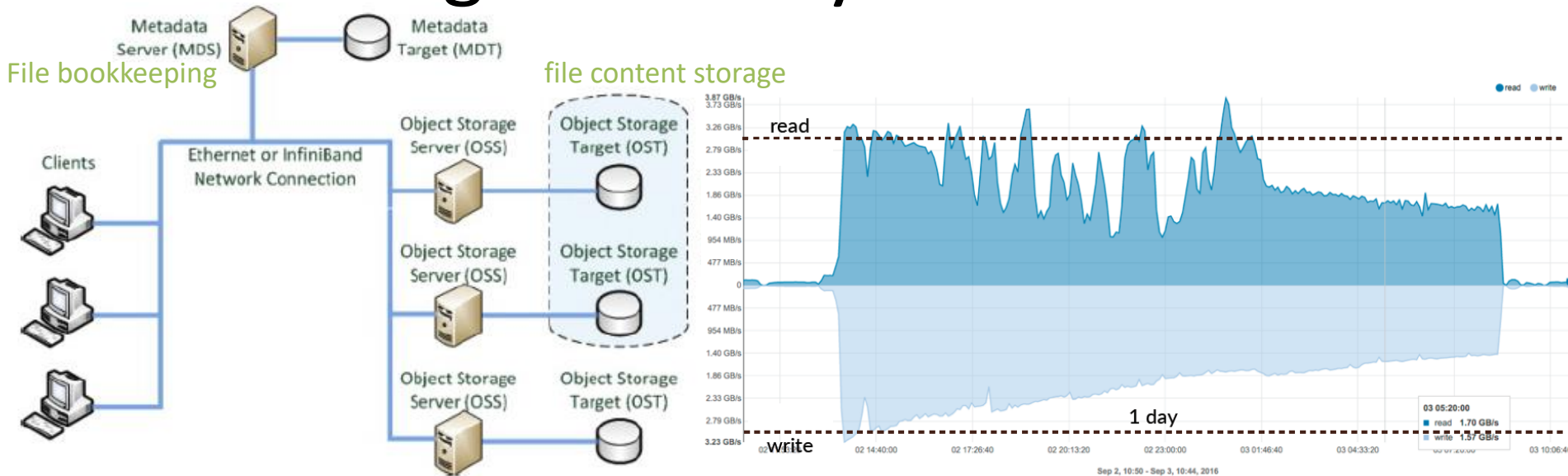


Nvidia  
Tesla T4

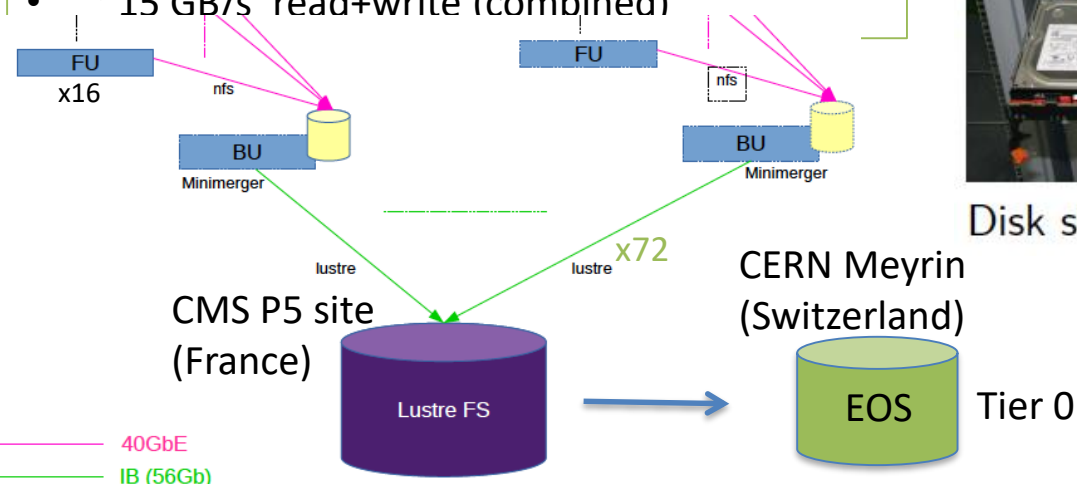
# High Level Trigger algorithms

- Start (*seeded*) with objects from L1
  - Muons, electrons, photons,  $\tau$ -jets, jets, missing- $E_T$ 
    - In HLT, availability of full detail detector data – including trackers (unlike L1)
    - Can apply [calibrations](#) and other [detector conditions](#)
    - Better discrimination of lepton fakes (jets), improved isolation (muons)
    - Exploit event topology and association between objects (mass cuts and similar)
  - Software-based:
    - No hard-coded hw. constraints
    - Can run complex algorithms: MVA (machine-learning based) selection etc.
    - Upgrades possible by adding more “COTS” computing power, bandwidth
    - Some algorithms more efficient on [GPUs](#)
      - CMS: Pixel tracking, ECAL and HCAL reconstruction and calibration
- HLT trigger menu
  - A collection of trigger selection criteria
    - tailored to physics priorities of the collaboration

# CMS global filesystem: Lustre



- Built from multiple metadata (MDS) and OST servers
- 1.2 PB usable space
- ~ 15 GB/s read+write (combined)



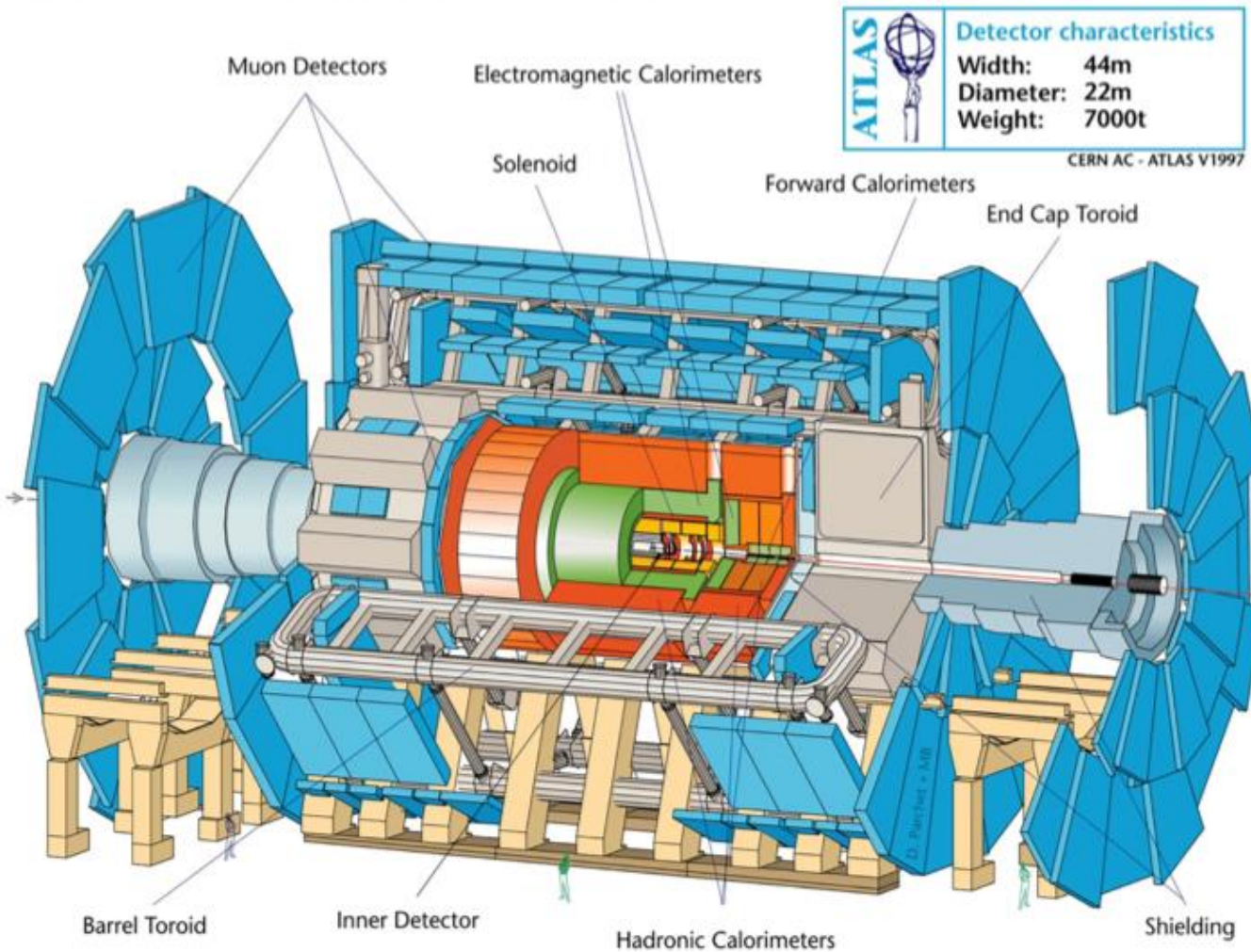
Disk shelves



Front OST

# ATLAS

## A Toroidal LHC ApparatuS



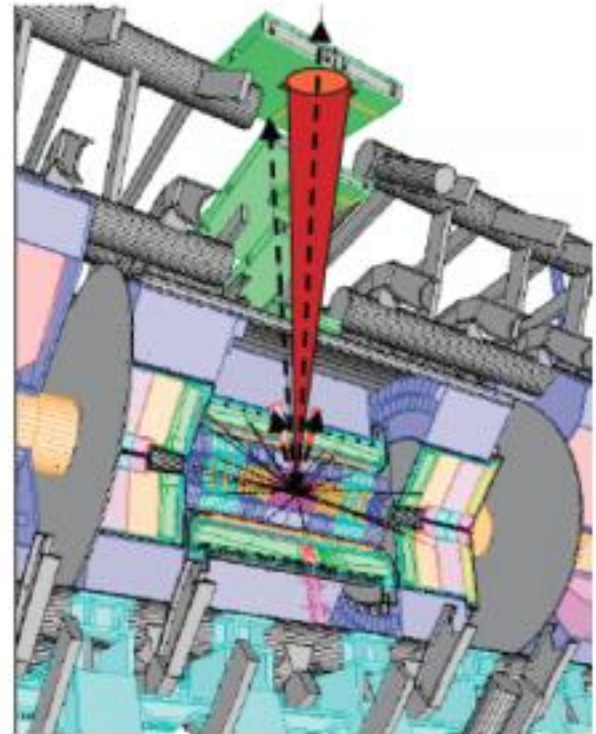
- Tracker and Pixel Detectors
  - 6 M channels: 80  $\square$ m x 12 cm
  - 100 M channels: 50  $\square$ m x 400  $\square$ m
  - space resolution  $\sim 15 \square$ m
- Solenoid 2T field momentum measurement
- Fine grained EM and hadronic calorimeters
- muon spectrometer (streamer tubes)
- 8 superconducting toroid magnets

### Run 3 – upgraded detector systems

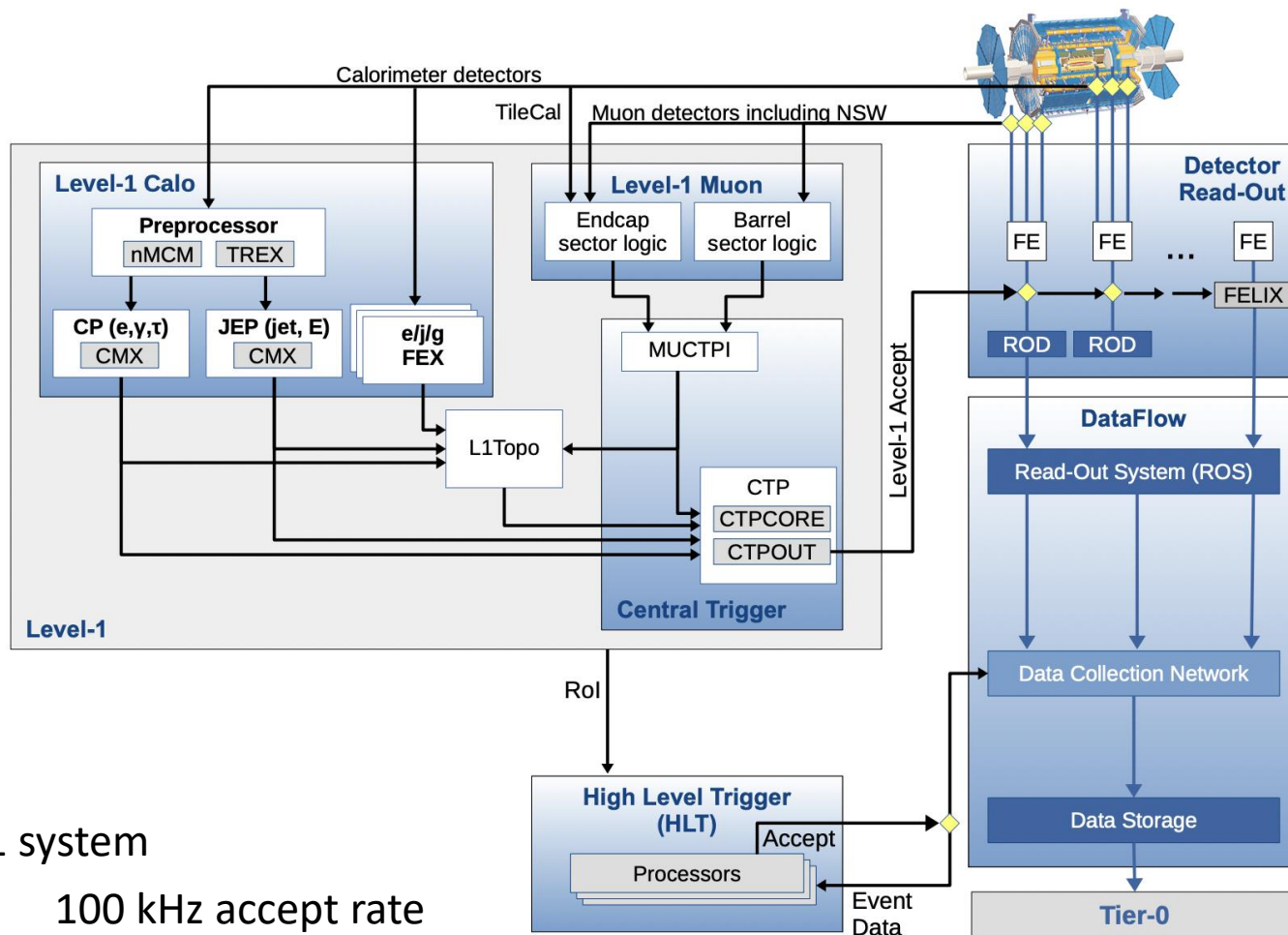
- Muon system: New Small Wheels, Inner Barrel RPC
- Calorimeters: Liquid Argon (LAr) digital readout, Tile run 4 demonstrator

# ATLAS Trigger-DAQ

- ATLAS operates on concept of Regions of Interest (ROIs) where data processing + readout proceeds in stages :
  - Level-1 trigger
    - Fast custom-hardware trigger, discrimination on trigger “primitive” data
    - defines **ROIs** for each L1-accepted event
  - High-Level trigger
    - Software initially seeded by ROI
    - Only ROI regions are read from detectors for HLT



# ATLAS TDAQ Run 3 diagram

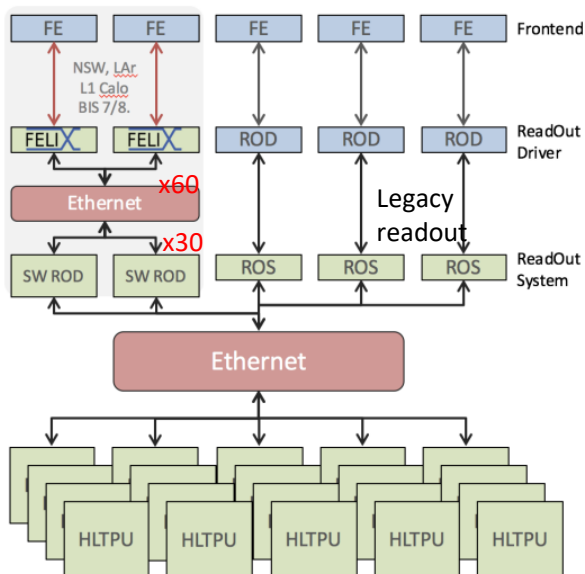


## L1 system

- 100 kHz accept rate
- L1 topo trigger
  - applying kinematic and angular requirements on electromagnetic clusters, jets, muons and total energy

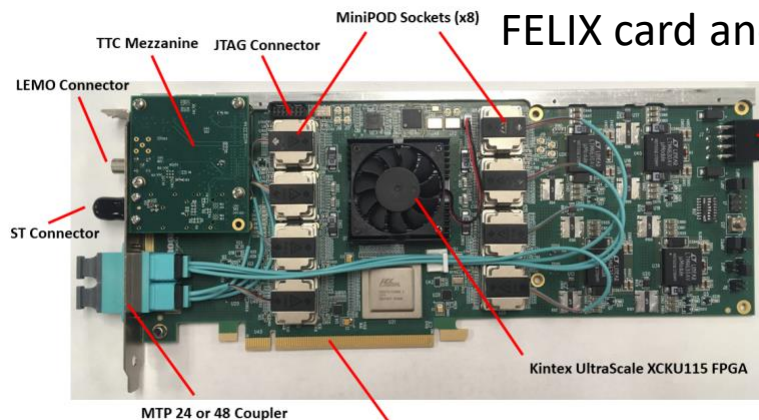
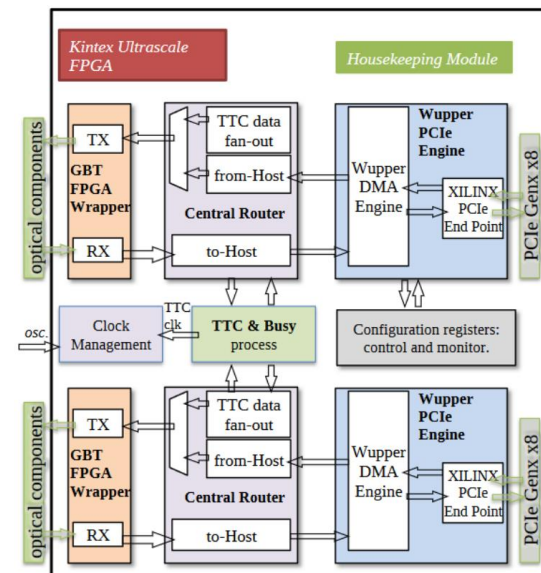
event size up to 2 MB

# ATLAS Readout



FELIX Host servers: Intel Xeon 8-core  
With 25/100 Gb/s Ethernet NIC

Legacy readout (run 1 and run 2)  
via PCIe cards to a PCs with 4 x 10 Gbit Ethernet



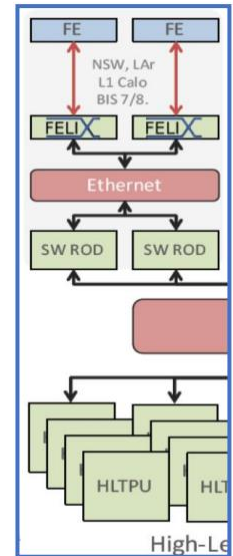
FELIX card and diagram

**FELIX** PCIe card (FLX-712) – for upgraded detectors in Run-3  
Readout from on-detector electronics:

- **24x** connections - 4.8 Gb/s(GBT) or 9.6 Gb/s  
Read into PCs → Ethernet (RDMA support!)  
Timing and BUSY logic

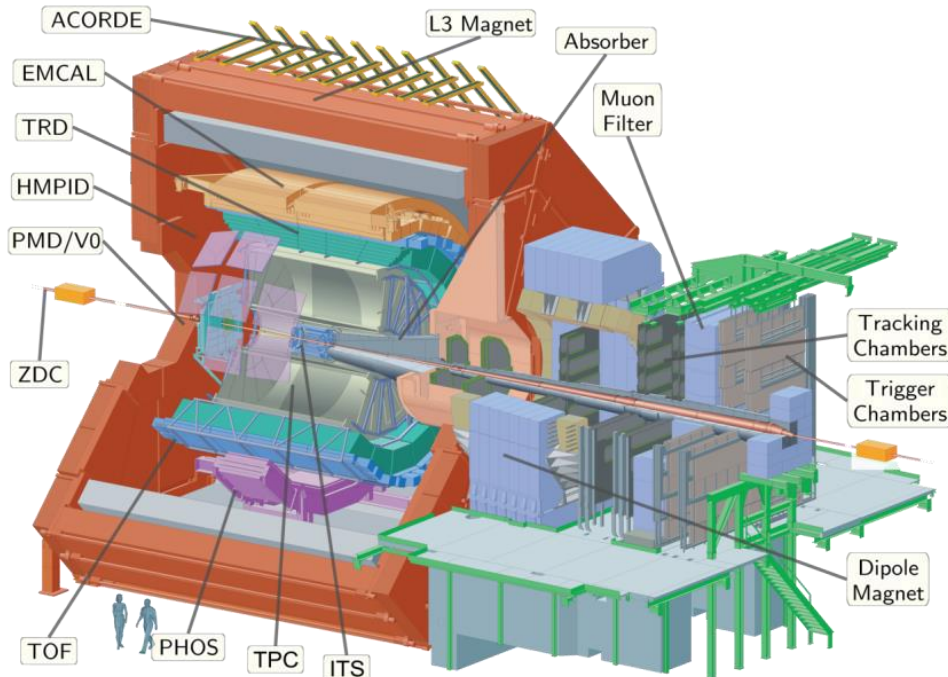
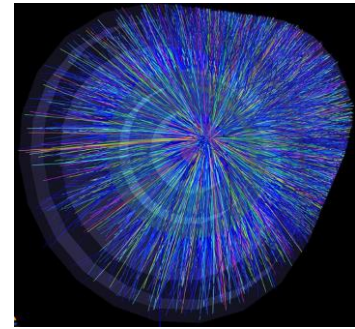
# ATLAS SW-ROD and HLT

- SW-ROD servers
  - Buffer readout data (FELIX) and transfer to HLT on demand
  - Working along the similar legacy system (ROS)
  - 2 x Xeon 16-core CPUs – 96 GB RAM
  - Input: **100 Gb/s** Ethernet NIC from FELIX
  - Output: **40 GB/s** Ethernet NIC to HLT
- HLT FARM
  - Large number of PCs / computer cores
  - 50,000 processing applications with 200 to 400 ms event processing time (on specific regions of the detector)
  - ROI based – readout-on-demand (progressively) until decision on accept/discard is reaches
  - Selecting about 3,000 events / seconds (output)
  - Software in Run-3 runs shared ATLAS software framework (AthenaMT)



# ALICE

- Dedicated mainly to heavy-ion physics
- Run 2: had a 3-stage hardware trigger
- DAQ handles two scenarios:
  - Large rate of very small events (pp)
  - Rare but very large events: Pb-Pb  $L = 10^{27} \text{ cm}^{-2}\text{s}^{-2}$ 
    - ➔ up to 1 GB for central collisions
    - ➔ software trigger **HLT** used for compression rather than rejection

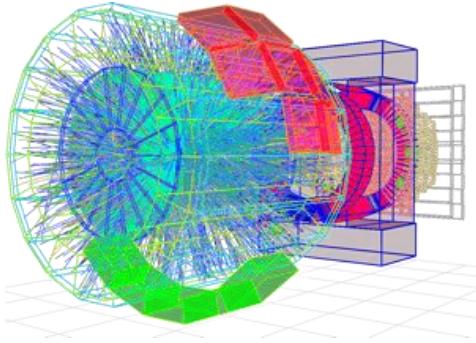


- Run II (2015):
  - Detector readout @ 17 GB/s - 1.4 to 3.5 kHz
    - 40/10/1 Gbit Ethernet based
  - 6 GB/s compressed data to disk sustained

# ALICE Run 3 (and Run 4) DAQ

- **Triggered event reading in Run-1 and Run-2**
- Run 3 - 50 kHz interaction rate of Pb-Pb collisions
- Upgraded detector (TPC with GEMs, new Inner Tracker, Muon system, improved trigger and readout)
- **Goals:**
  - Measurement of rare probes at low  $p_T$  which cannot be selected with a trigger (focus on charm physics)
  - Read-out all Pb-Pb interactions at a maximum rate of 50kHz  
→ **continuous (triggerless) readout of Time Projection Chamber (TPC) and Inner Tracking System (ITS)**
  - Instead of event readout - the output is **Time Frames** (1000 events in one, ~20ms) - bandwidth reduction
  - keeping trigger readout of calorimeters and muon systems
- **Target: recorded Pb-Pb luminosity - factor 100 in statistics over the Run1 + Run2 programme.**

# ALICE O<sup>2</sup>



↓ **3.5 TByte/s**  
into PC farm

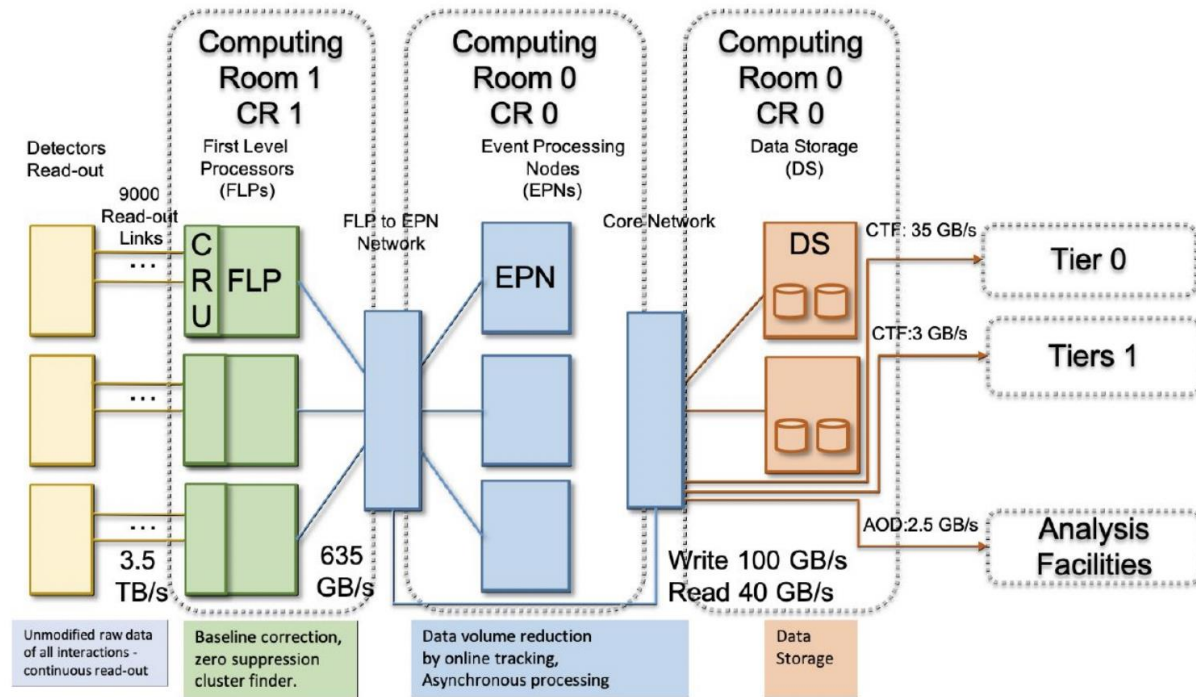
## O<sup>2</sup> (Online Offline) System

Joined Trigger and offline analysis (partial)  
Including calibration and reconstruction  
online, data compression

↓  
**90-100 GB/s**

**STORAGE**

Stored event data rate:  
Pb-Pb 50 kHz pp and  
up to 200 kHz p-Pb



• A

O2/FLP

9000 fibers

O2/EPN

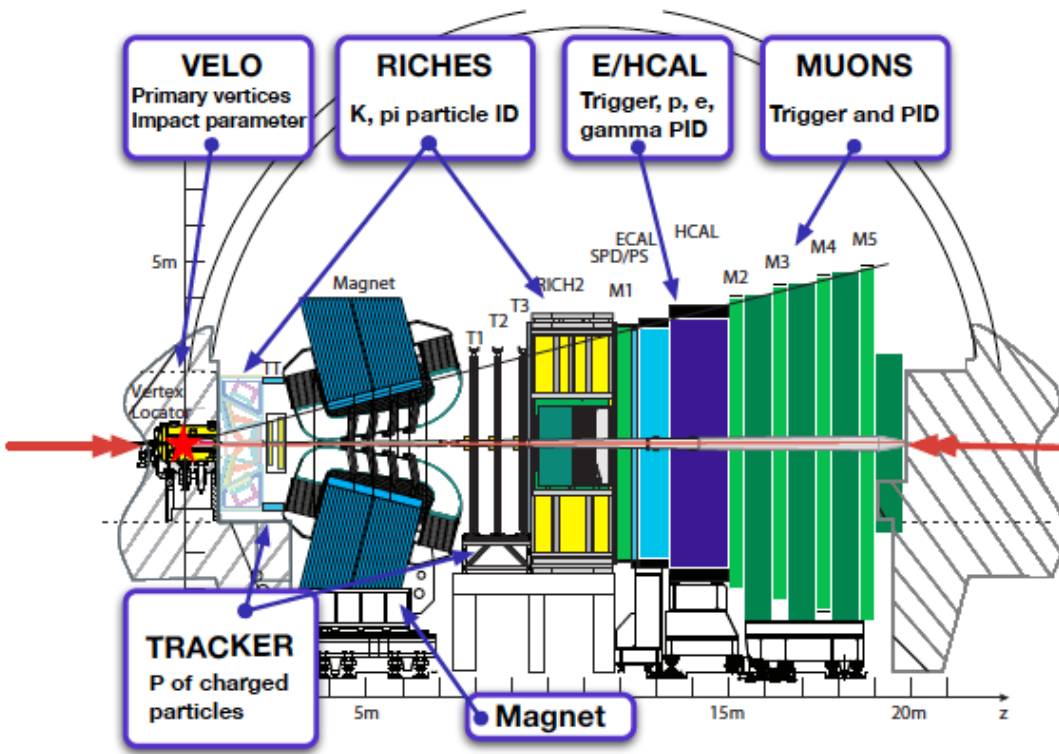
O2/PDP

Year		Year + 1											
N	D	J	F	M	A	M	J	J	A	S	D	N	D
S. RECO													
CALIB													
		A. RECO											
		SIM											
									A. RECO				
									SIM				

**PbPb Run Synchronous**

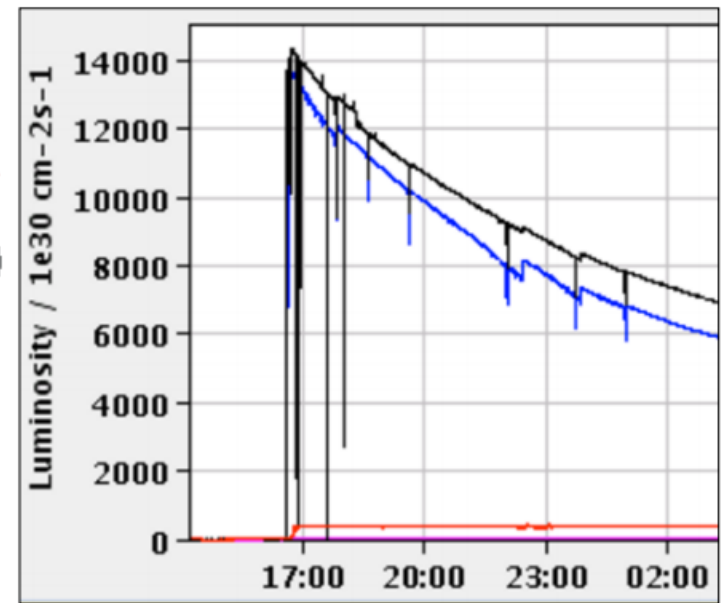
## Asynchronous 2

# LHCb



At 13 TeV and  $\mathcal{L} = 4 \times 10^{32} \text{ cm}^{-2} \text{ s}^{-1}$ :  
 $\sim 45 \text{ kHz } b\bar{b}$  pairs and  $\sim 1 \text{ MHz } c\bar{c}$  pairs

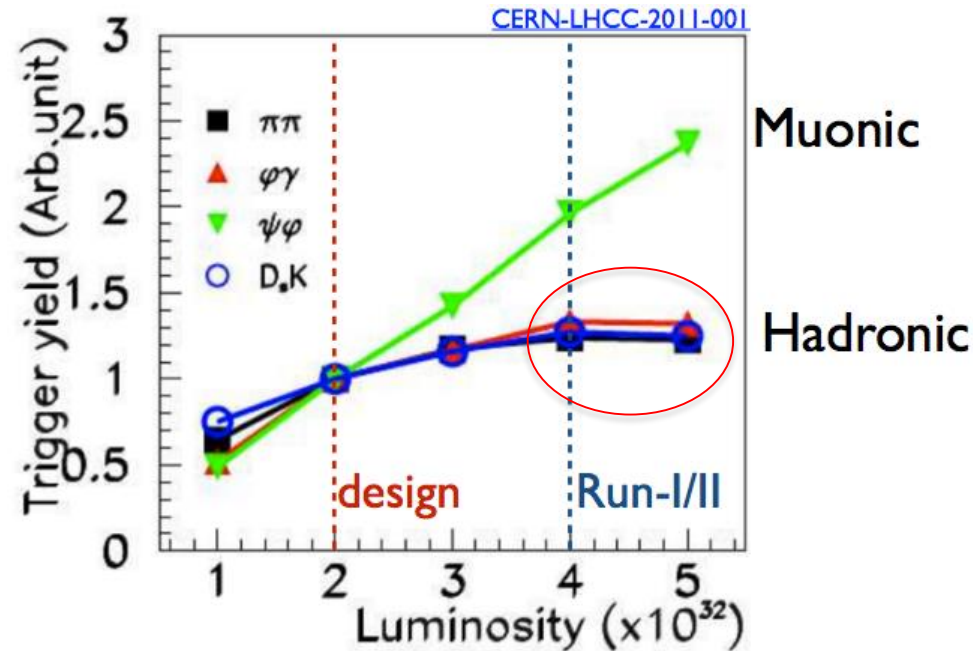
## LHC luminosity



LHCb

# Run 3 LHCb triggering

- Handle x5 higher inst. luminosity than in run I - II ( $2 \times 10^{33} \text{ cm}^{-2}\text{s}^{-1}$ )
- Hadronic triggers loose efficiency at higher luminosity



- Software trigger coping better with it
  - Scalability: enough CPU power can do track reconstruction at full rate
  - More relaxed latency constraints
  - Cheaper: commercial off the shelf PC hardware
  - 2x higher efficiency than HW trigger

# Run 3 LHCb DAQ architecture

- No HW trigger
- All data read out at full rate (40 MHz)
  - 10000 x 350 m optical links (detector to surface)
  - 500 readout boards
    - 4.8 Gbit/s per link

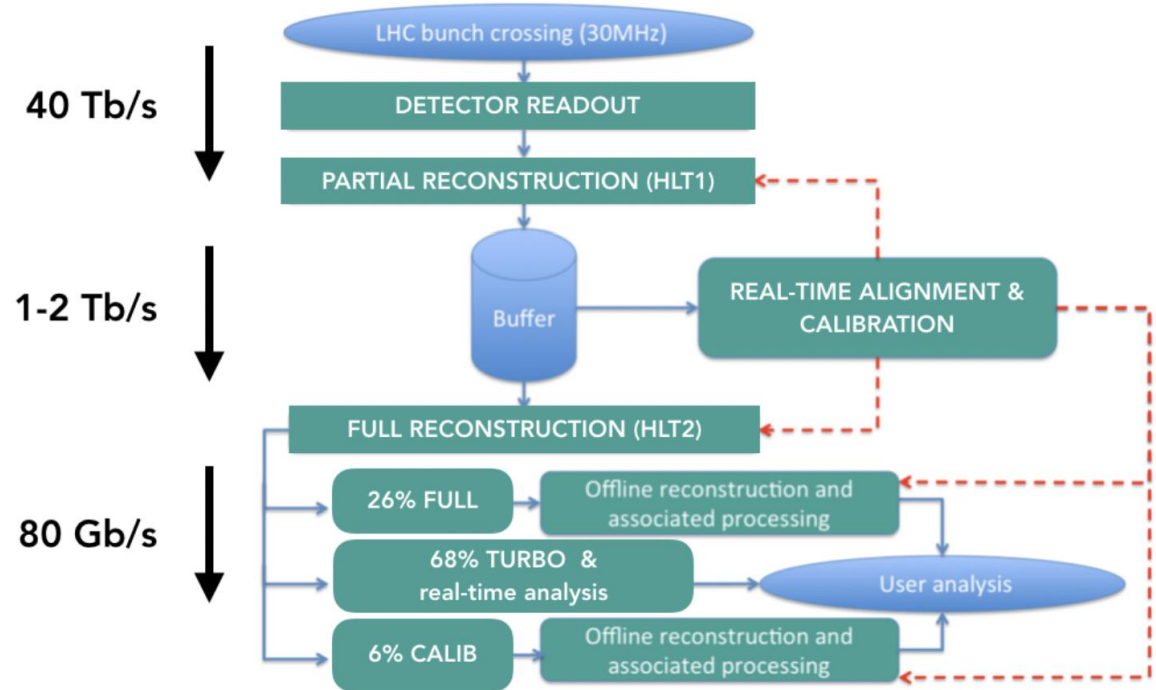
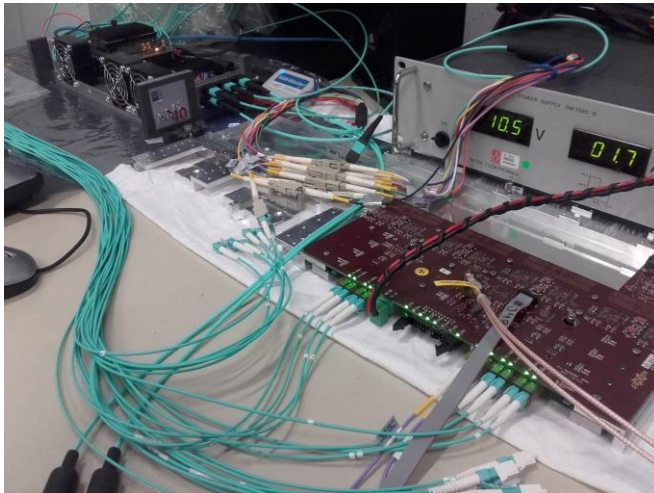
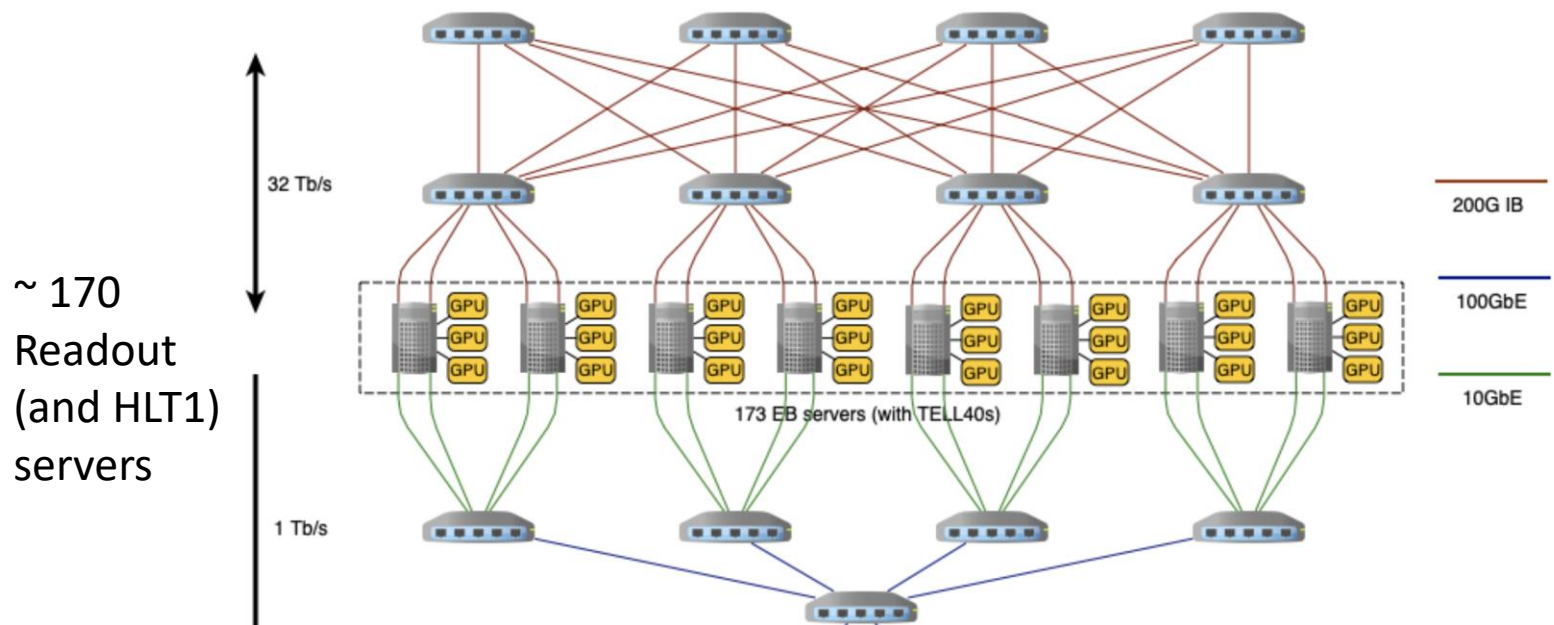
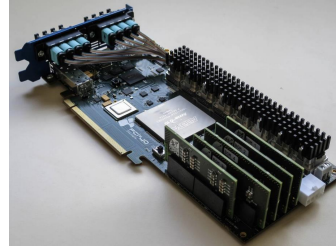


Figure 1: Dataflow in the upgraded LHCb detector.

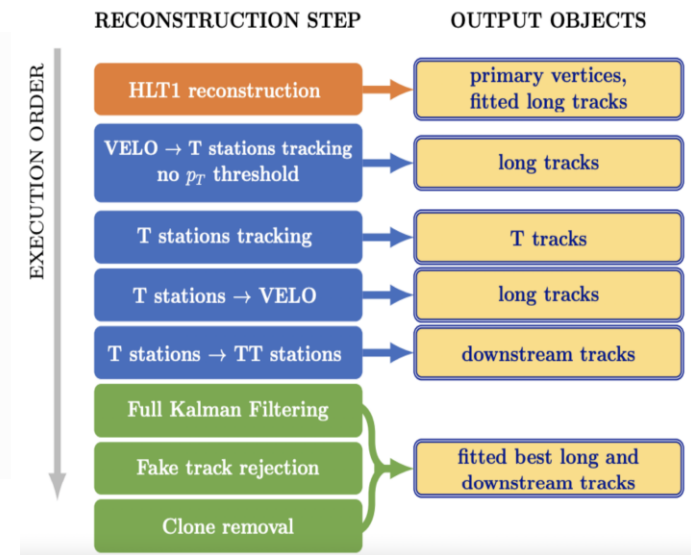
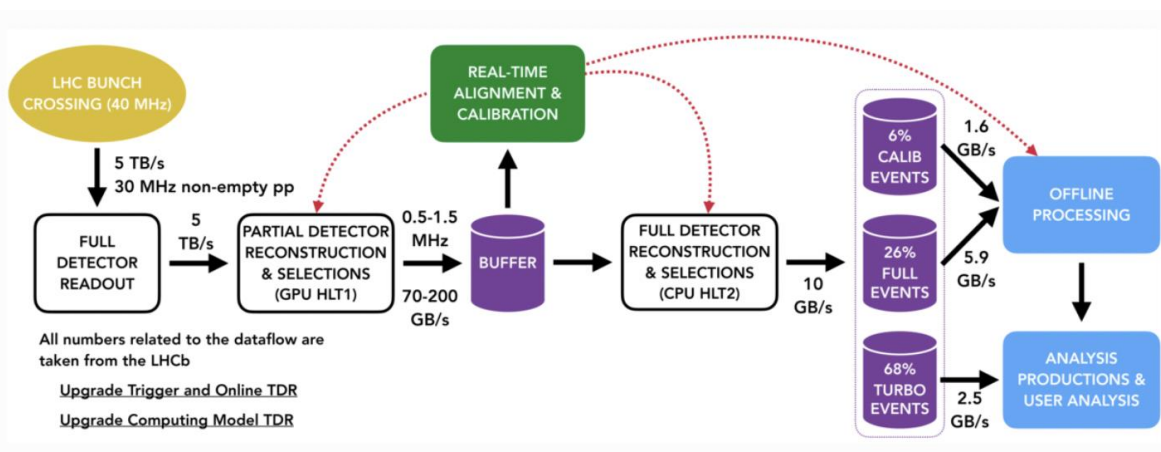
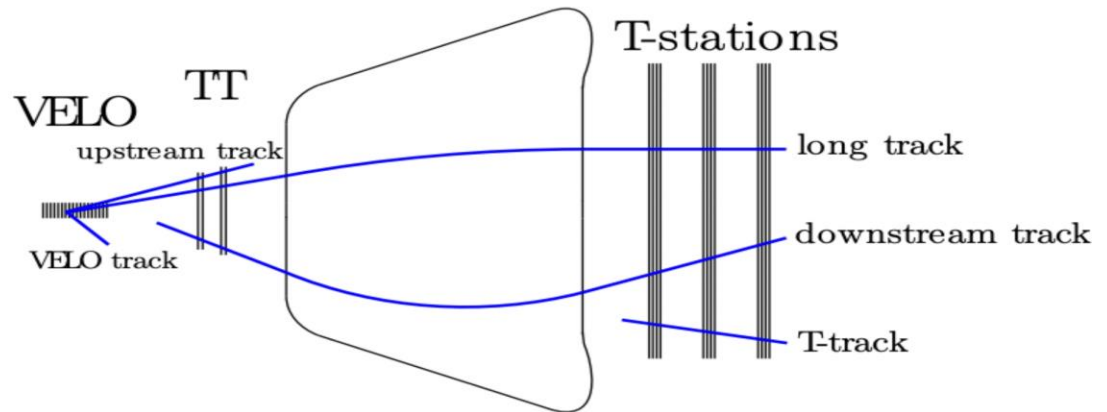
# LHCb DAQ System

- Readout - **PCIe40**
  - common LHCb and **ALICE** board
- Large FPGA (>1m cells)
- 48 x 10 Gbit/s bidirectional links
- Sustained 112 Gbits/s interface with CPU through PCIe
- Installed in PCs → network interfaces
  - Infiniband / Ethernet

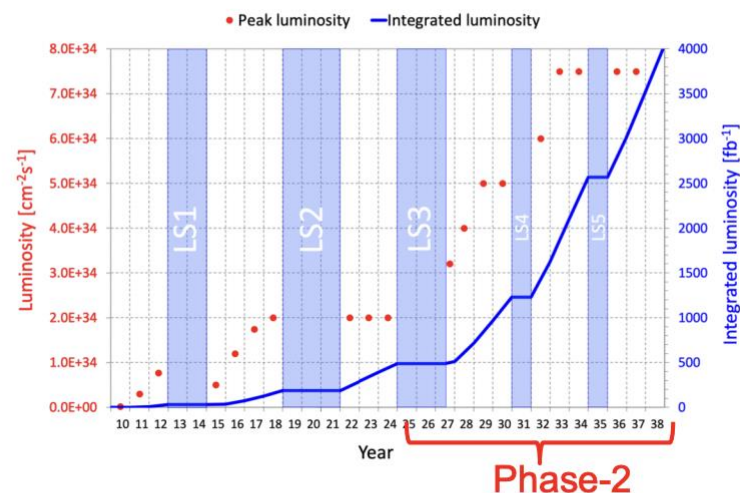
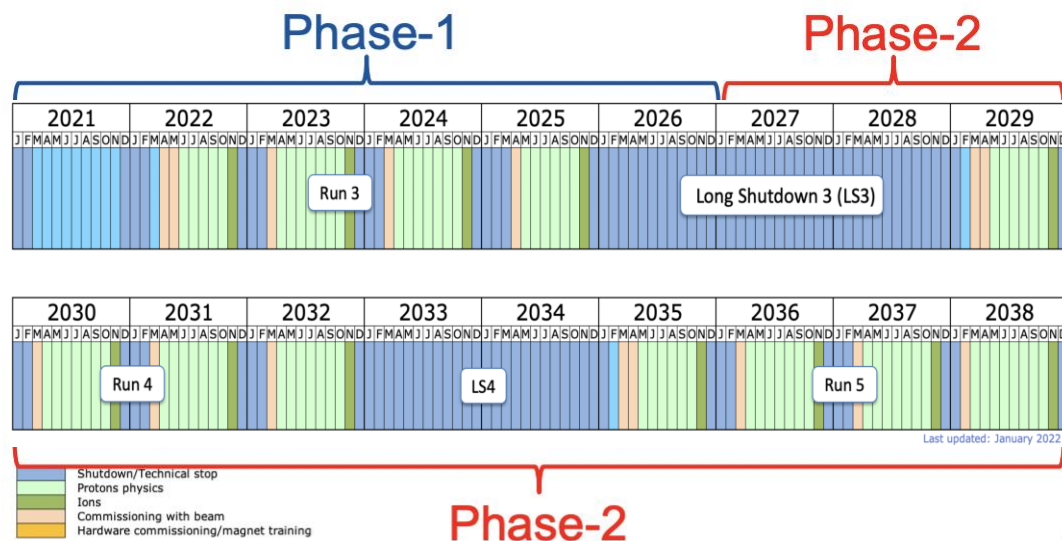


# LHCb HLT

- 2 HLT levels
- Performing offline quality reconstruction
  - Including alignment and calibrations



# High-Luminosity LHC (Phase 2)



	$\mathcal{L}$	$\langle \text{PU} \rangle$	Vertex Density	$\int \mathcal{L} / \text{year}$
Baseline	$5 \cdot 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$	140	0.8 / mm	$250 \text{ fb}^{-1}$
Ultimate	$7.5 \cdot 10^{34} \text{ cm}^{-2} \text{ s}^{-1}$	200	1.2 / mm	$> 300 \text{ fb}^{-1}$

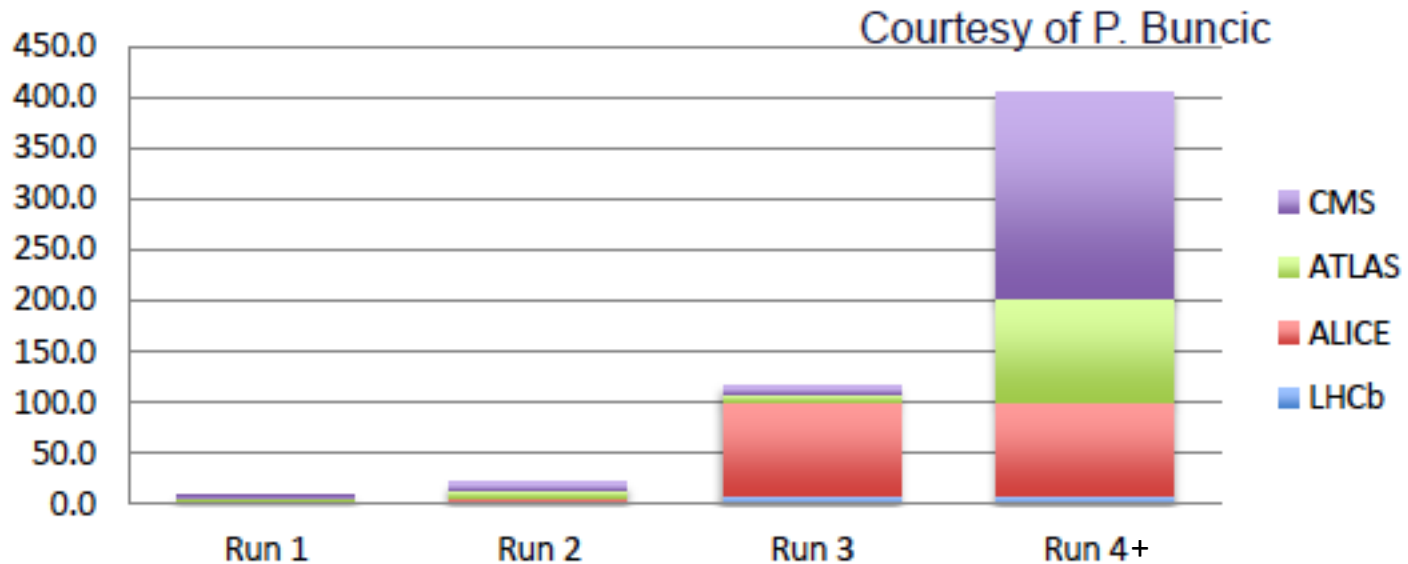
Accelerator upgrade path target:

$\geq 3000 \text{ fb}^{-1}$

Large luminosity and pileup increase wrt.  
Run-3 (~60)

- New paradigms for HEP experiments to fully exploit HL-LHC luminosity

# Data volume in Phase-2



- CMS and ATLAS upgrade for Run 4
  - Event rate x 10 and big increase in data (event) volume
  - Aided by evolution in computer and network hardware
- LHCb and ALICE had big upgrades for Run 3
  - Expected to cover Phase-2

# CMS Phase-2 upgrade

*If there is time...*

~50000  
FE optical links



ATCA modular  
electronics



FPGAs with ~100  
High-speed serial  
transceivers



## Barrel Calorimeters

- ECAL crystal granularity readout at 40 MHz with precise timing for  $e/\gamma$  at 30 GeV
- ECAL and HCAL new Back-End boards

## Muon systems

- DT & CSC new FE/BE readout
- RPC back-end electronics
- New GEM/RPC  $1.6 < \eta < 2.4$
- Extended coverage to  $\eta \approx 3$

## Beam Radiation Instr. and Luminosity, and Common Systems and Infrastructure

## MIP Timing Detector

Precision timing with:

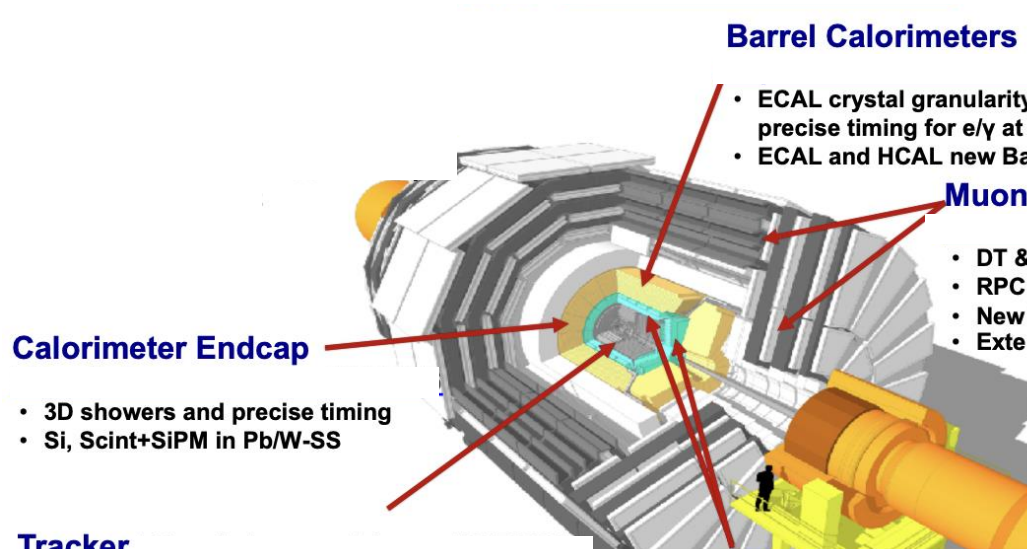
- Barrel layer: Crystals + SiPMs
- Endcap layer: Low Gain Avalanche Diodes

## Calorimeter Endcap

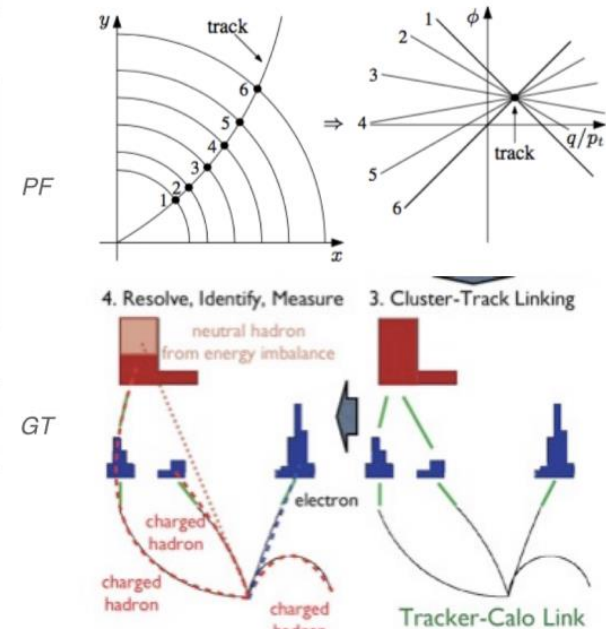
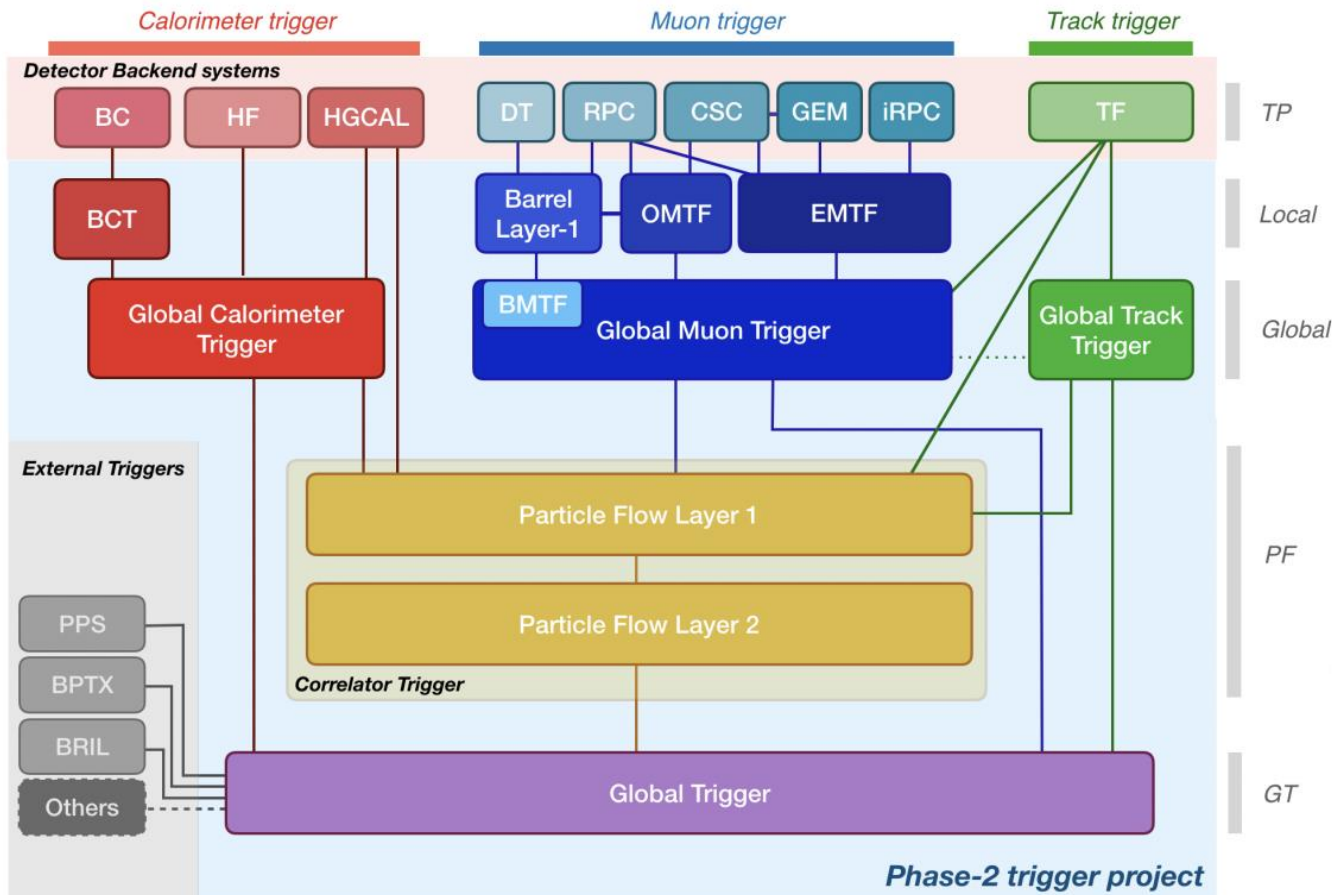
- 3D showers and precise timing
- Si, Scint+SiPM in Pb/W-SS

## Tracker

- Si-Strip and Pixels increased granularity
- Design for tracking in L1-Trigger
- Extended coverage to  $\eta \approx 3.8$

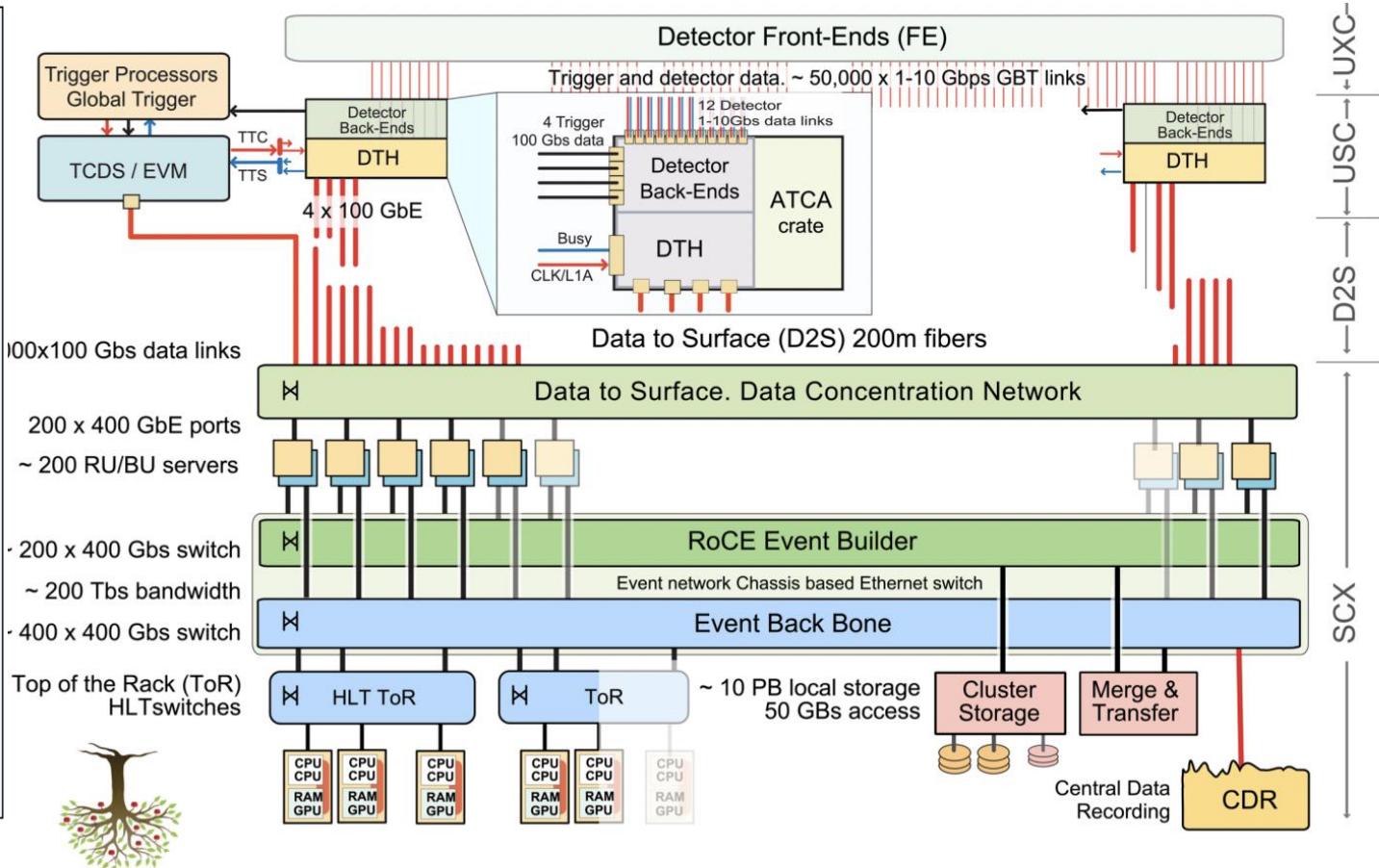
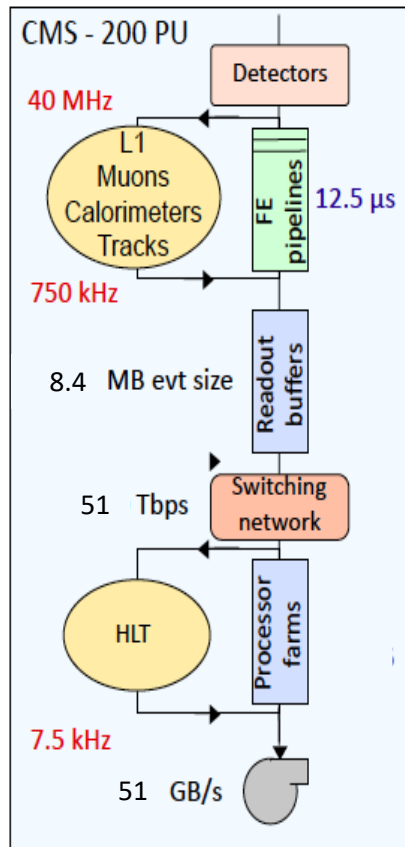


# Run 4 CMS Trigger



- Up to **750 kHz** ( $\leftarrow$  100 kHz Run-3) –  $\sim 12 \mu\text{s}$  latency
- New:
  - Track trigger
  - Particle flow layers
    - application of an algorithm building a global event description from all parts of detector (adopted from CMS analysis and HLT)

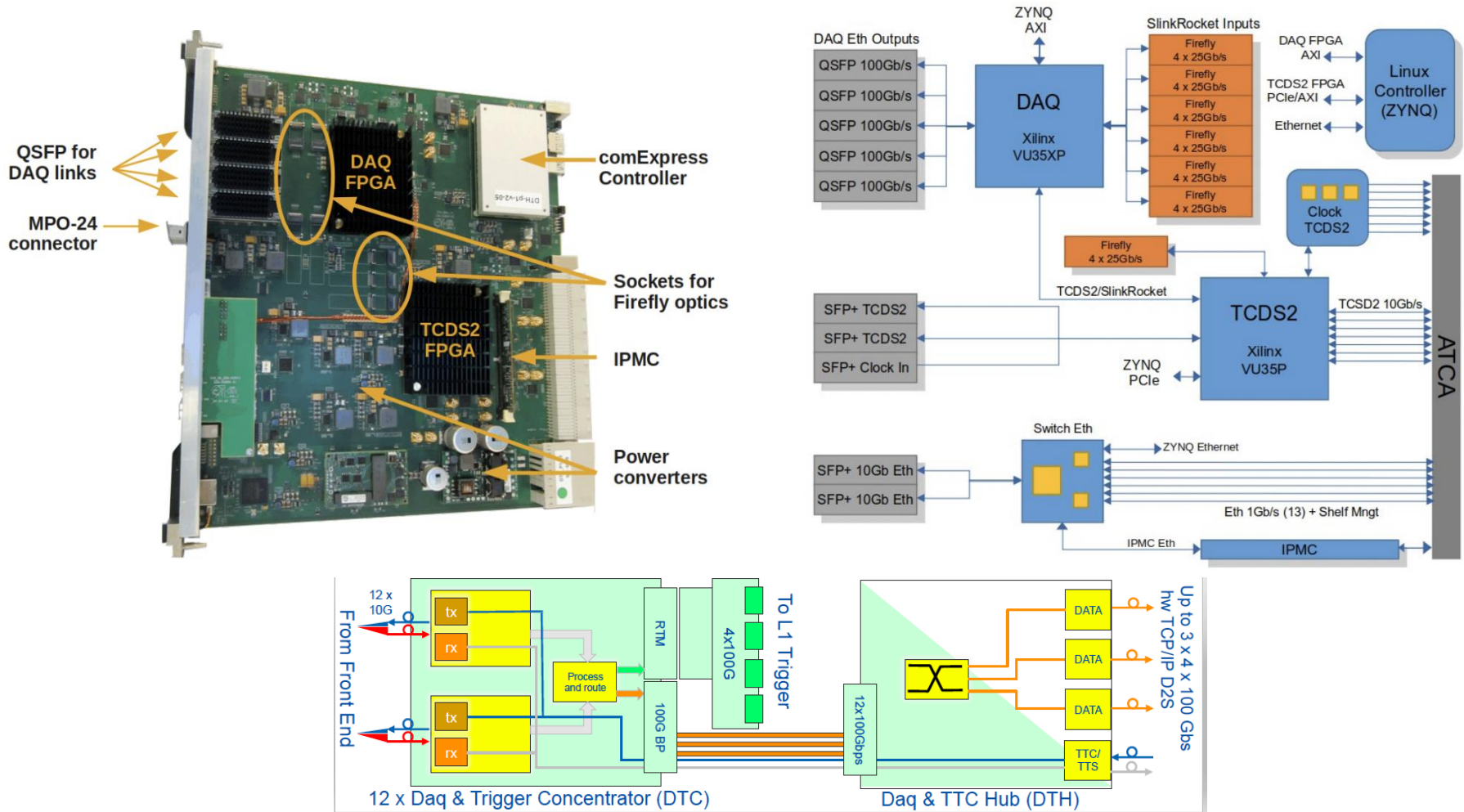
# Run4 CMS DAQ: scale view



- Run-3  $\rightarrow$  Run 4/5 requirements
  - 5 – 7.5 times L1 rate
  - > 4 times event size
  - > 30 times readout bandwidth
  - 50 times HLT computing power  $\rightarrow$  PU x trigger rate & new detectors!
  - 15 times storage (3 times bandwidth)

# HL-LHC – CMS readout

- DTH-400 (800) – high-bandwidth readout to 100 Gb/s Ethernet links
- Clock distribution and trigger-throttling states (BUSY logic etc.)
- installed in ATCA crate with detector backend electronics

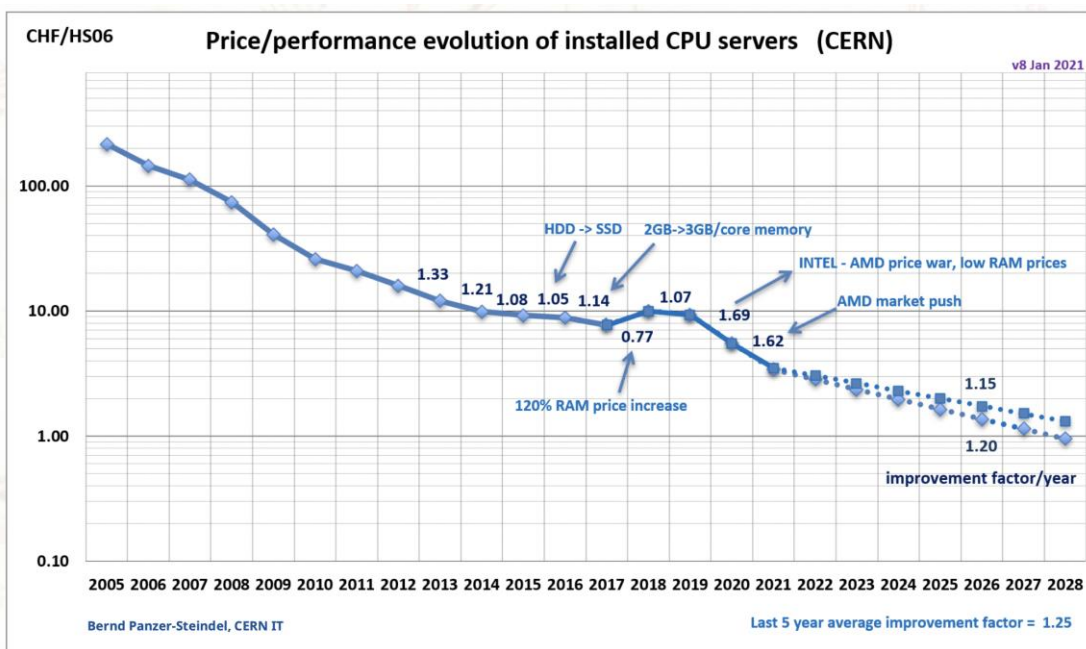


# CMS Run 4/5 HLT

- Large challenge – 50x computing power increase needed (run-5)
  - Very large cost (100 million CHF now!)
  - Also very large electric power load (MWs)
- Due to:
  - High pileup and high L1 rate
  - Upgraded and new detectors (e.g. HGICAL endcaps)

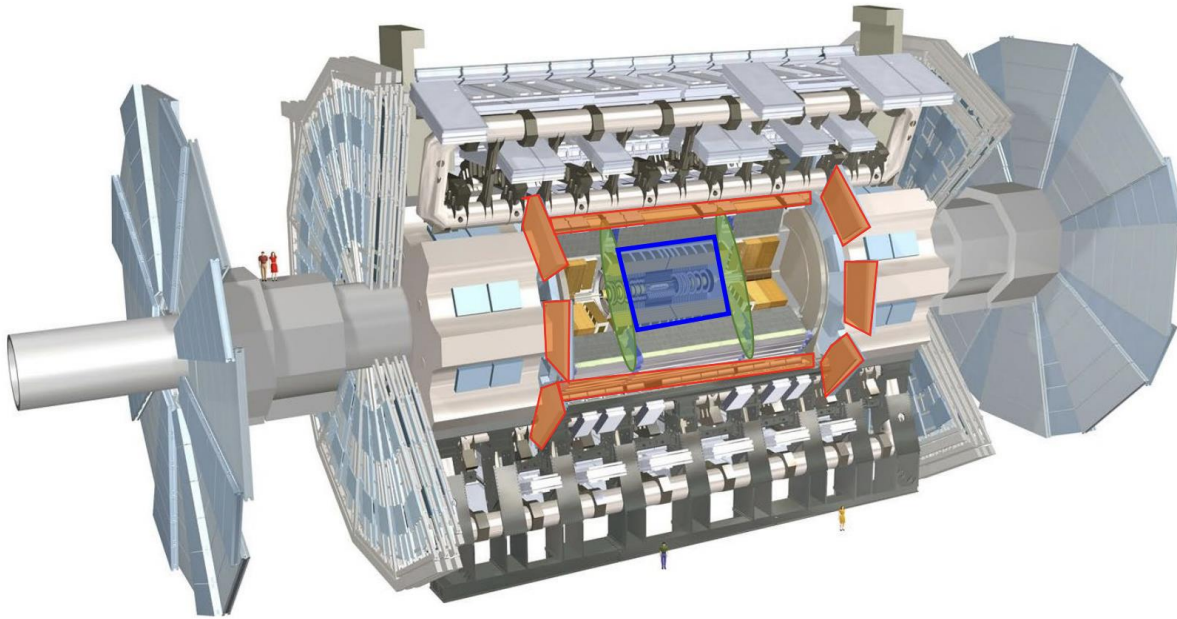
## Strategy:

- Technology evolution of CPUs (cost reduction)
- Coprocessor (GPU etc.) offloading
  - If more cost effective!
  - Increase offloading fraction:
    - 25% (now)
    - → 50% (run-4)
    - → 80% (run-5)
- Algorithmic improvements in HLT
- Many uncertainties !



# ATLAS upgrades HL-LHC *If there is time...*

- Significant detector upgrades
- Electronics + Trigger & DAQ upgrade



## New Muon Chambers

- Inner barrel region with new RPCs, sMDTs, and TGCs
- Improved trigger efficiency/momentum resolution, reduced fake rate

## New Inner Tracking Detector (ITk)

- All silicon with at least 9 layers up to  $|\eta| = 4$
- Less material, finer segmentation

## Electronics Upgrades

- On-detector/off-detector electronics upgrades of LAr Calorimeter, Tile Calorimeter & Muon Detectors
- 40 MHz continuous readout with finer segmentation to trigger

## High Granularity Timing Detector (HGTD)

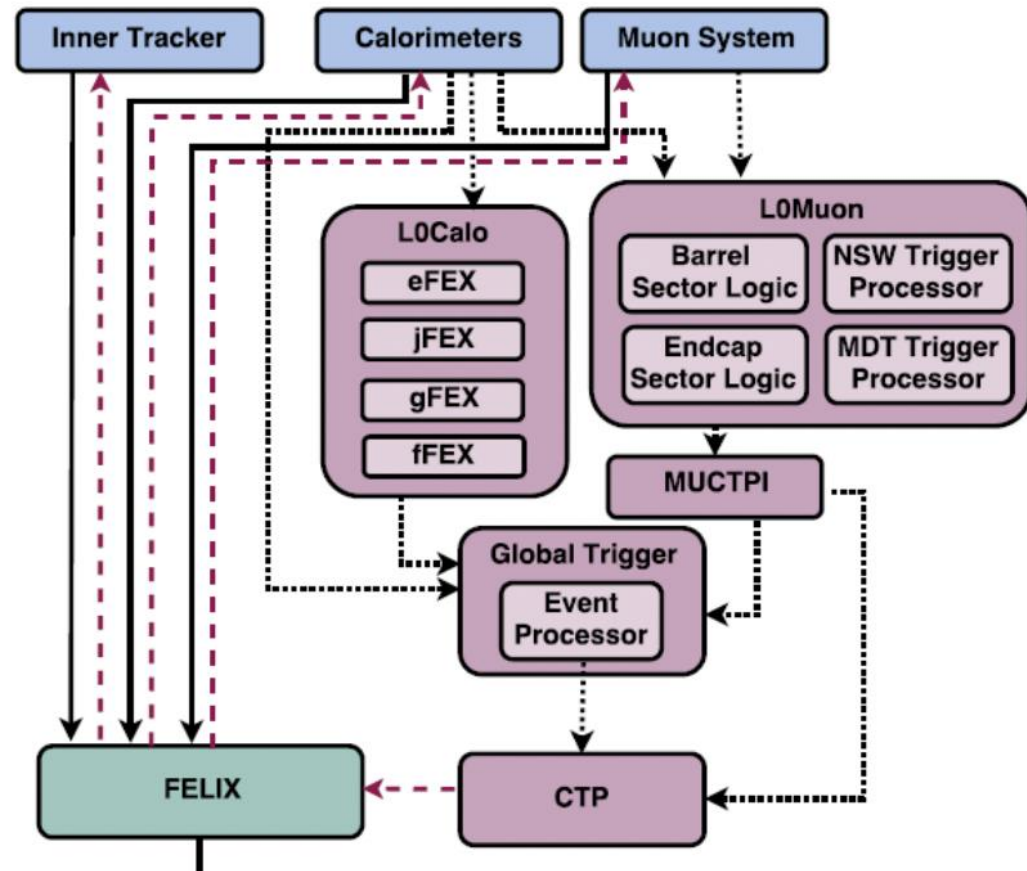
- Precision time reconstruction (30 ps) with Low-Gain Avalanche Detectors (LGAD)
- Improved pile-up separation and bunch-by-bunch luminosity

## Additional small upgrades

- Luminosity detectors (1% precision)
- HL-ZDC (Heavy Ion physics)

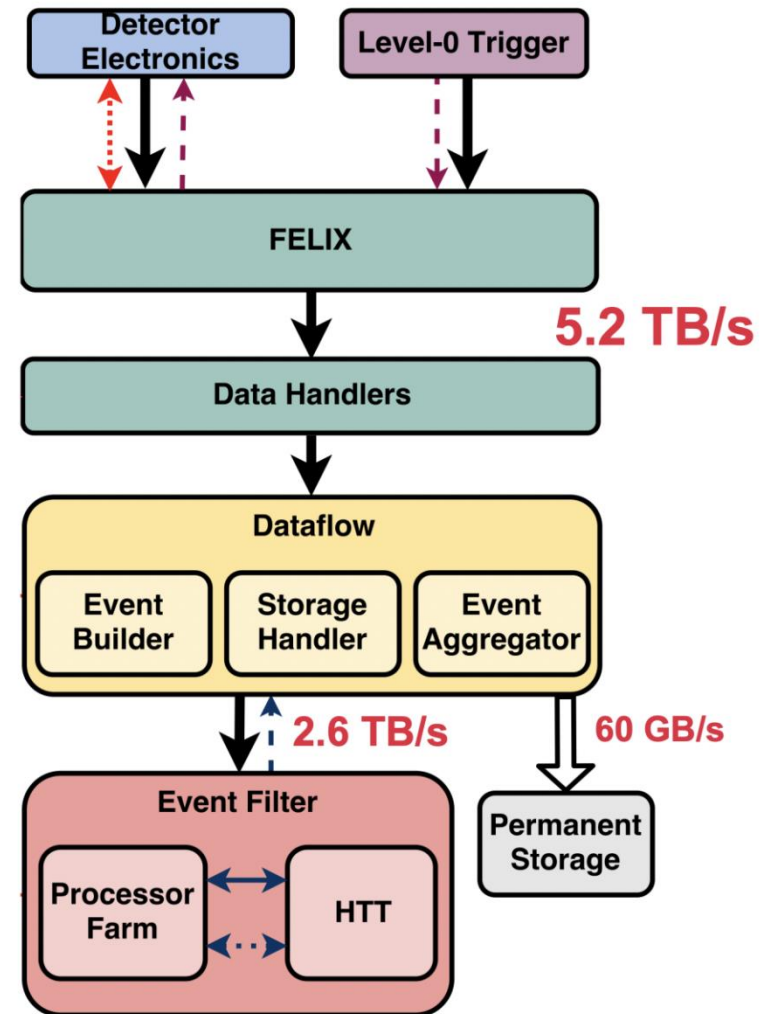
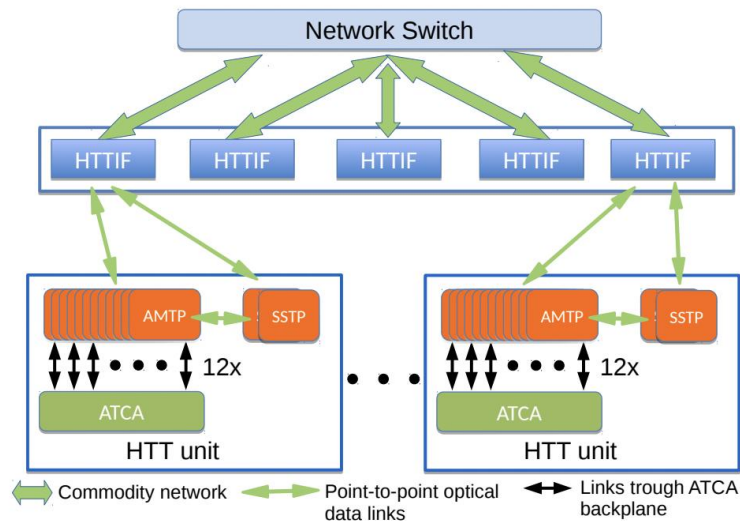
# ATLAS L1 Trigger in Run 4+

- Two level trigger architecture (L0 and HLT)
- L0 Trigger
  - 1 MHz accept-rate
  - 10  $\mu$ s latency
- Calorimeter and Muon trigger
- Adding/extending triggering for new detectors



# HL-LHC ATLAS DAQ and HLT overview

- Comparison with Run 3
  - 10 x trigger rate (1 MHz)
  - 20 x readout rate (5.2 TB/s)
  - FELIX deployed for all detector
  - HLT – ROI approach kept
    - Upgraded CPU farm
    - Dedicated hardware for track triggering (HTT) – ASICs, associative memory



# Summary

- Trigger and DAQ systems of LHC experiments
  - Perform readout at up to 100 kHz
  - or full 40 MHz in triggerless mode (LHCb)
  - Large scale network systems used to retrieve, transfer data and build events (order of 200 GB/s or more)
- Coupled to several levels of triggering used for data reduction:
  - hardware implementations (initial levels)
  - Software HLT farms (higher levels)
- Combination of custom electronics and “COTS” components
- Large storage requirements
- HL-LHC (Phase 2):
  - More information included in triggering (tracking)
  - Much higher rate and event size for readout
  - Demanding on HLT performance
  - Will take advantage of technology developments in networking, CPUs and coprocessors

# BACKUP

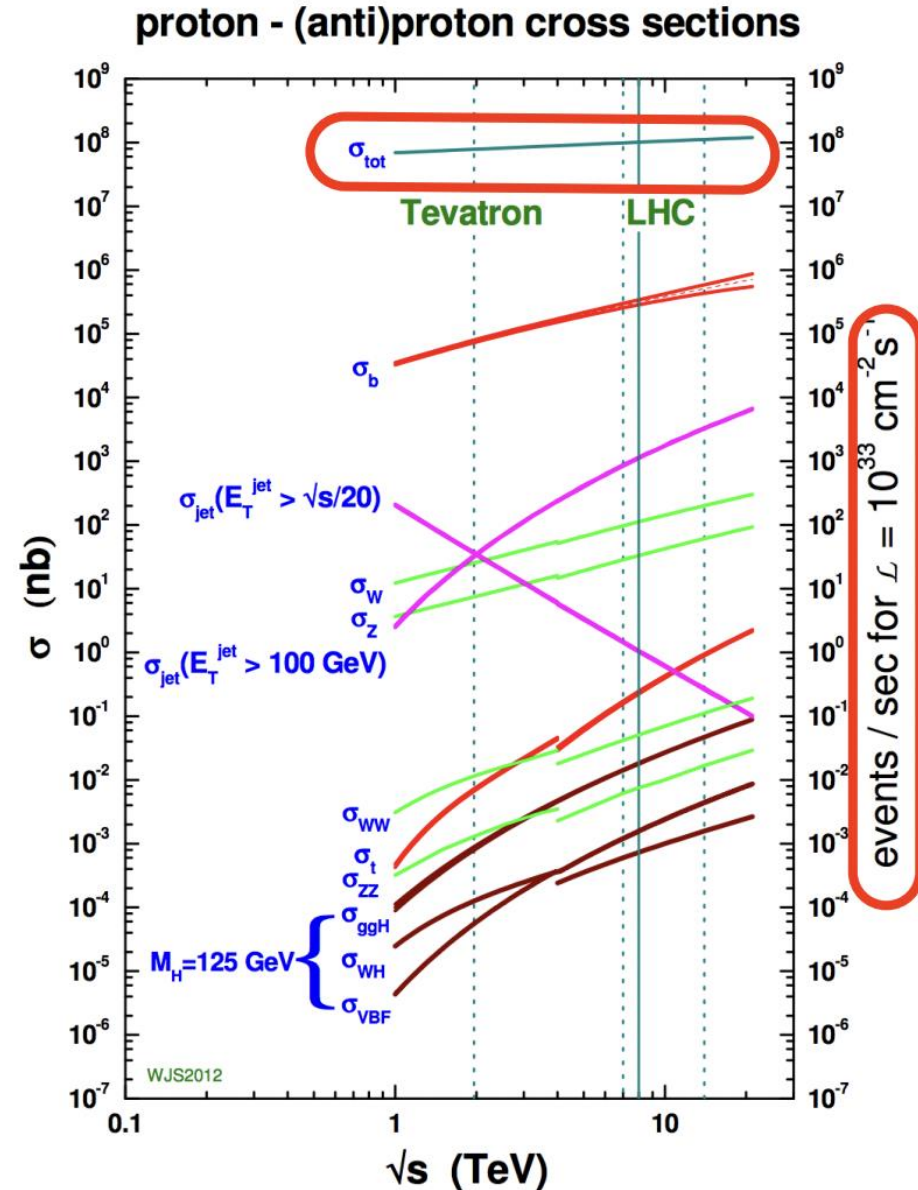
# Challenges for DAQ at large experiments

- built for even more rare physics  
Higgs production: 1 in a billion pp collisions ( 13 - 14 TeV)
  - But also, a lot of background that is hard to reject (especially by a live system)
  - → inevitably, more common physics will be selected in the mix (and later either filtered out, or removed by detailed analysis)

Saving all data is often not useful

→ DAQ systems, Storage systems are scaled to be smaller (and cheaper!) with the assumption of saving a fraction of more useful detector data

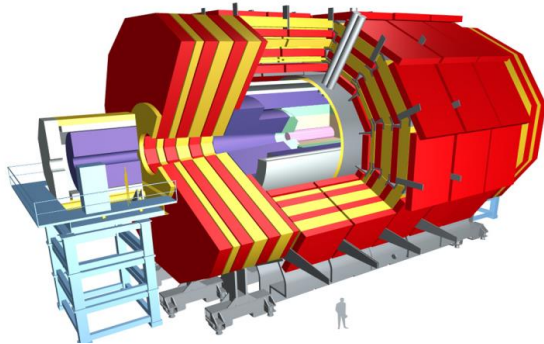
- More specialized experiments (LHCb, ALICE) differ from more general purpose (ATLAS, CMS) in respect to selection (rates) of useful physics data



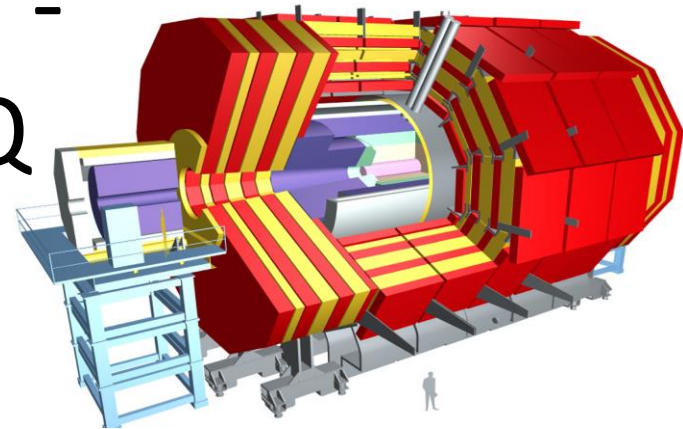
RUN 1

# Evolution - CMS DAQ

RUN 2/3

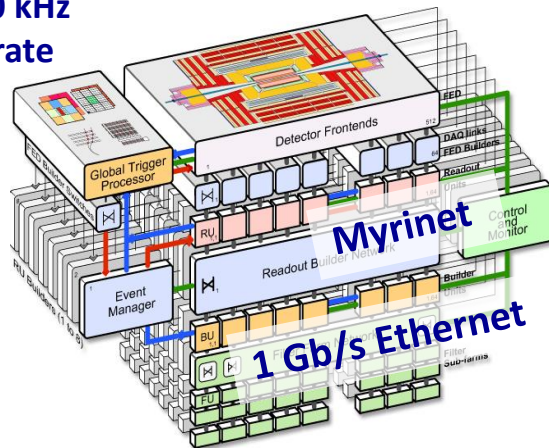


Event size up to 1MB



Event size up to 2MB

100 kHz  
L1 rate



100  
GB/s

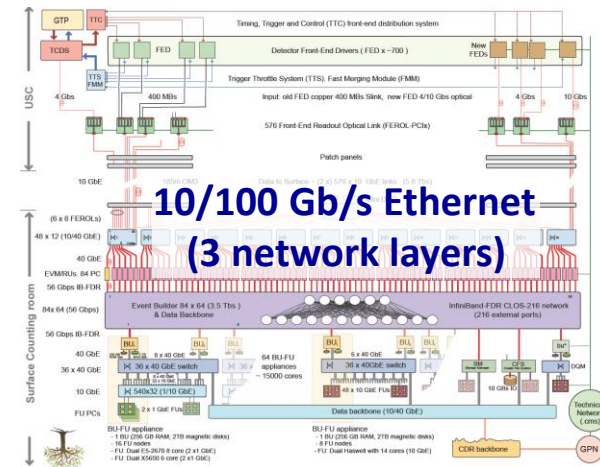
8 slices

**CMS DAQ 1**

13000 core,  
1260 host  
filter farm

max. 1.2 GB/s to storage

100 kHz  
L1 rate



10/100 Gb/s Ethernet  
(3 network layers)

~200  
GB/s

1 slice

**Latest iteration:  
CMS DAQ 3**

Filter Farm  
25000 cores,  
200 PCs

~ 15 GB/s to storage

# CMS DETECTOR

Total weight : 14,000 tonnes  
Overall diameter : 15.0 m  
Overall length : 28.7 m  
Magnetic field : 3.8 T

STEEL RETURN YOKE  
12,500 tonnes

SILICON TRACKERS  
Pixel ( $100 \times 150 \mu\text{m}^2$ )  $\sim 1 \text{ m}^2 \sim 66\text{M}$  channels  
Microstrips ( $80\text{--}180 \mu\text{m}$ )  $\sim 200 \text{ m}^2 \sim 9.6\text{M}$  channels

SUPERCONDUCTING SOLENOID  
Niobium titanium coil carrying  $\sim 18,000 \text{ A}$

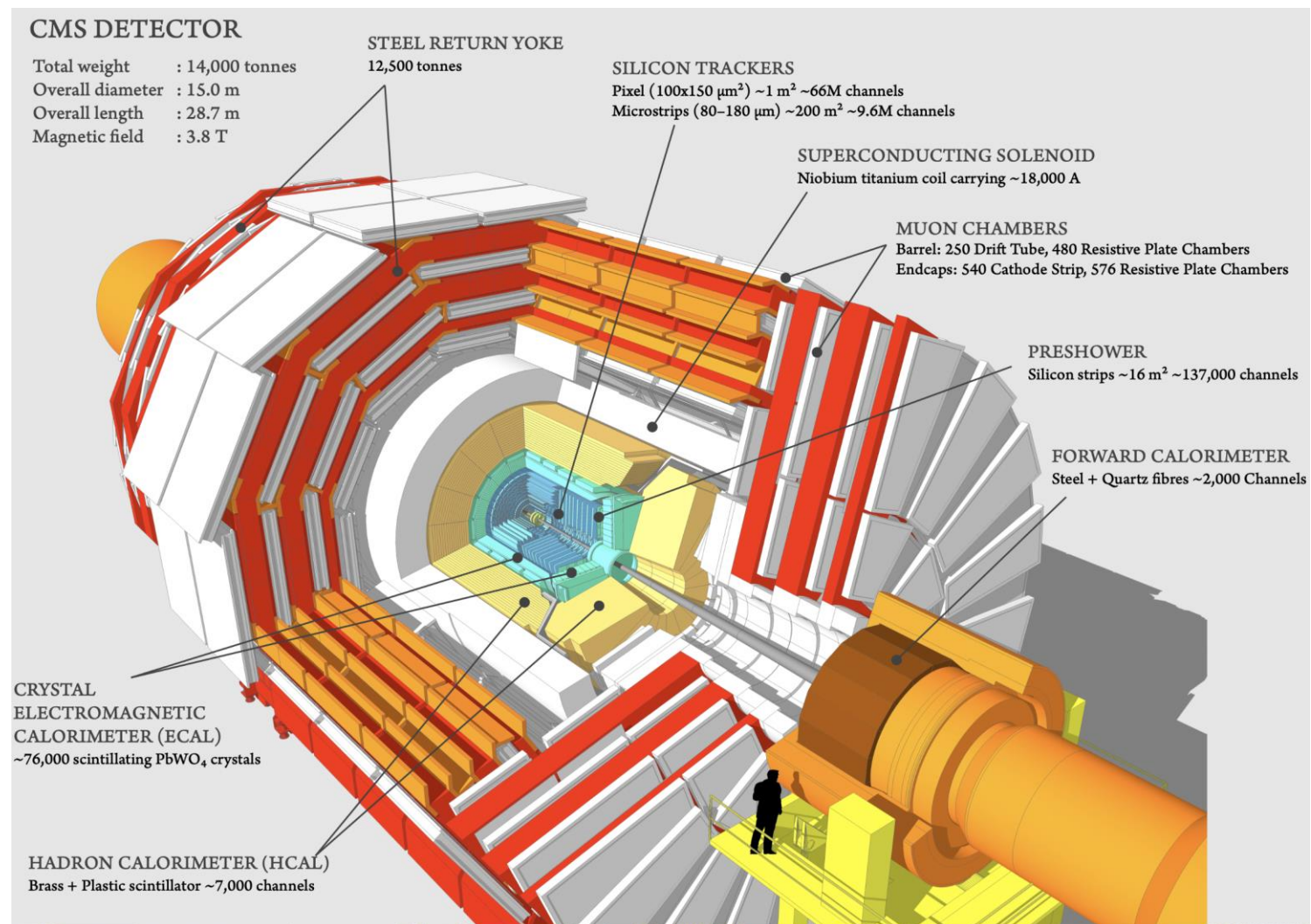
MUON CHAMBERS  
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers  
Endcaps: 540 Cathode Strip, 576 Resistive Plate Chambers

PRESHOWER  
Silicon strips  $\sim 16 \text{ m}^2 \sim 137,000$  channels

FORWARD CALORIMETER  
Steel + Quartz fibres  $\sim 2,000$  Channels

CRYSTAL  
ELECTROMAGNETIC  
CALORIMETER (ECAL)  
 $\sim 76,000$  scintillating  $\text{PbWO}_4$  crystals

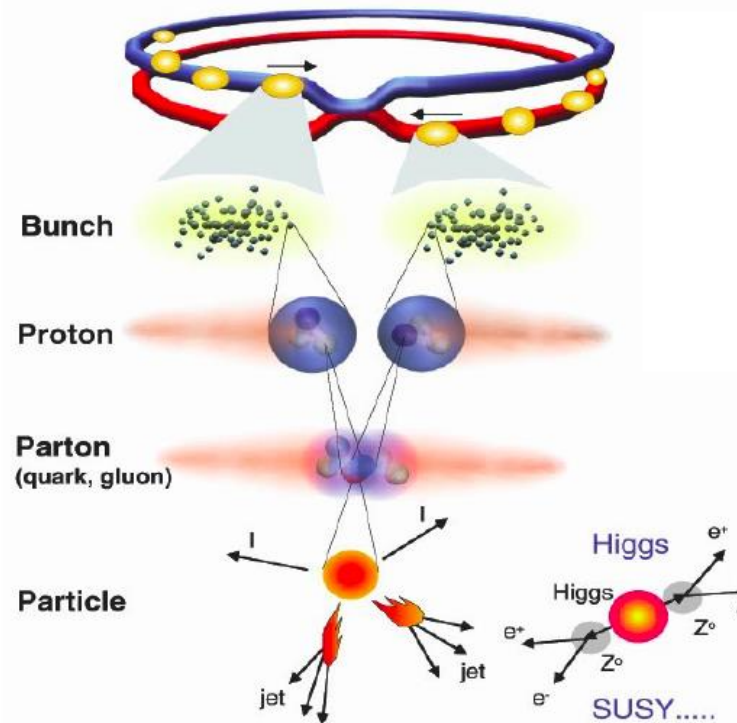
HADRON CALORIMETER (HCAL)  
Brass + Plastic scintillator  $\sim 7,000$  channels



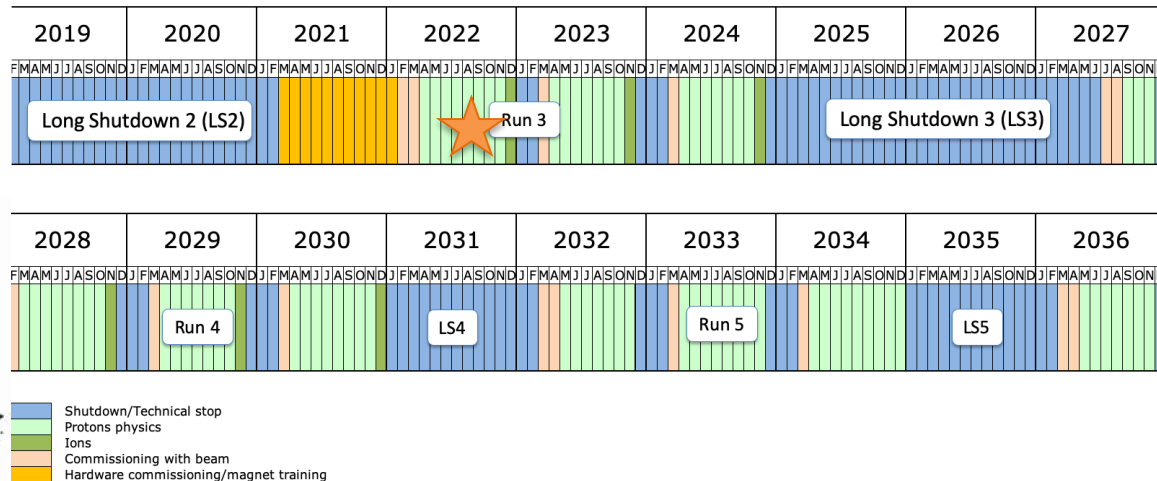
# BACKUP Introduction

- LHC in Run 3: proton-proton collisions @ 13.6 TeV

LHC run3	Beam Energy [TeV]	Protons/bunch	Colliding Proton bunches/beam	Luminosity [ $\text{cm}^{-2}\text{s}^{-2}$ ]	Bunch spacing [ns]
2022	7.8	$1.2 \times 10^{11}$	2400	$2 \times 10^{34}$	25

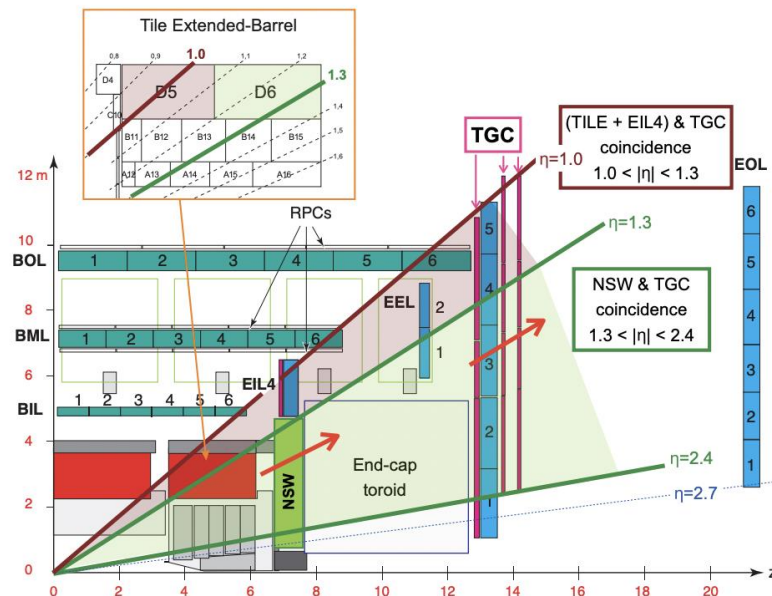


2022:  $\int L dt 10 \text{ fb}^{-1}$  delivered (ATLAS / CMS) so far



# ATLAS TDAQ Level-1 Muon Trigger

- Phase-1 NSL board
  - Includes New Small Wheel triggering input



6 optical inputs (6.4 Gbps) from NSW  
SFP+ with GTX RX in FPGA

CPLD (XC2C256-7PQ208C  
for VME control

6 optical inputs  
for other detectors

2 optical outputs  
(6.4 Gbps) from MUCTPI  
SFP+ with GTX TX  
10 optical outputs  
for spares

BPI (PC28F256P30TF)  
for FPGA configuration

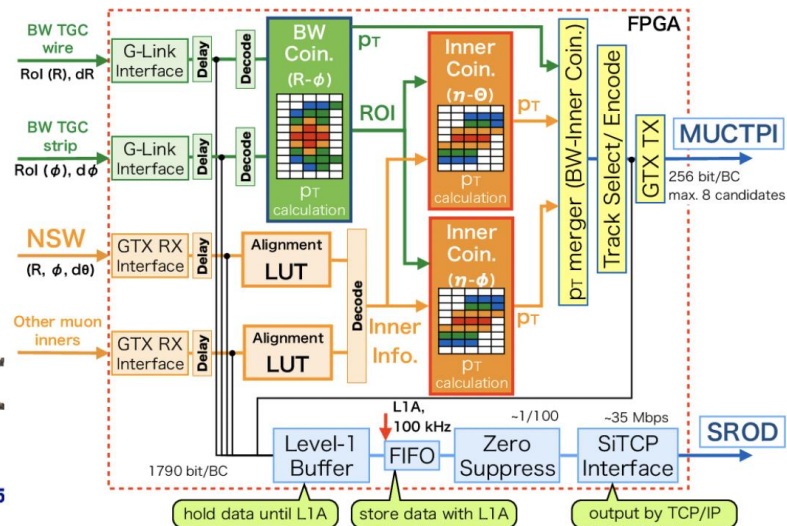
RJ45 connector  
for readout

16-pin connector  
for TTC receive

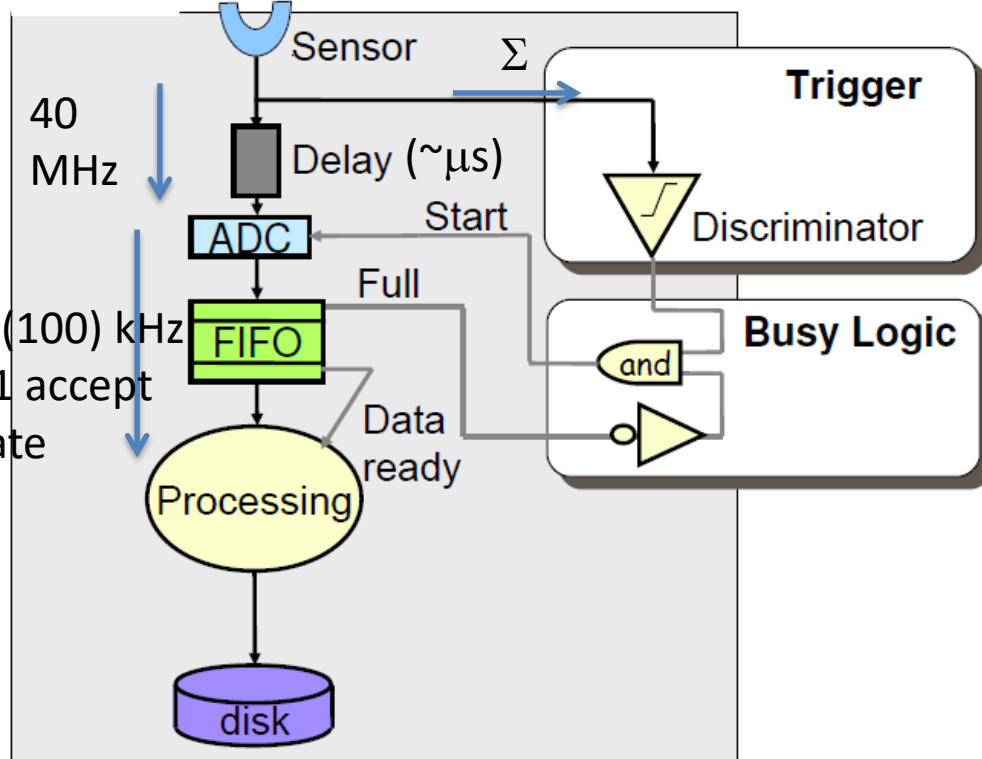
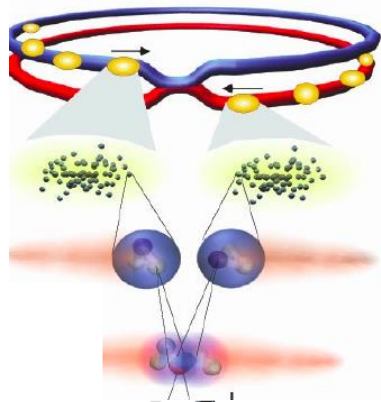
LEMO IN/OUT

14 optical inputs (800 Mbps) from BW-TGC  
SFP RX + G-Link RX chip

FPGA (Xilinx Kintex-7 XCK325  
G-Link receiver chips (Agilent HDMF-1034A)



# BACKUP - 1<sup>st</sup> level Trigger and DAQ - LHC

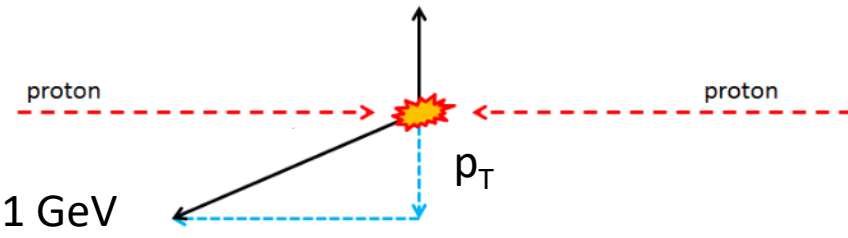


- L1 trigger
  - Synchronized with 40 MHz LHC collision clock (25 ns intervals)
  - Decision based on “trigger primitive” data (small-size physical quantities calculated in real-time by detectors)
- Delay/buffer
  - analog pipeline or a digital buffer
  - accommodates for trigger latency
  - expensive, radiation-hard on-detector electronics
- FIFO (derandomizer)
  - Accommodates fluctuations of accepted trigger rate (while data is processed)
  - Reduces likelihood of readout BUSY on next accepted event

# Pileup

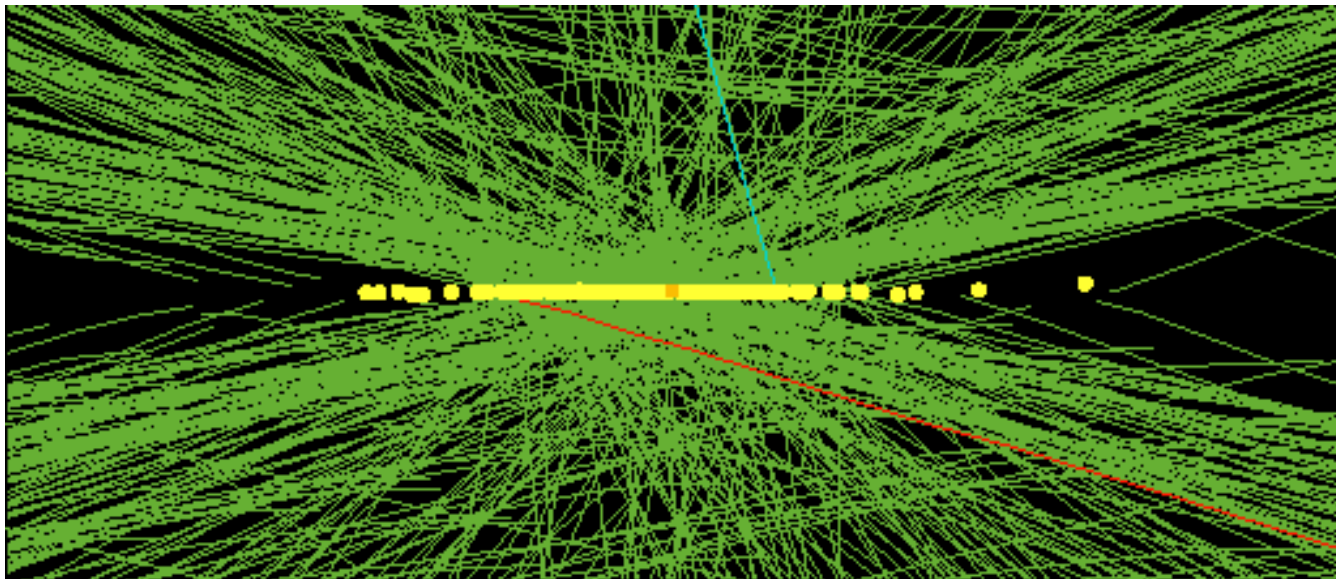
- Presence of multiple interactions in collision of bunches from opposing beams (“bunch crossing”)
- Run 3: over **50** interactions/bx
  - mostly soft scattering (pp inelastic collisions)

→ Very low outgoing transverse momentum,  $\sim p_T < 1 \text{ GeV}$



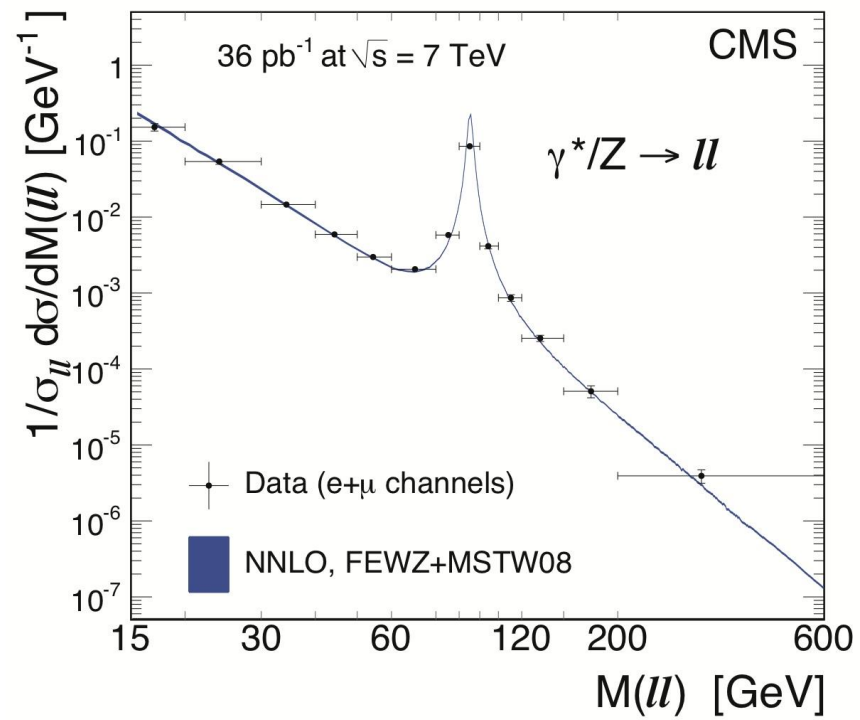
Hard scattering (interesting physics) : can have high  $p_T$  momentum

- Challenge: filtering out interesting data!
- Effect of pileup on resolution, filtering efficiency...

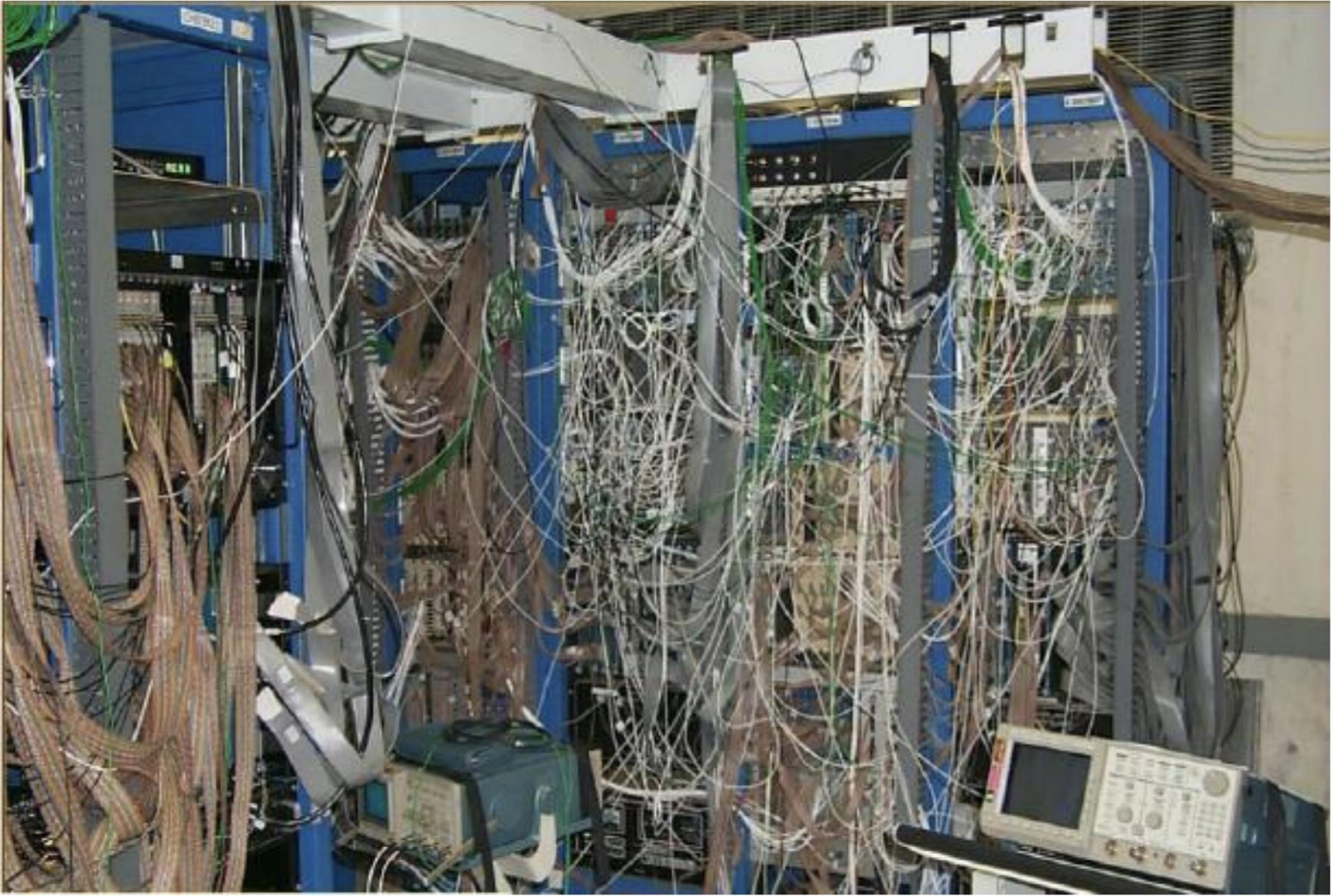


‘Event’ = record  
of a bunch  
crossing

Visualisation of *event* with 78 reconstructed collision vertices in CMS



# Limitations of all-to-all approach



# Example – CMS Readout hardware

- Many **Run 1** detectors remains in use
- ➔ Readout electronics based on VME bus
- Several detectors / online-systems upgraded to cope with higher luminosity
- ➔ New readout electronics based on  $\mu$ TCA bus

Run 2 – Run 3

- 2014: New Trigger Control and Distribution System
- 2014: Stage-1 calorimeter trigger upgrade
- 2014/15: new HCAL readout electronics
- 2016: Full trigger upgrade
- 2017: New pixel detector and readout electronics



Sender card  
plugged onto VME electronics

Fragment size 1..4 kB

SLINK copper  
cable  
400 MB/s



Frontend-  
Readout  
Link

640 Legacy links



$\mu$ TCA electronics

Fragments 2..8 kB

Optical SLINK-  
express  
5 Gb/s - 10 Gb/s



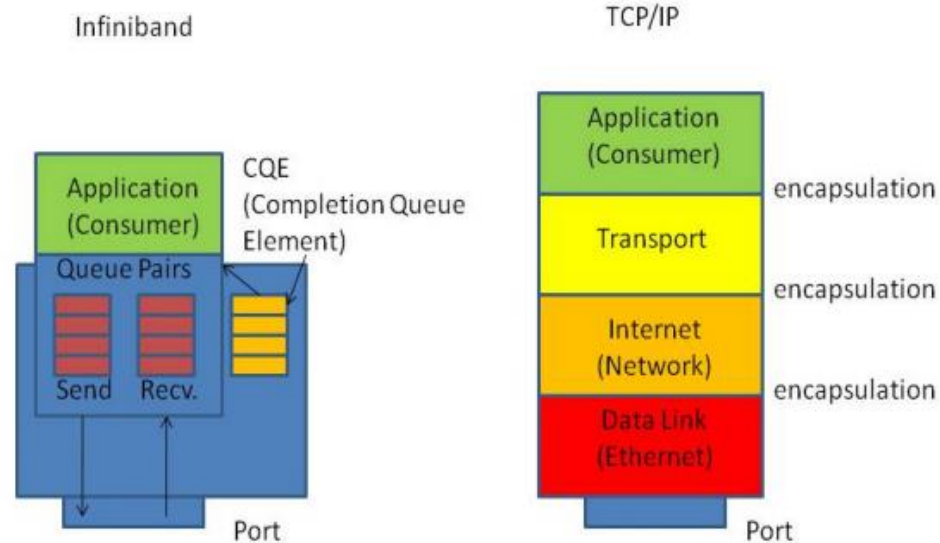
Frontend-  
Readout  
Optical  
Link

50 links (+170 for new pixel ???)

# Comparison of Infiniband and TCP/IP

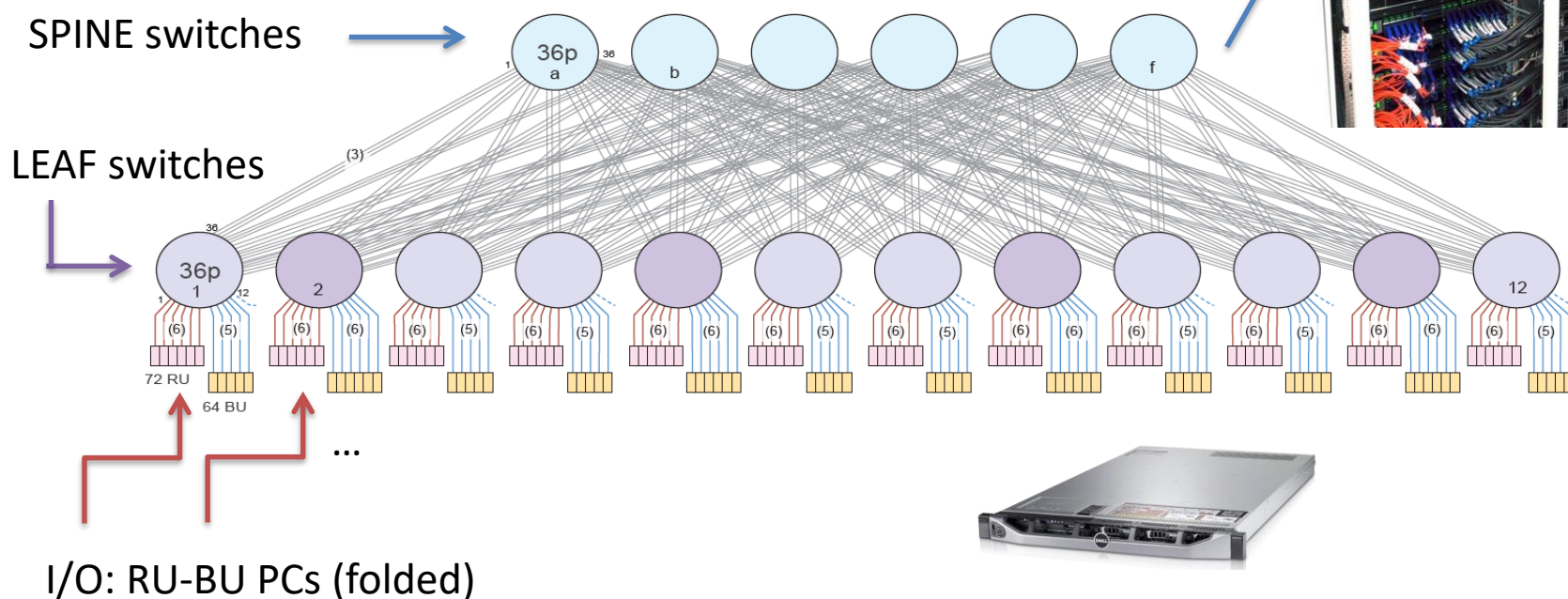
## Infiniband

The protocol is defined as a very thin set of **zero copy** functions when compared to thicker protocol implementations such as TCP/IP

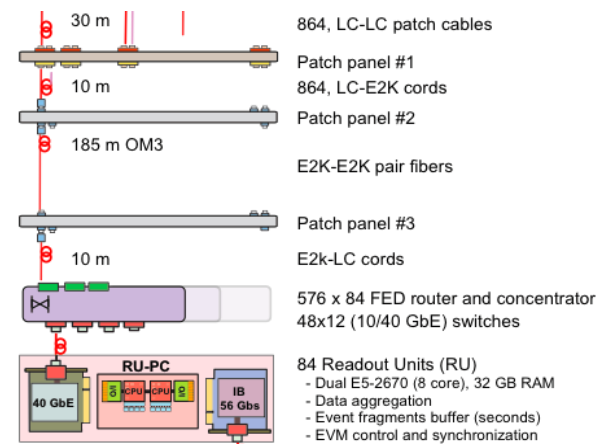


# Event building network in CMS (DAQ2)

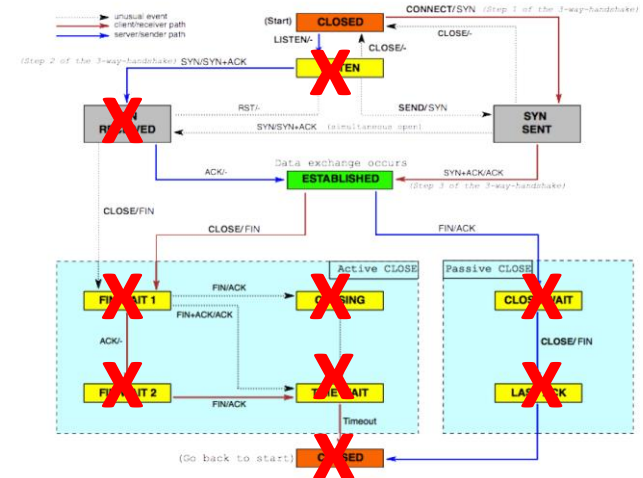
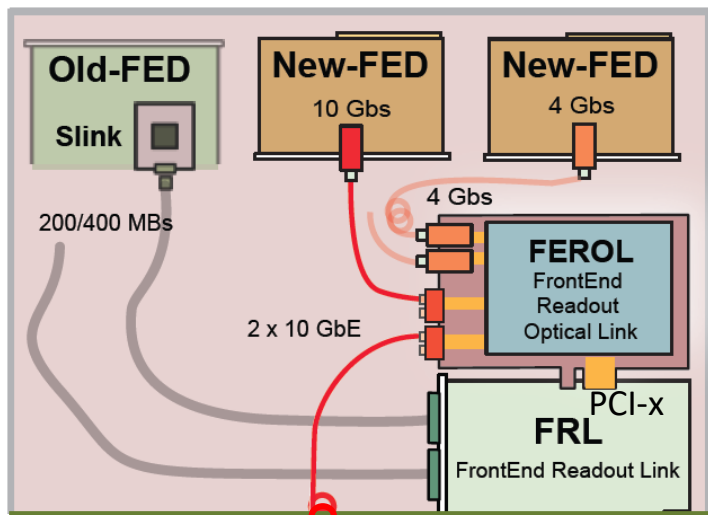
- RU superfragments assembled into complete events
- Builder-unit (BU) machines receive superfragments and combine them into full events
- Based on 56 Gb/s FDR Infiniband CLOS network
  - (12 leaf + 6 spine) switches
- 108 x 72 Event Builder
- 200 GB/s total throughput (RUs→BUs)



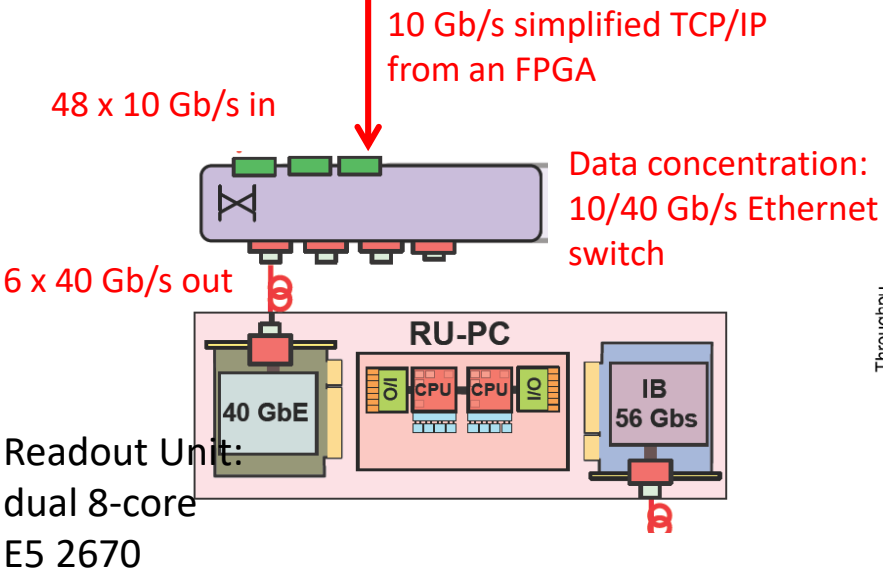
- 10/40 Gb Ethernet
- 576 FEROLs → 108 Readout-Units (max.)
- Fat Tree structure



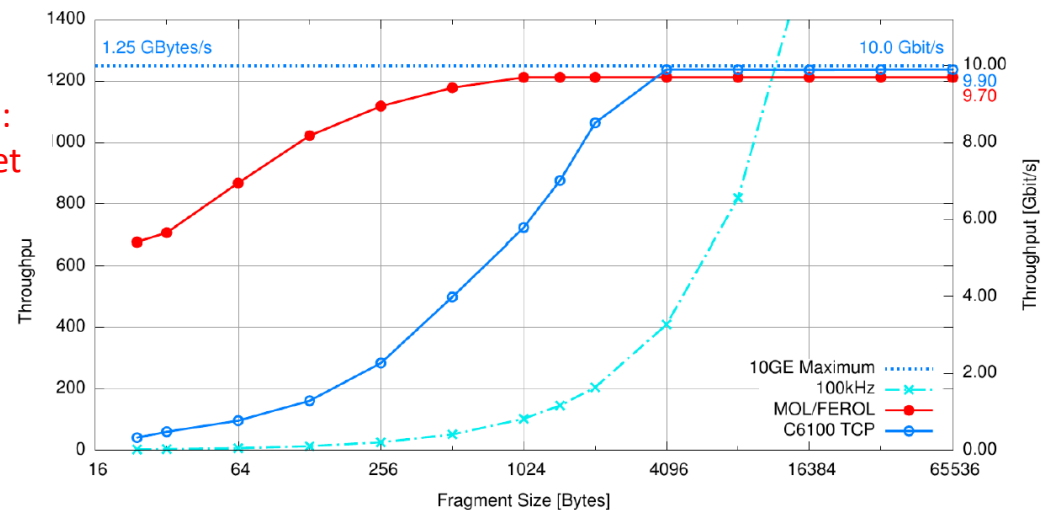
# FEROL TCP/IP



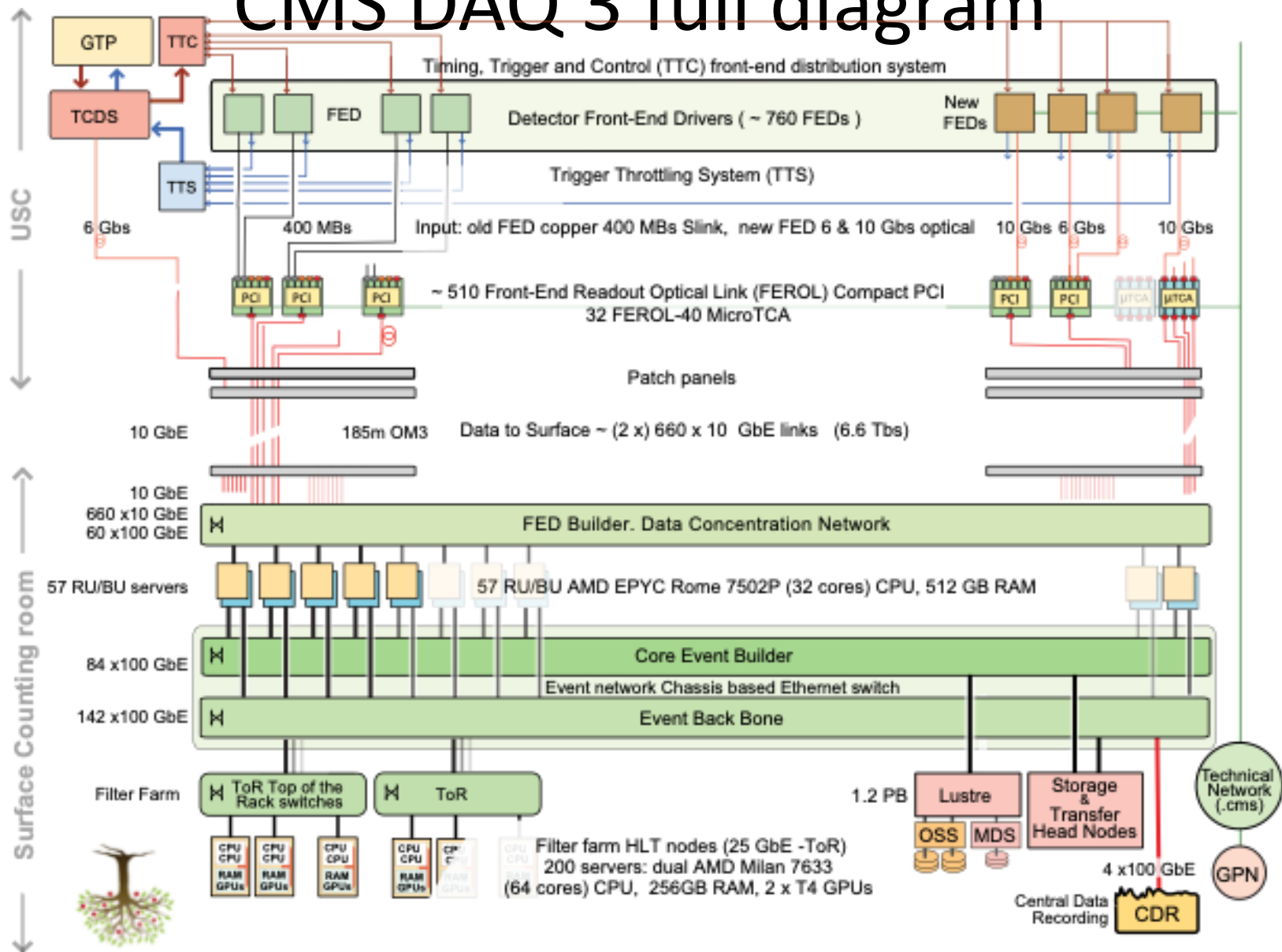
Simplified unidirectional TCP/IP implemented in FPGA



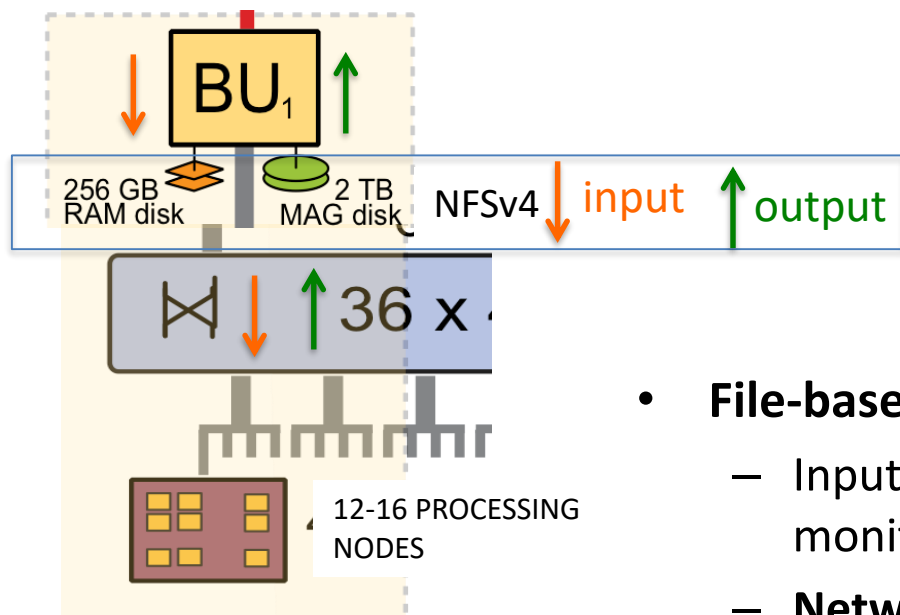
Point to point link performance: 9.7 Gb/s for fragments > 1 kB



# CMS DAQ 3 full diagram



# CMS Filter Farm implementation

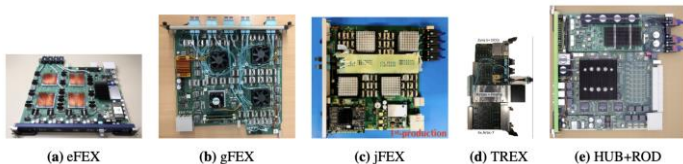


- **File-based Filter Farm**
  - Input, output of event and non-event data, monitoring and logging through **files**
  - **Network filesystem** used as transport (and resource arbitration) protocol
- Reduced coupling between DAQ and HLT software platforms
  - HLT uses standard CMS offline software (DAQ-specific code implemented as modules)
  - DAQ is built on custom Online framework (XDAQ)
    - separate release cycles, simplified development, maintenance and debugging

# BACKUP: ATLAS TDAQ Level-1 Calo Trigger

L1-calo trigger boards (phase I)

System	modules	FPGAs	Function
eFEX	24	4+1	electrons, photons, taus
gFEX	1	3+1 Soc	large-R jets, MET, $\text{Sum}E_T$
jFEX	6	4+1 Soc	large/small-R jets, MET, $\text{Sum}E_T$ , taus
TREX	32		digitizes trigger towers
HUB+ROD	8	1+1	clock source and data buffer for e/jFEX
FOX	6	N/A	routes 7.5k + 1.5k fibers to/from FEXs

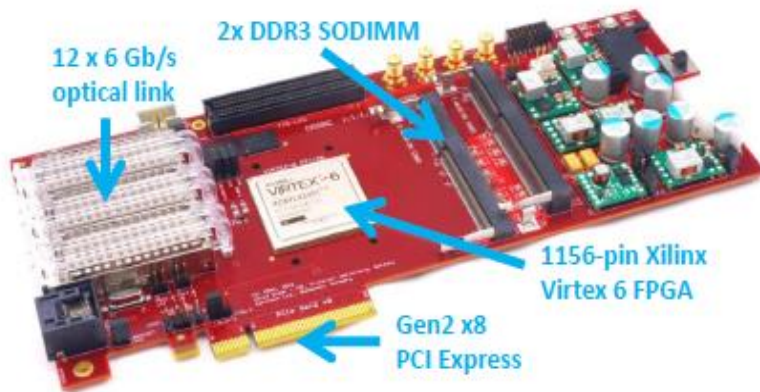


- Analogous structure to CMS trigger
  - FPGAs heavily used

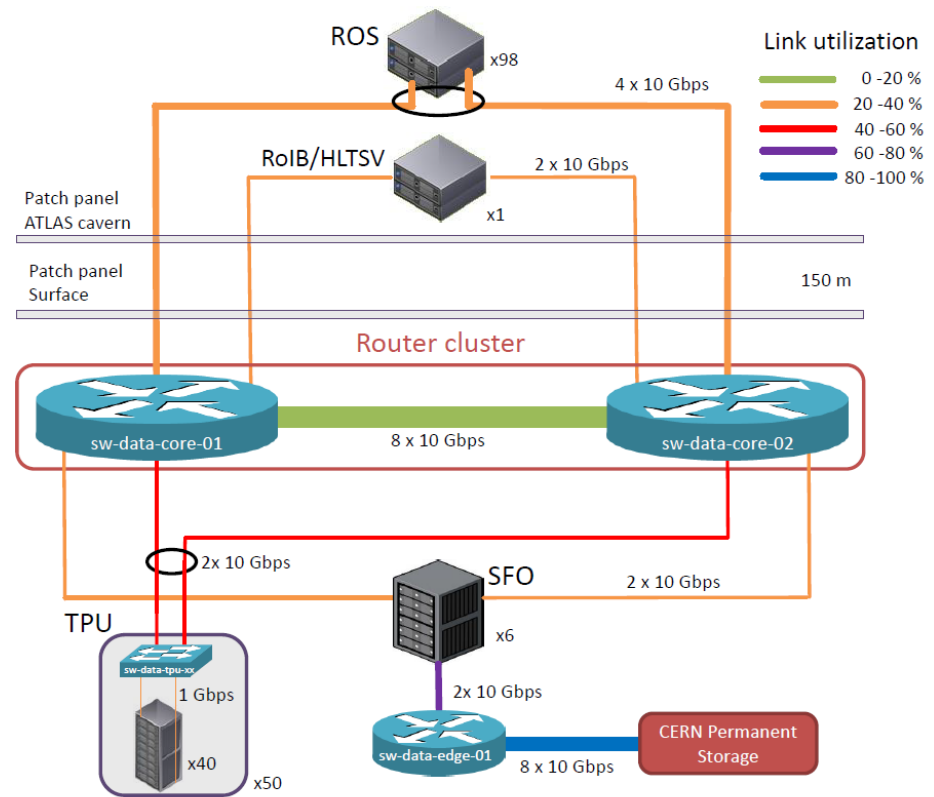
# ATLAS – readout and network Run II

## Readout

- PCI-e cards on host (ROS) PCs
  - Joint R&D with ALICE
- Host PC reads data, outputs over 4 x 10 Gbit Ethernet links

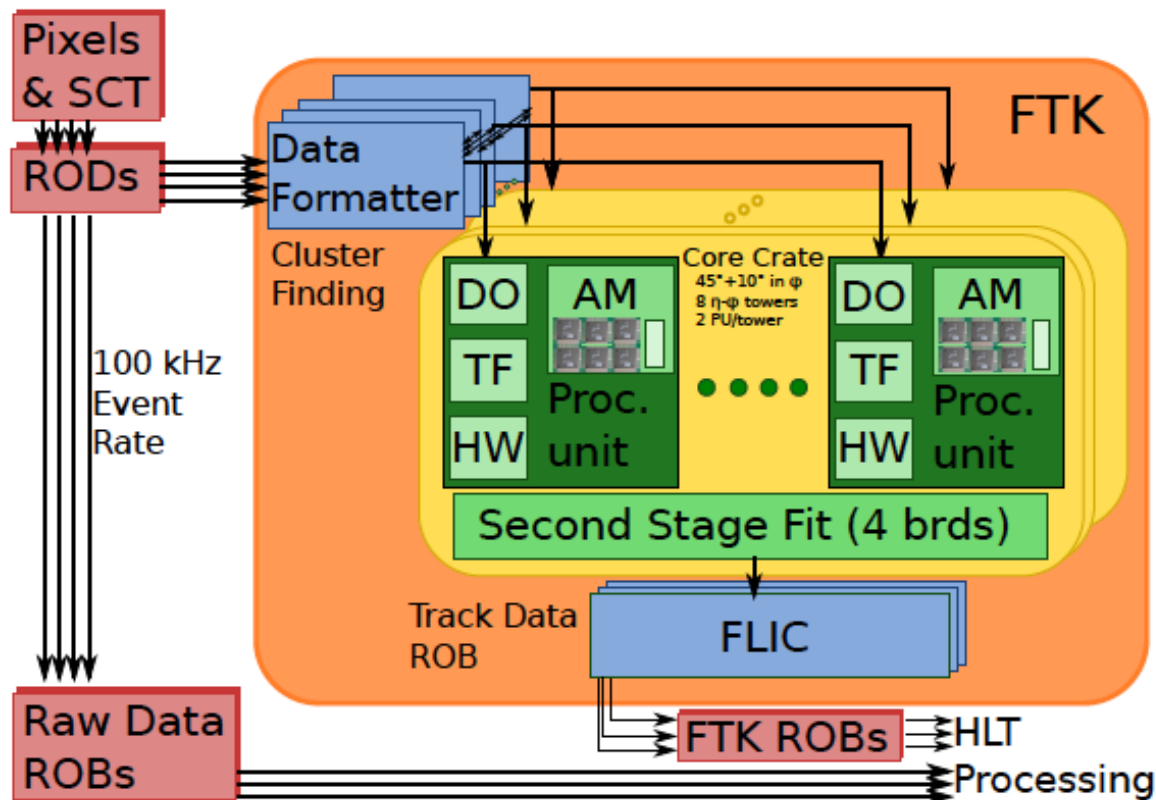


## Data collection network diagram



# ATLAS FTK (Run 2)

- Specialized hardware for tracking in Pixel and SCT (tracker)
- Post-L1 trigger event processing, replaces part of CPU-intensive tasks (track seeding and tracking) now done in HLT

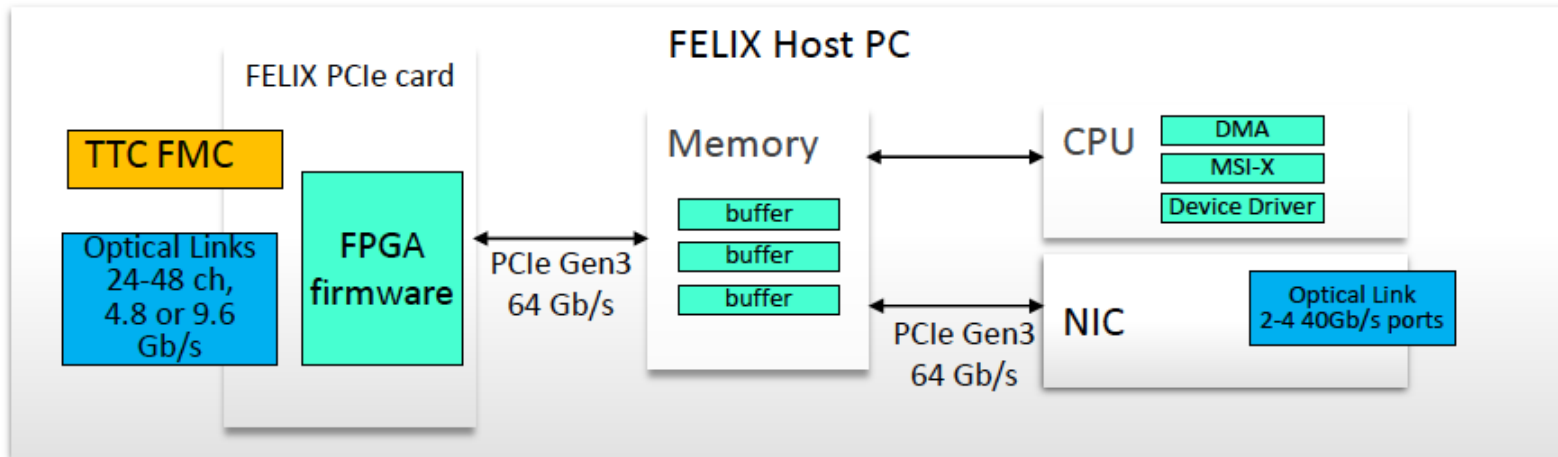


Hardware pattern matching:

- 16400 associative memory (AM) chips
- 200 FPGAs for other functions

# ATLAS readout upgrade

- FELIX



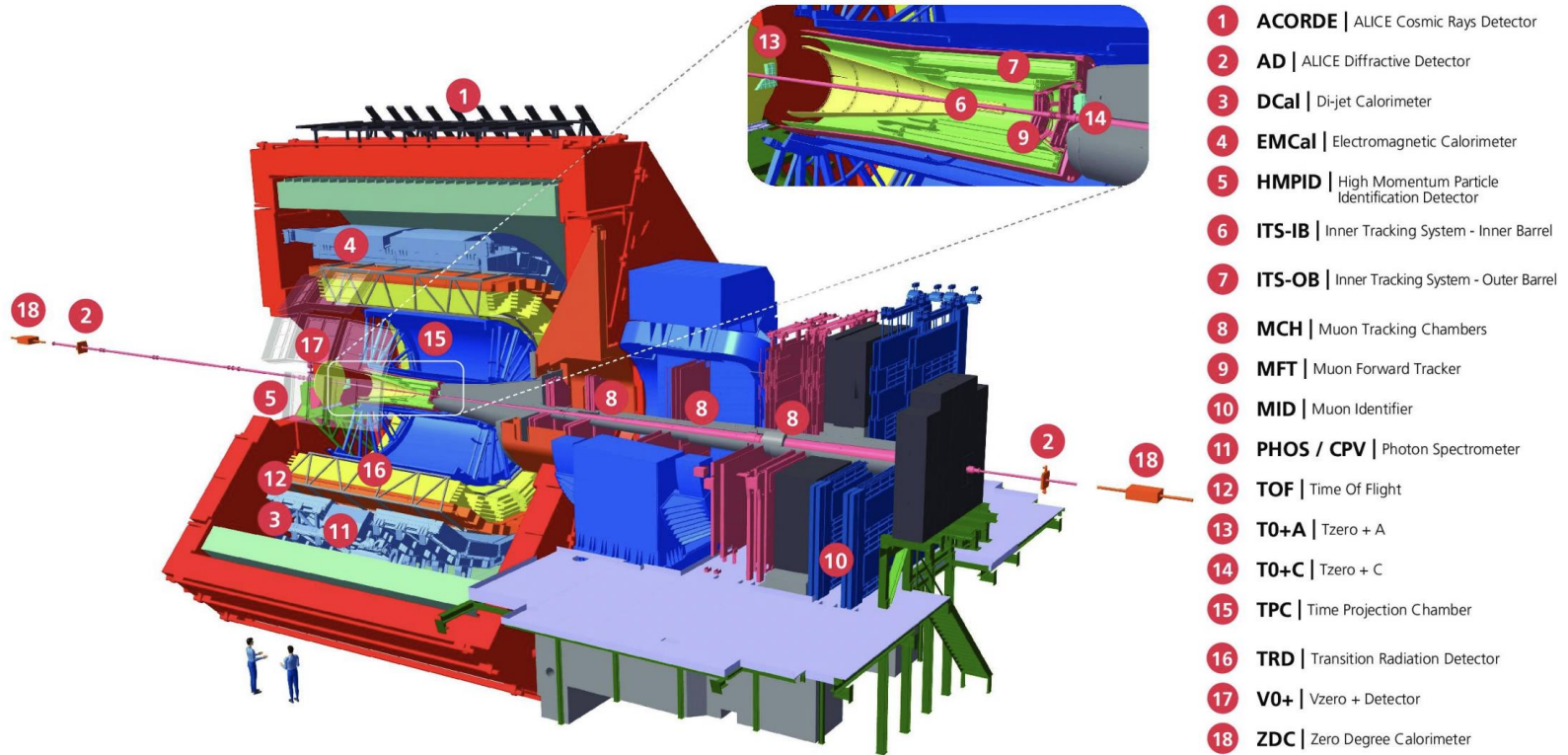
PCIe card with FPGA chip + Host PC + NIC

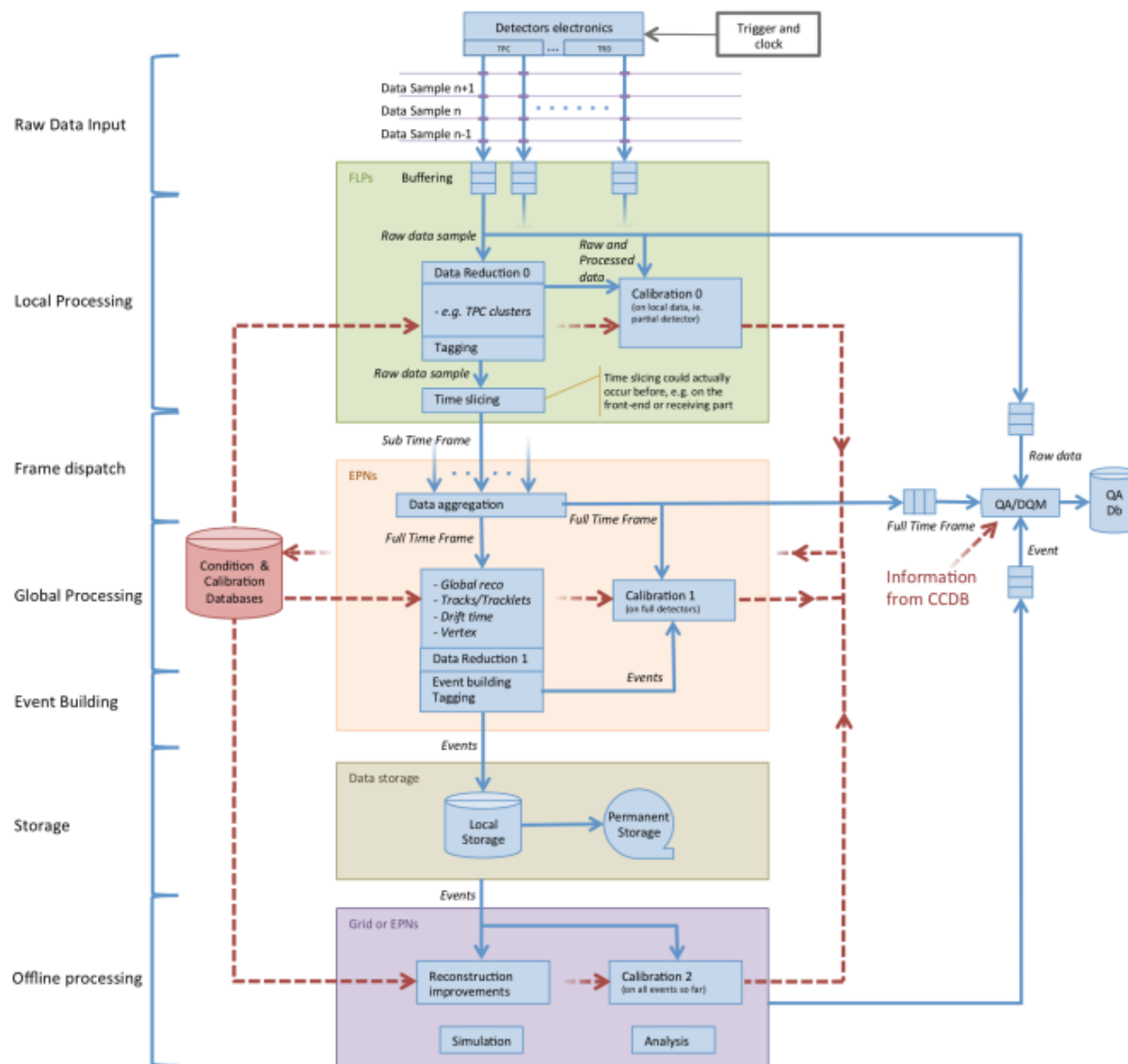
- Replaces custom readout hardware used now
- Replaces custom link standard used now (with “GBT”)
- Handles routes clock and trigger information to front-ends
- Receives event fragments from front-ends

**Router between serial/synchronous links and high level network links (40 GBE, InfiniBand...)**

- Start deployment in Run 3

# ALICE in Run 3



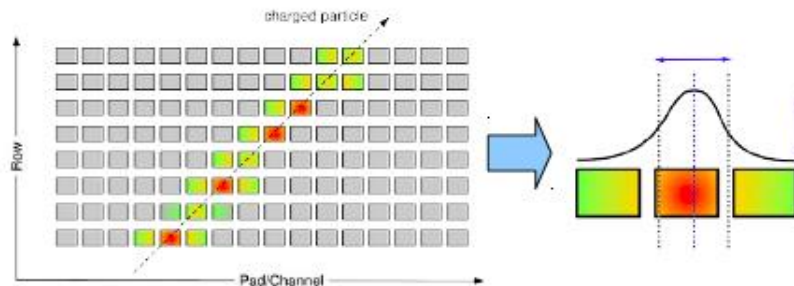


**Figure 4.** Outline of the data flow and processing in the ALICE O<sup>2</sup> computing system.

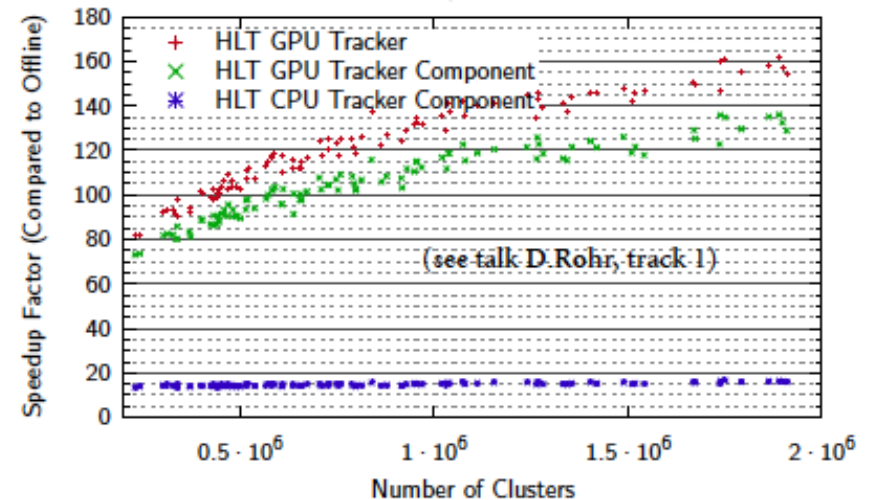
# ALICE High level Trigger – Run 2

- Use of GPUs and FPGAs

- Online reconstruction and data compression facility.
- 180 worker nodes, 8640 HT cores.
- Efficiency through use of hardware acceleration.



- FPGA clusterfinder.
- 1 FPGA board ~ 125 XEON cores.



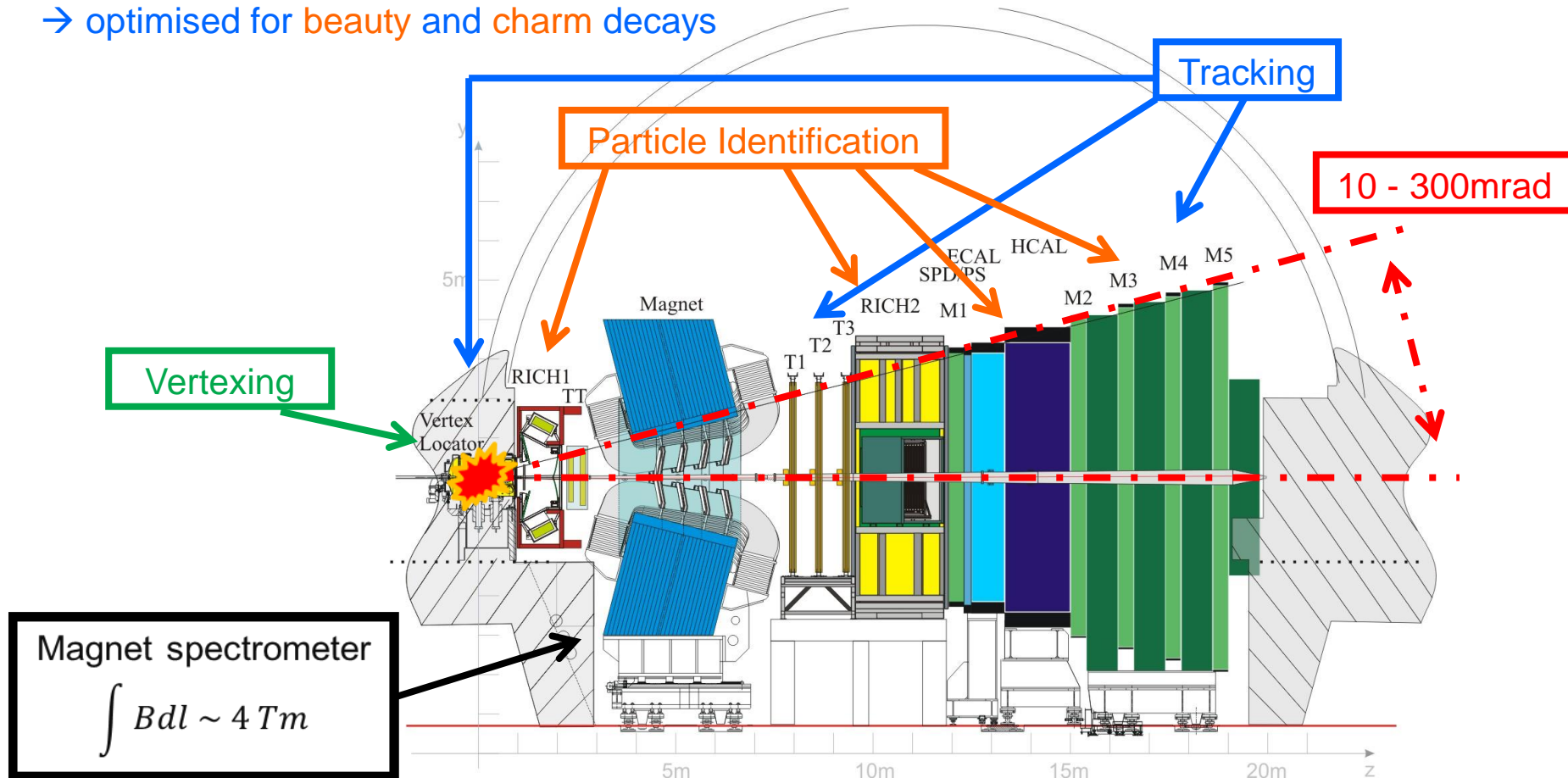
GPU tracking.

# LHCb

Dedicated flavour physics experiment

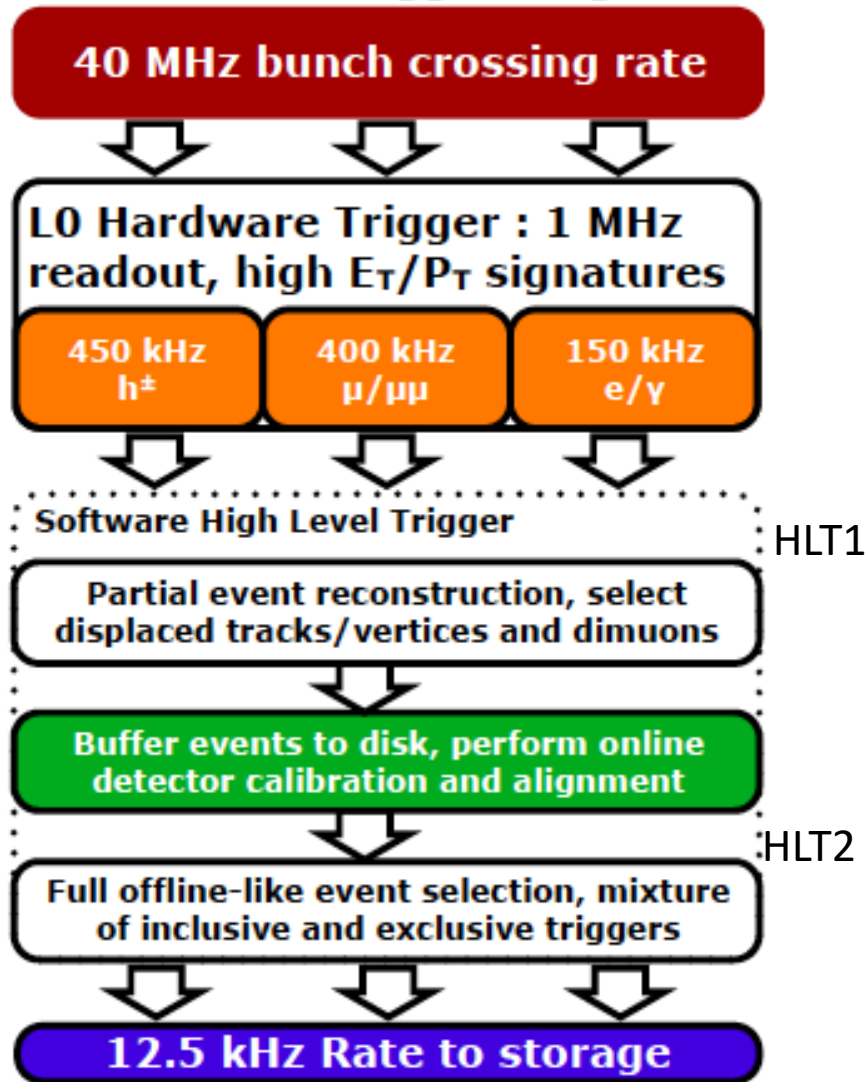
→ forward precision spectrometer

→ optimised for beauty and charm decays

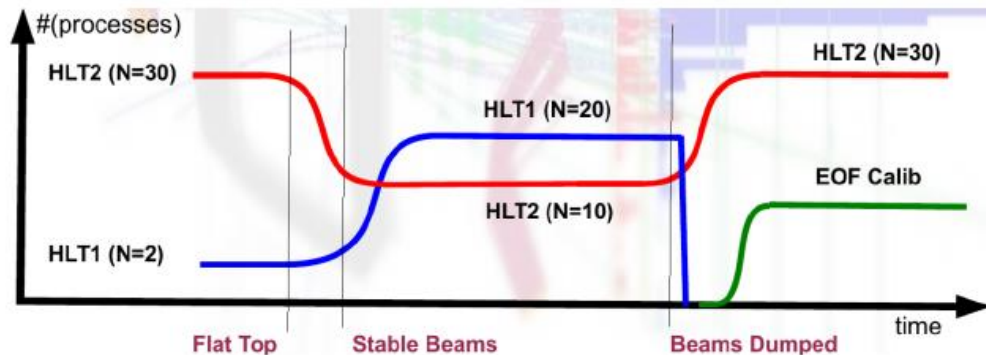


# LHCb Run II DAQ data flow

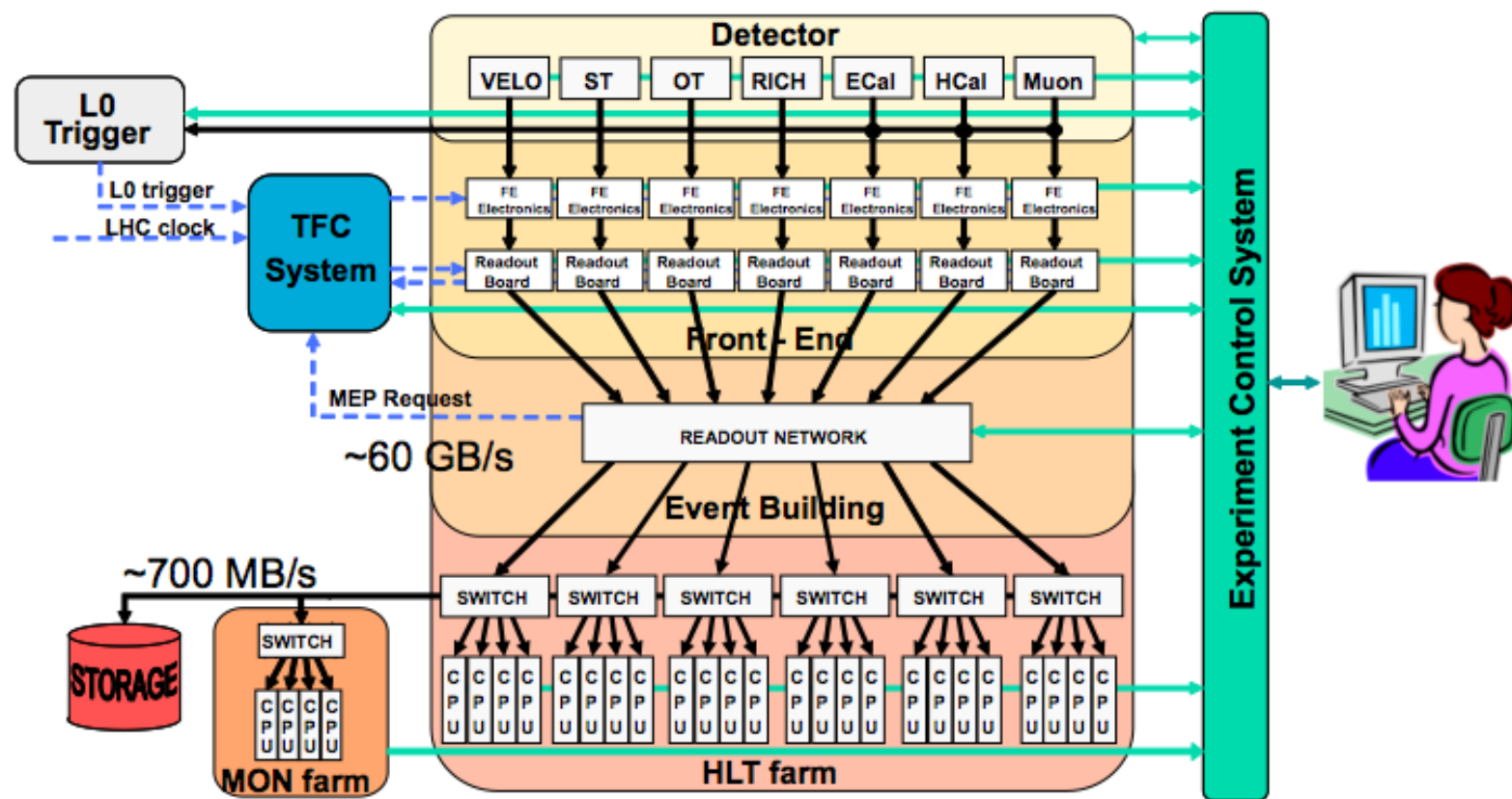
## LHCb 2015 Trigger Diagram



- Hardware trigger (L0)
  - Based on multiplicity, calorimeters and muon detectors
  - Fixed latency of  $4\ \mu\text{s}$
  - Accept rate 1 MHz before readout
- Software trigger (HLT)
  - HLT Split in two stages
  - Events buffered to disk after HLT1
  - Output rate 12.5 kHz
  - 62 subfarms: 1780 nodes (27000 CPU cores) via 12 or 2x10 Gbit Ethernet
  - 10 PiB disk space
- HLT2 is fully asynchronous:



# LHCb Trigger and DAQ - run 2

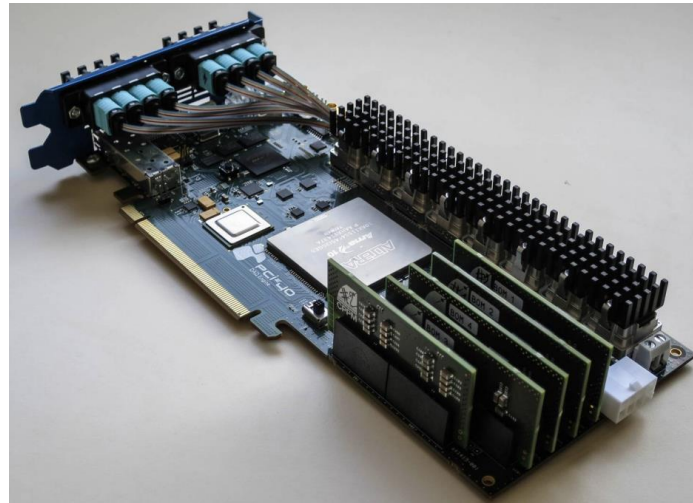


Gbit Ethernet based  
readout / event  
building network  
 $\sim 1500$  ports

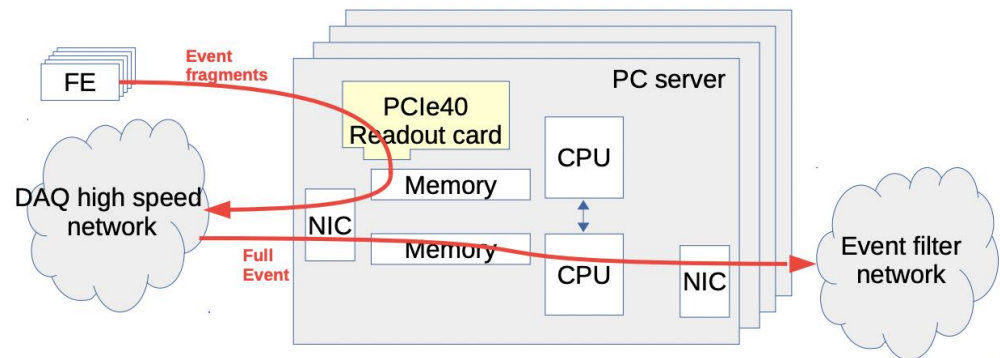
Average event size 60 kB  
Average rate into farm 1 MHz  
Average rate to tape  $\sim 12 \text{ kHz}$

# LHCb DAQ – Readout in Run 3

- PCIe40
  - common LHCb and ALICE readout board
- Large FPGA (>1m cells)
- 48 x 10 Gbit/s bidirectional links
- Sustained 112 Gbits/s interface with CPU through PCIe

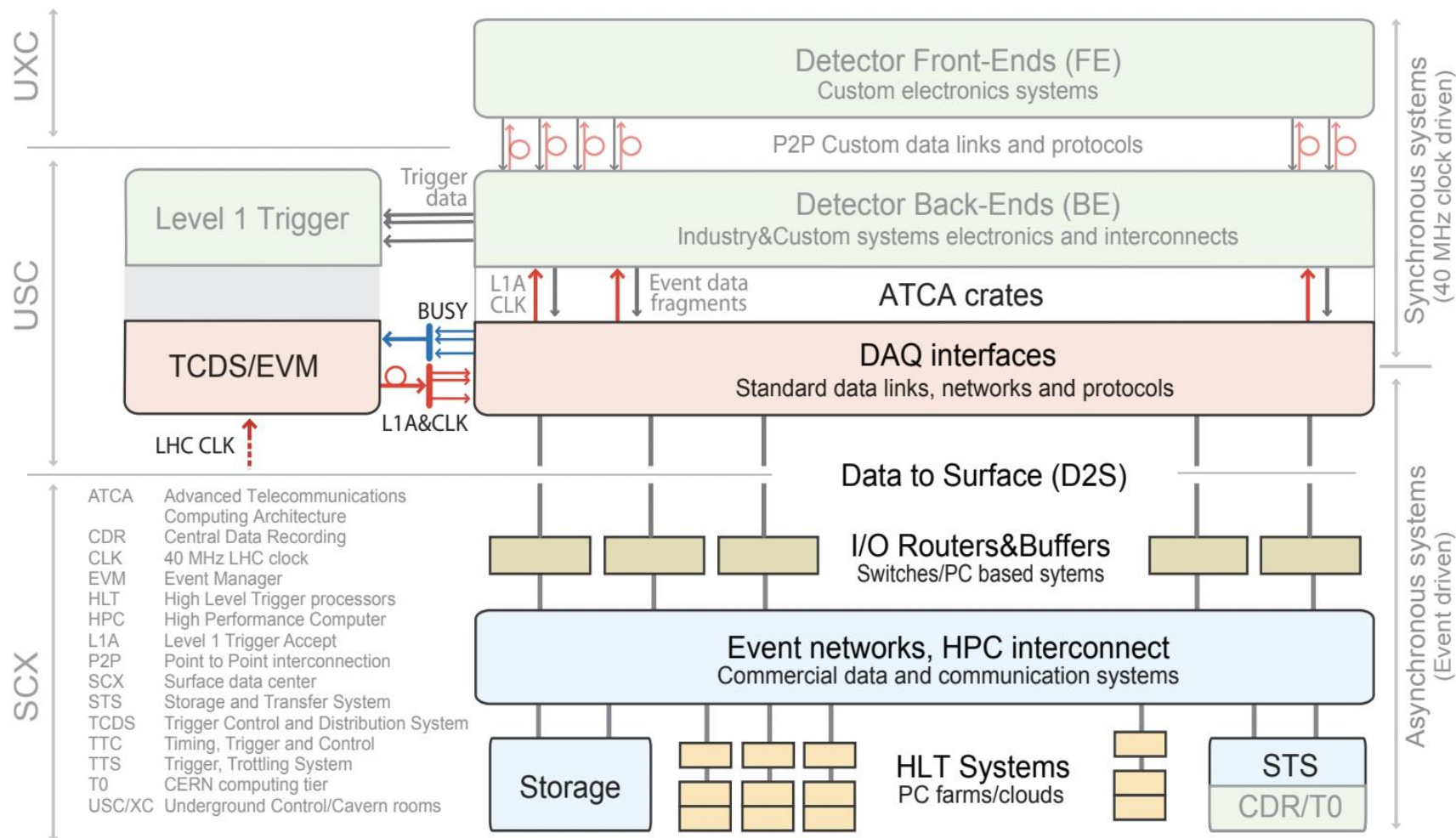


PC → network interfaces



# CMS DAQ for HL-LHC

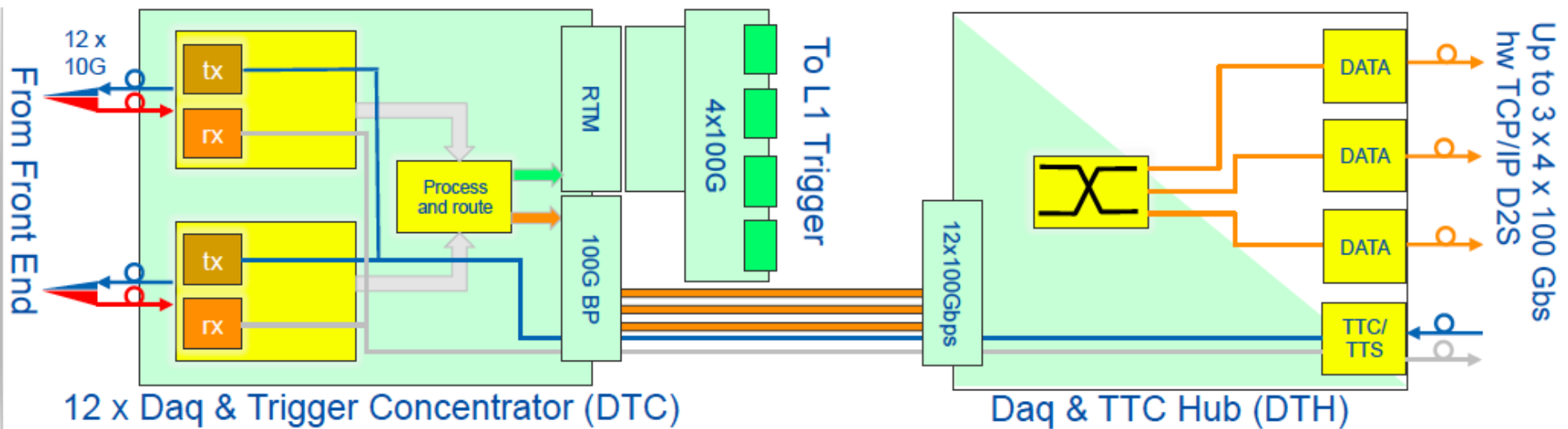
- Functional view



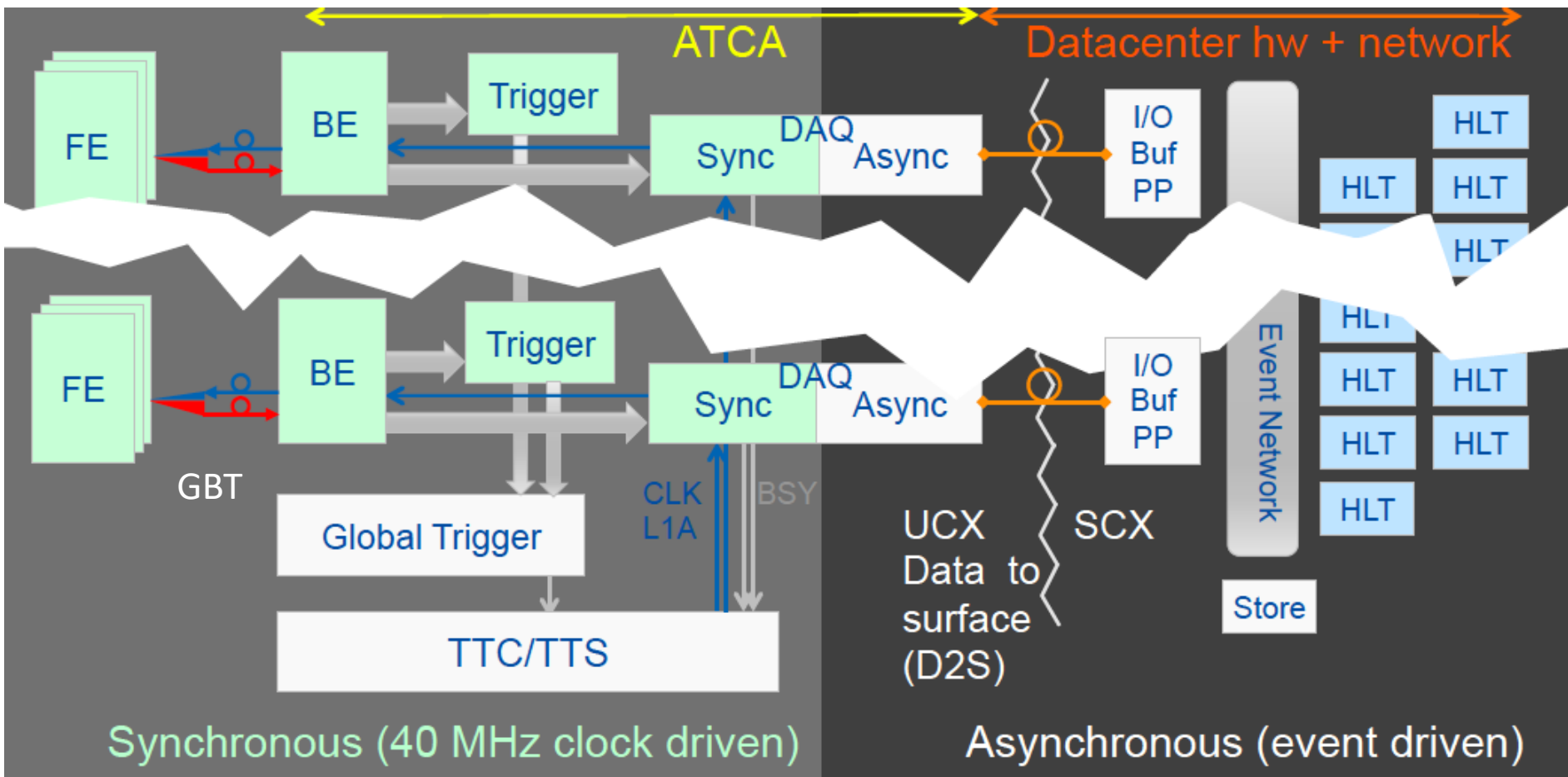
# Run 4 readout and DAQ input

## Readout:

- About 50k point-to-point bidirectional optical links (GBT) on-to-off-detector with varying fractions devoted to trigger data
- Read out central (barrel) calorimeter and muon systems in untriggered (continuous / streaming) mode



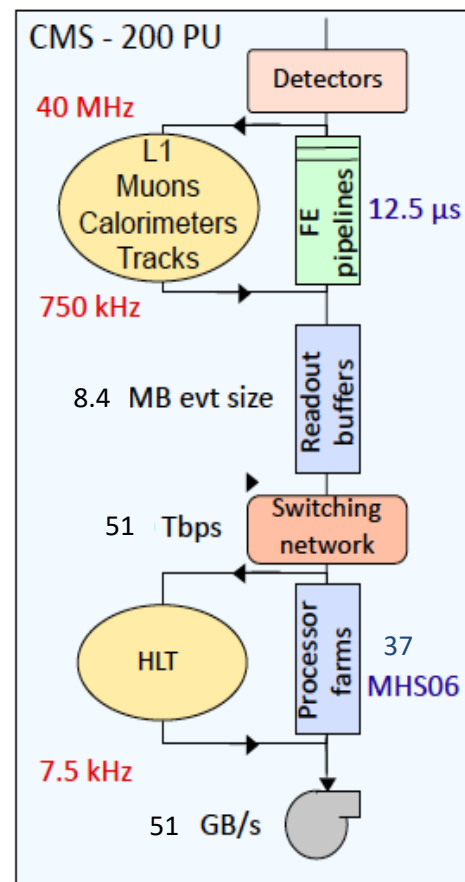
# Run 4 readout and DAQ



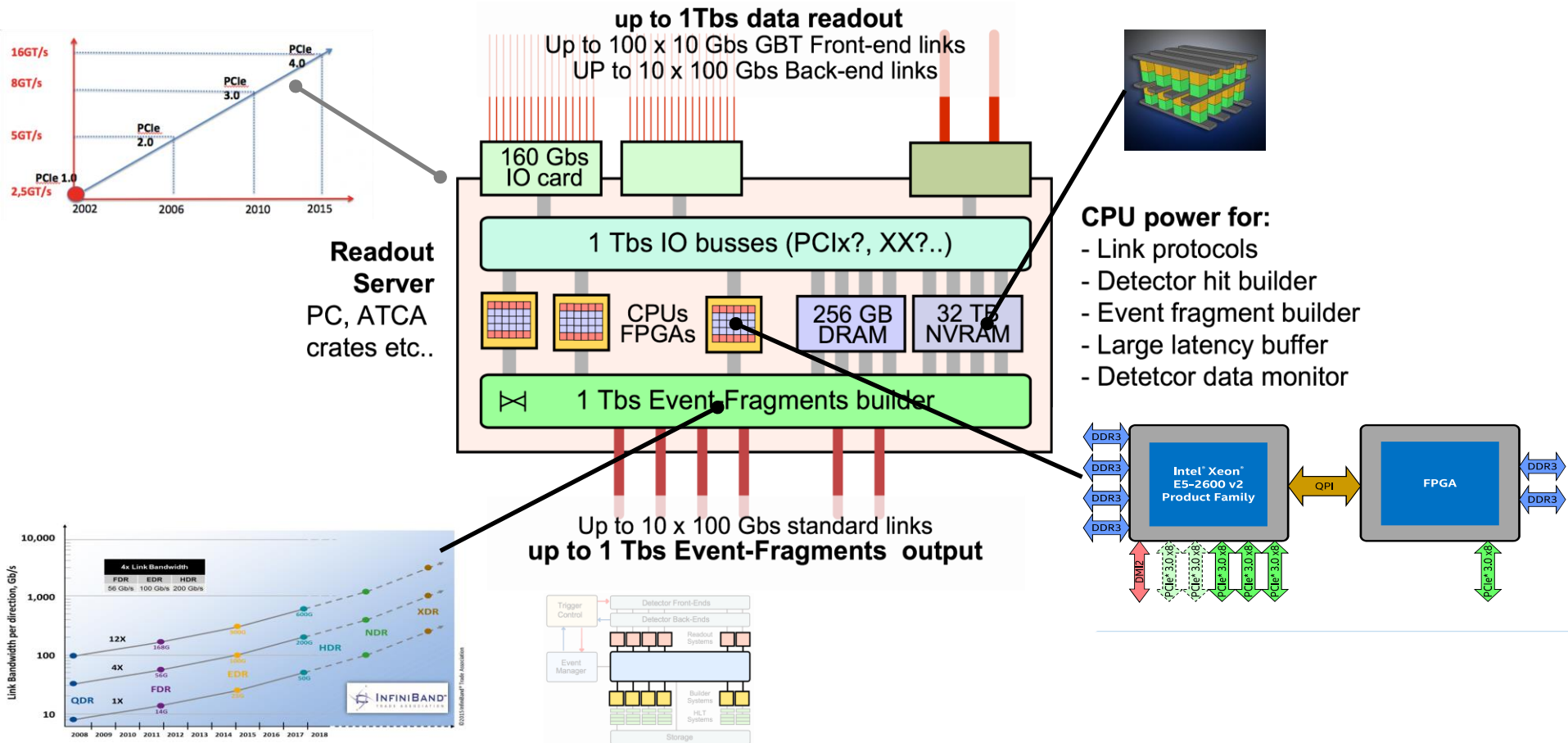
# CMS DAQ requirements for HL-LHC

- Run-3 → Run 4/5 requirements
  - 5 – 7.5 times L1 rate
  - 3 – 4 times event size
  - 31 times readout bandwidth**
  - 50 times HLT computing power → PU x trigger rate & new detectors!
  - 15 times storage (3 times bandwidth)

CMS detector Peak $\langle$ PU $\rangle$	LHC Phase-1 60	HL-LHC Phase-2	
		140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size at HLT input	2.0 MB <sup>a</sup>	6.1 MB	8.4 MB
Event Network throughput	1.6 Tb/s	24 Tb/s	51 Tb/s
Event Network buffer (60 s)	12 TB	182 TB	379 TB
HLT accept rate	1 kHz	5 kHz	7.5 kHz
HLT computing power <sup>b</sup>	0.7 MHS06	17 MHS06	37 MHS06
Event Size at HLT output <sup>c</sup>	1.4 MB	4.3 MB	5.9 MB
Storage throughput <sup>d</sup>	2 GB/s	24 GB/s	51 GB/s
Storage throughput (Heavy-Ion)	12 GB/s	51 GB/s	51 GB/s
Storage capacity needed (1 day <sup>e</sup> )	0.2 PB	1.6 PB	3.3 PB



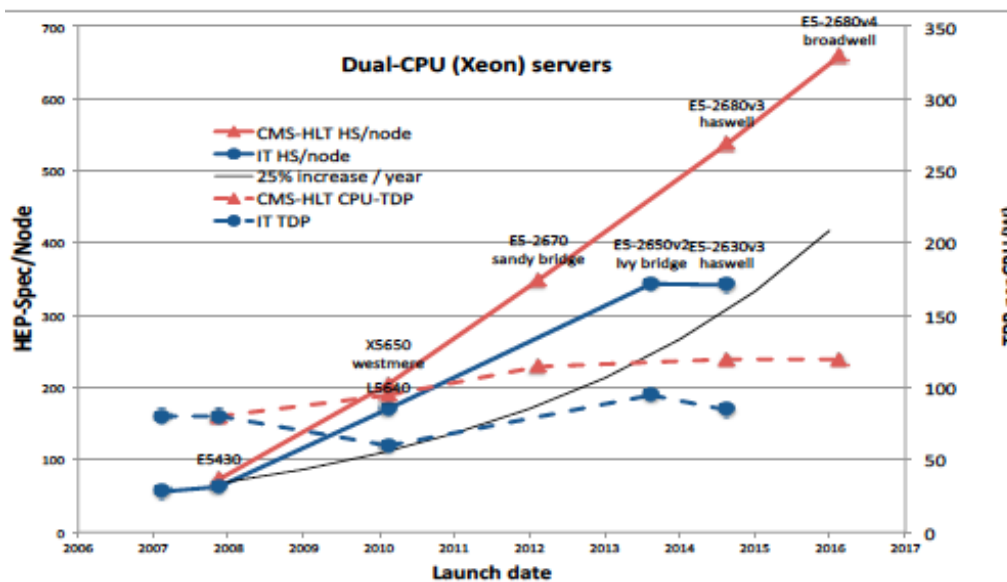
# Possible 2027 CMS DAQ RU/BU PC



# ATLAS upgrade HLT projections (2016)

Parameter	L0/L1	L0	Run2
Filtering Rate	400 kHz	1 MHz	100 kHz
Overall Compute Power	11 MHS06	>11 MHS06	0.8 MHS06
Computer Power excluding tracking	5 MHS06	5 MHS06	—

- Highly depends on scaling of future PC platforms



HEP-SPEC06 – benchmark designed to scale with performance of High Energy Physics code on a similar machine

# Units

$N$  – number of interaction events

$\sigma$  – interaction cross section

Instantaneous  
luminosity

$$L = \frac{1}{\sigma} \frac{dN}{dt}$$

$$\eta \equiv -\ln \left[ \tan \left( \frac{\theta}{2} \right) \right]$$

Integrated  
luminosity

$$L_{\text{int}} = \int L \, dt.$$