# Software & Computing in ATLAS Trigger and Data-Acquisition

W.Vandelli
CERN Experimental Physics Department/ADT

on behalf of
ATLAS Collaboration

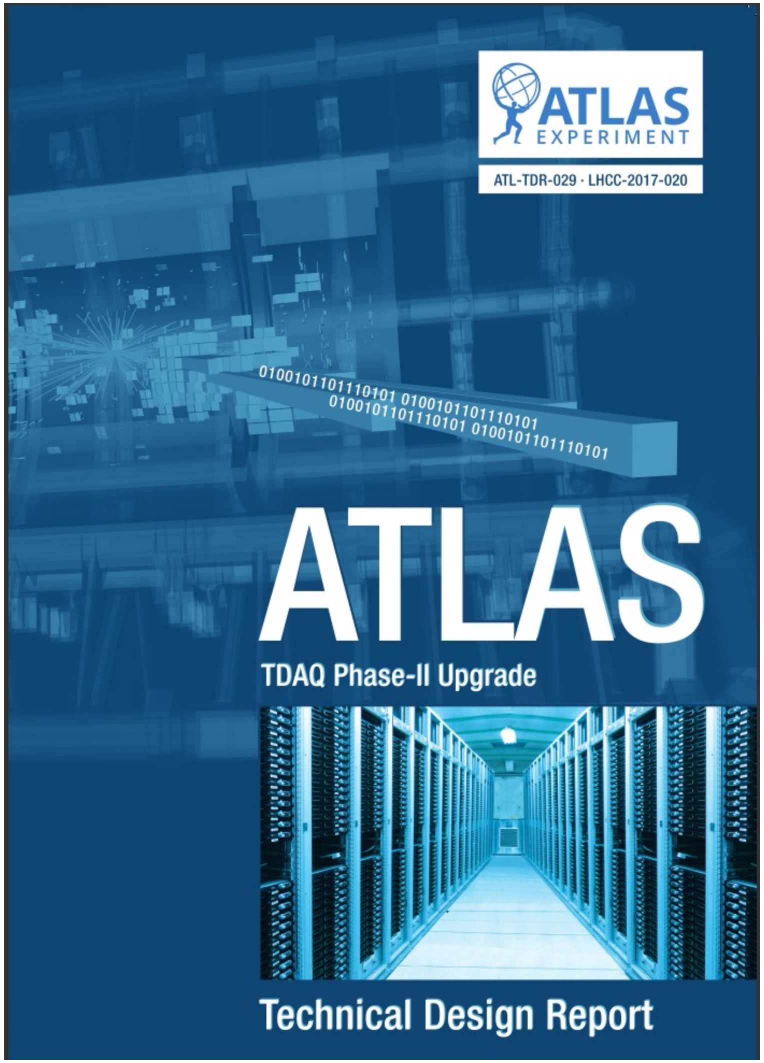
# Introduction

ATLAS TDAQ is a large mission-critical heterogeneous computing infrastructure

It operates in-house developed software and the DAQ team is responsible for all aspects from hardware procurement and installation to operation and upgrade

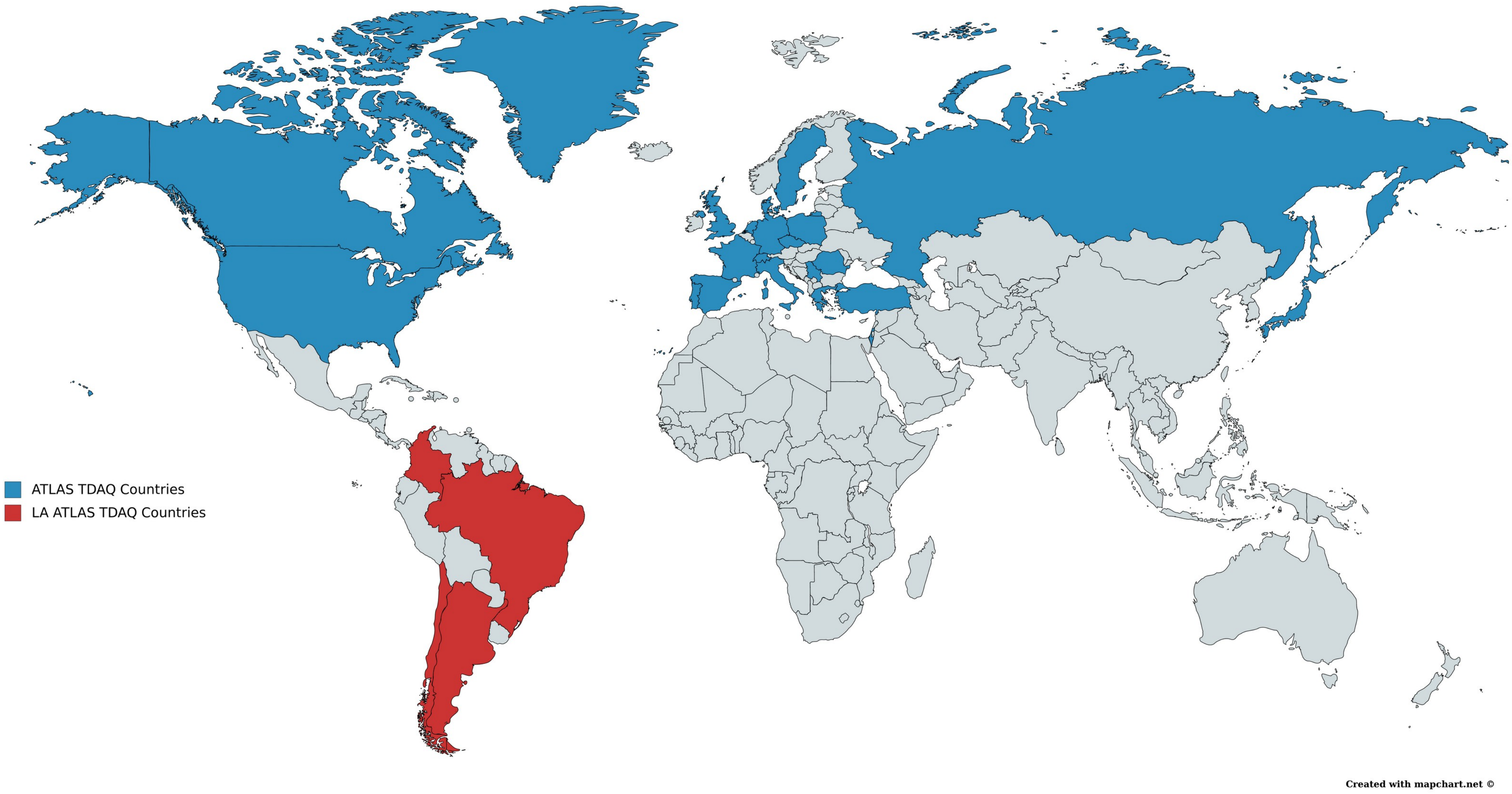
I will mainly focus on the challenges associated with the Phase-II/Run 4 upgrade (2024), even though Phase-I/Run 3 (2021) operation is not given

| | Phase-I/Run 3 TDAQ |
|---|---|
| Processing Servers | ~2000 |
| Processing Appl. | ~50000 |
| Local Storage (TB) | ~800 |
| 10 GbE Links | ~500 |

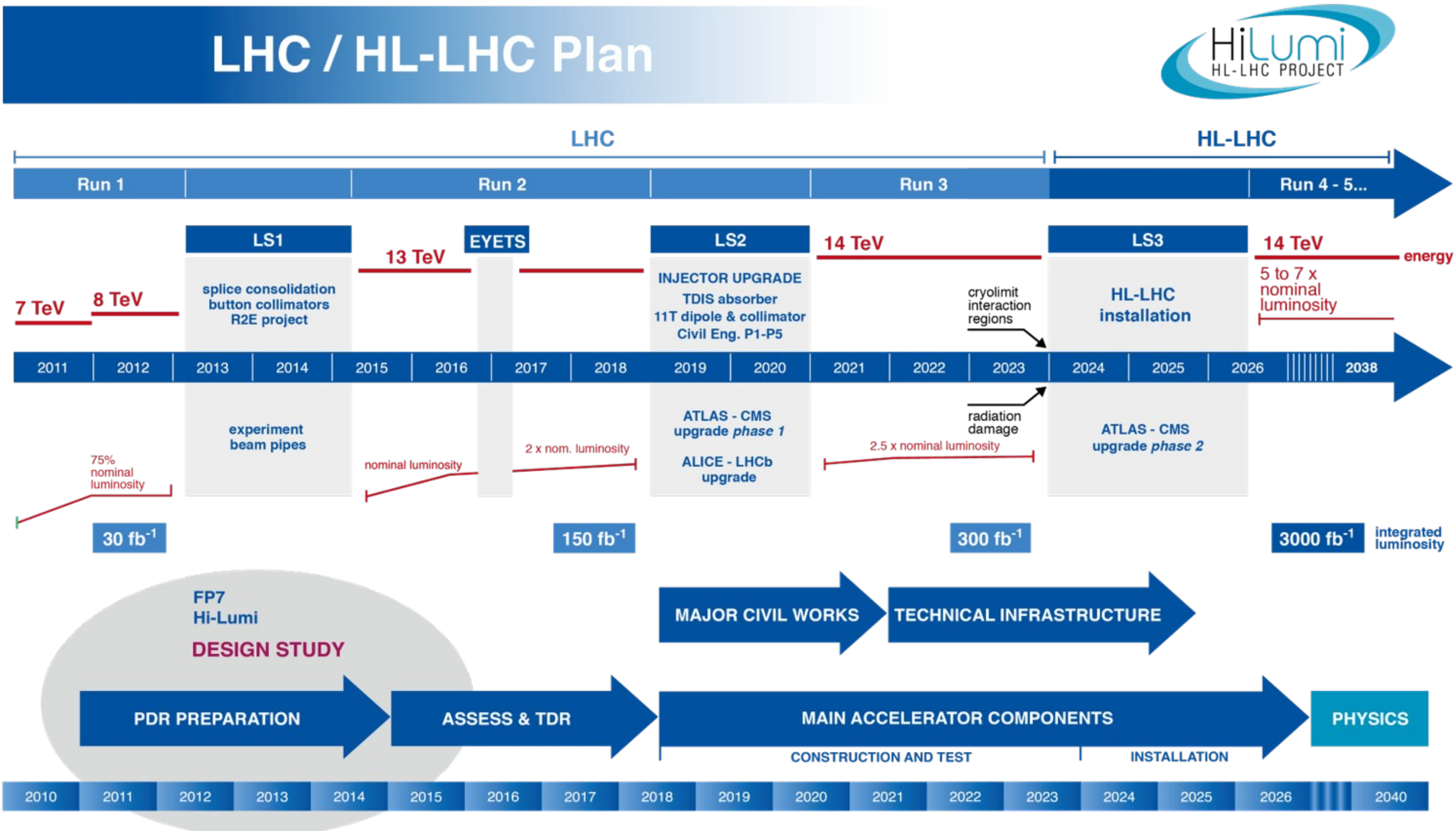



ATLAS Collaboration
Technical Design Report for the Phase-II Upgrade of the ATLAS TDAQ System
https://cds.cern.ch/record/2285584

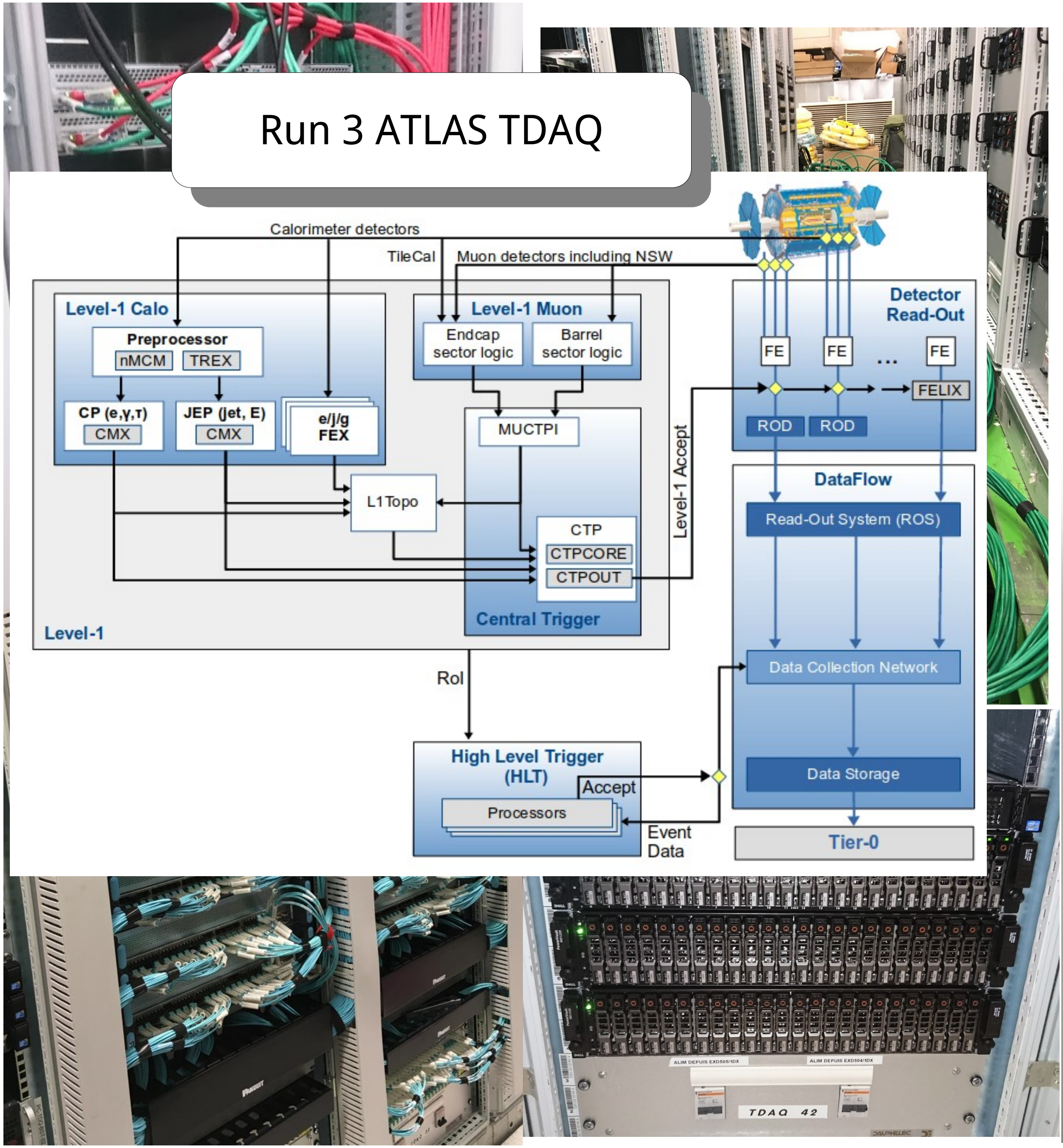
# Latin America in ATLAS TDAQ

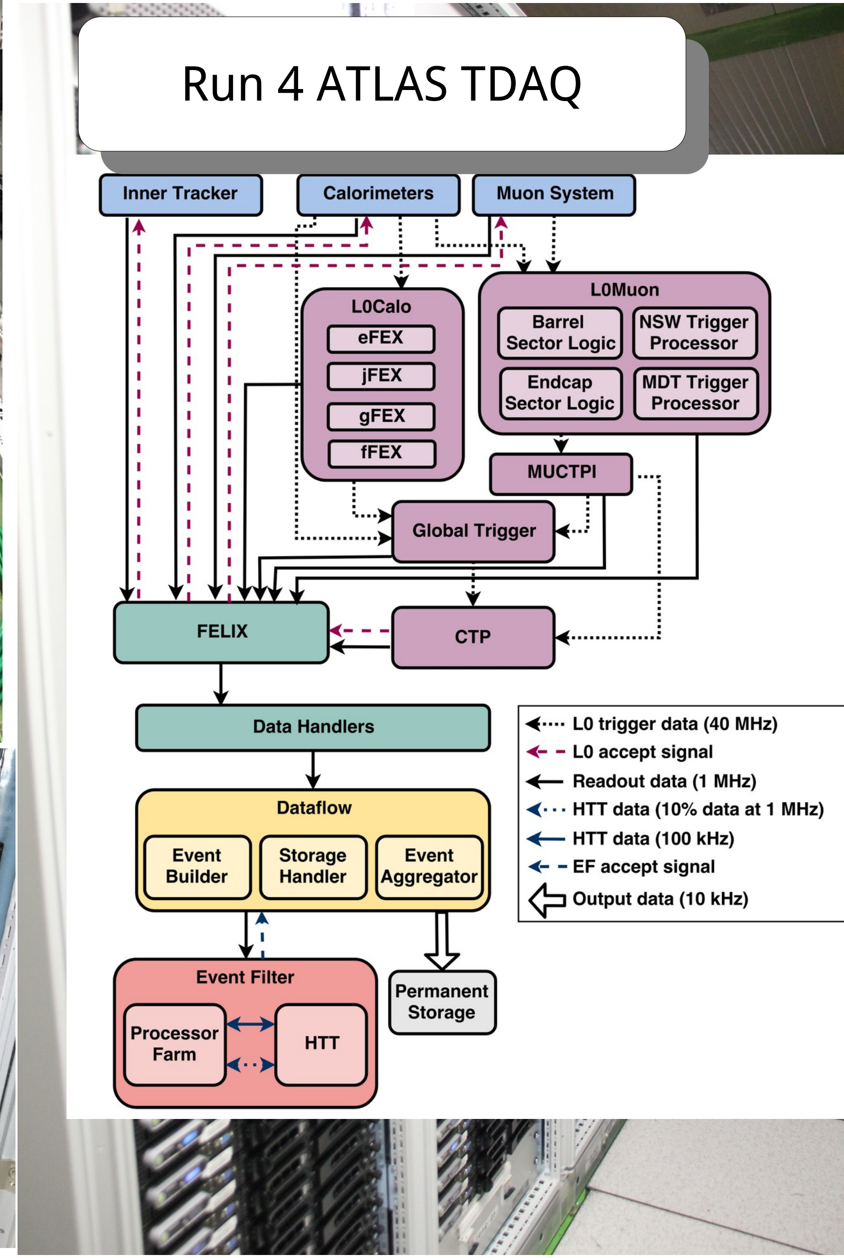


# LHC Schedule



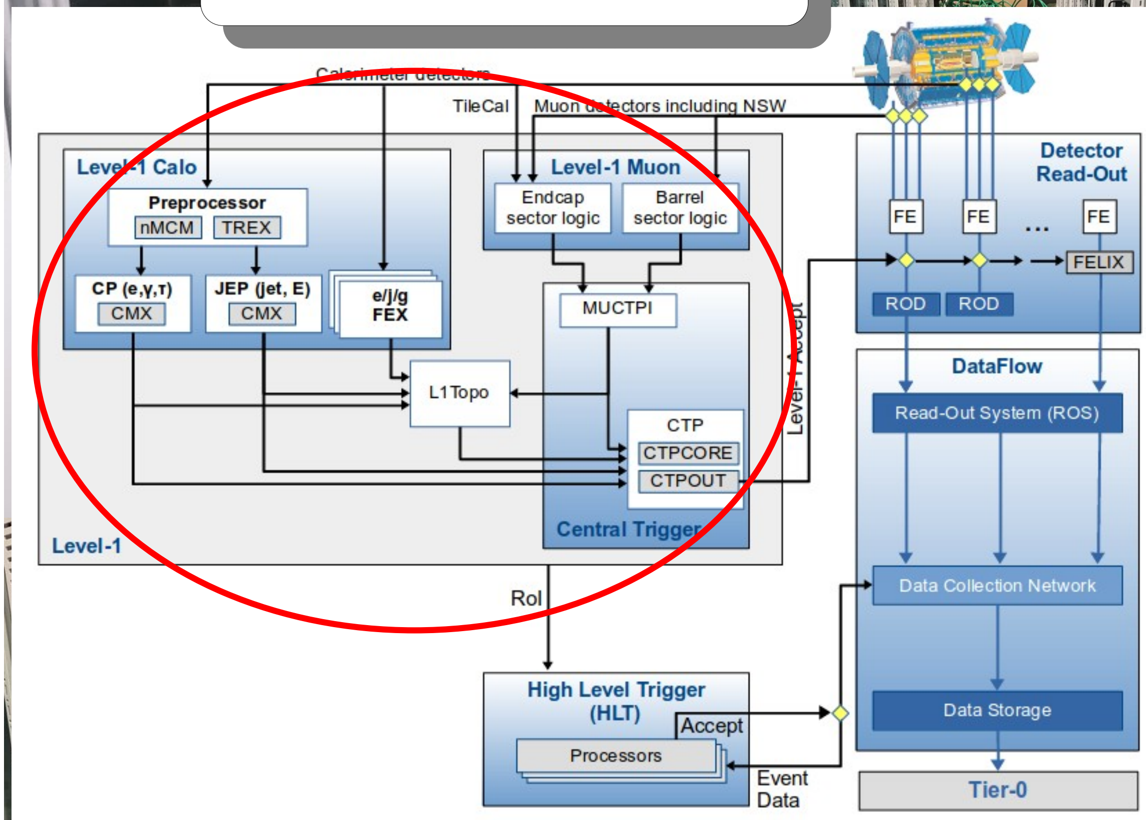
# ATLAS TDAQ Architecture



# ATLAS TDAQ Architecture

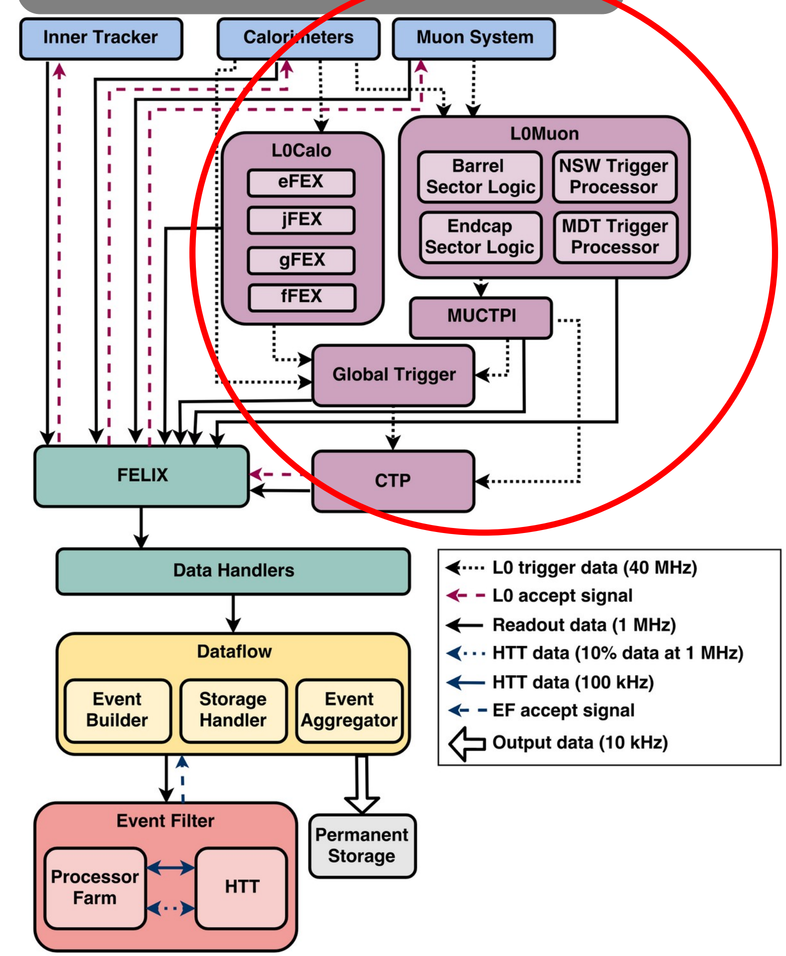
## Run 3 ATLAS TDAQ




Realtime first-level trigger

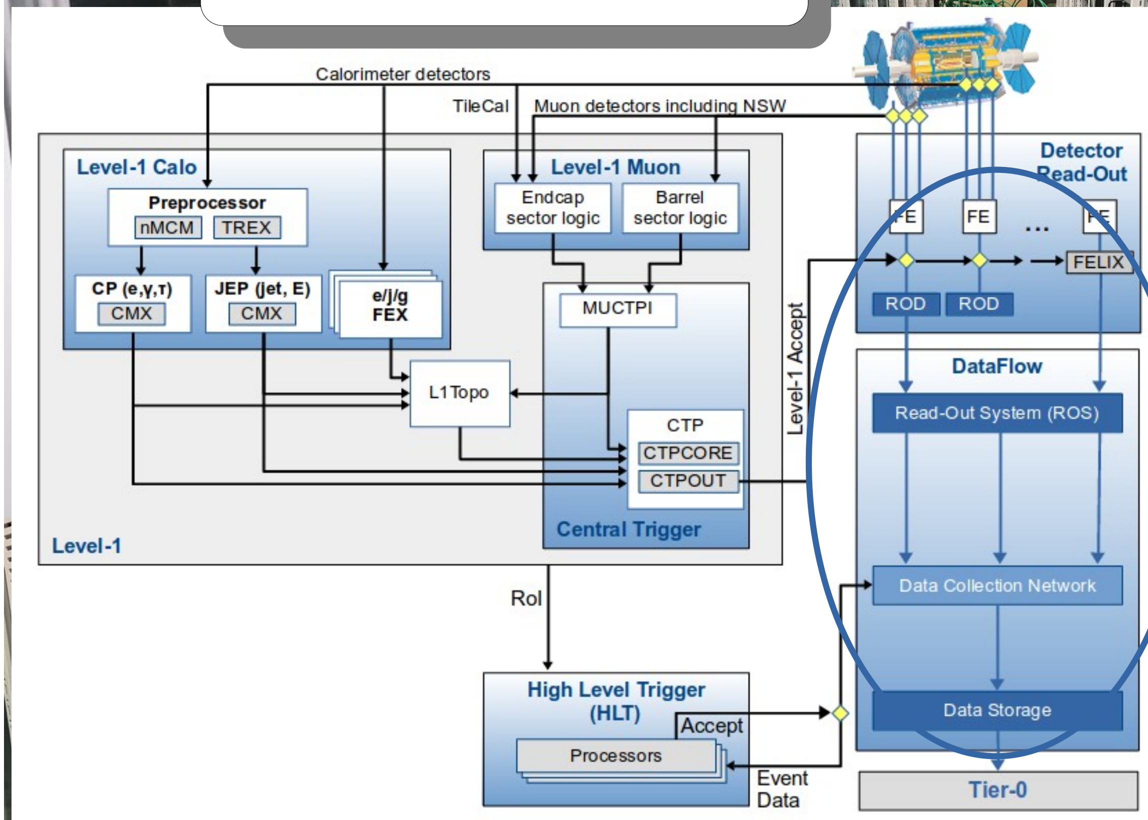
Implemented with custom electronics and connections

## Run 4 ATLAS TDAQ

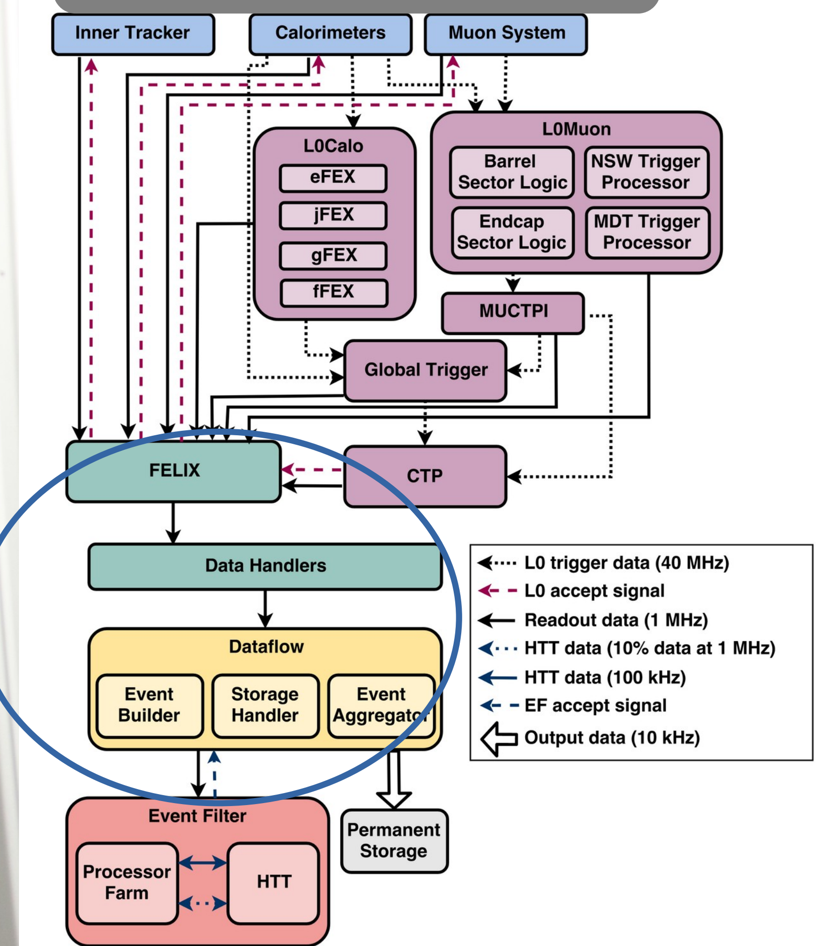


# ATLAS TDAQ Architecture

## Run 3 ATLAS TDAQ




## Run 4 ATLAS TDAQ

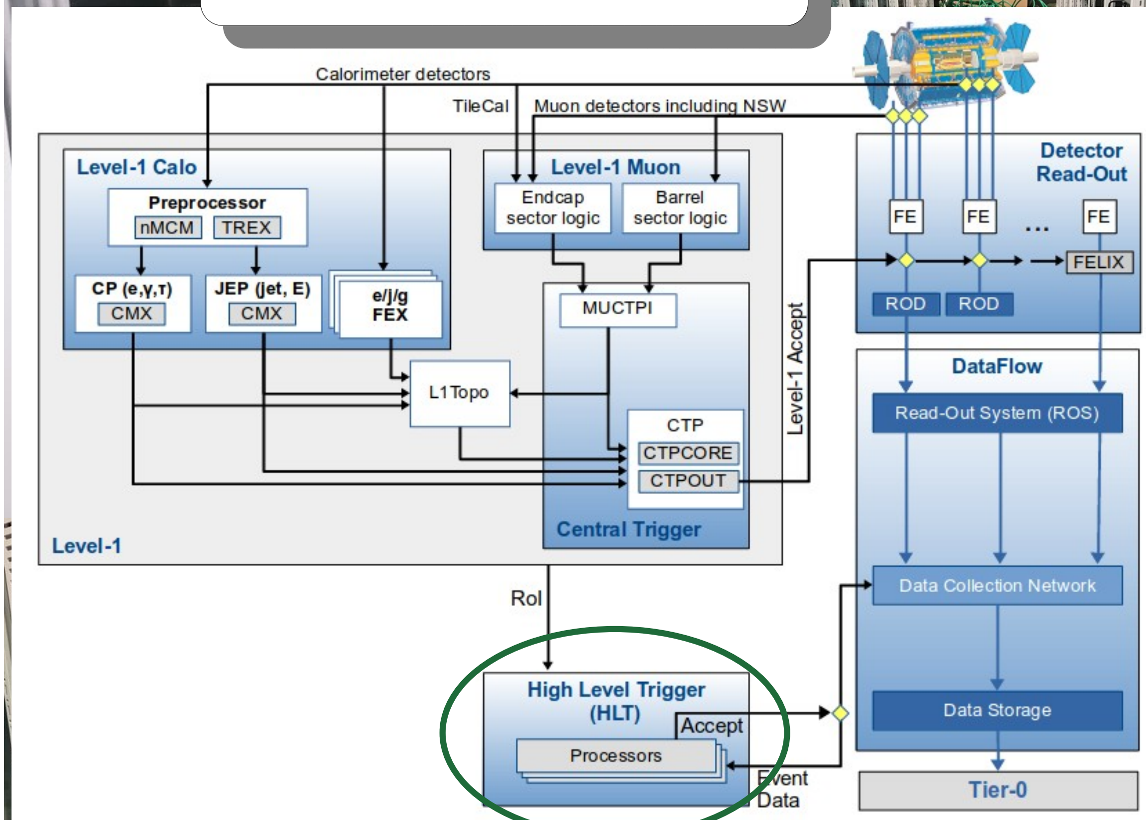



Data-flow infrastructure: conveys, buffers and compresses data

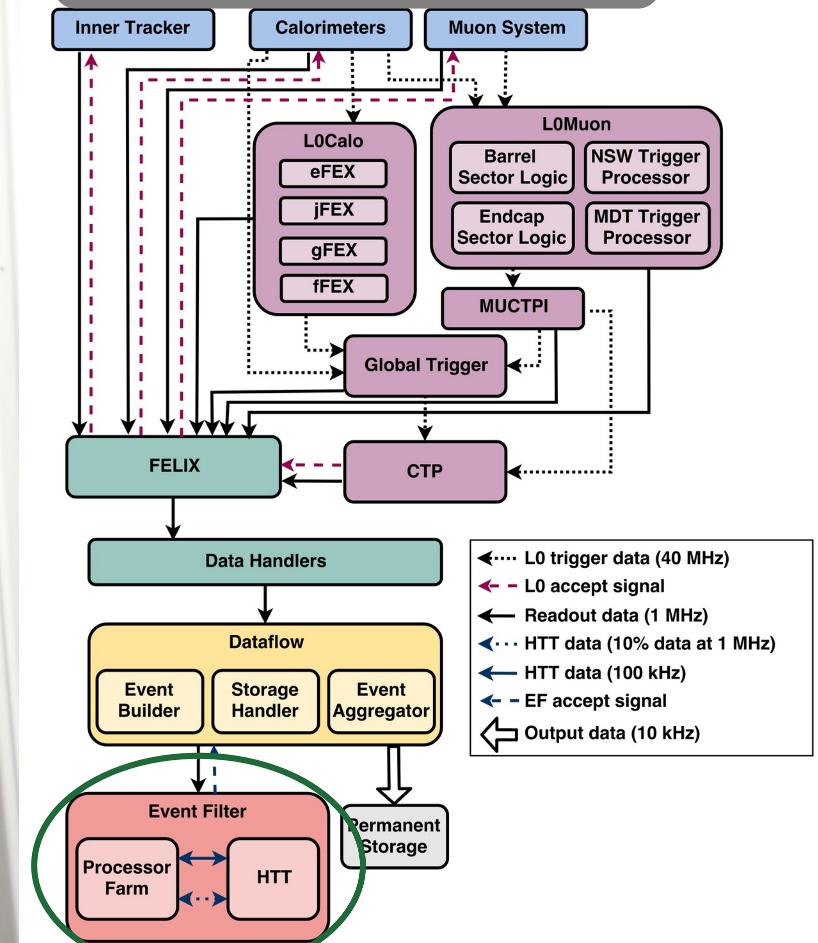
Implemented mainly with COTS servers and networks

# ATLAS TDAQ Architecture

## Run 3 ATLAS TDAQ



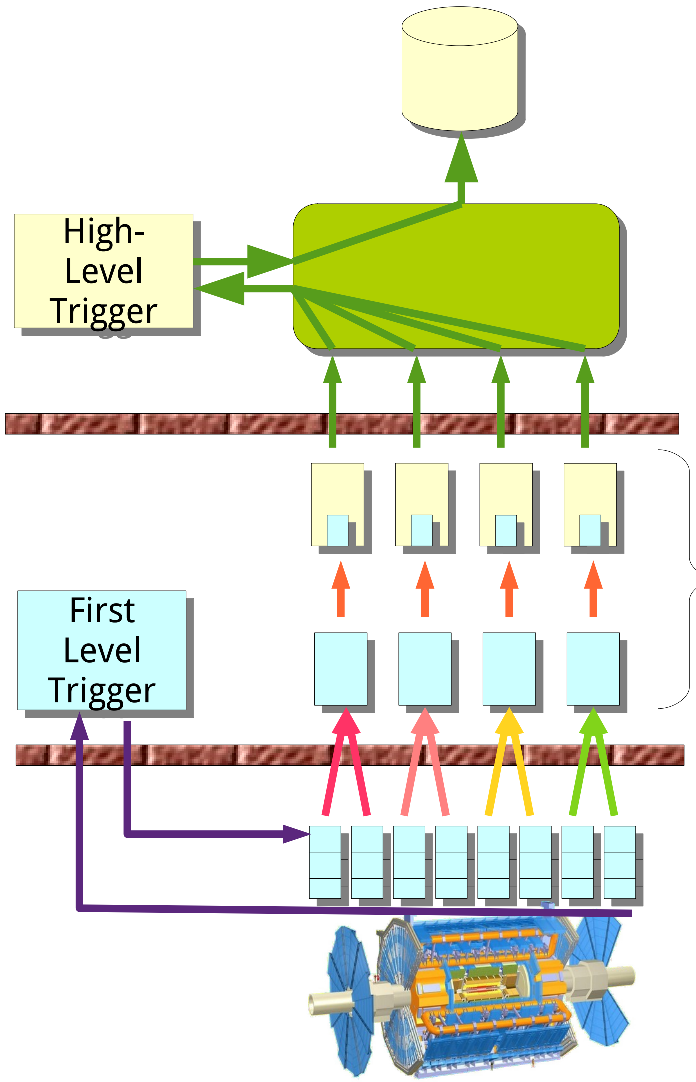

## Run 4 ATLAS TDAQ




Event Filter: quasi-realtime event selection

Large computer cluster, possibly aided by coprocessors

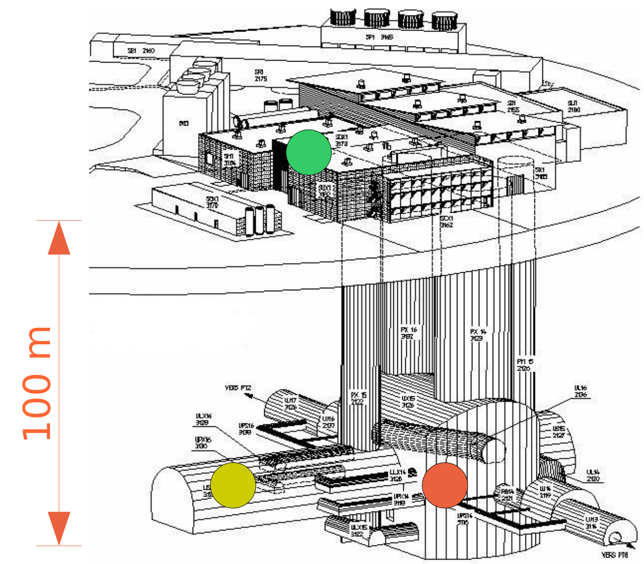
# Distributed in space



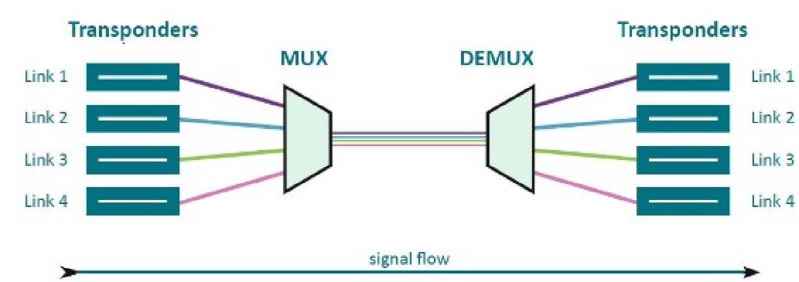

Surface

Service cavern

Experimental cavern

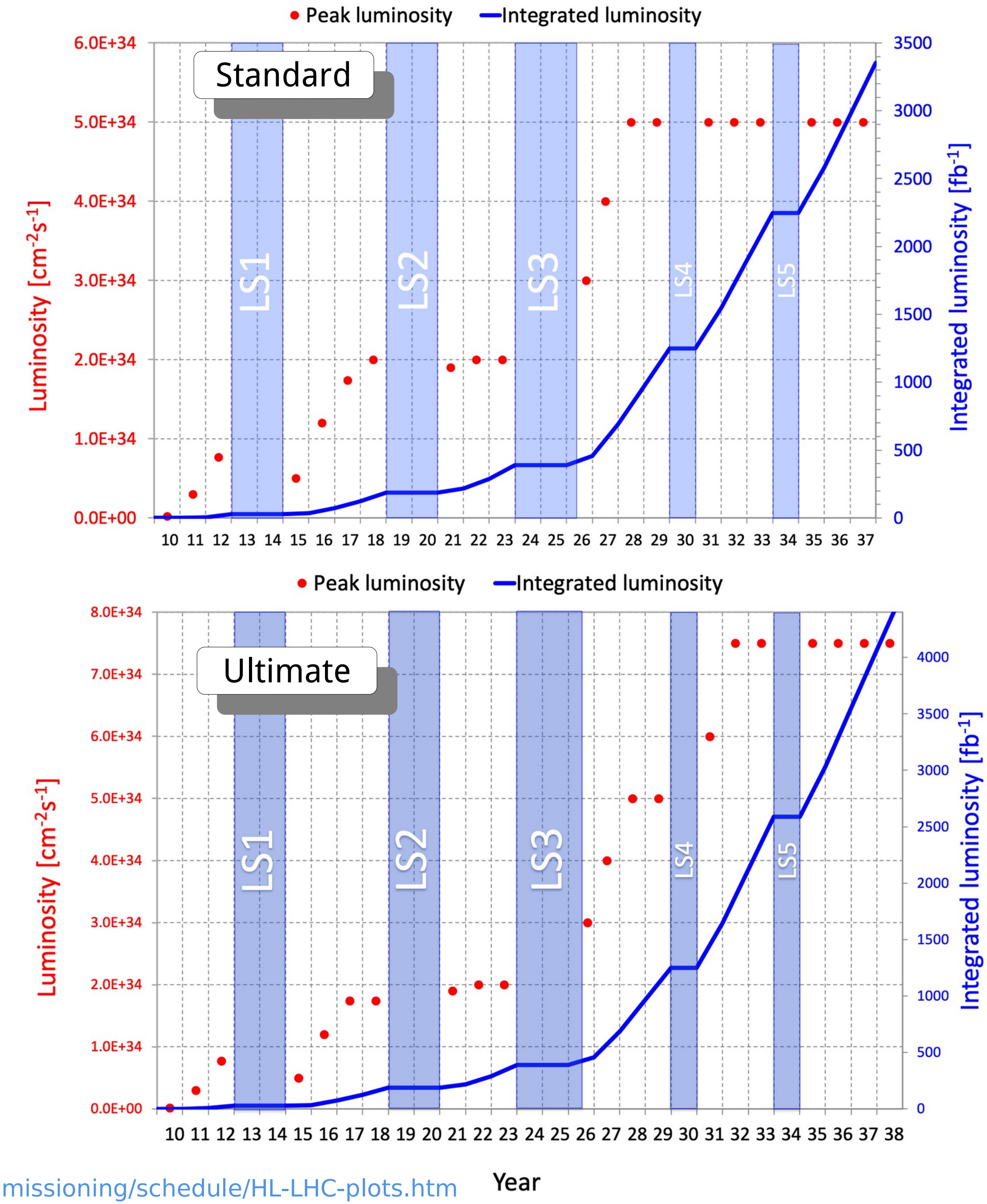



- e.g. considering WDM (wavelength division multiplexing) to reduce the number of long range fibres
  - trade-off like fibres vs transceivers cost



# HL-LHC Operation Point

- HL-LHC programme aims at a total integrated luminosity of at least 3000 fb$^{-1}$
  - ten-fold increase wrt Run 1/2/3 aggregate
- Corresponding increase in peak instantaneous luminosity
  - $\mathscr{L}$~5·10$^{34}$ cm$^{-2}$s$^{-1}$ (ultimate 7.5·10$^{34}$ cm$^{-2}$s$^{-1}$)
  - achieved mainly via pileup <μ>: 140 (ultimate 200)
- For reference Run 3 operation point:
  - $\mathscr{L}$~2·10$^{34}$ cm$^{-2}$s$^{-1}$ - <μ>~50




https://lhc-commissioning.web.cern.ch/lhc-commissioning/schedule/HL-LHC-plots.htm

# Physics Motivations

- The challenging and broad HL-LHC programme requires trigger thresholds comparable with the current ones, e.g.:
  - electroweak scale requires low $p_T$ leptons
  - searches for new physics with low $\Delta m$
  - HH measurements requires low $p_T$ jets /b-jets
- At fixed threshold, trigger rates scale with peak luminosity
  - worsened by pileup environment

**Major increase in readout and recording rates**

| Trigger Selection offline threshold (GeV) | Run 1 | Run 2 | HL-LHC |
|---|---|---|---|
| Isolated single e | 25 | 27 | **22** |
| Isolated single $\mu$ | 25 | 27 | **20** |
| Di-$\gamma$ | 25, 25 | 25, 25 | **25, 25** |
| Di-$\tau$ | 40, 30 | 40, 30 | **40, 30** |
| Four-jet w/ b-jets | 45 | 45 | **65** |
| $H_T$ | 700 | 700 | **375** |
| MET | 150 | 200 | **200** |

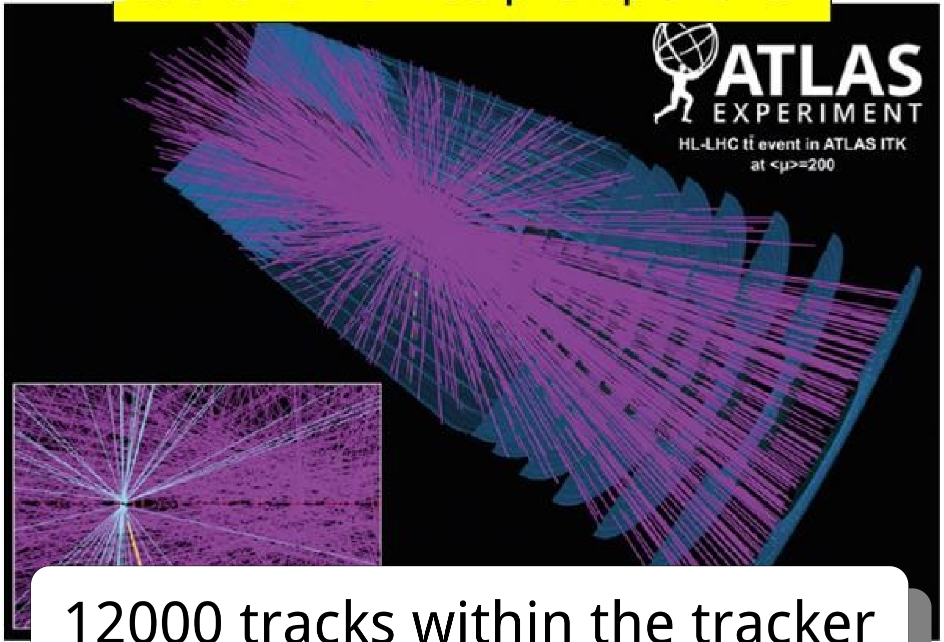
| | Run 3 | Run 4 |
|---|---|---|
| Readout rate (MHz) | 0.1 | 1 (4) |
| Recording rate (kHz) | 1.5 | 10 |

# ATLAS Phase-II Upgrade

- High-granularity to cope with pileup
  - both for readout and trigger
  - complete replacement of inner detector → ITk

Larger event size

$t\bar{t}$ event with 200 pile-up events




12000 tracks within the tracker

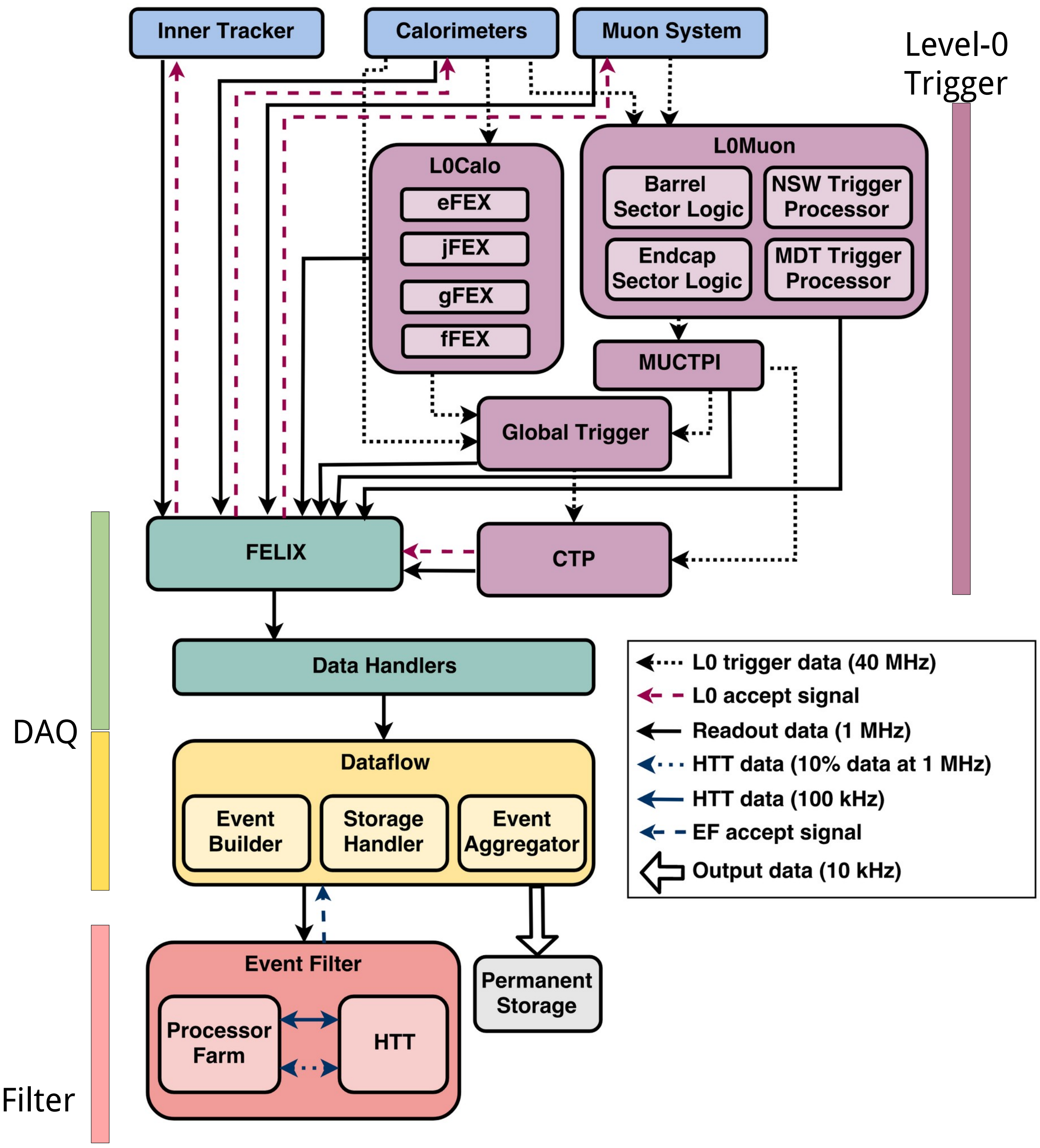
- Higher readout rate needs overhaul of detector front-end electronics
  - occasion to increase first level-trigger latency
    - *currently limited by on-detector buffer depths*
  - adopt unified readout link technology
    - *GBT/Versatile*

| | Run 3 | Run 4 |
|---|---|---|
| First-level trigger latency (μs) | 2.5 | 10 |
| Event size (MB) | 2.5 | >5 |

# Phase-II TDAQ Architecture

- Two-Level Trigger and Data Acquisition System
  - hardware-based L0 trigger system
  - software-based Event Filter, aided by dedicated tracking accelerator

- Storage-based data-flow infrastructure
  - decouple realtime domain from software processing
  - enable advanced data processing strategies

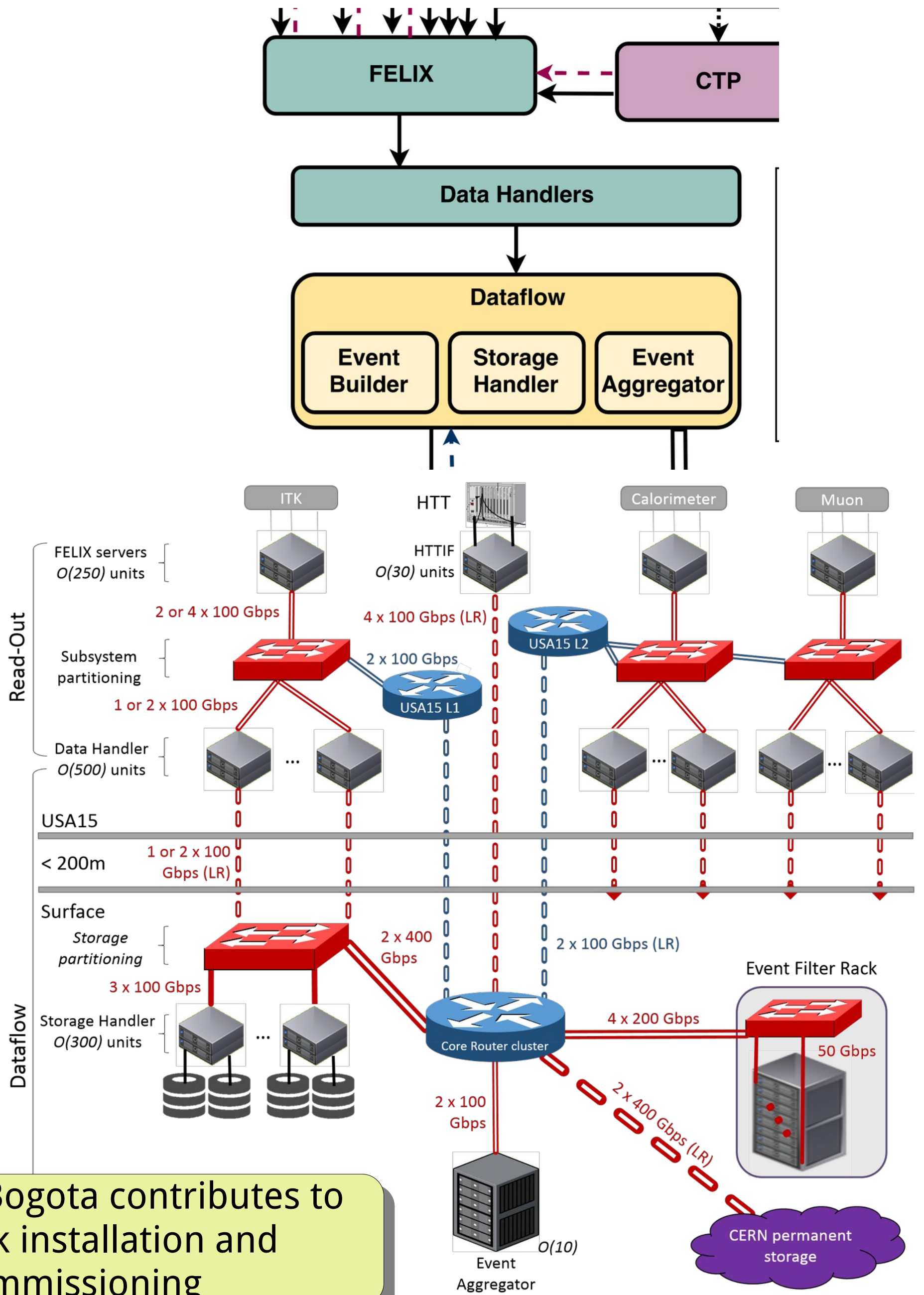


# DAQ: data transport and management

- DAQ infrastructure responsible for
  - interfacing the detector readout links to a commercial network domain
  - buffering the data and serving them to the Event Filter processors
  - discarding rejected events and formatting selected data for offline transfer

- Largely implemented with commodity off the shelf hardware

- Backbone is a multi-layered sliced network
  - baseline design based on Ethernet, do not exclude HPC technologies
  - ~2500x 100 Gbps
  - ~200x 200 Gbps
  - ~70x 400 Gbps

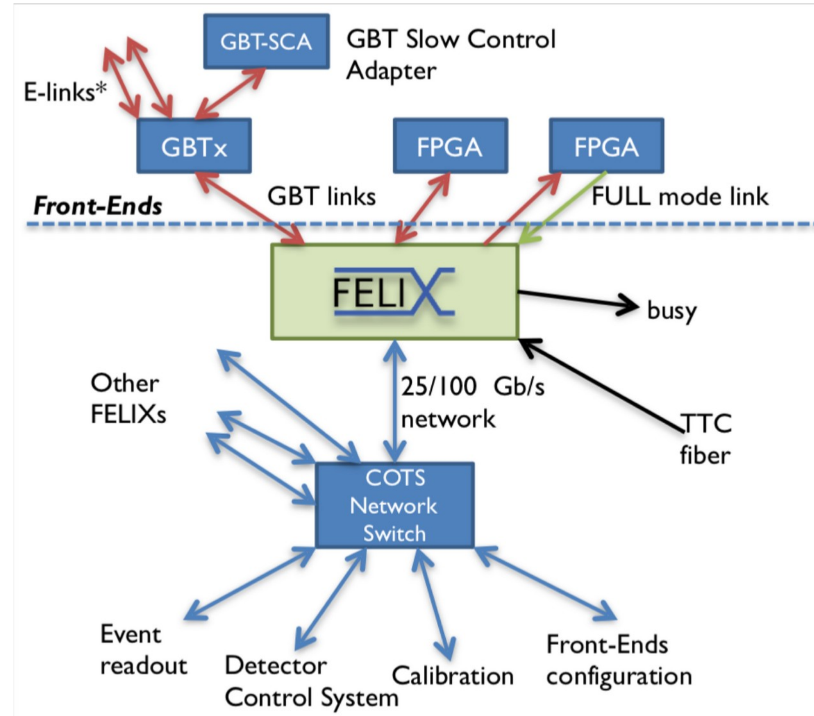



Network Design, Installation, Operation, Simulation

Colombia/Bogota contributes to network installation and commissioning

# DAQ: Detector Interface

- Detector interfacing relies on a concept being deployed for Run 3
  - extended to the whole ATLAS
- Front-end Link Exchange (FELIX) acts a heterogeneous router
  - translates between network and serial links
  - distributes timing and trigger signals
  - as detector-agnostic as possible
    - *still provision for detector specific functions*
- Implementation based on commercial servers equipped with custom FPGA-based PCIe interfaces
  - plan for 48 10Gbps links per card
  - ~550 cards serving almost 20000 links
- In Run 3 a single FELIX server will handle 40 MHz of data frame rate at 15 Gbps
  - challenging software operating close to the hardware
    - *6 core 3 GHz CPU → ~500 clock cycles per frame*
  - expect at least factor 10 increase in Phase-II



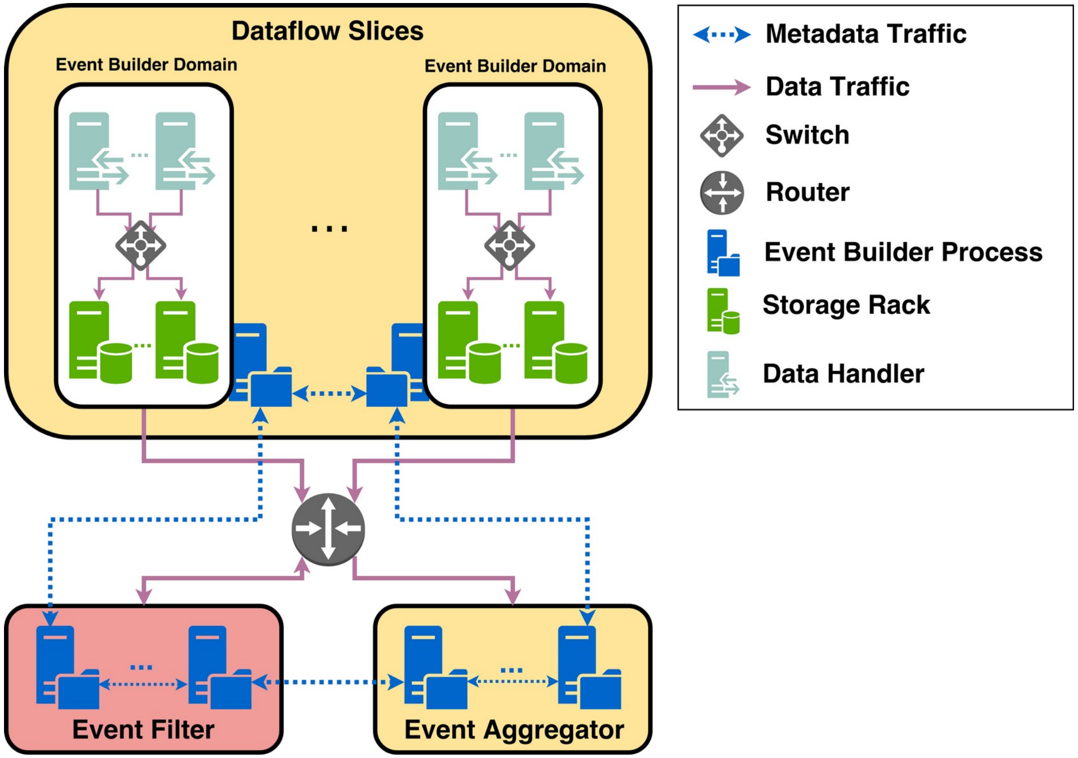

Phase-I FELIX Interface Card

High-Performance I/O Software, Linux Drivers

# DAQ: Storage & Data management



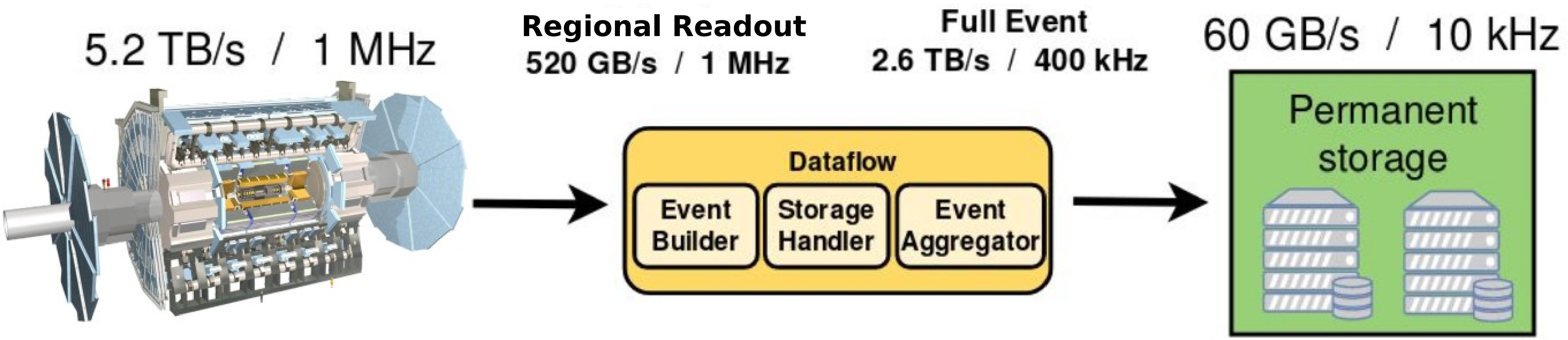
- Extend the DAQ buffering capabilities using a large storage infrastructure
  - decouple realtime domain (Level-0) and software domain (Event Filter)
  - enable delayed processing or fail-over scenarios
- Event Filter computer farm may be operated similarly to a batch system
  - quasi-realtime data stream required for online physics and detector monitoring




| Component Connection | | Traffic |
|---|---|---|
| Detector Front-ends to FELIX | | 5.2 TB/s |
| FELIX to Data Handlers | | 5.2 TB/s |
| Data Handlers to Event Builder/Storage Handler | | 5.2TB/s |
| Storage Handler to Event Filter | | 2.6 TB/s |
| Event Filter to HTTIF | Event Filter to rHTT | 175 GB/s |
| | Event Filter to gHTT | 560 GB/s |
| Event Filter to Event Aggregator and Permanent Storage | | 60 GB/s |

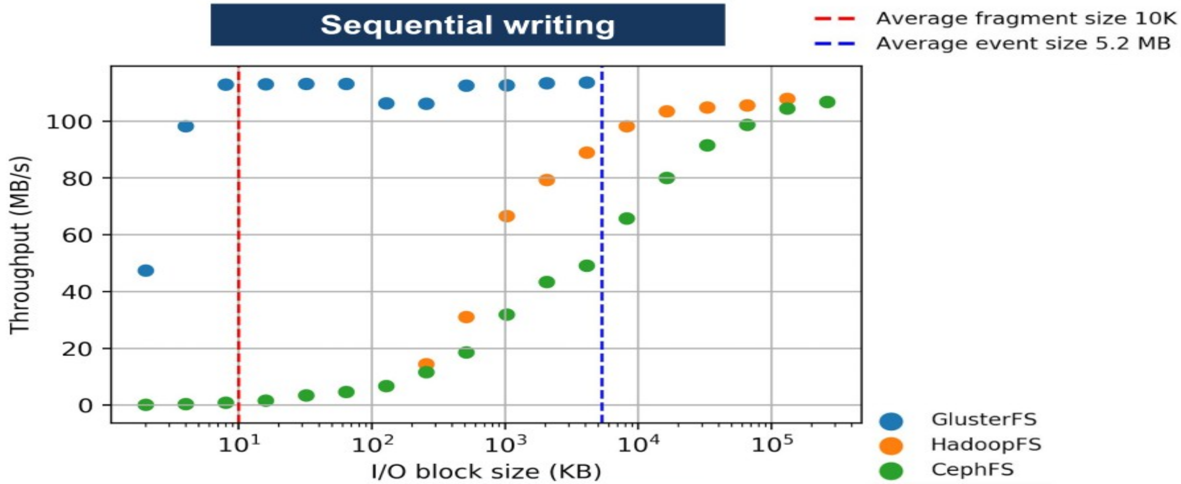
Storage, Databases, Simulation

# Storage Investigations




- Understand feasibility of commodity hardware and software for the storage infrastructure
  - hardware infrastructure: novel solid state technologies, hierarchical storage, ...
  - storage software: distributed file system, distributed hash-tables, ...
  - operation model and interfaces: trade-off between compute, networking, storage

Argentina/UBA simulations of data movements

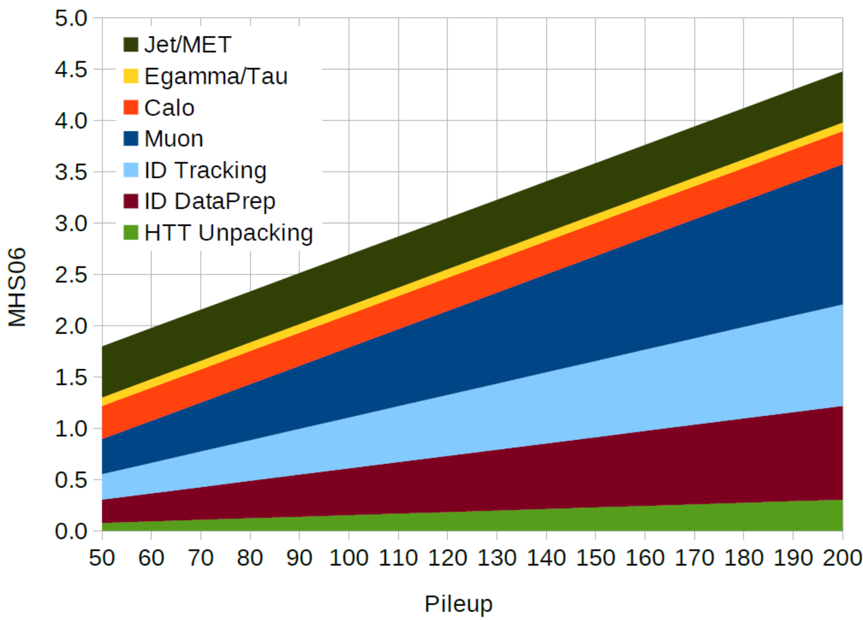


# Event Filter

- Similar to Run 3 → large computer farm
  - aided by a dedicated tracking system
  - performs the last level of selection from 1 MHz to 10 kHz
- In high pileup environment tracking is key to recover algorithms performance and maintain low thresholds
  - separation of electrons and background jets
  - calculation of global event quantities like $E_T^{miss}$
  - jet energy resolution
- Event Filter baseline implementation is based on CPUs
  - in parallel investigations of accelerators (GPGPU & FPGA) and associated dedicated algorithms

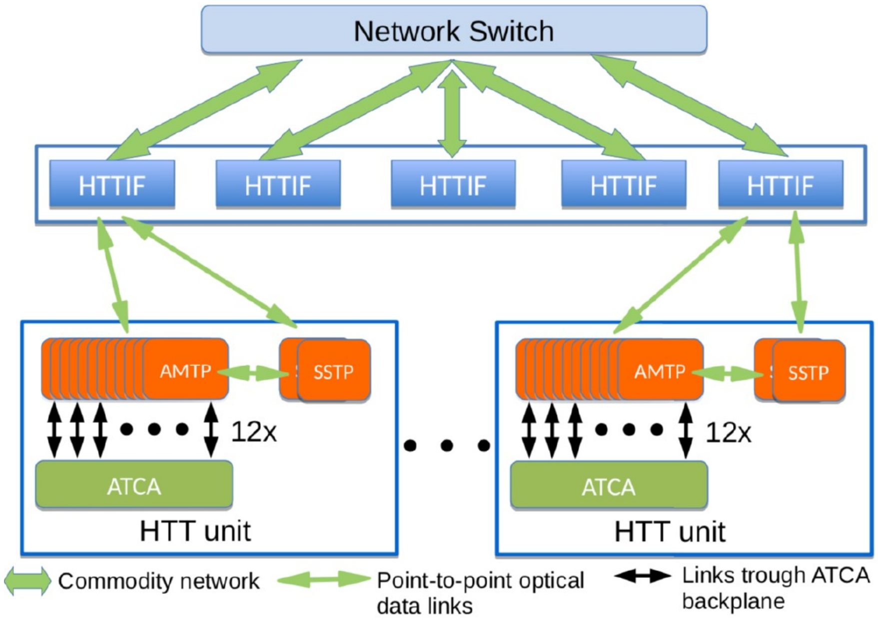
High-Performance Software, AI/ML, GPU/FPGA Programming

Brazil AI-based e/γ identification

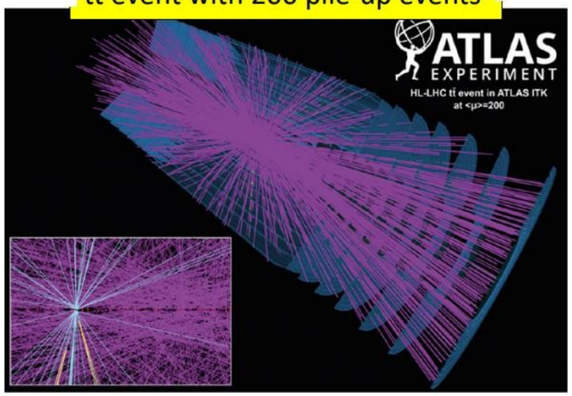


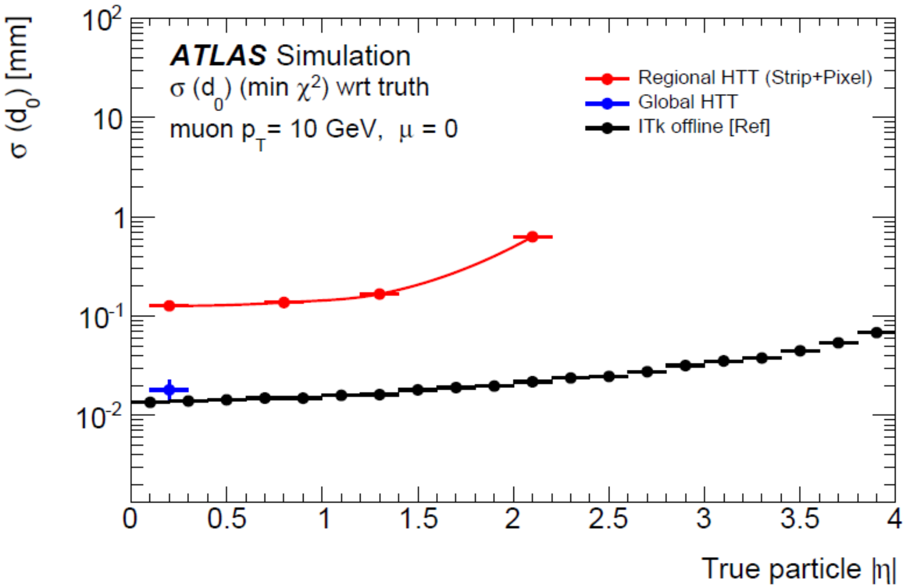
# Hardware Track Trigger

- ITk tracking software → 10 times larger computer farm would be required
  - based on current tracking software
  - ongoing software optimisations potential to significantly reduce this estimate
- HTT (Hardware Track Trigger) massively parallel device
  - custom electronics using Associative Memories (AM ASICs) for pattern recognition and FPGAs for fitting
  - driven by the Event Filter requests




tt̄ event with 200 pile-up events



# Track Trigger: COTS Alternatives

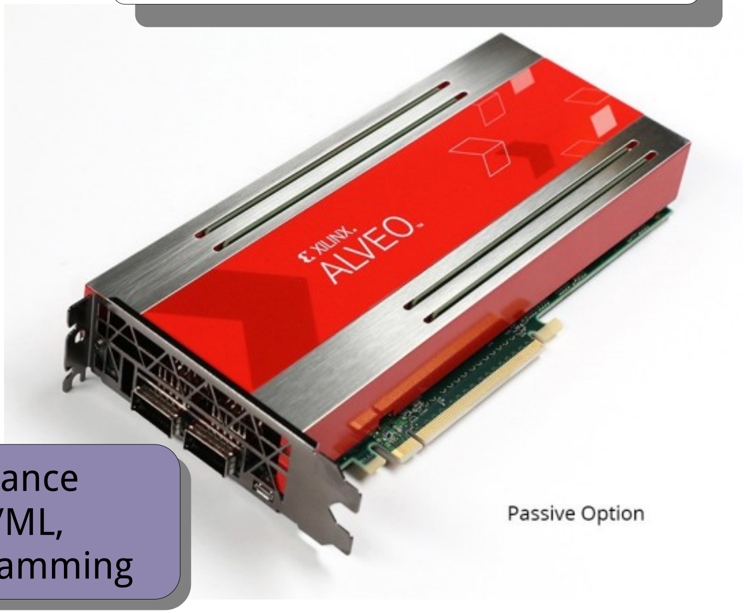
- Active investigations in COTS alternative to tracking in custom hardware

- Based on
  - CPU-based software implementation
  - coprocessor and specialised algorithms

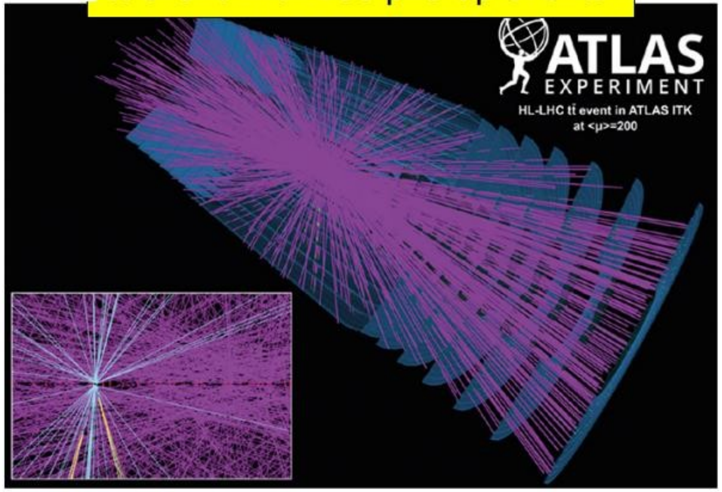
- Studying both algorithmic and implementation improvements

GPGPU


FPGA




$t\bar{t}$ event with 200 pile-up events

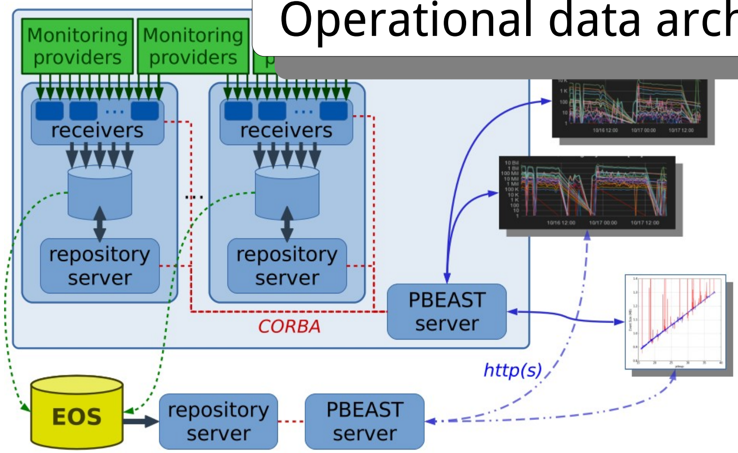



High-Performance Software, AI/ML, GPU/FPGA Programming

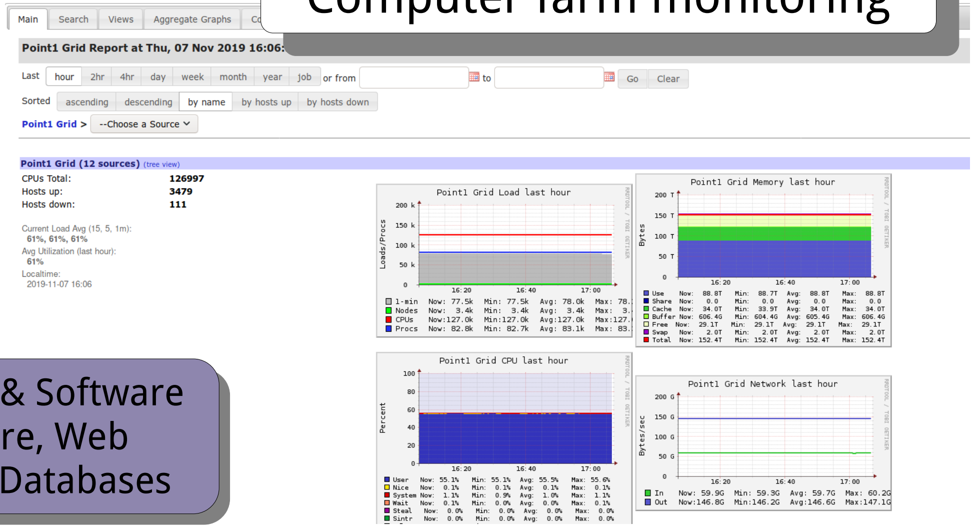
# Control and Monitoring Infrastructure

- Managing, monitoring and configuring 50000+ applications
  - on a heterogeneous cluster
  - on a mission-critical duty
- Storage system enables staged filtering
  - mechanism to control applications in different domains (realtime) and with different lifetimes
- Orchestration systems current answer in cloud environment
  - applicability to data-acquisition to be evaluated
- More in general, require to scale all aspects of monitoring
  - operational and hardware
  - detector
  - trigger and physics


Operational data archiving

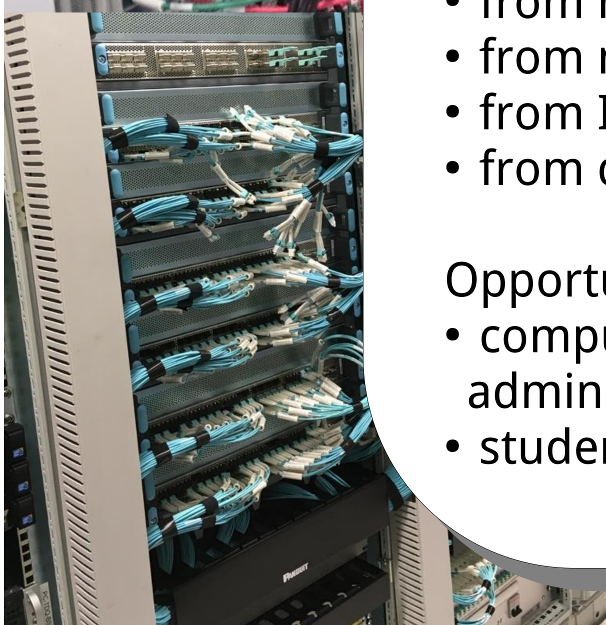


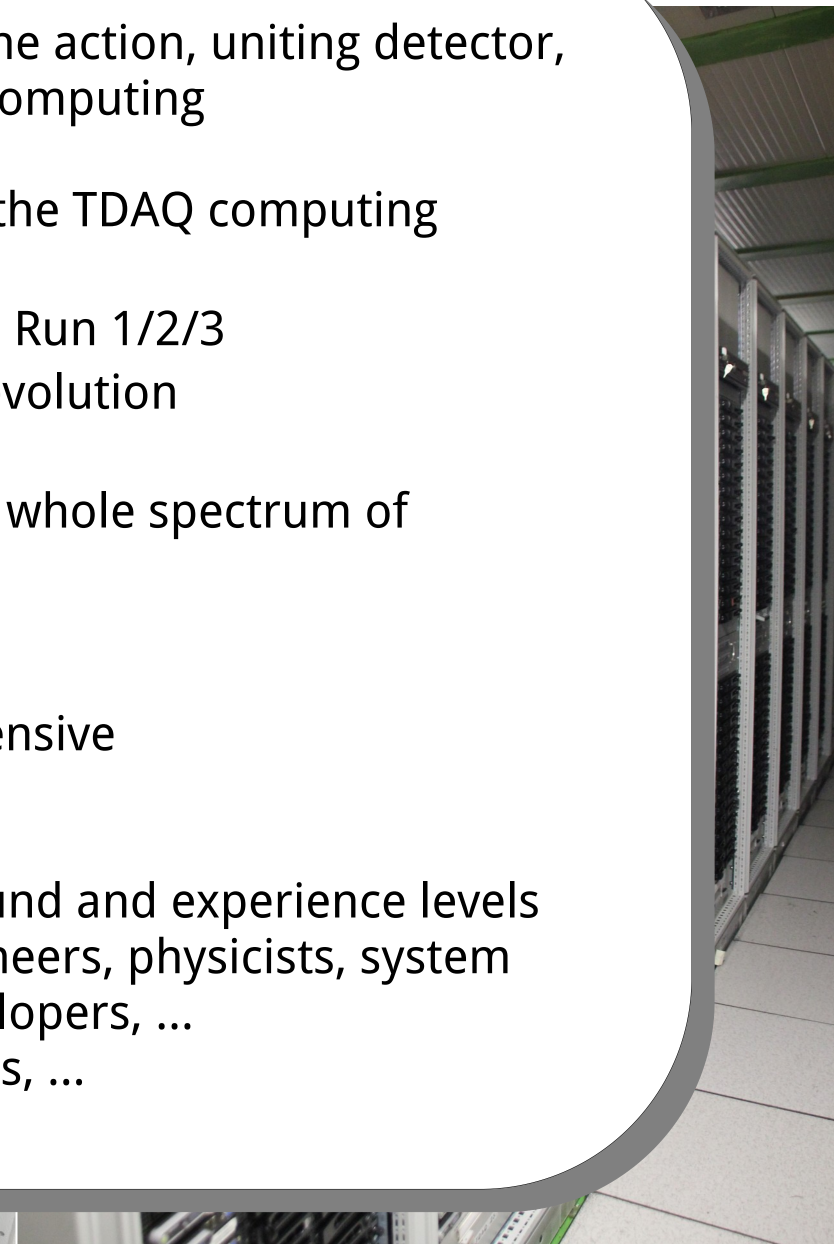

Computer farm monitoring




HPC Hardware & Software Infrastructure, Web development, Databases

# Outlook





ATLAS TDAQ sits at the centre of the action, uniting detector, physics, off-line computing

Phase-II upgrade major scale-up of the TDAQ computing infrastructure

- apply lessons and experience from Run 1/2/3
- take advantage of the technology evolution

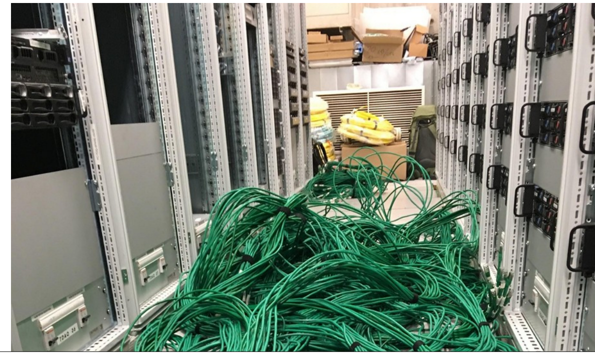
In general ATLAS TDAQ team covers whole spectrum of computing aspects

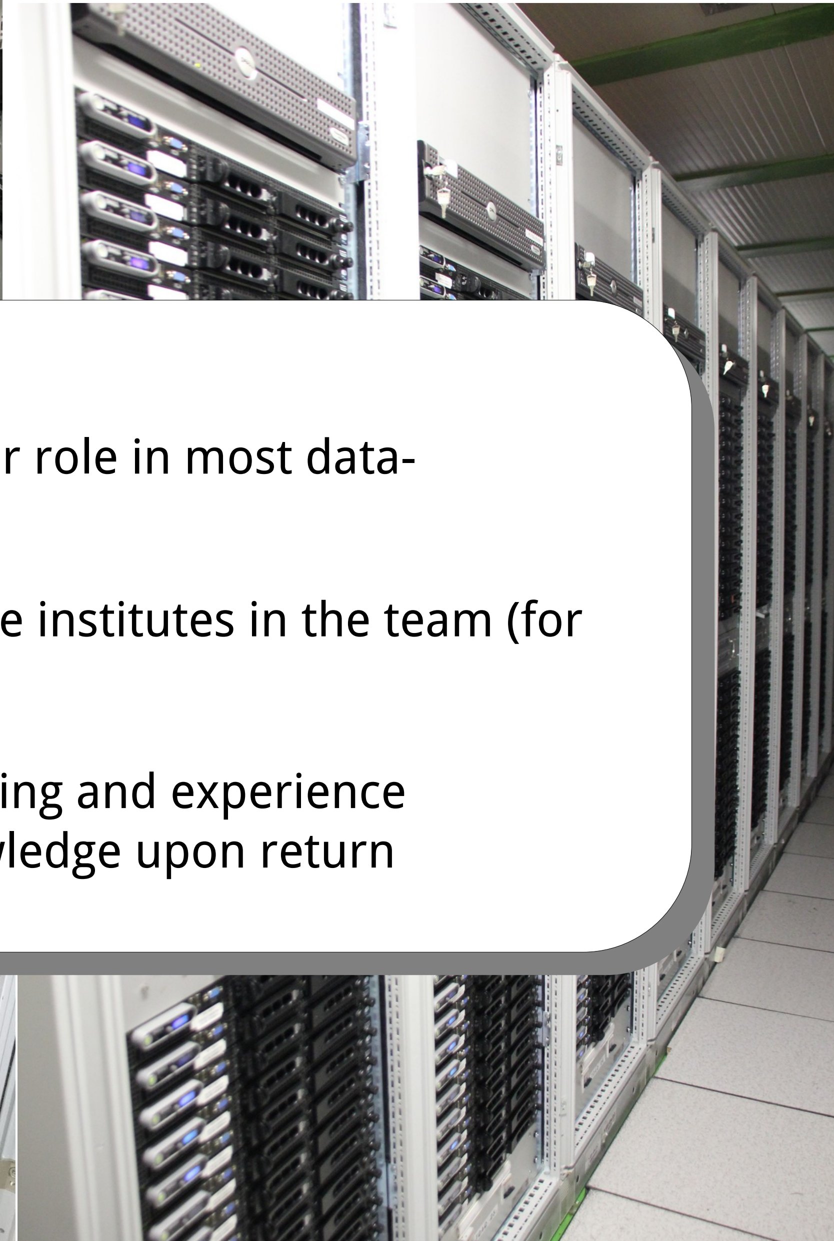
- from hardware to software
- from networking to storage
- from I/O intensive to compute intensive
- from design to operation

Opportunities for different background and experience levels

- computer scientists, network engineers, physicists, system administrators, software/web developers, ...
- students, technicians, professionals, ...

# CERN ATLAS TDAQ Team

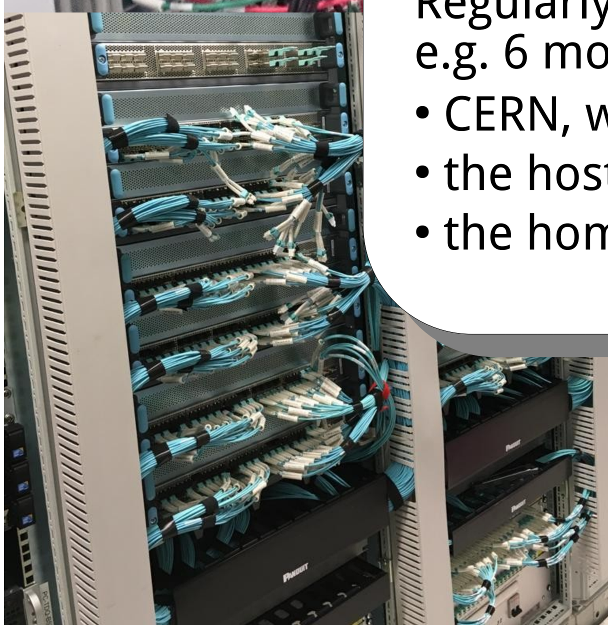




CERN ATLAS TDAQ Team plays a major role in most data-acquisition domains

Regularly host colleagues from remote institutes in the team (for e.g. 6 months). Benefits for

- CERN, we can profit from the work
- the hosted colleagues through learning and experience
- the home institutes, spreading knowledge upon return

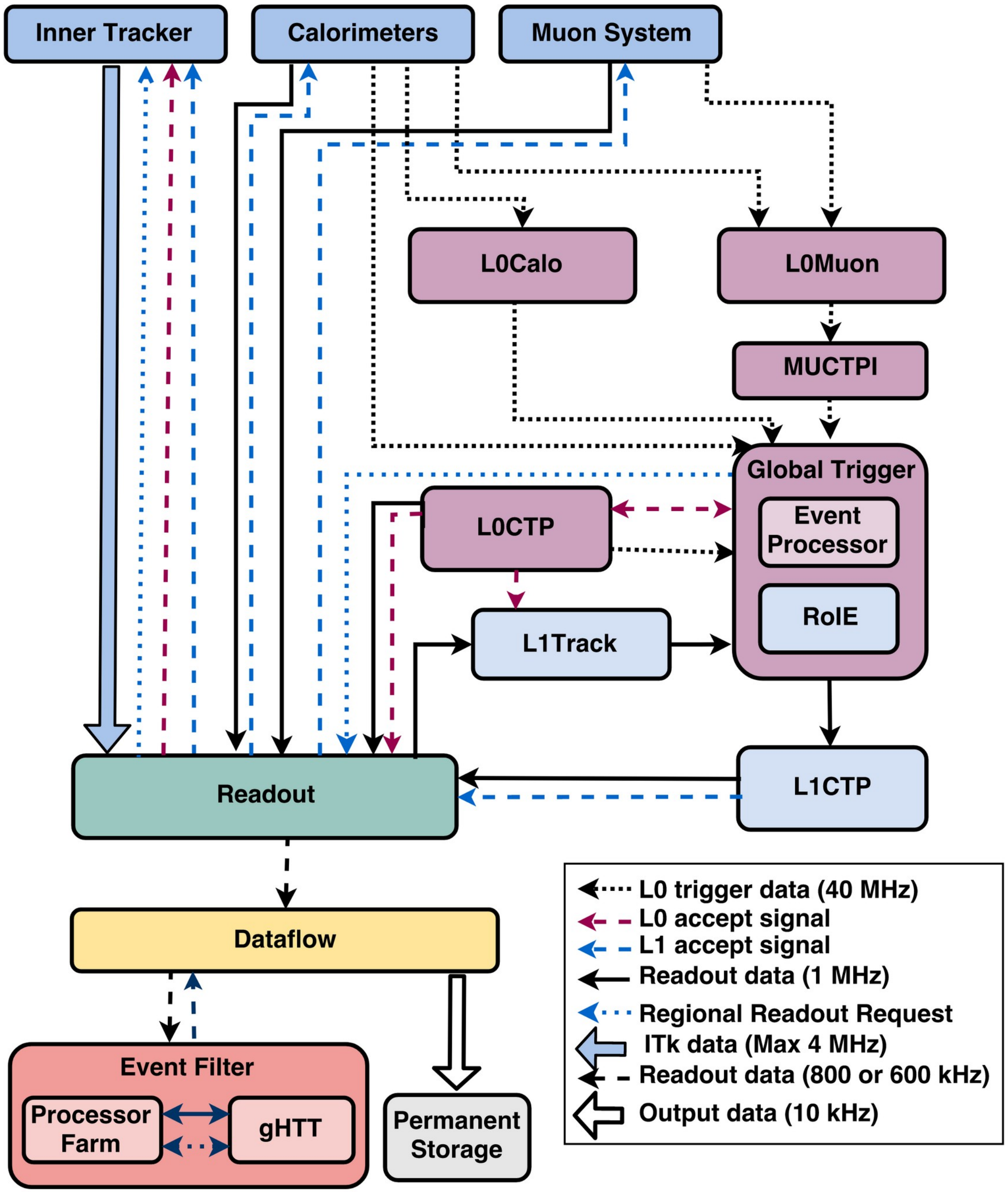


# Bonus

# Phase-II TDAQ Evolution Architecture



- Evolution path to a two-level hardware trigger included in the design
  - L0 – 4 MHz
  - L1 – 1 MHz
  - Event Filter – 10 kHz
- Possible transition from baseline to evolution driven by physics requirements
  - hadronic trigger rates
  - occupancy of inner layers of ITk
- Avoid the baseline TDAQ implementation restricting the trigger menu at the ultimate HL-LHC operating conditions
- Level-1 Trigger combines L0 objects with track information from a dedicated subsystem to discriminate against pileup in the calorimeter



# Outlook