The ATLAS Data Acquisition System in LHC Run 2

Eukeni Pozo Astigarraga (CERN)

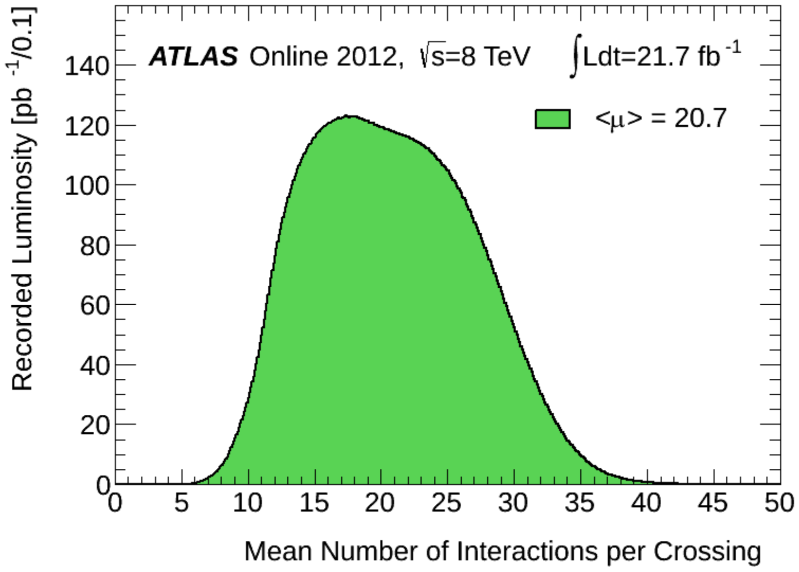
on behalf of the ATLAS collaboration

eukeni.pozo@cern.ch

Outline

- From LHC Run 1 to Run 2
- High level overview of trigger systems
- The ATLAS Data Acquisition System:
  - Data Flow components
  - Data acquisition network
  - Control and Monitoring
- Conclusions
- Future work

From LHC Run 1 to Run 2

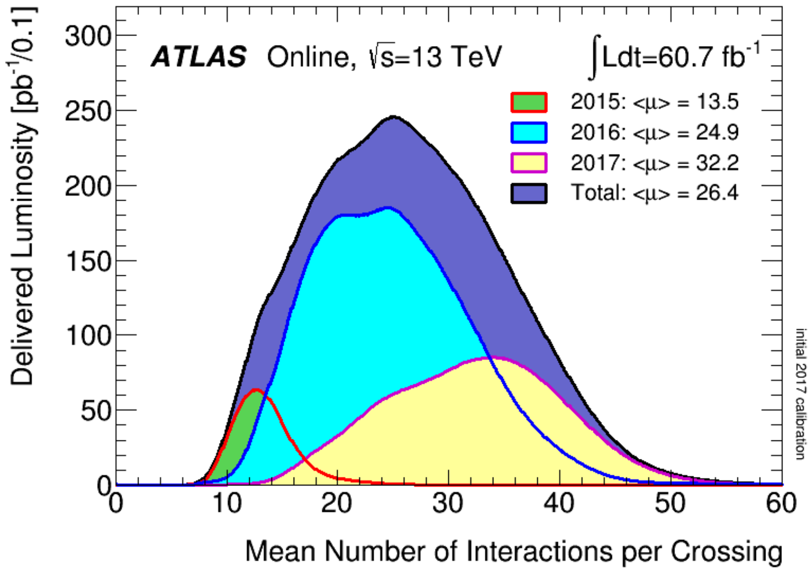


Run 1 conditions:

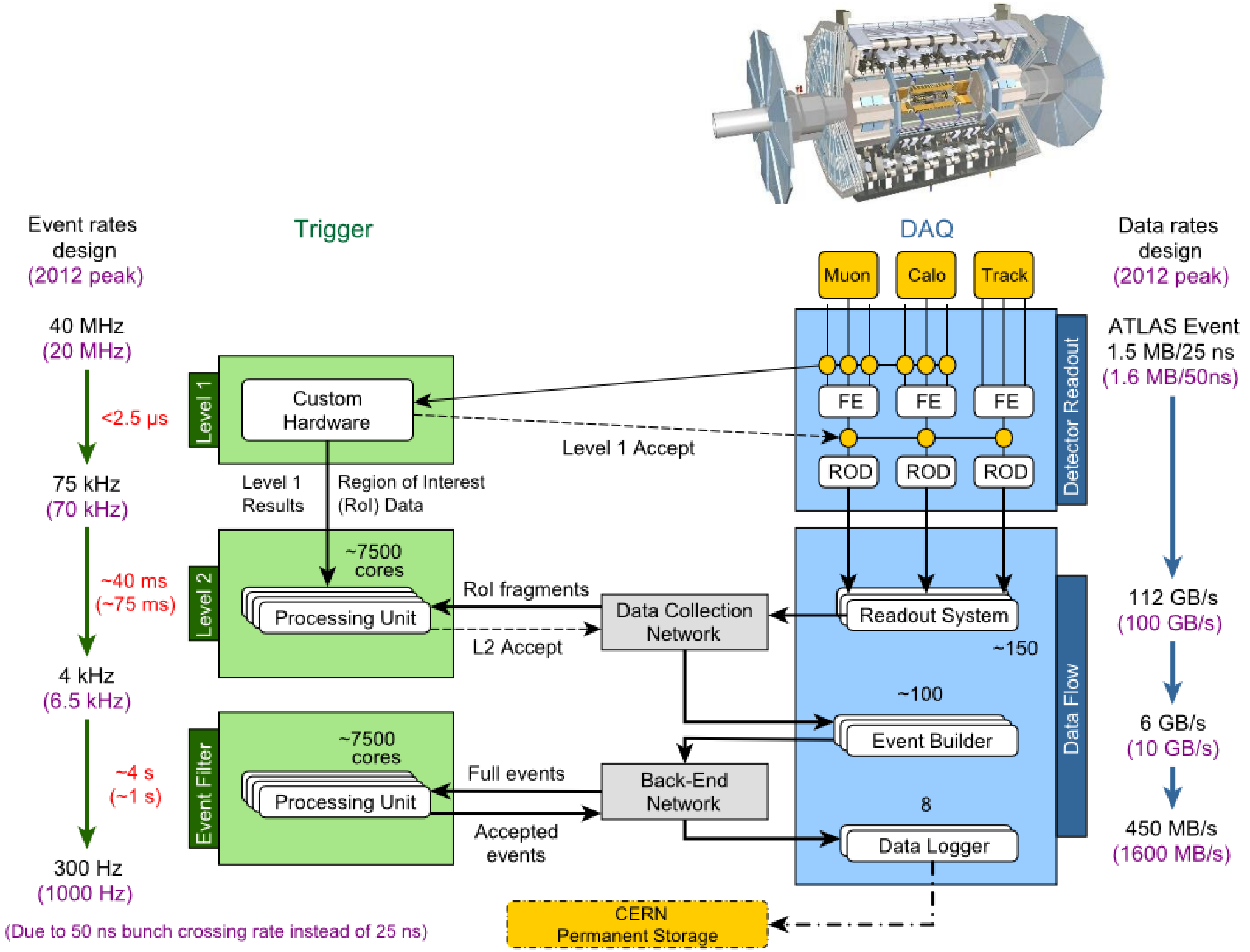
- 8 TeV centre-of-mass energy
- Peak Lumi $= 7 \times 10^{33} cm^{-2} s^{-1}$
- Average pileup = 20.7 (peak = 40)
- 70 kHz L1 trigger rate (peak 2012)

New conditions in Run 2:

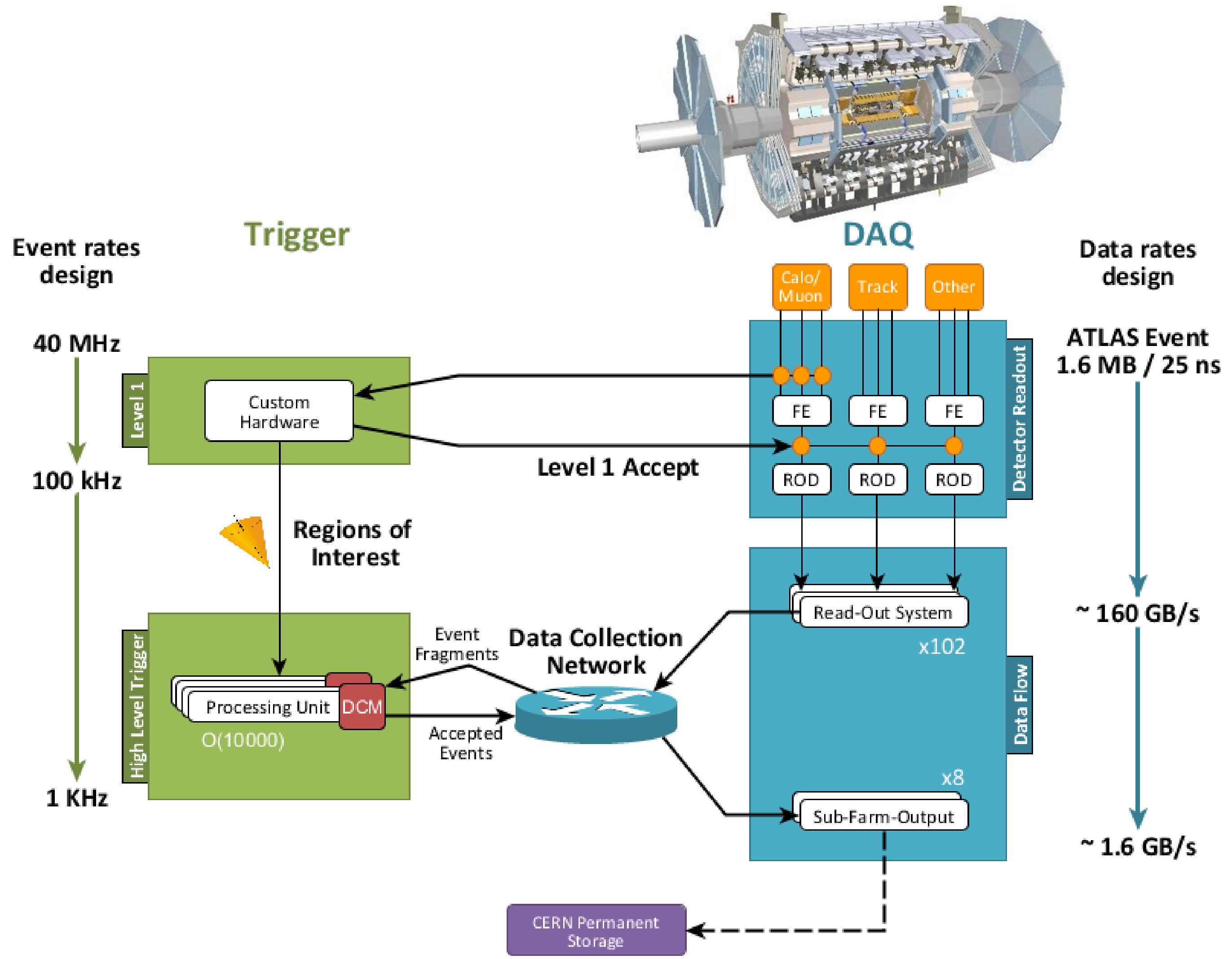
- 13 TeV centre-of-mass energy
- Peak Lumi $= 1.74 \times 10^{34} cm^{-2} s^{-1}$
- Average pileup = 32.2 in 2017 (peak = 60)
- 100 kHz L1 trigger rate (97 kHz peak 2016)



The ATLAS Trigger and Data Acquisition System in Run 1

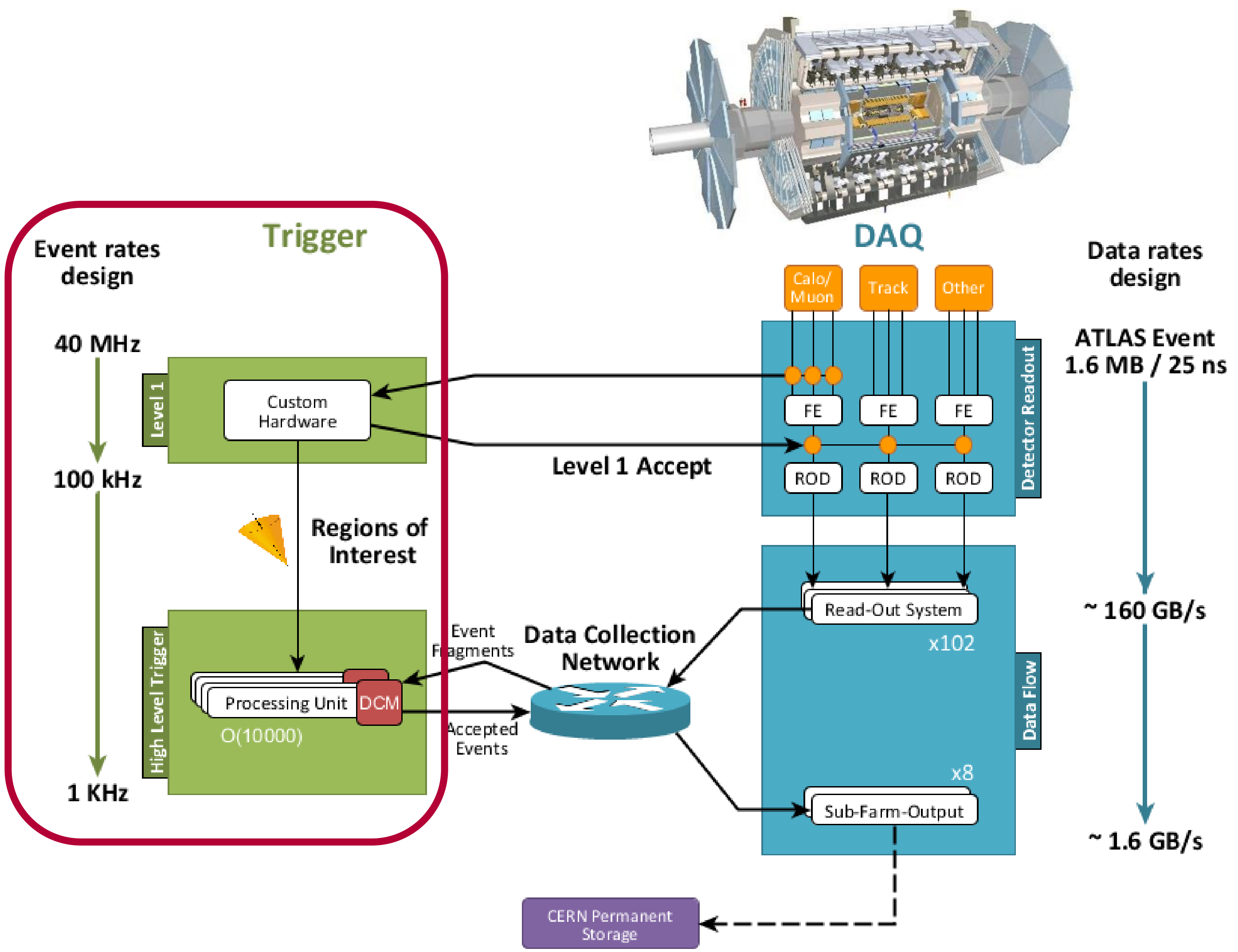


The ATLAS Trigger and Data Acquisition System in Run 2



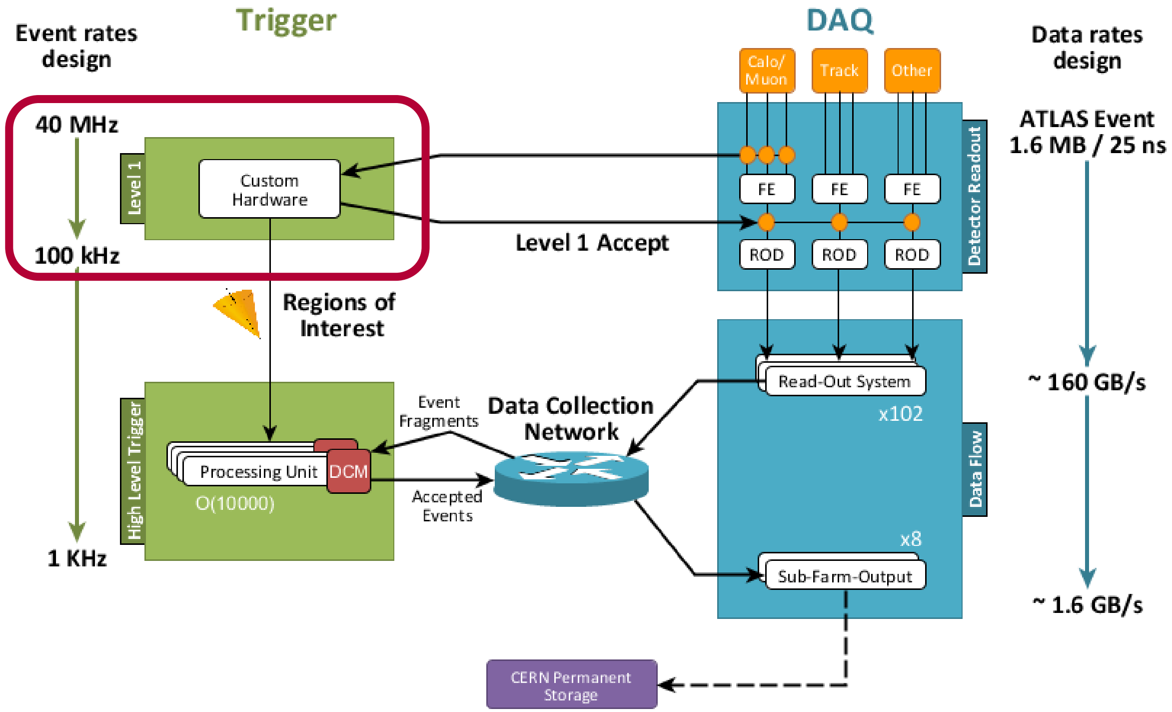
The ATLAS Trigger

A high level overview



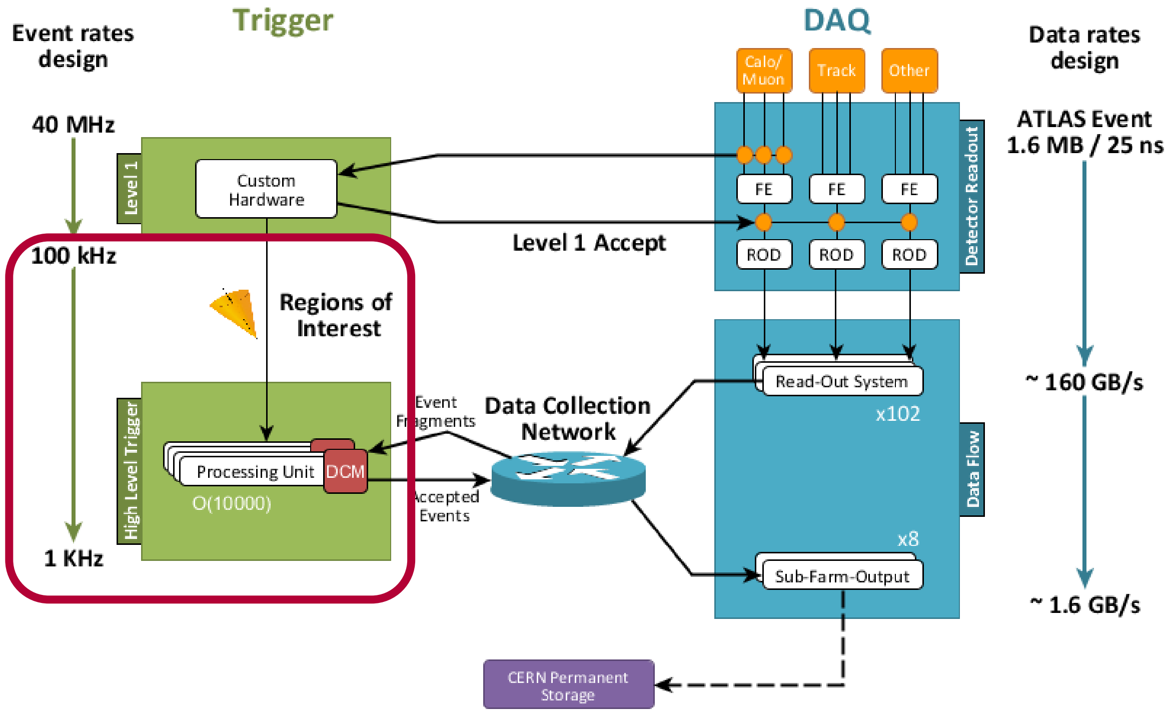
Level 1 Trigger

- Rejects or accepts an event at 40 MHz
- Reduces the event rate down to 100 kHz (peak)
- Deterministic decision time: 2.5 μs
- Implemented in custom electronics:
  - ASICs and FPGAs

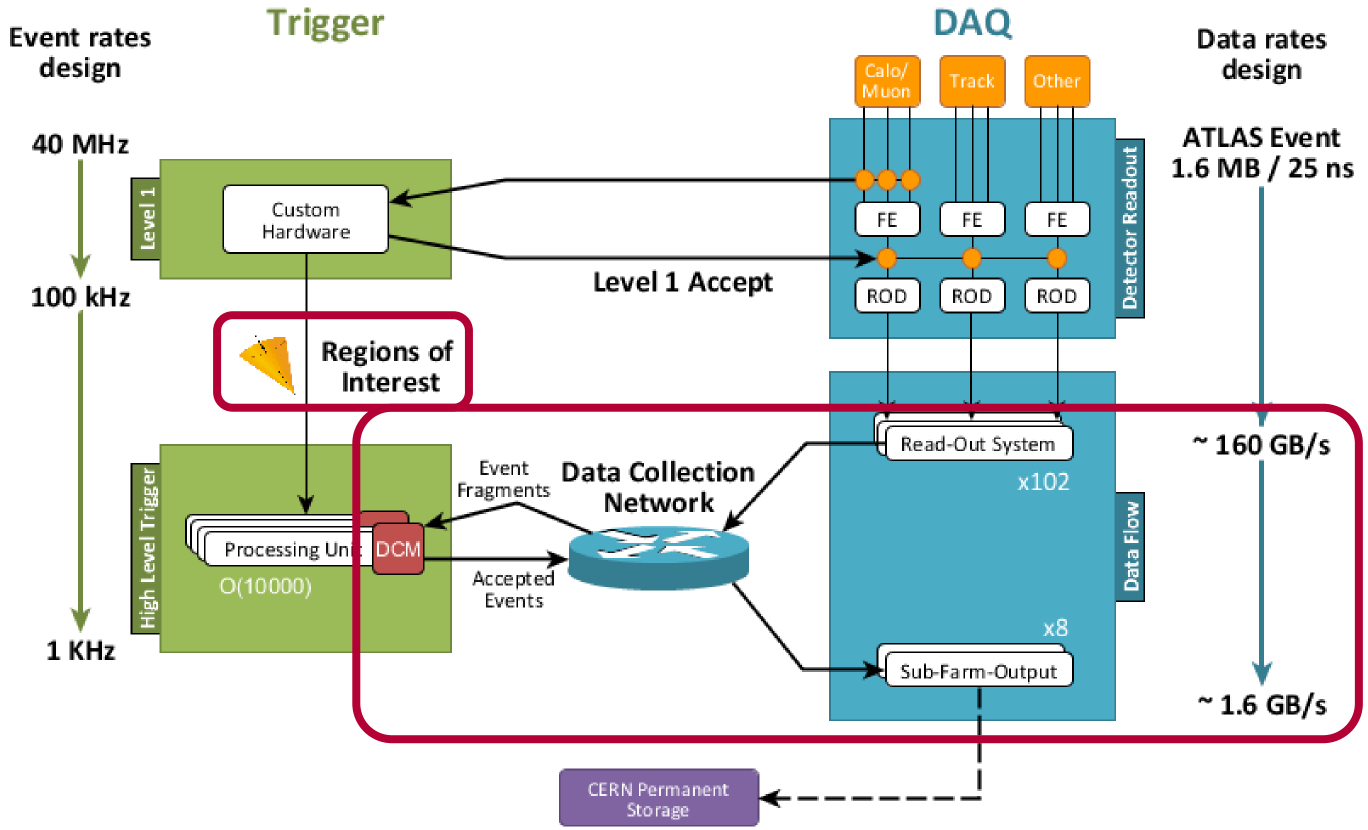


High Level Trigger

- Reduces the event rate from 100 kHz (peak) to few kHz
- Incremental data collection and filtering
  - Average decision time: $\sim 300\ ms$ at $<\mu> = 30$
- Server farm:
  - ~2000 nodes, ~40000 cores

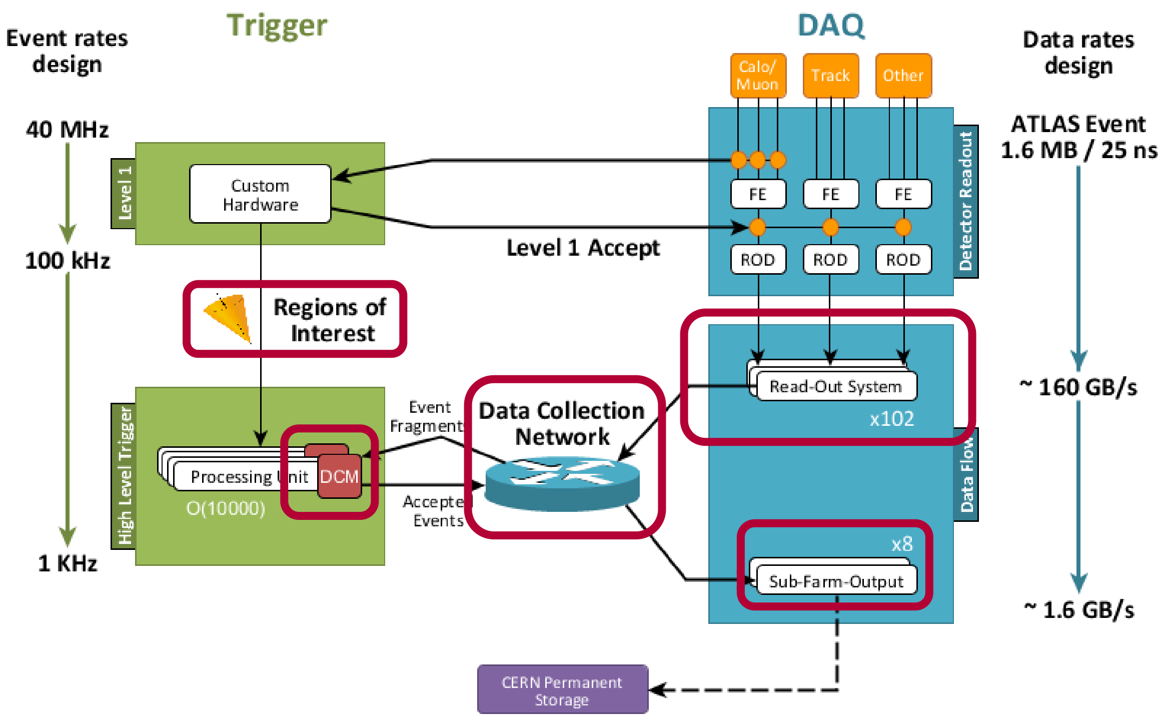


The ATLAS Data Acquisition System in Run 2



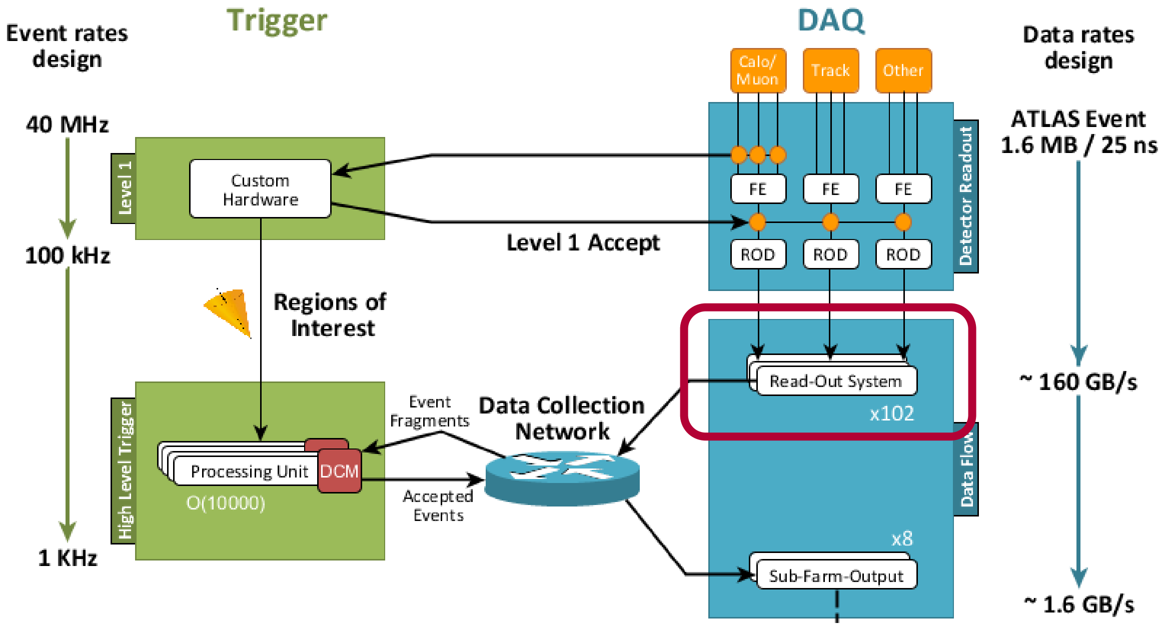
DAQ system responsibilities

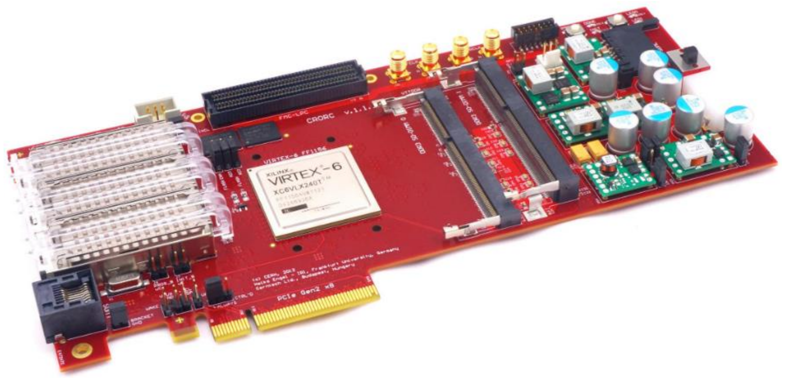
- Read out and buffer 160 GB/s from custom electronic devices (Read Out Drivers)
- Transport necessary event data (>30 GB/s) to the HLT farm for event filtering
- Provide temporary storage for selected events before copying to permanent storage
- Orchestrate all the previous tasks
- Monitor the system and recover from bad states and error conditions



The Read-Out System

- Receives and buffers event fragments from detector specific electronic boards (RODs) and sends them to the HLT farm upon request
- New read-out card: RobinNP
  - Based on ALICE C-RORC card
    - C-RORC ≡ Common Readout Receiver Card
    - Custom ATLAS FW
  - 12 Readout Links
  - 8 GB of RAM
- 1900 Read-Out Links connected to 102 ROS PCs
  - 2 RobinNPs / PC
  - Data management done by CPU of host PC
    - Contrary to Run1 card where processing was done on-board
  - Data collection network connectivity: 4 x 10 GbE
    - For redundancy

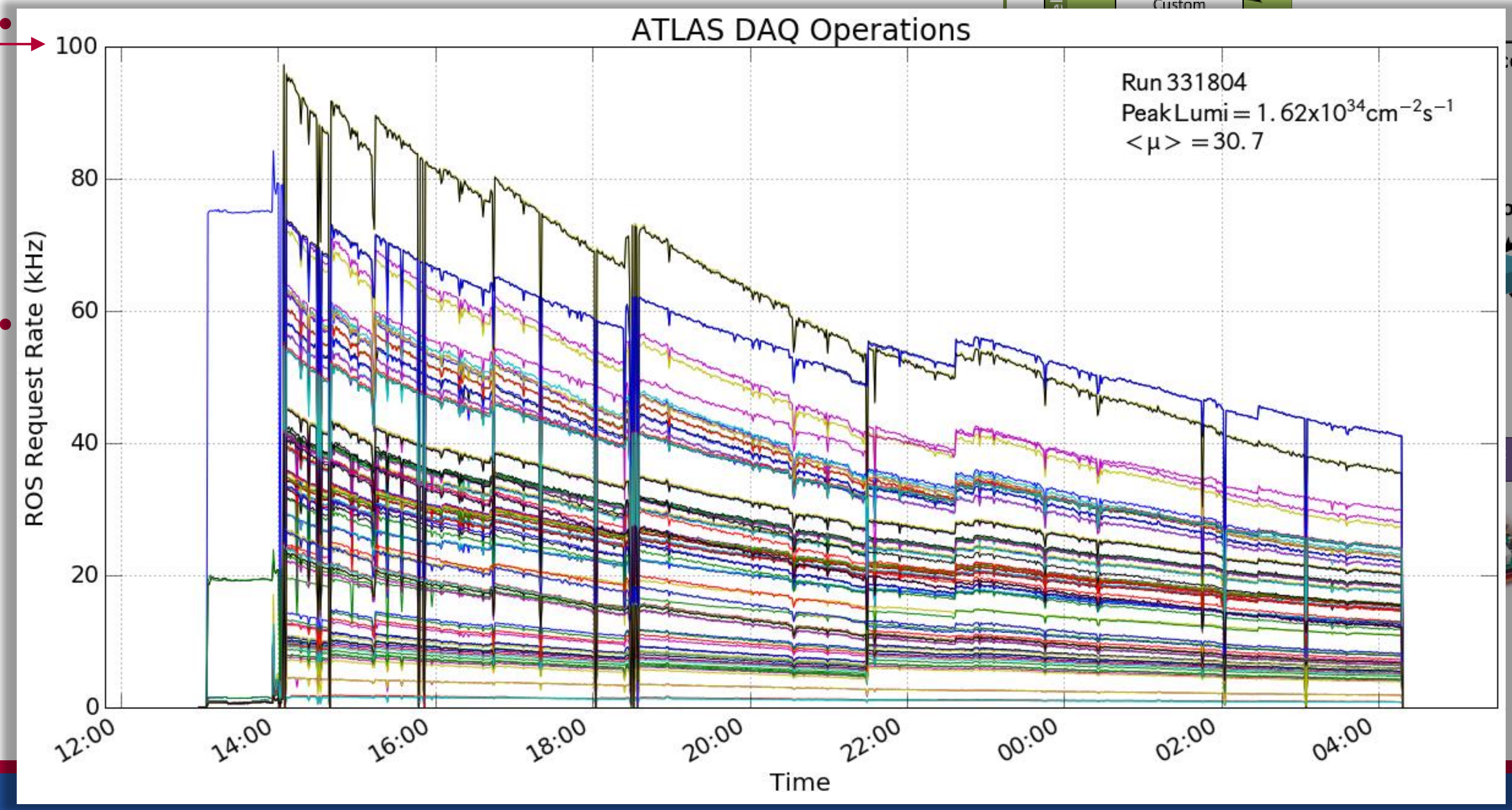


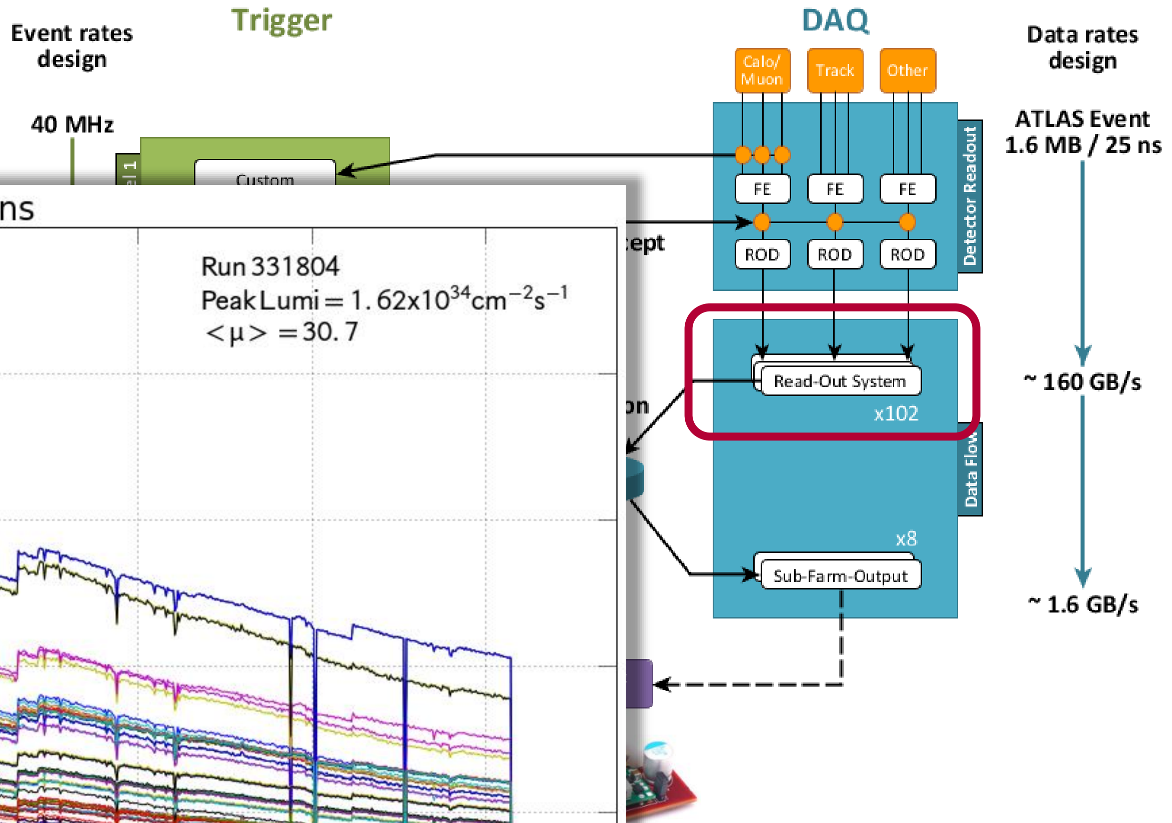


The Read-Out System

- Receives event fragments from detector specific electronic boards (RODs) and sends them to the HLT farm upon request

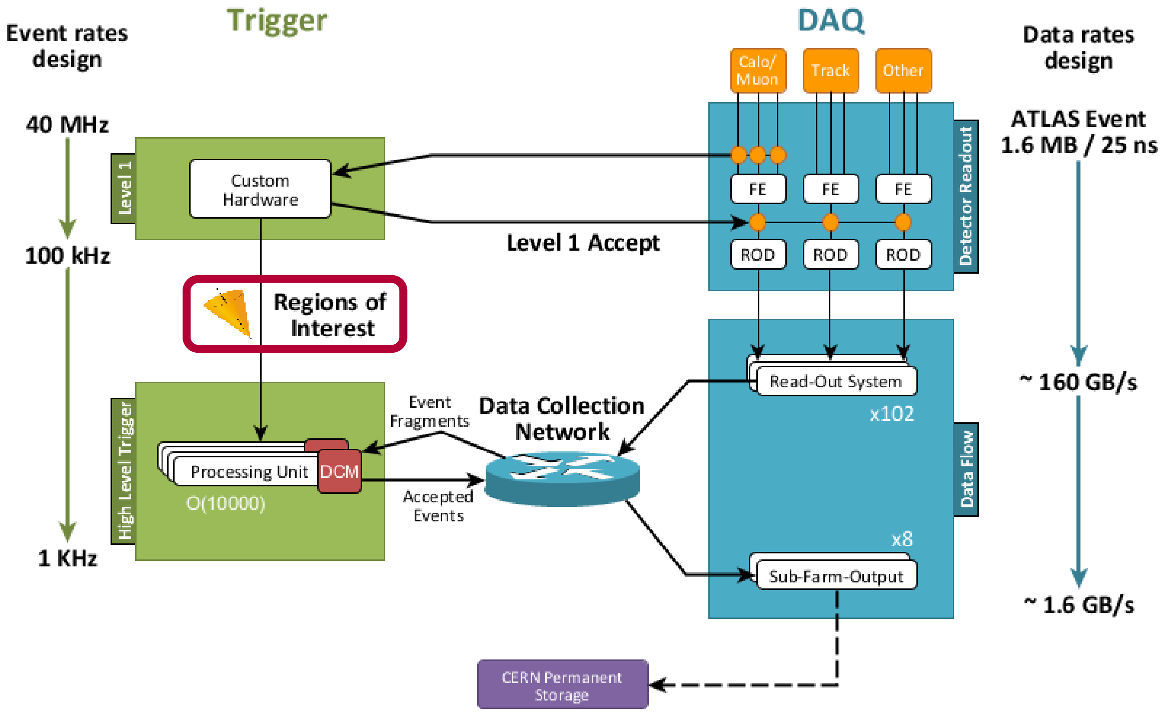
100 kHz





Region of Interest Builder and Supervisor

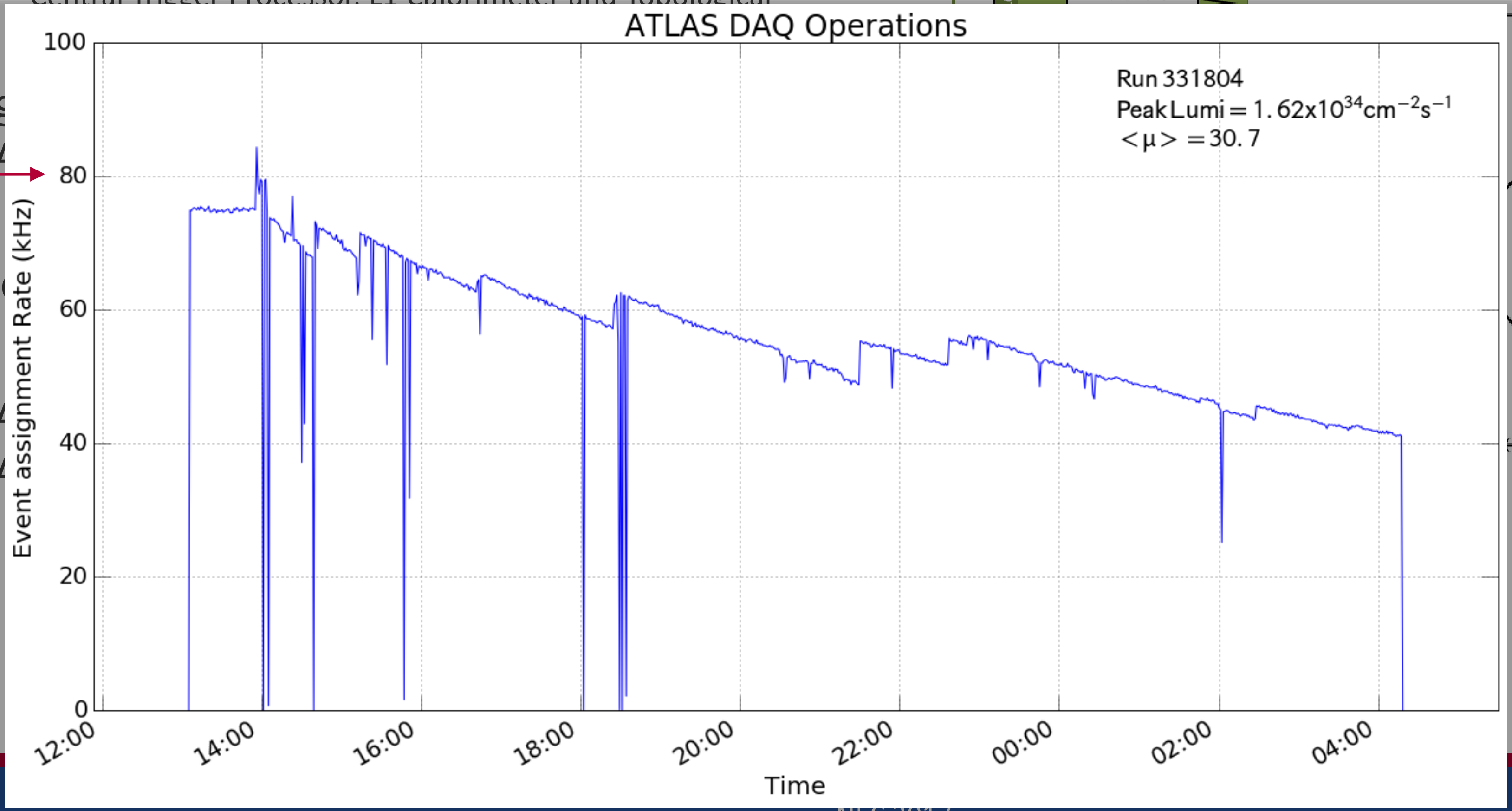
- Region of Interest Builder (RoIB)
  - Assembles Regions of Interest (RoI) from fragments produced by L1 sources:
    - Central Trigger Processor, L1 Calorimeter and Topological Triggers and L1 Muon
- High Level Trigger SuperVisor (HLTSV)
  - Assigns events to HLT nodes
  - Clears events from the ROS buffers
  - Re-assigns events that time out during processing
- In Run 1 the RoIB was implemented in custom hardware in VME crate
- In Run 2 the RoIB and the HLTSV run on a commodity server
  - A ROS PC with a RobinNP card
  - Achieve over 100 kHz event assignment rate

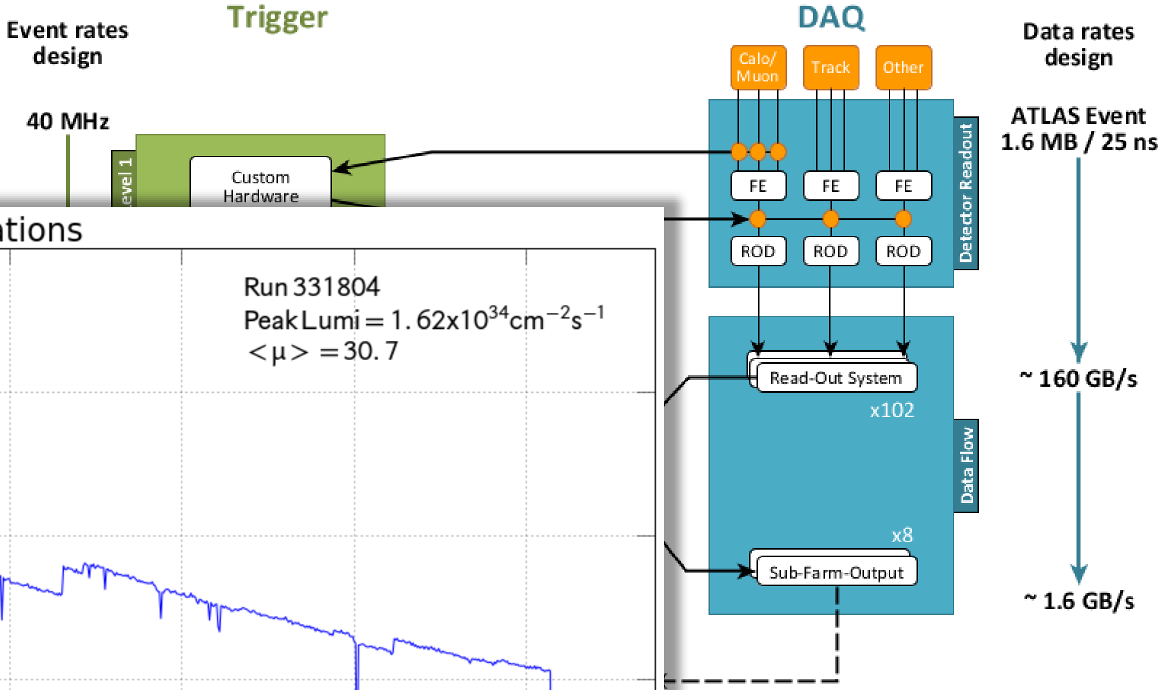


Region of Interest Builder and Supervisor

- Region of Interest Builder (RoIB)
  - Assembles Regions of Interest (RoI) from fragments produced by L1 sources:
    - Central Trigger Processor, L1 Calorimeter and Topological
- Hig
  - A
- In
har
- In
  - A
  - A

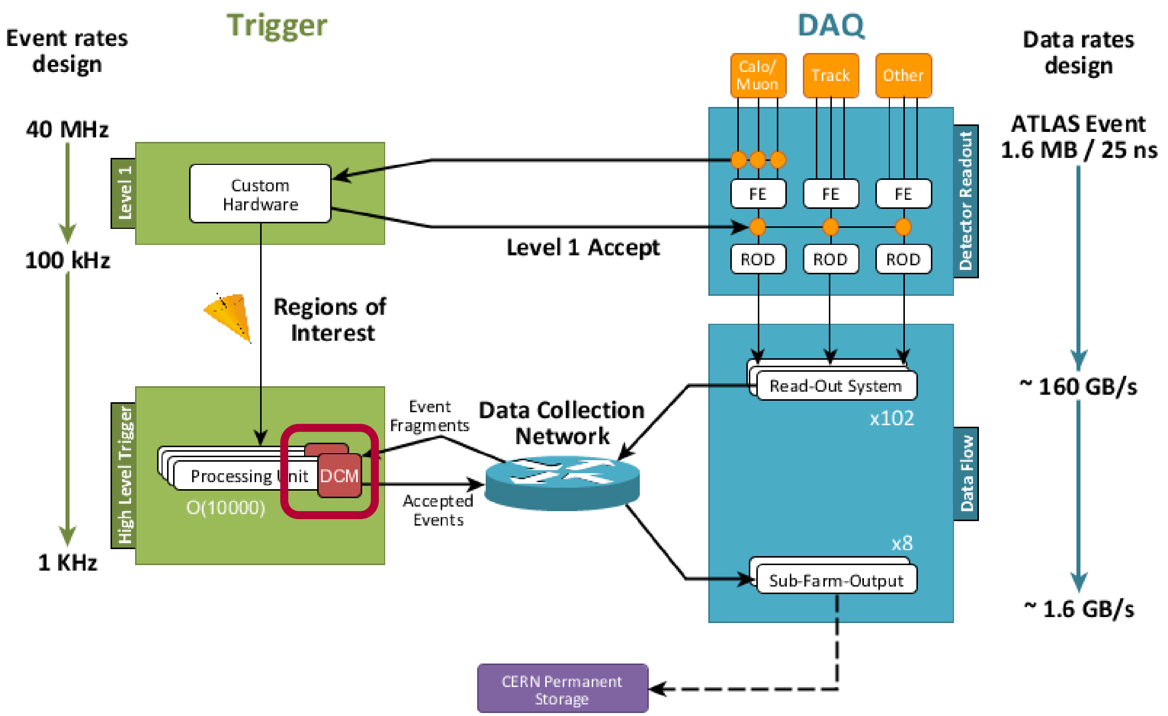
80 kHz





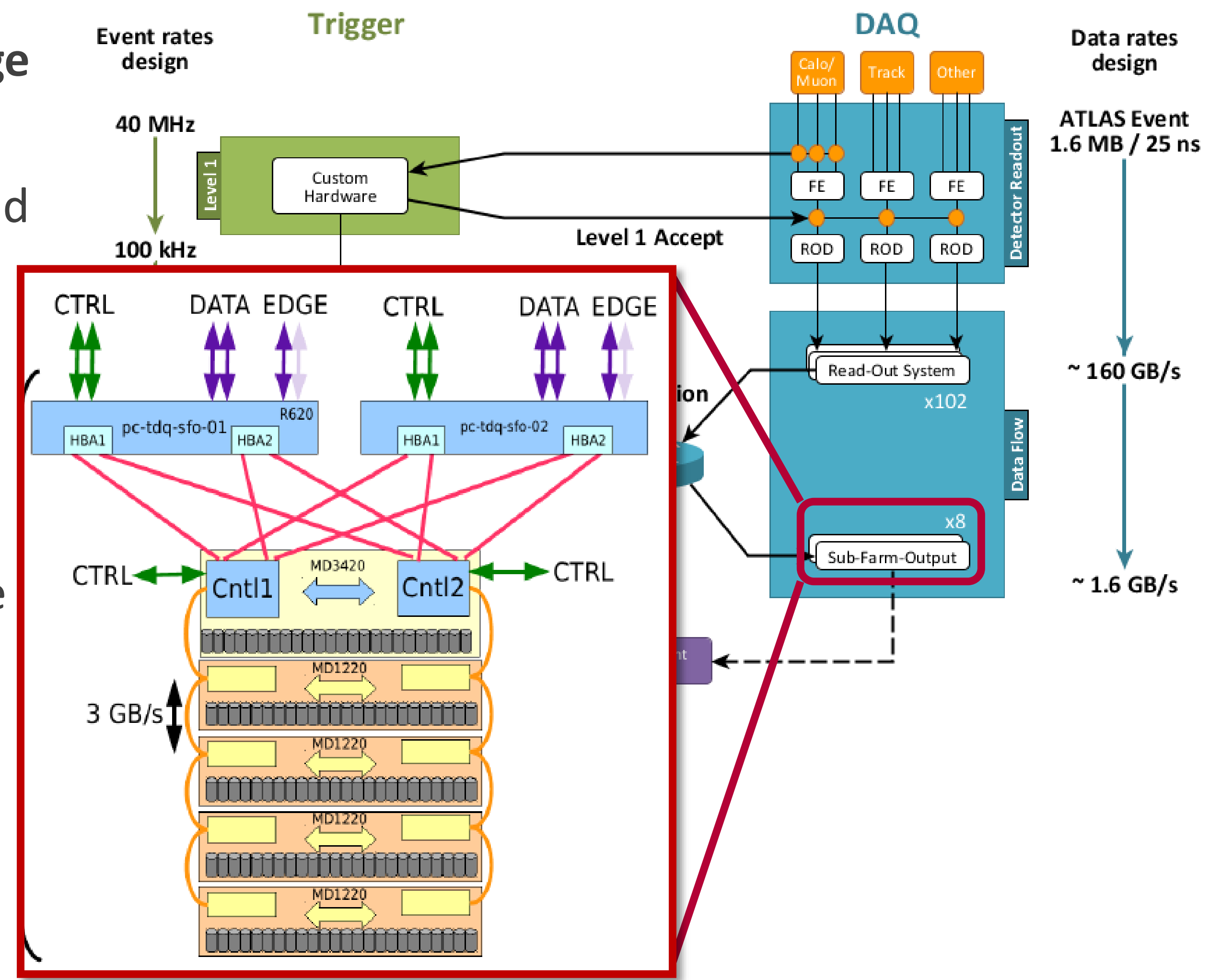
Event Building: Data Collection Manager

- Requests the event fragments from the ROS on behalf of the Processing Units
  - Single-threaded instance based on *Boost ASIO*
  - Shared memory communication with Processing Units based on *Boost Interprocess*
- A credit based traffic shaping mechanism used to avoid instantaneous network saturation
- In addition, it supports:
  - Duplication of accepted events going to different output streams (e.g. for calibration purposes)
  - Data compression before event logging



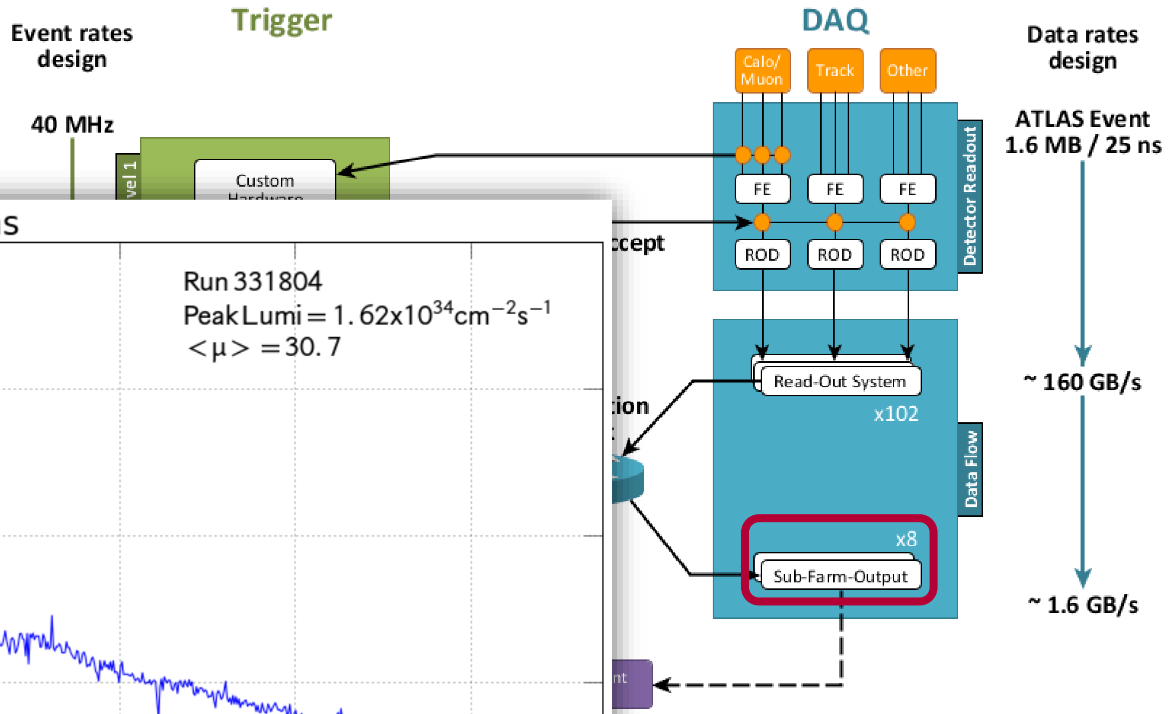
Data logging: The Sub-Farm Output

- The SFOs provide 48h of **temporary storage**
- Direct Attached Storage unit with multiple redundant data paths for fault tolerance and resilience
  - Maximum I/O: 6.5 GB/s
  - Average I/O: ~1.6 GB/s
- Background jobs copy the files to permanent storage, deleting them on the local disk only when they are safely on tape

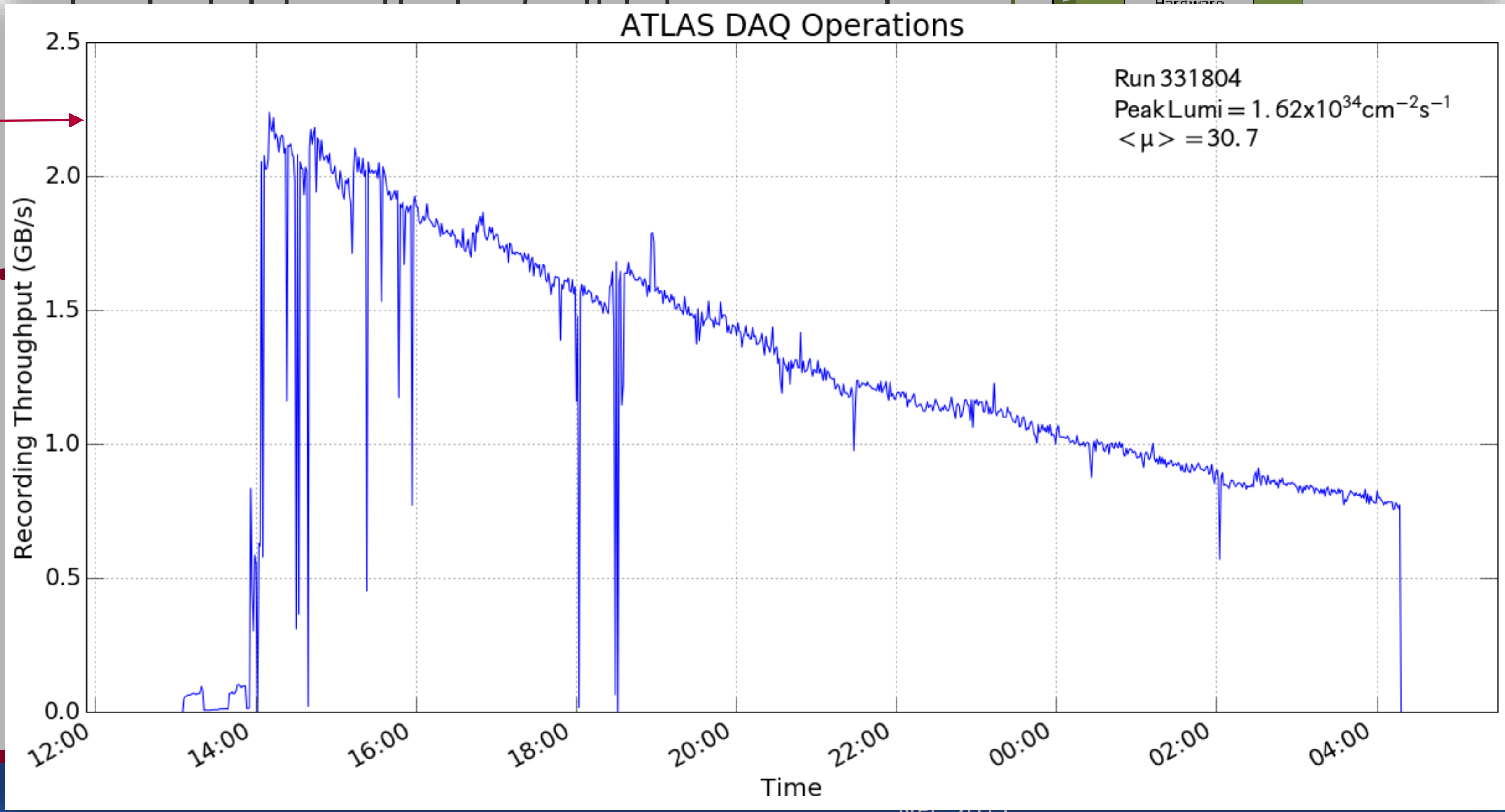


Data logging: The Sub-Farm Output

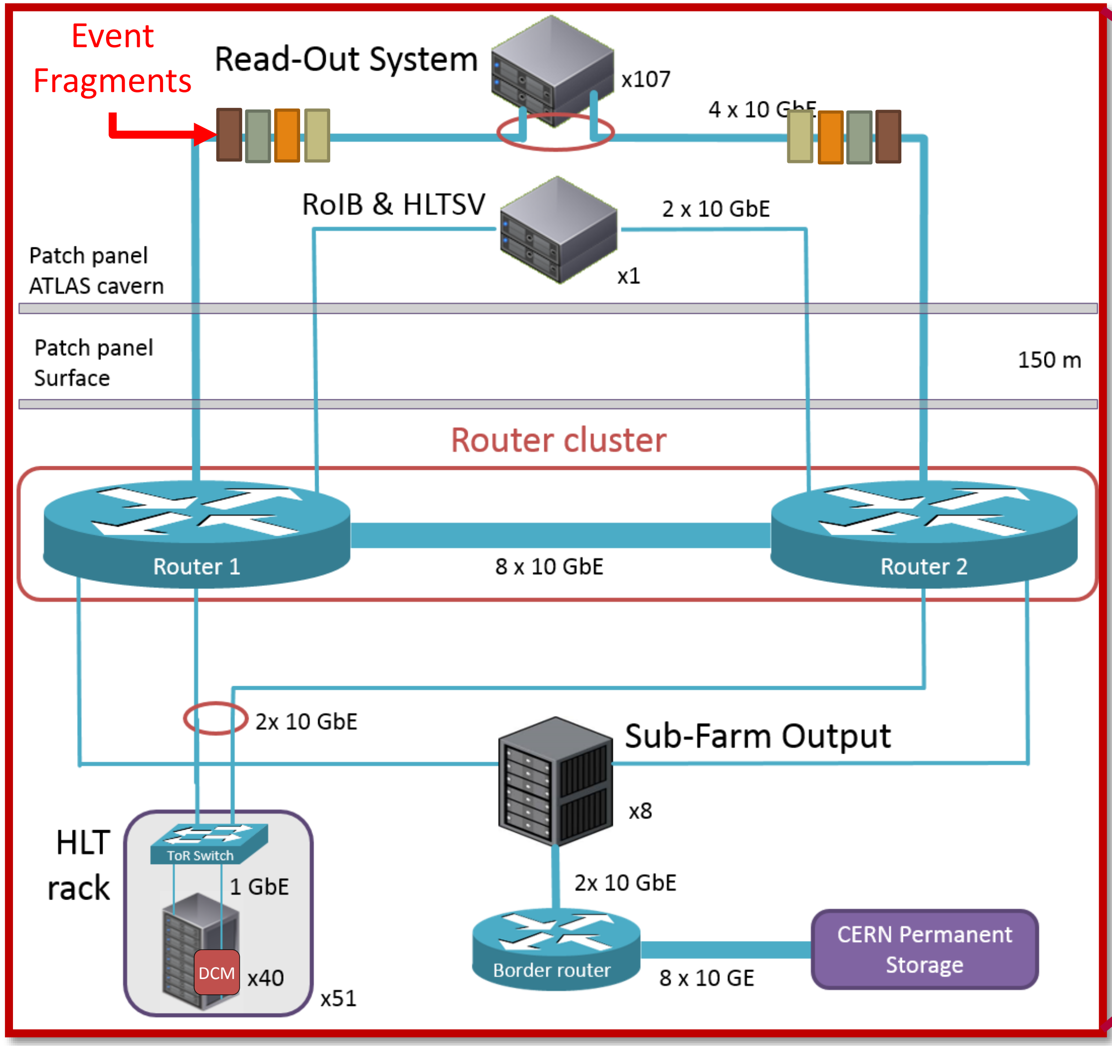
- The SFOs provide 48h of **temporary storage**
- Direct Attached Storage unit with multiple

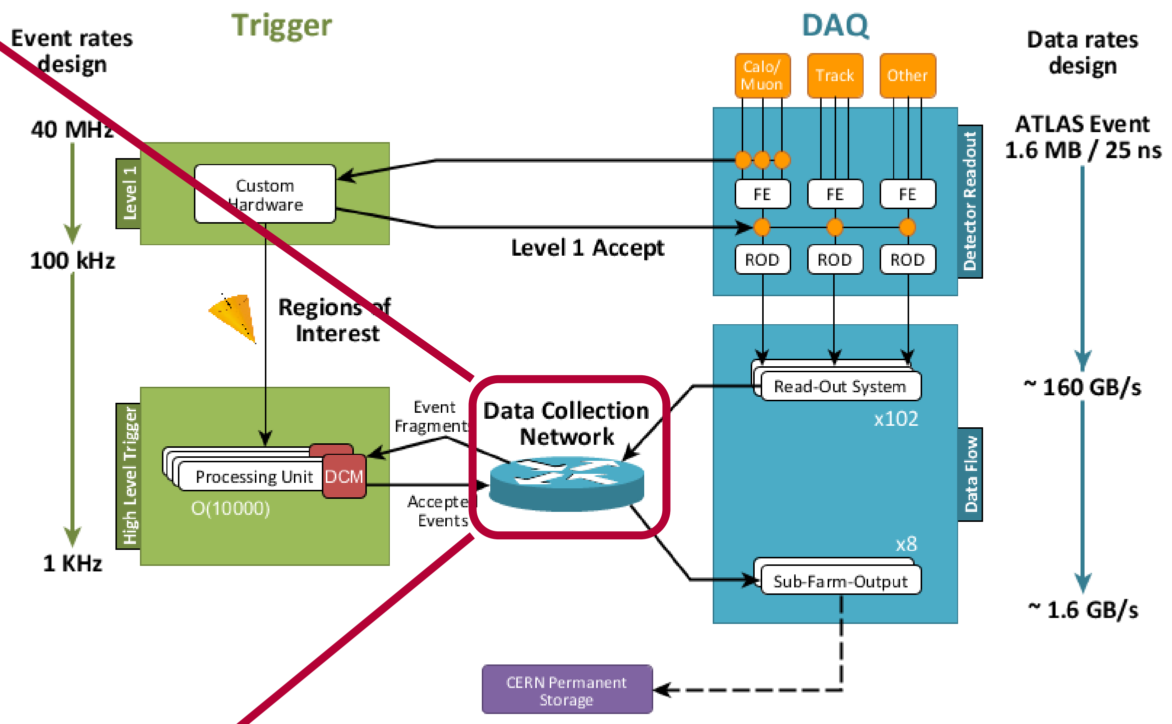


2.3 GB/s



The Data Acquisition Network

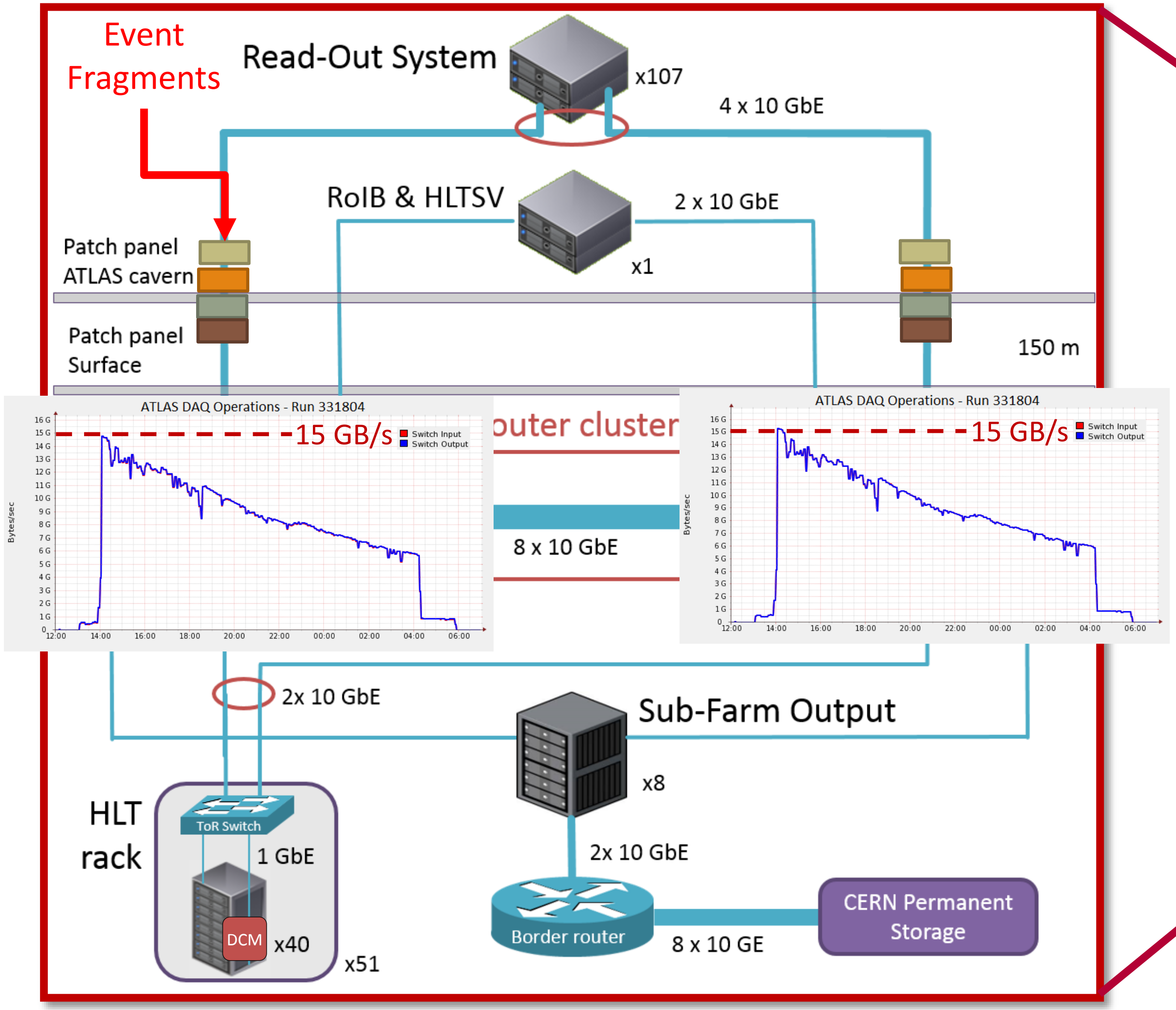


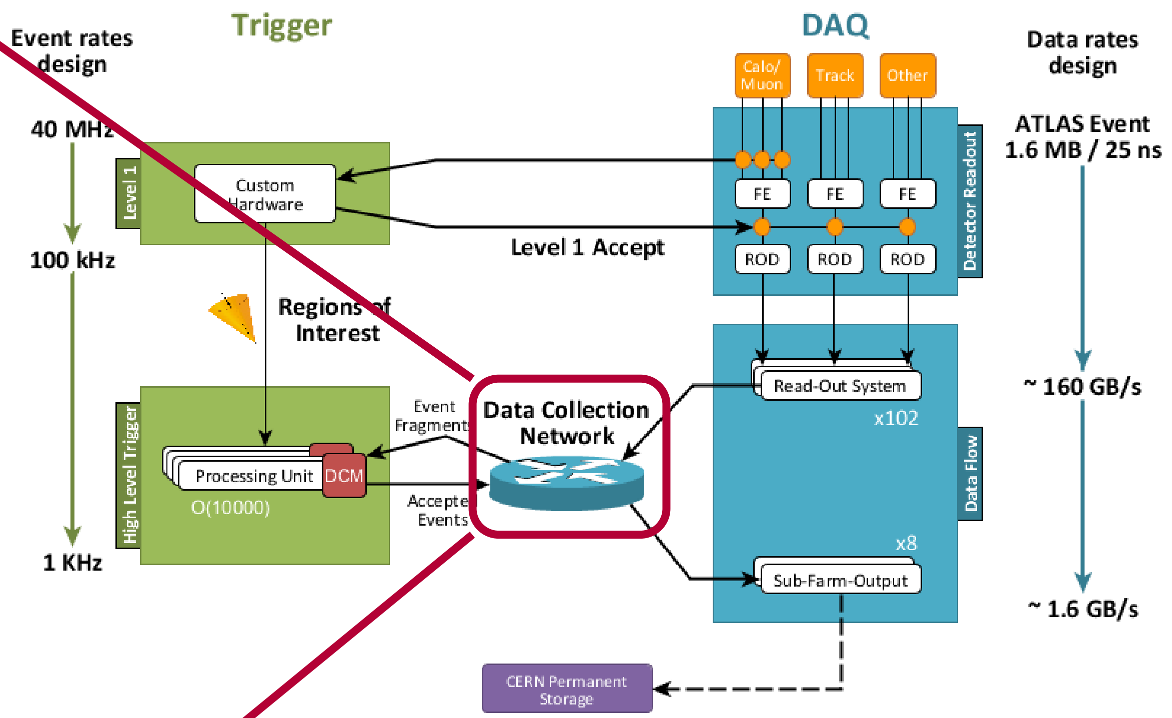


High throughput and high availability network

- More than 500 x 10 GbE ports
- 2 Brocade MLXe devices in cluster mode
- Built-in redundancy can deal with any single HW failure

The Data Acquisition Network

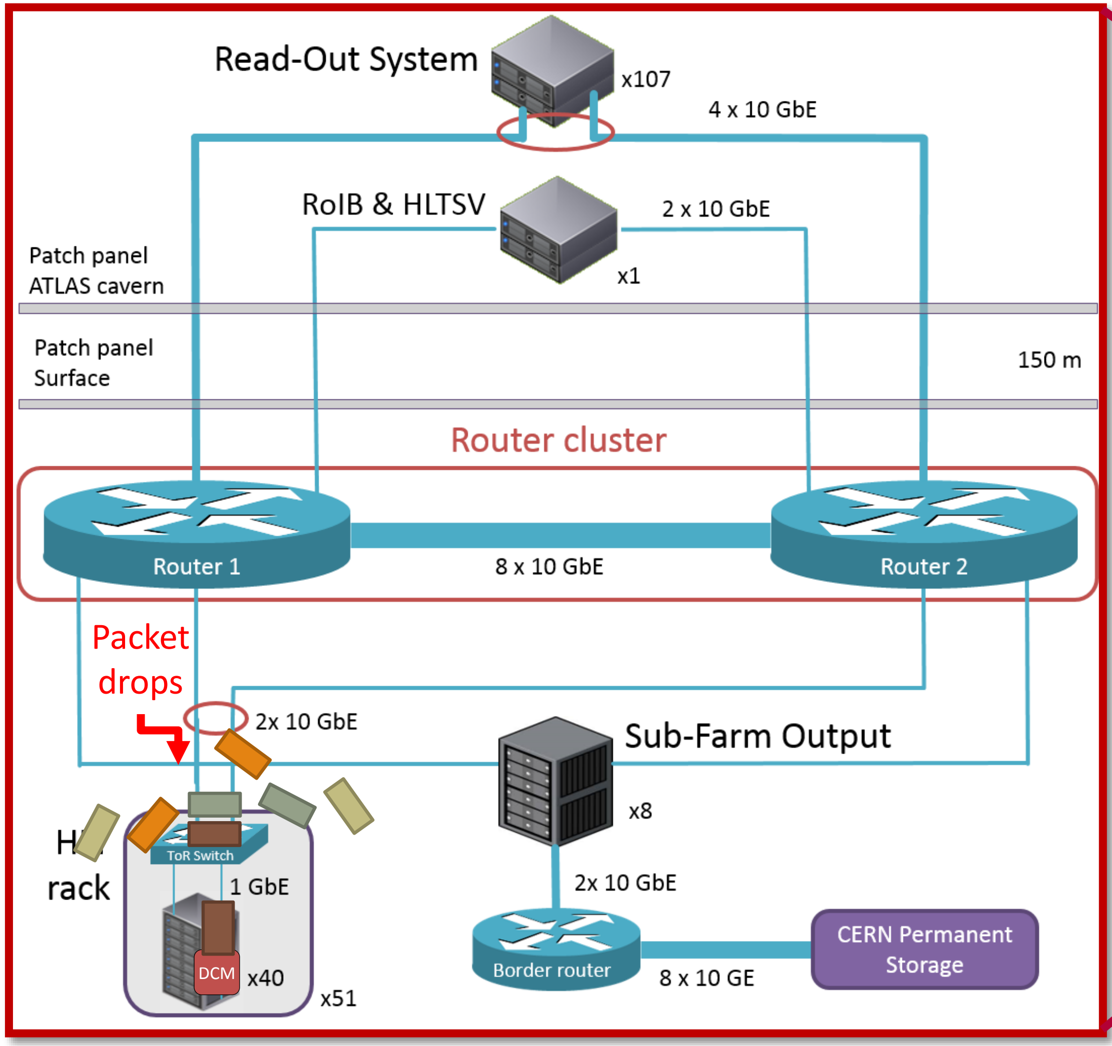


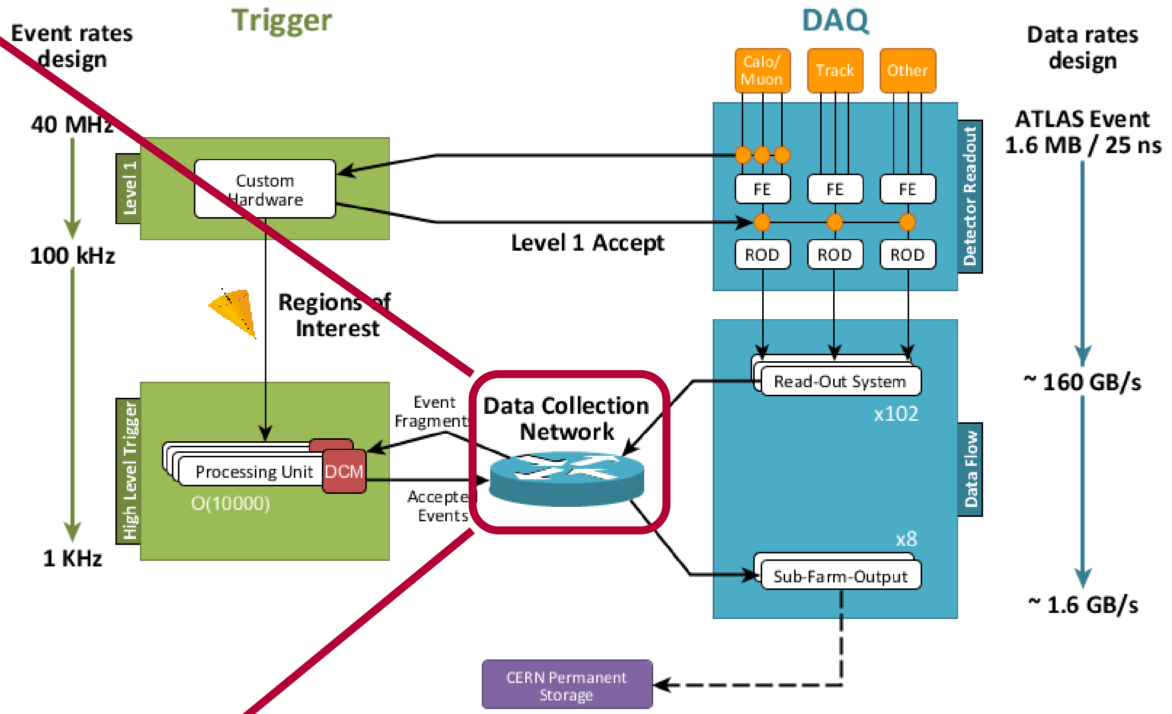


High throughput and high availability network

- More than 500 x 10 GbE ports
- 2 Brocade MLXe devices in cluster mode
- Built-in redundancy can deal with any single HW failure

The Data Acquisition Network

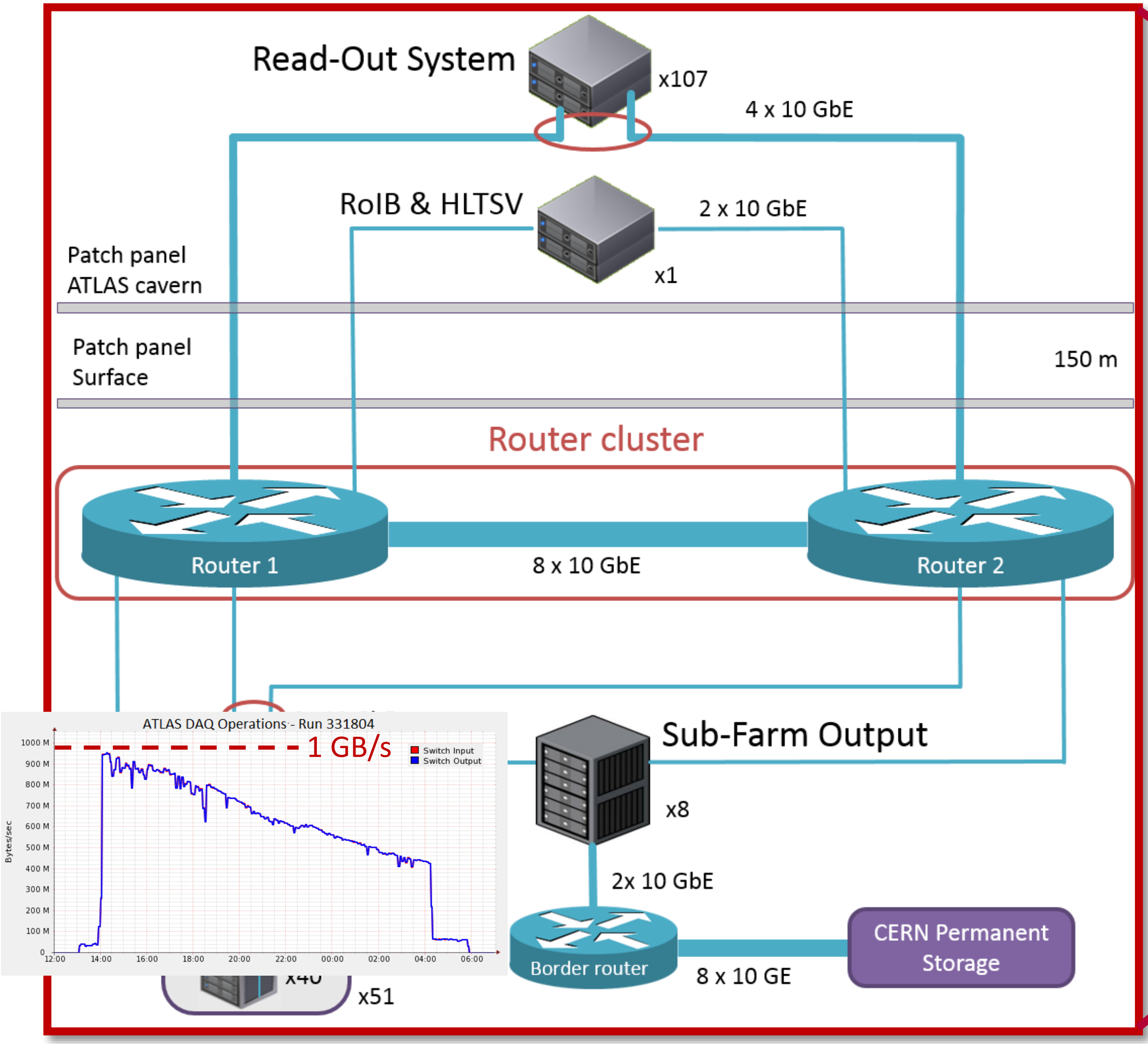


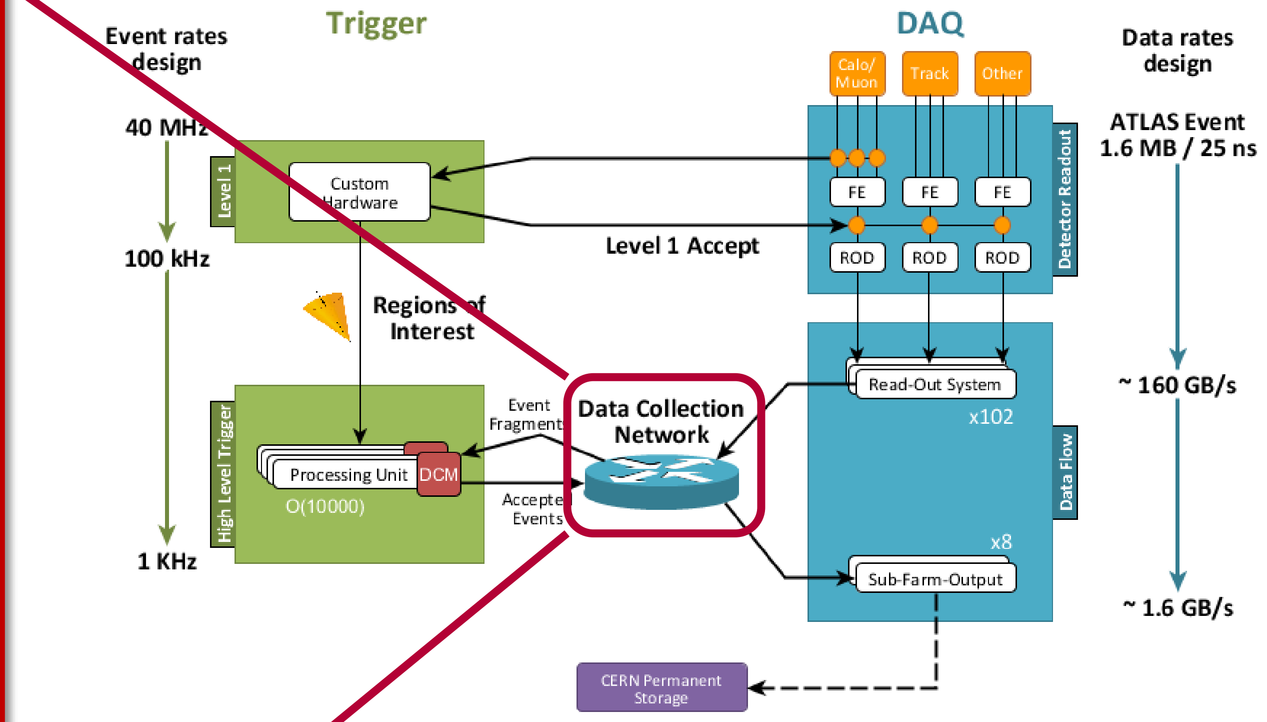


Top-of-rack switches suffer from instantaneous oversubscription. Solutions:

- Application-level congestion-control mechanism
- Fixed buffer allocation on deep-buffer switches (1.5 GB)
- Dynamic buffer allocation on shallow buffer switches (12.5 MB)

The Data Acquisition Network

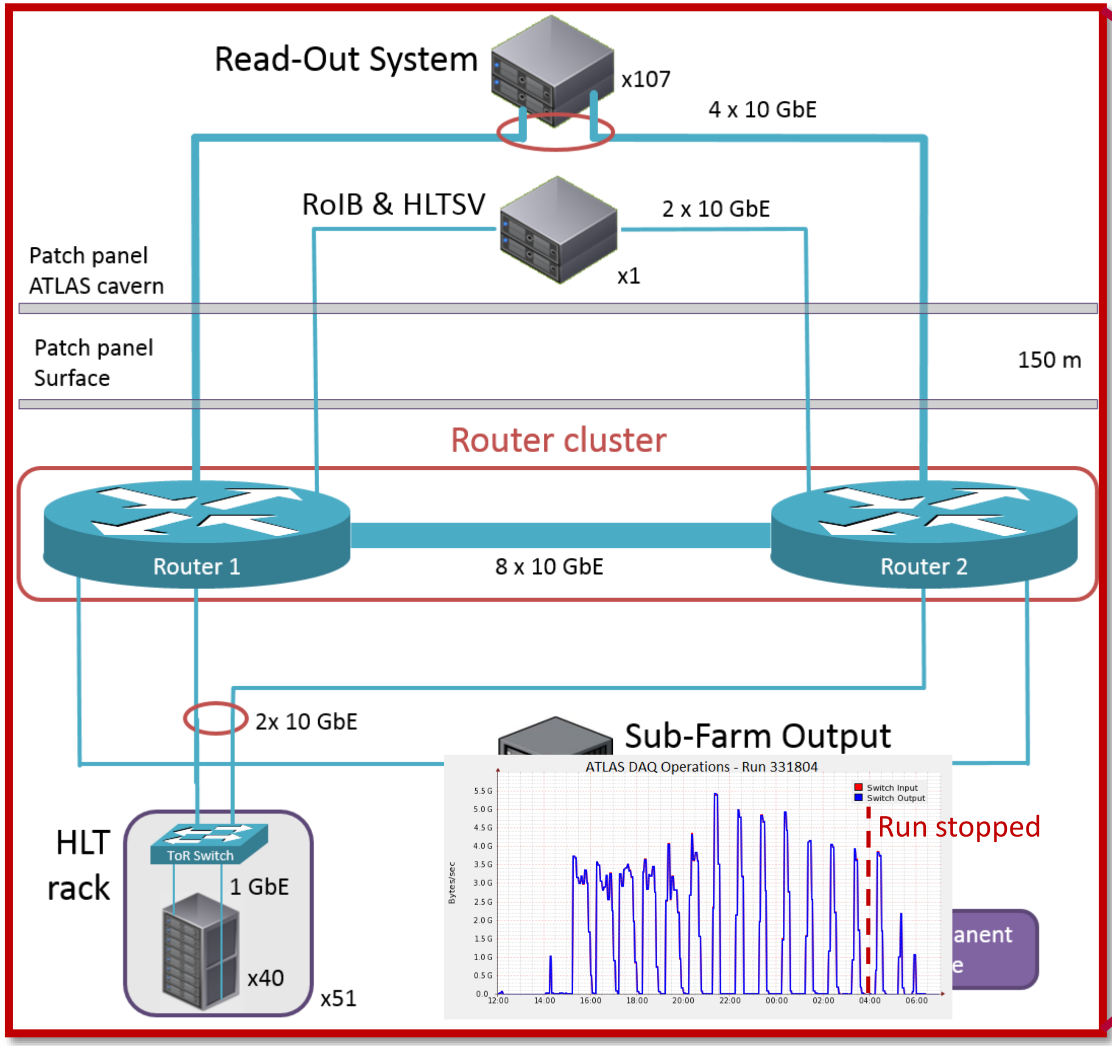


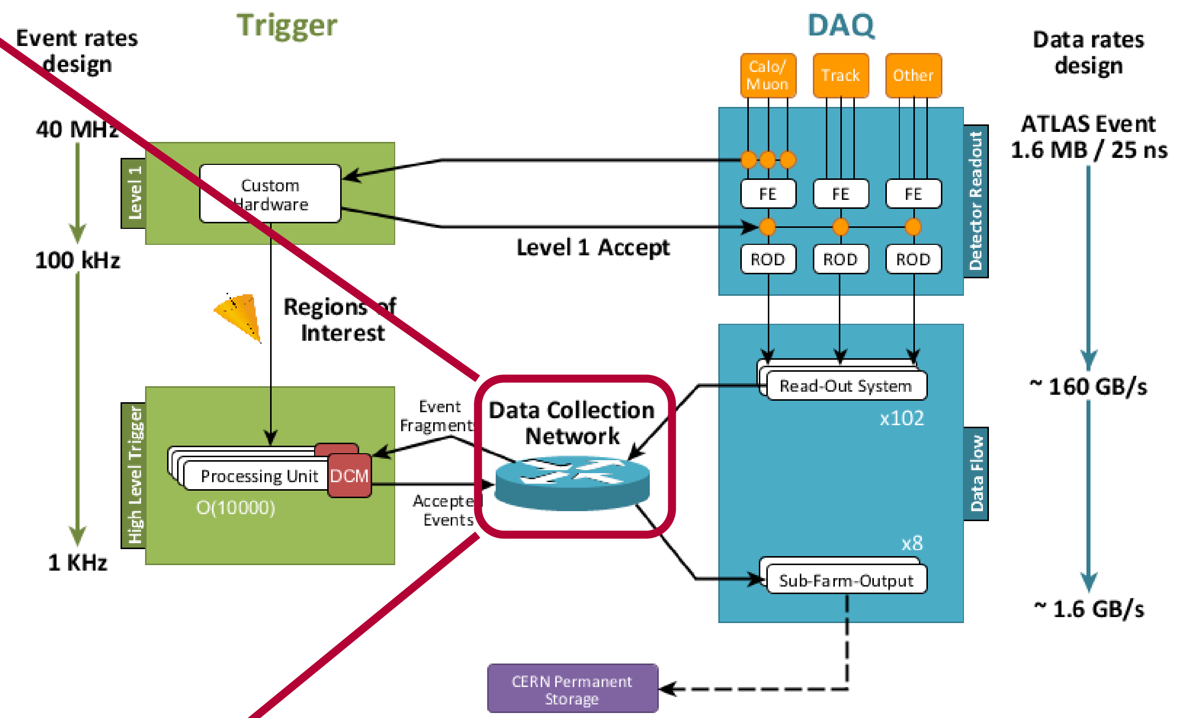


Top-of-rack switches suffer from instantaneous oversubscription. Solutions:

- Application-level congestion-control mechanism
- <u>Fixed buffer allocation on deep-buffer switches (1.5 GB)</u>
- Dynamic buffer allocation on shallow buffer switches (12.5 MB)

The Data Acquisition Network

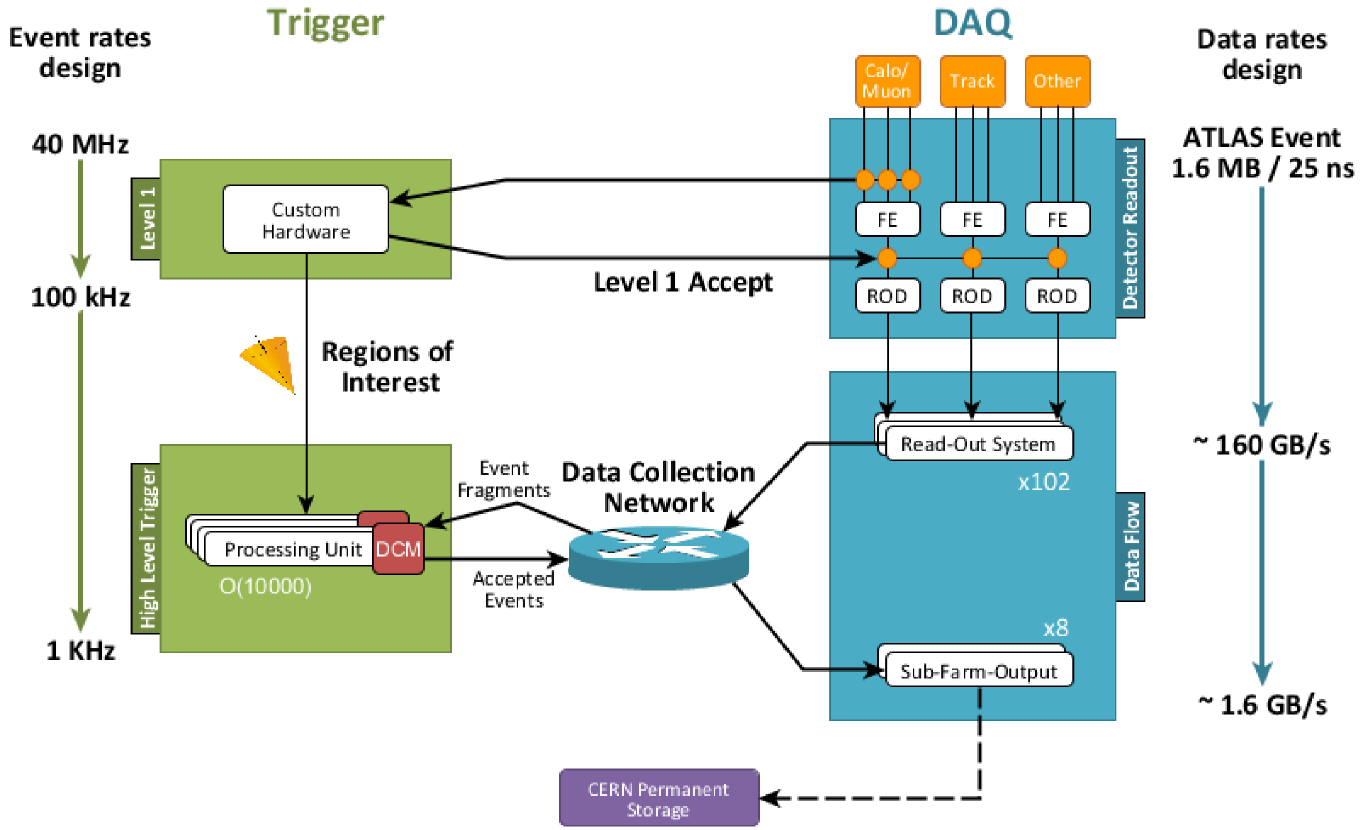




8 x 10 GbE connectivity to permanent storage

- Offline file transfers decoupled from ATLAS runs

Control and monitoring in the ATLAS DAQ system



Control and monitoring in the ATLAS DAQ system

Tasks

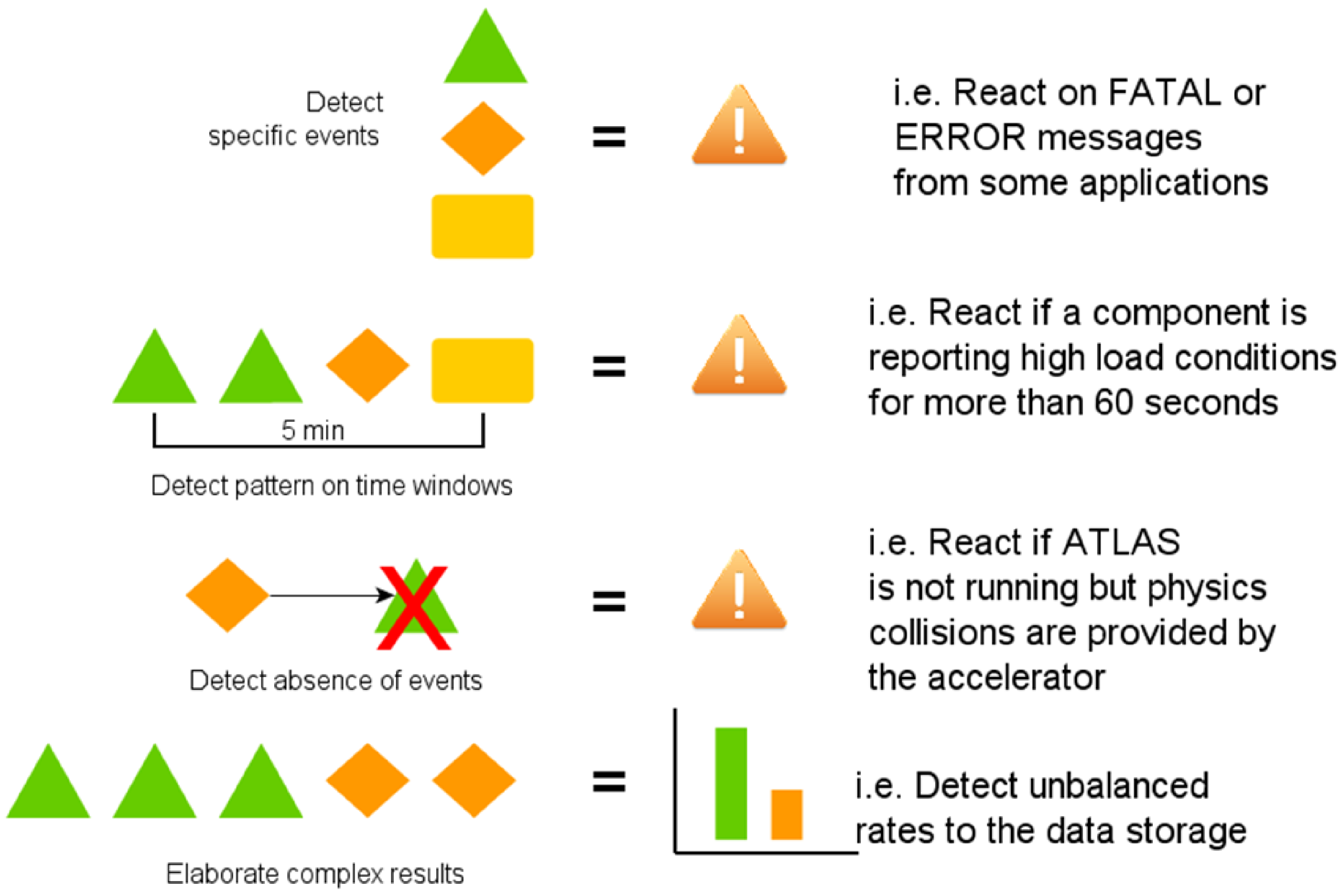
- **Coordinate** more than **30.000 applications** used for the detector control and data taking
  - Custom software supported by heterogeneous development teams
- System **robustness** is mandatory
  - Hardware and software failures happen frequently
  - Impact on data taking must be minimized
- The DAQ system provides all the needed infrastructure for:
  - Run control
  - Process management
  - Resource management
  - System configuration
  - ...and much more
- In Run 2 two Complex Event Processing (CEP) engines have been introduced:
  - CHIP and SA

CEP: Complex Event Processing

- Methods for finding complex patterns in monitoring data:
  - Correlation
  - Aggregation
  - Causality
  - Sliding time window

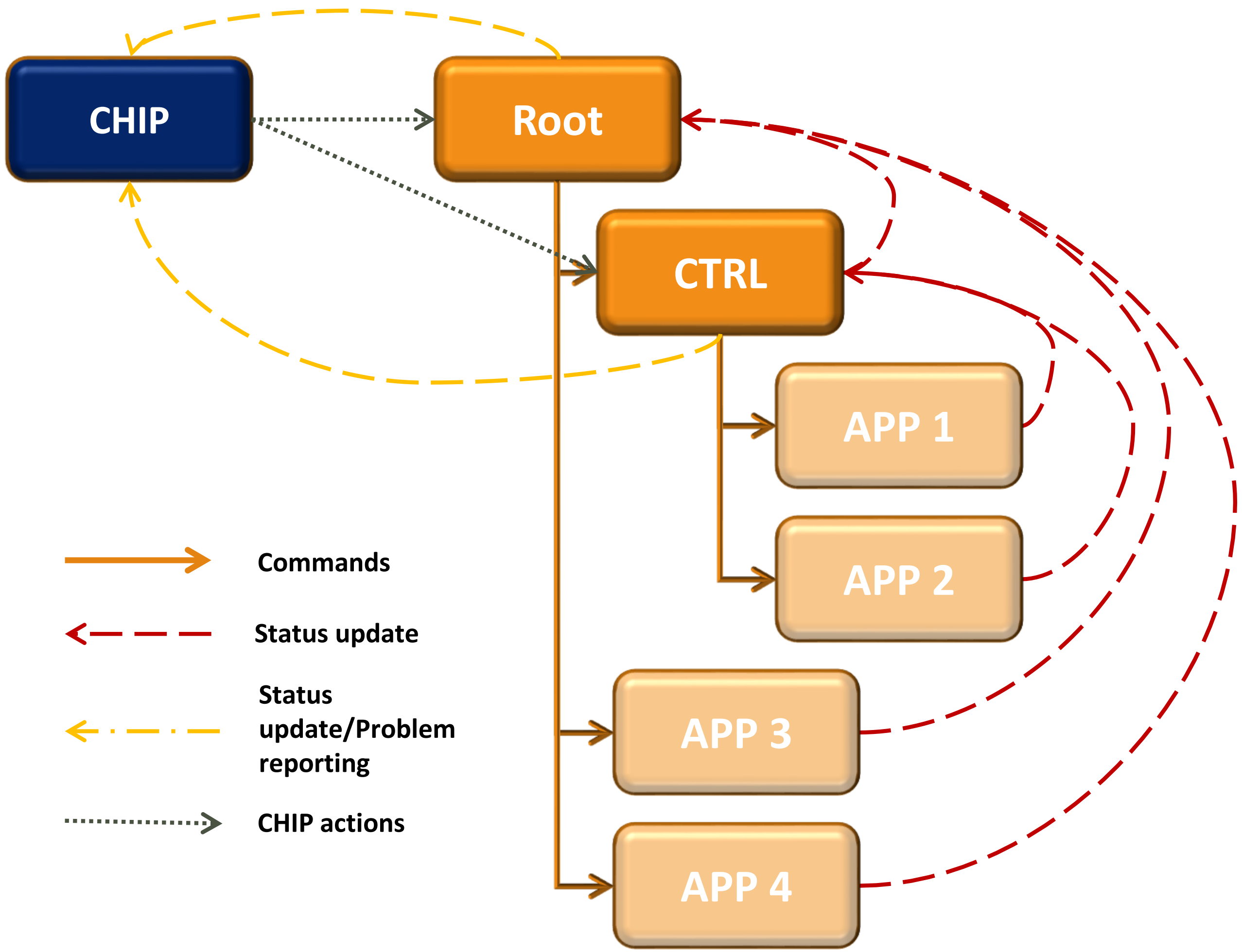
…and take corresponding actions

- ESPER: CEP engine which uses an SQL-like syntax for defining rules
  - It allows working on continuous streams of data
  - Simple syntax to define actions



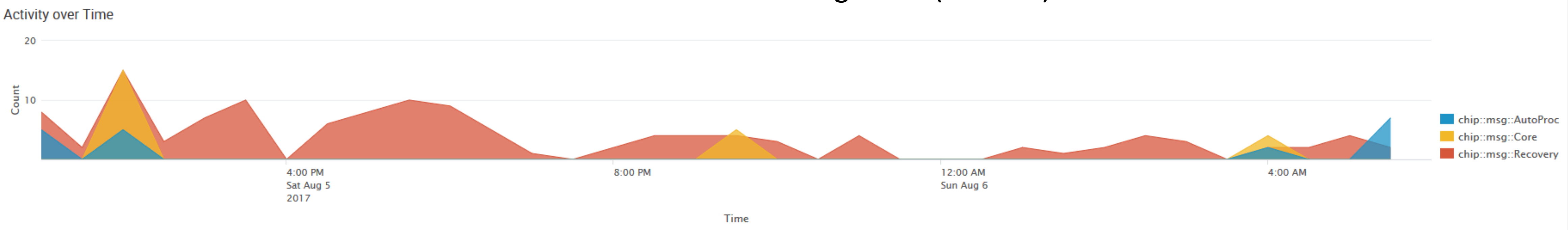
CHIP: *Central Hint and Information Processor*

- Is the brain of the Run controller application
  - Error management
  - Anomaly detection
- Maximizes efficiency:
  - Reacting fast and effectively to errors
  - Reducing the need of human interventions
- Optimizes manpower resources:
  - Reduces workload on the operator
  - Formalizes expert knowledge

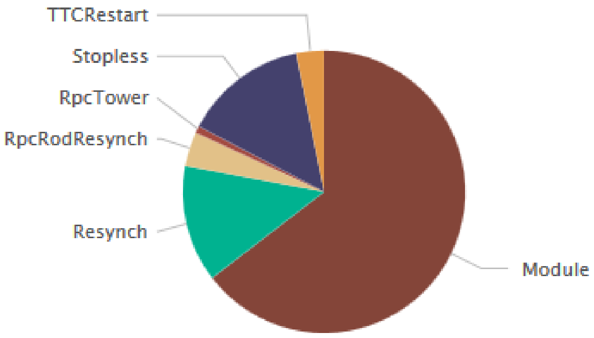


CHIP: *Central Hint and Information Processor*

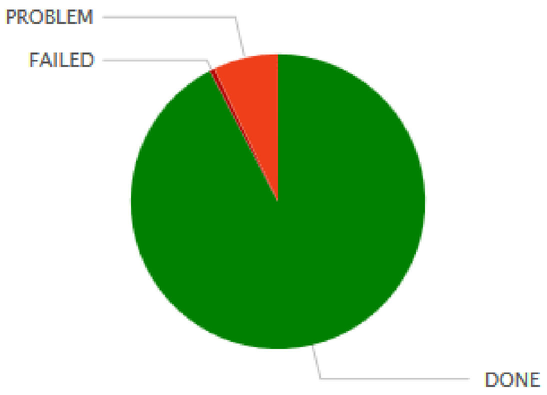
200 actions taken in a single run (331804)



Action type

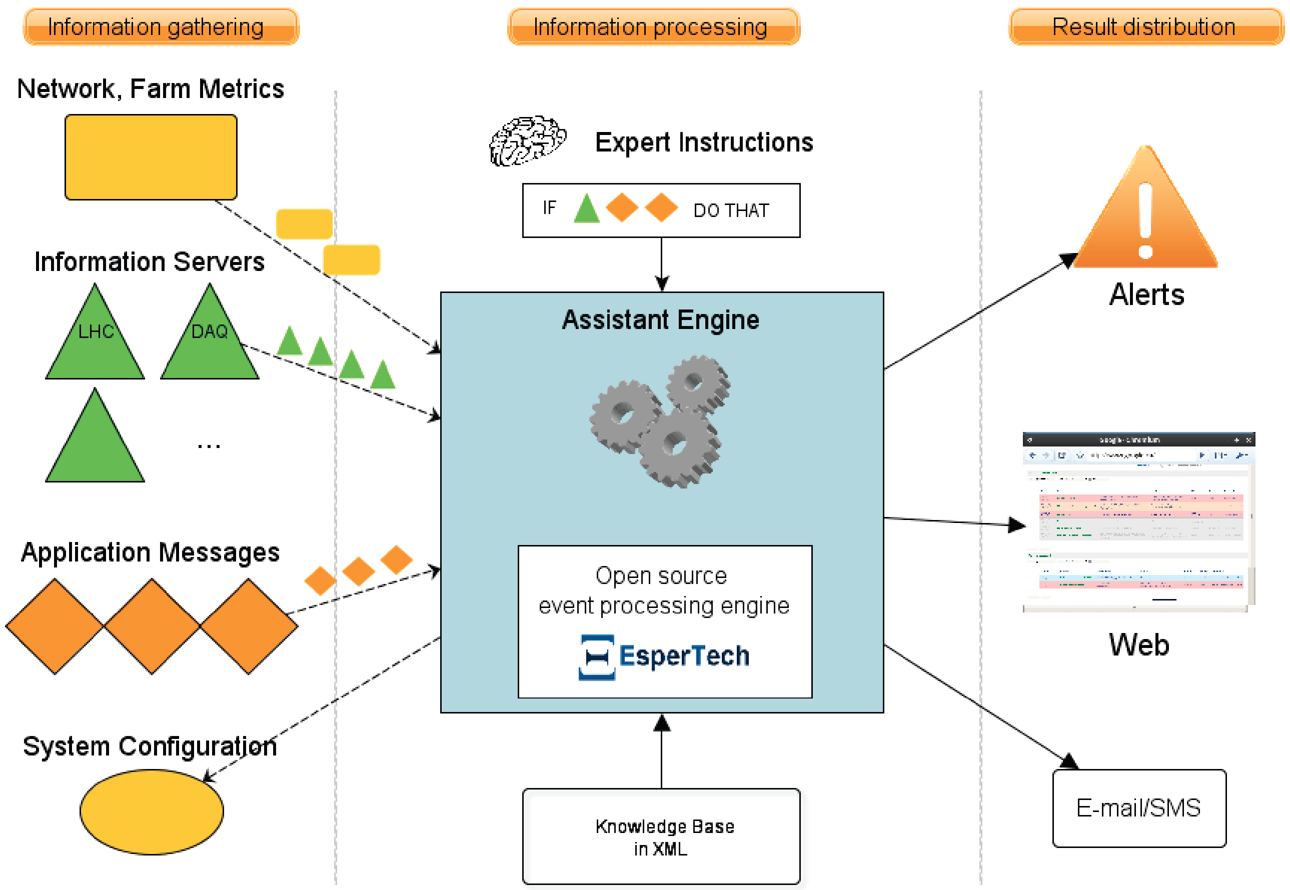


Action status



Shifter Assistant

- Another CEP engine to assist ATLAS shifters
  - Promptly notify shifters of problems and failures
  - Pertinent information provided taken from different data sources
  - Reminds the shifters to (not) take action

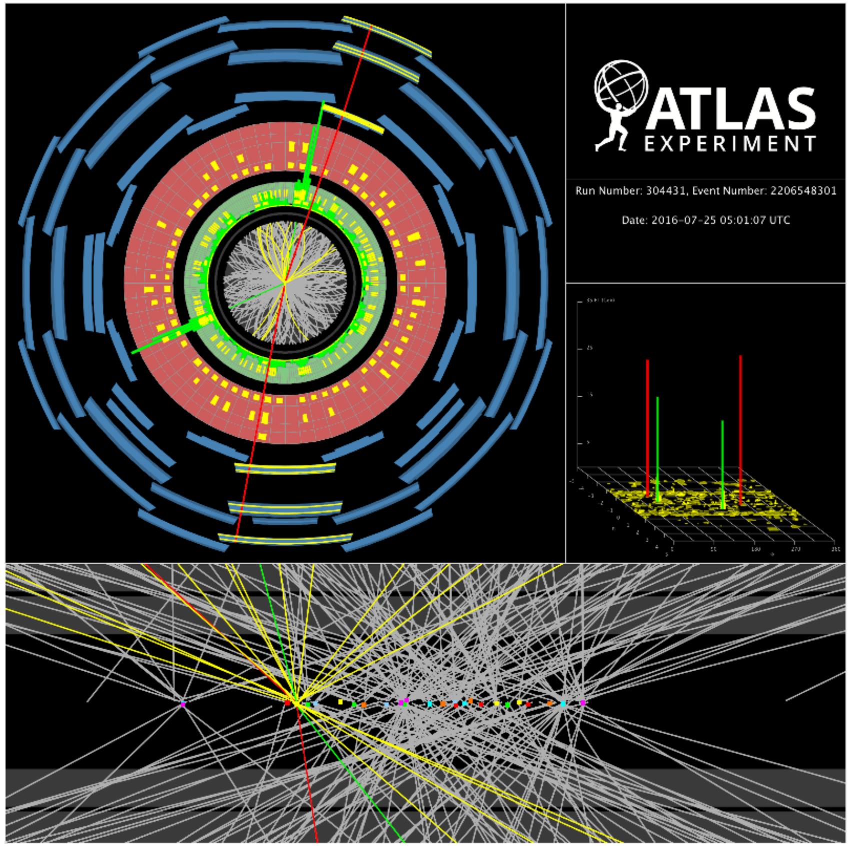


Example:

Name	ERROR **trigger-left-on-hold** Created on 09/18/2017, 6:14:09
Message	The expert system failed completing automated recoveries Stopless-REMOVAL on component TRTEndcapC_E-06/TRTBusyChannel_rod341501 (detector TRT_ENDCAP_C_SIDE). The trigger was left on hold 1 times.
Action	Alert the relevant TRT_ENDCAP_C_SIDE desk, if you are sure it will be harmless and it is appropriate, you should resume the trigger.

System Monitoring

- **Monitoring** the system is necessary for its correct functioning:
  - System health
  - Operating point monitoring
  - Physics rates
  - Data quality
- The DAQ monitoring framework offers a wide set monitoring tools to fulfill these requirements:
  - Online monitoring
  - Error reporting
  - Histogram publication
  - Event sampling
  - …
- In Run2 a new tool was introduced for persistent storage of the monitoring data:
  - PBEAST: Persistent Back-End for AtlaS Tdaq

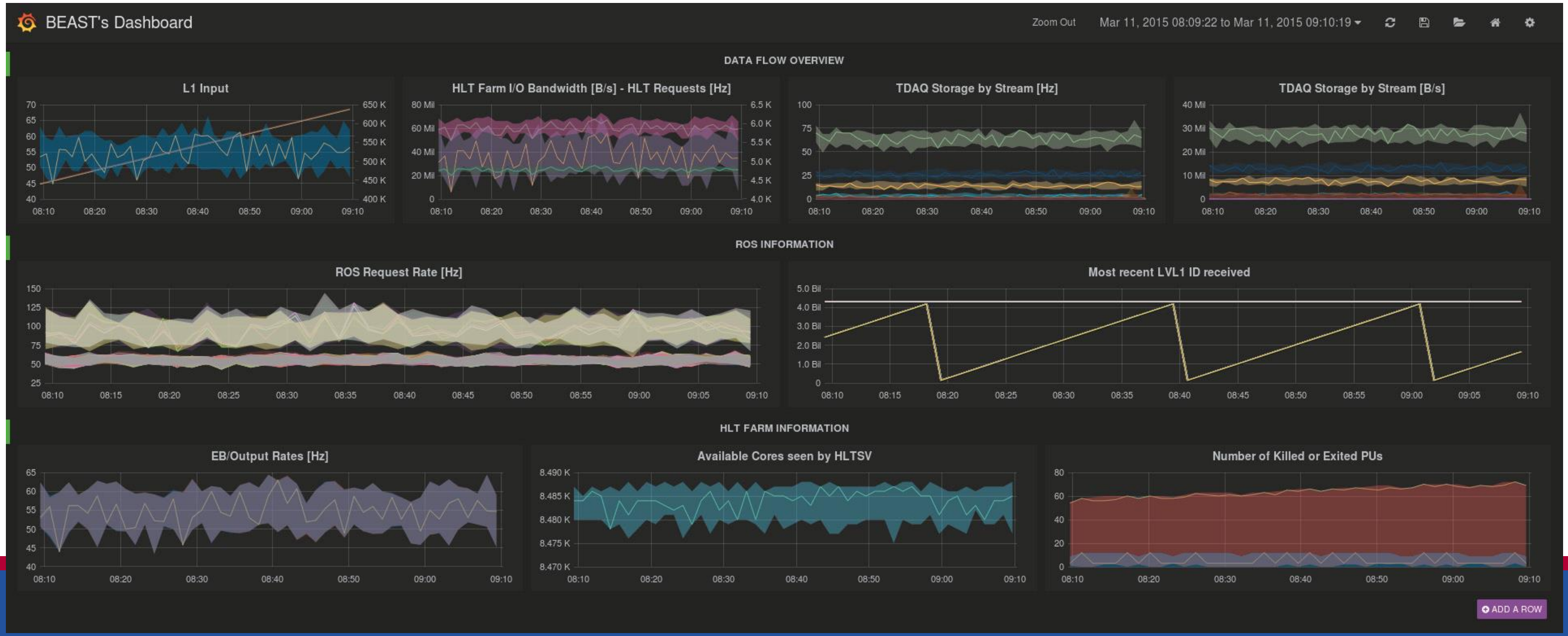


PBEAST: Persistent Back-End for AtlaS Tdaq

- Time-series database to store monitoring information
- Stores an important fraction of operational data from ATLAS
  - Up to 500 kHz attributes refresh rates
  - 1 TB/month in 2015 and 1.5 TB/month in 2016
    - All raw data of Run 2 are available
- PBEAST provides several programming interfaces:
  - Data insertion: via Online Monitoring API or REST API
  - Data retrieval: C++, Python and REST
- Operates on two nodes (2015q4):
  - Dual 12 cores CPU Xeon E5-2680V3 @ 2.5GHz, 256 GB RAM, 8x4TB RAID
- Custom implementation based on low level primitives of Google Protocol Buffers for data interoperability, compaction and compression
  - Cassandra and Splunk prototypes tested during Run 1 -> Unsatisfactory outcome so completely reimplemented by start of Run 2

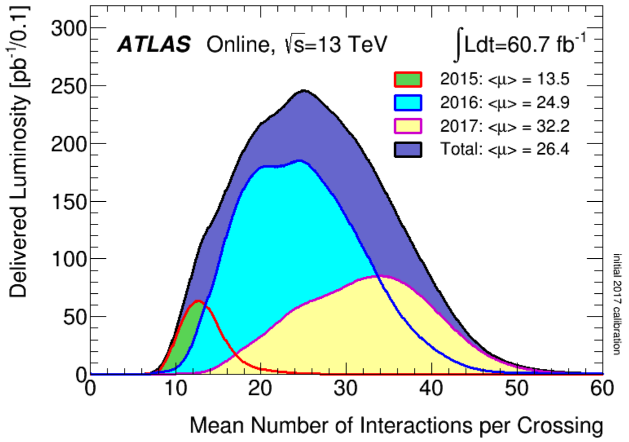
PBEAST: Grafana dashboard

- **Grafana** is an open source metric analytics & visualization suite mostly for time-series data
- Custom PBEAST plugin developed:
  - Minimize the amount of transferred data
  - Minimize the required post-processing in the browser
- Perform persistent down-sampling on PBEAST server side

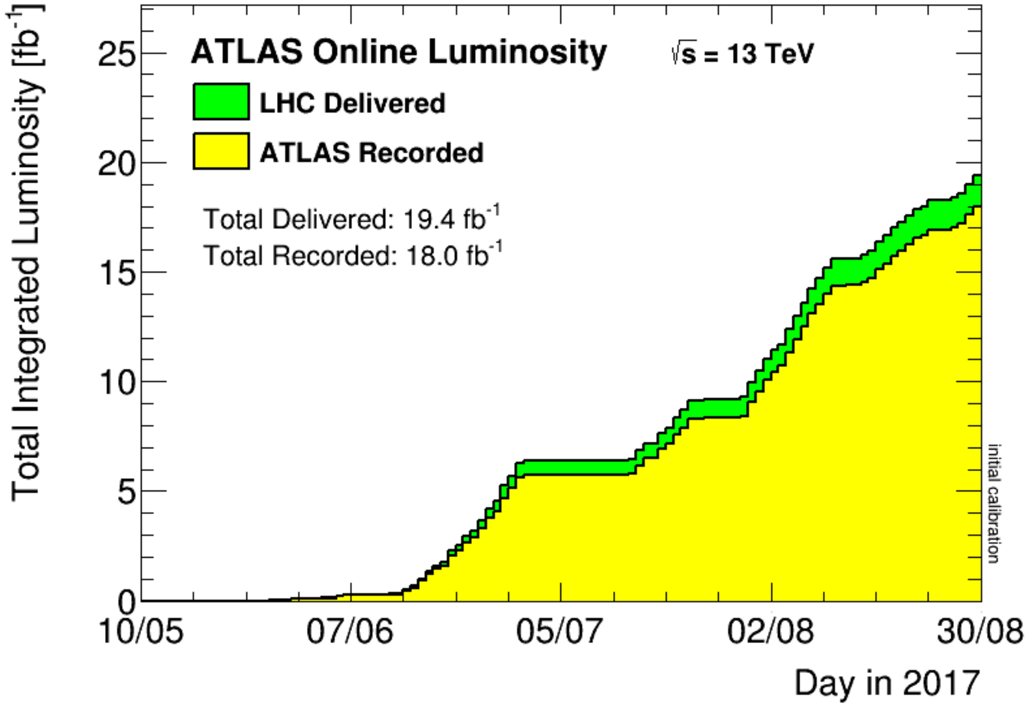


ATLAS data taking efficiency in 2017

- The Data Acquisition System has positively contributed to the high ATLAS efficiency in 2017 despite the more challenging conditions:

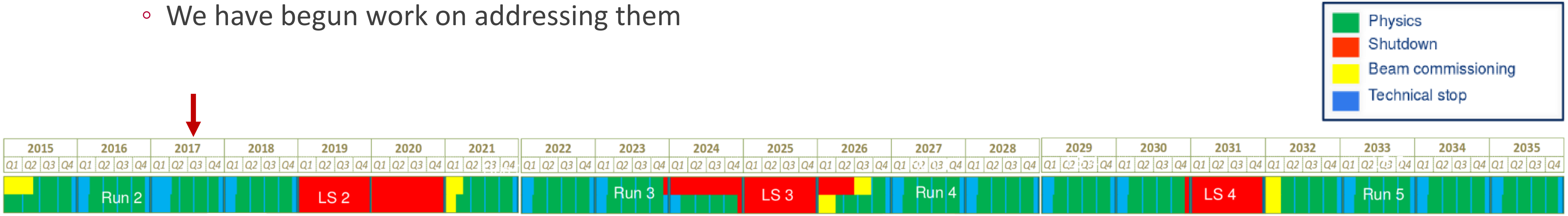


Efficiency: 92.8%



Conclusions

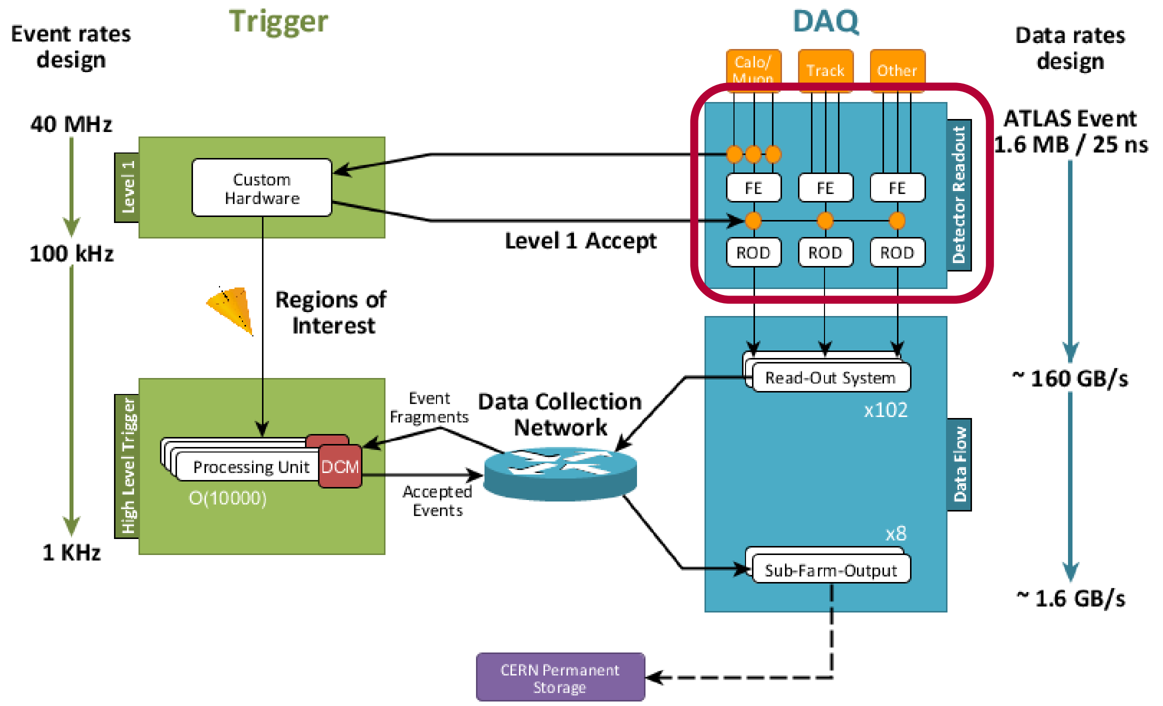
- Data Acquisition is a complex engineering problem making use of the latest technology
  - Electronics, computing, networking, storage…
  - Based on Run 1 experience, ATLAS DAQ has evolved to address the new challenges of Run 2
- Controlling and Monitoring data taking in ATLAS is a complex task
  - CEP engines help automate many of them, maximizing the efficiency of the system
  - New monitoring tools deployed
    - PBEAST system had to be developed for persistent storage of monitoring data
- New challenges will arise as LHC increases luminosity
  - We have begun work on addressing them



Next challenges for the DAQ system

LHC Phase I: 2019-2024

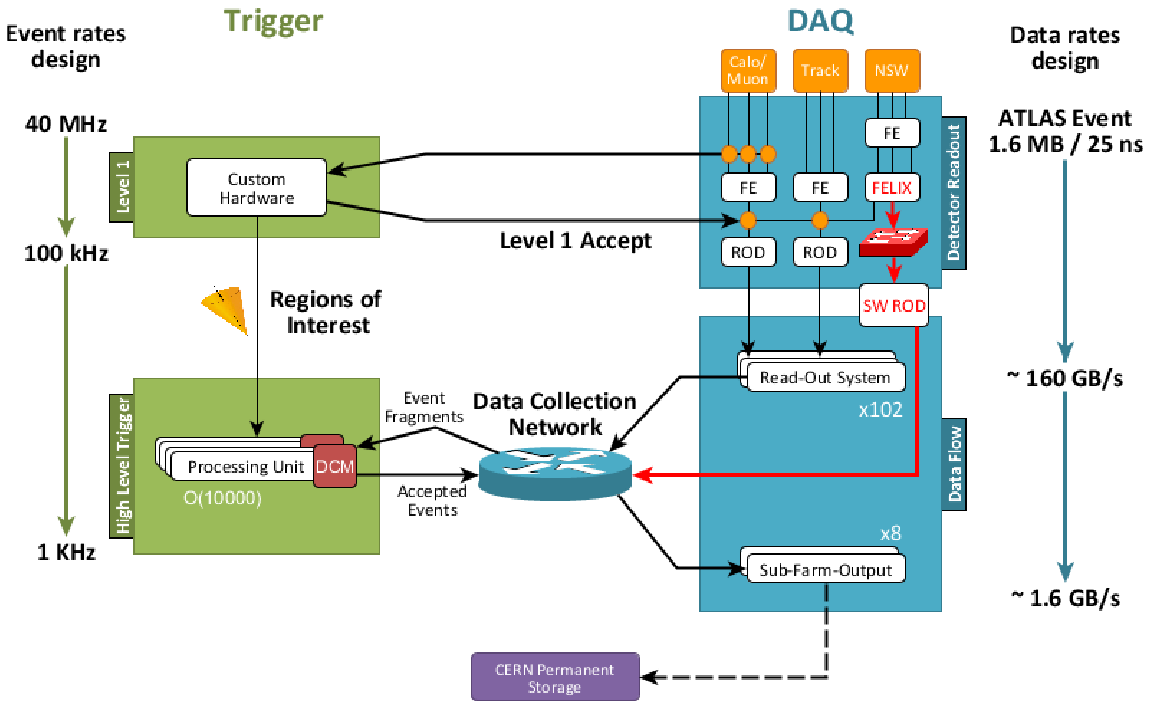
- Bring Commercial Off-The-Shelf technology closer to the subdetectors
  - Servers and network
- Use common read-out hardware
  - New common read-out board: FELIX
  - Implement ROD functionality in SW
  - Only few subdetectors in a first instance



Next challenges for the DAQ system

LHC Phase I: 2019-2024

- Bring Commercial Off-The-Shelf technology closer to the subdetectors
  - Servers and network
- Use common read-out hardware
  - New common read-out board: FELIX
  - Implement ROD functionality in SW
  - Only few subdetectors in a first instance



Next challenges for the DAQ system

LHC Phase II: 2026-2035

- High-Luminosity LHC
  - Target: $7.5 \times 10^{34} cm^{-2} s^{-1}$
- **A new DAQ is needed**
  - 6 TB/s read-out
  - FELIX system expanded to all subdetectors
  - Decouple detector read-out and software trigger processing using a high throughput distributed storage system