ICHEP2016CHICAGO
38th INTERNATIONAL CONFERENCE ON HIGH ENERGY PHYSICS
AUGUST 3 - 10, 2016
CHICAGO

The ATLAS Dataflow System in Run2: Design and Performance

Othmane Rifki
University of Oklahoma

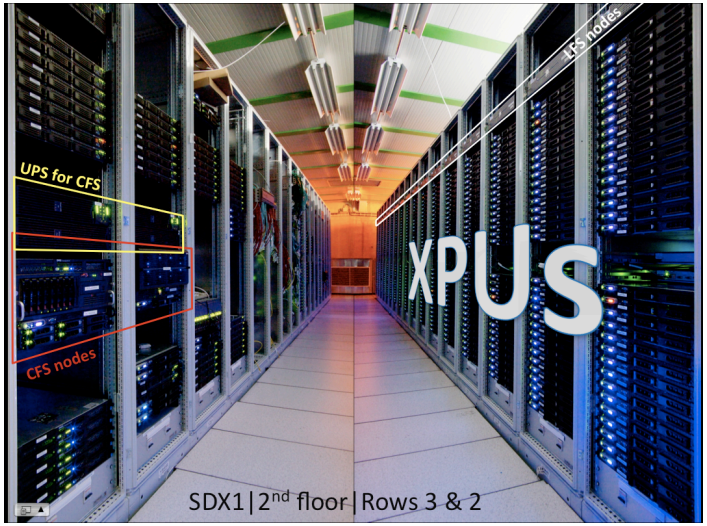
On behalf of the ATLAS Collaboration

The University of Oklahoma

ATLAS EXPERIMENT

Outline

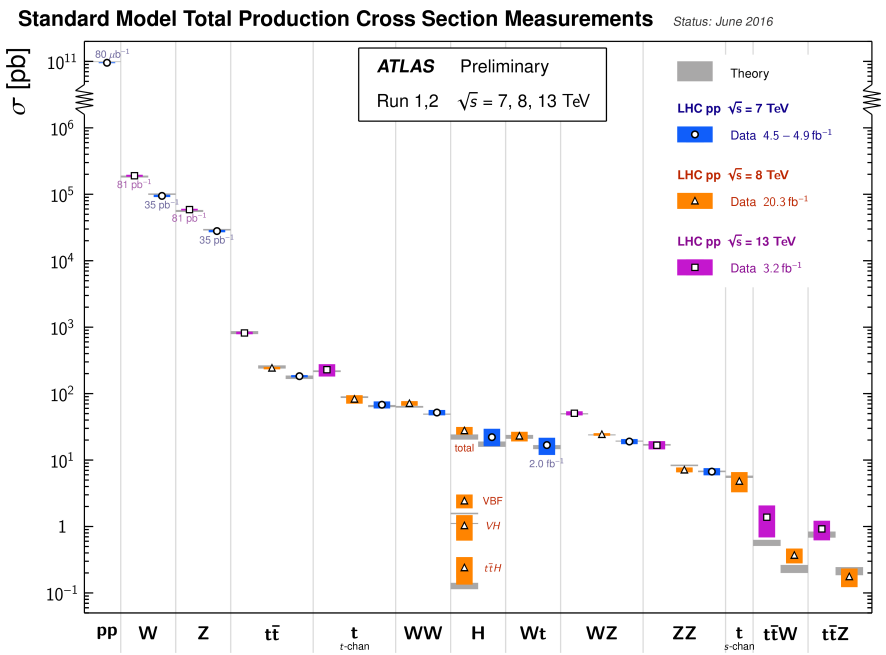
- Motivation
- Run 2 Challenges and Evolutions
- ATLAS Dataflow Design
- Region of Interest Builder
- Dataflow Network and Software
- Performance in Run 2



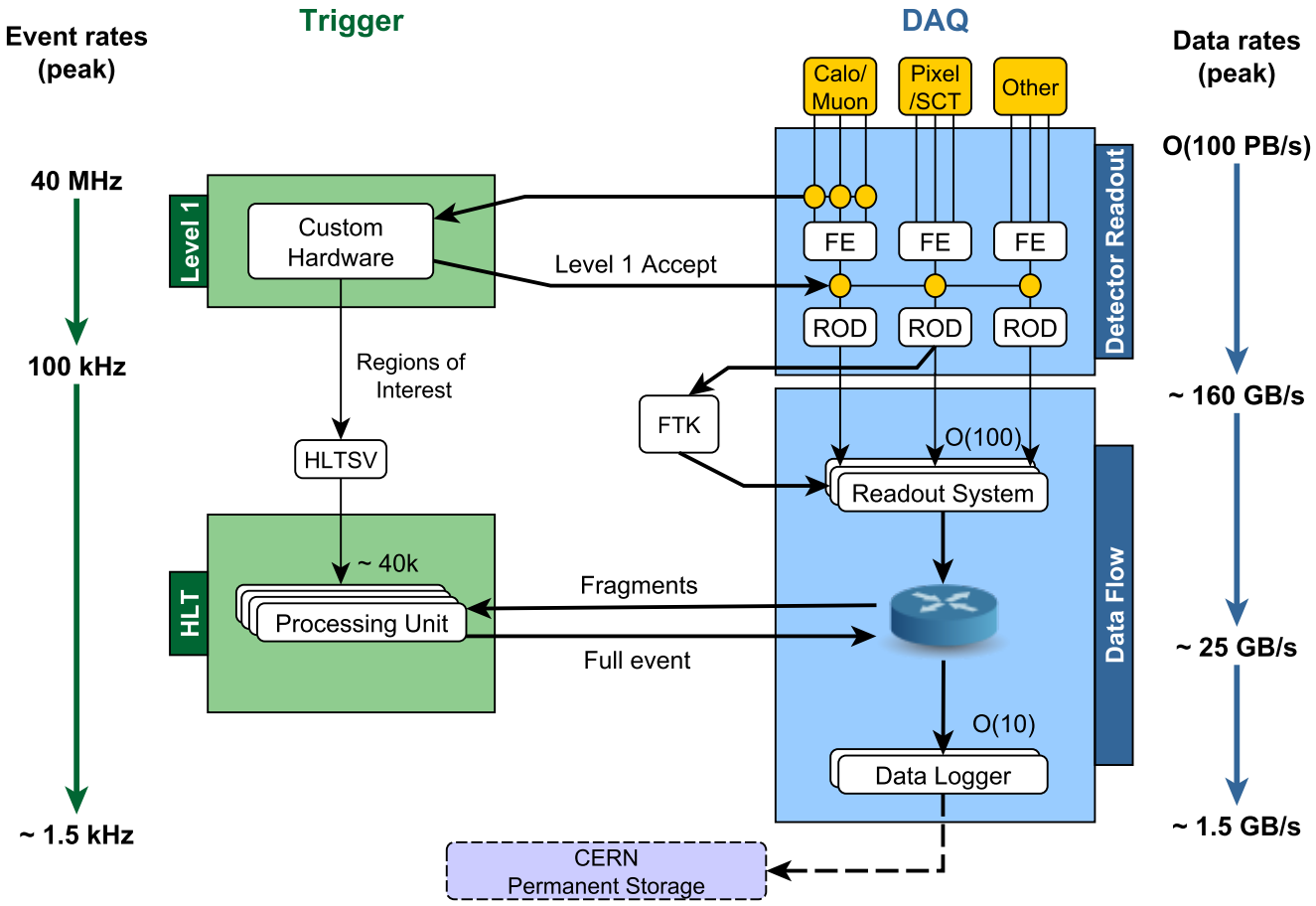
Motivation

- The data acquisition (DAQ) system is designed to buffer, transport and record events triggered by the ATLAS detector
- The goal is to achieve a high data collection efficiency during collisions
  - Dead-time and event losses must be low
- Performance is maximized with respect to the available software and hardware technologies

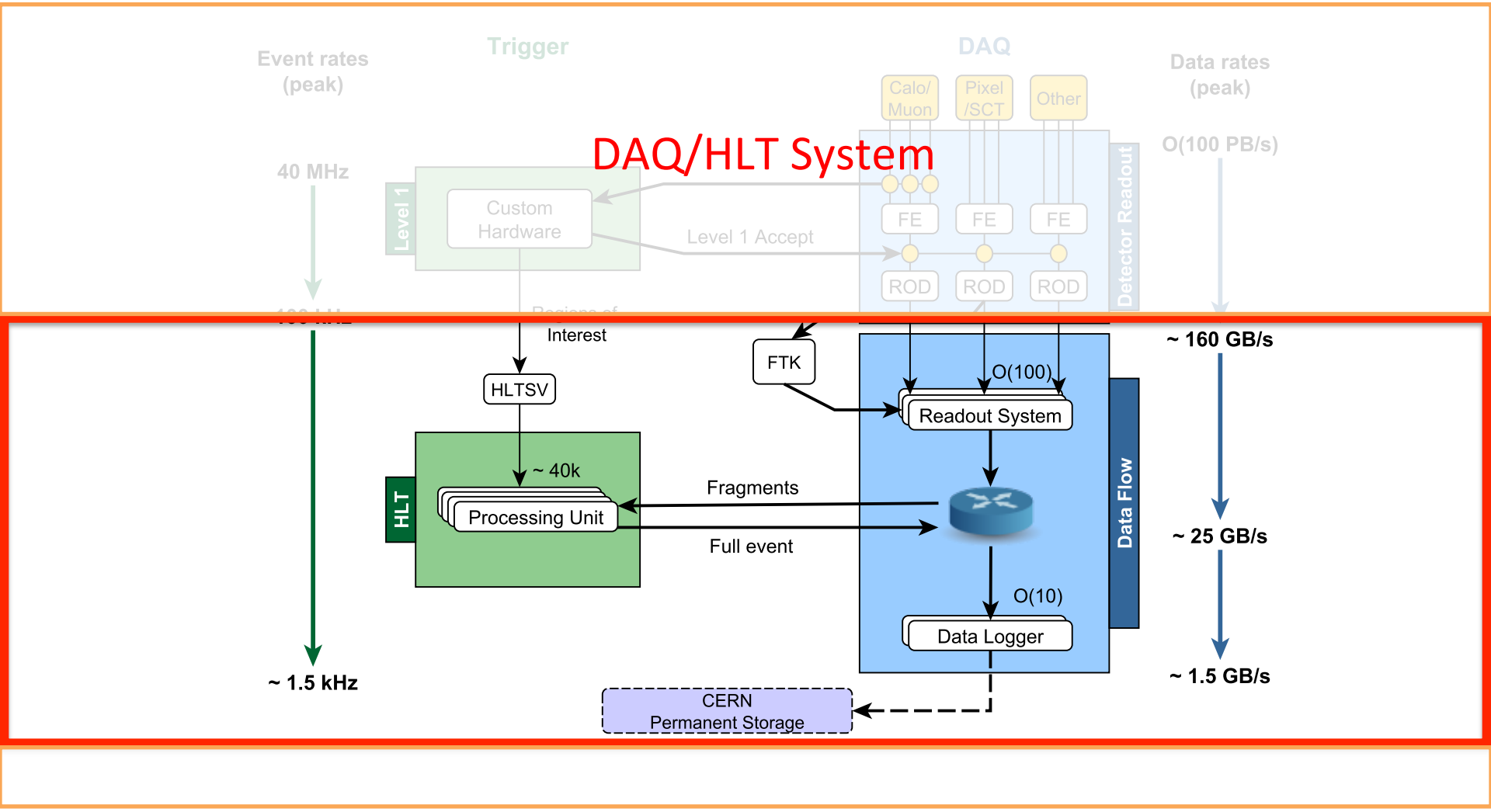
High data taking efficiency is crucial to the ATLAS physics program!



ATLAS TDAQ in Run 2

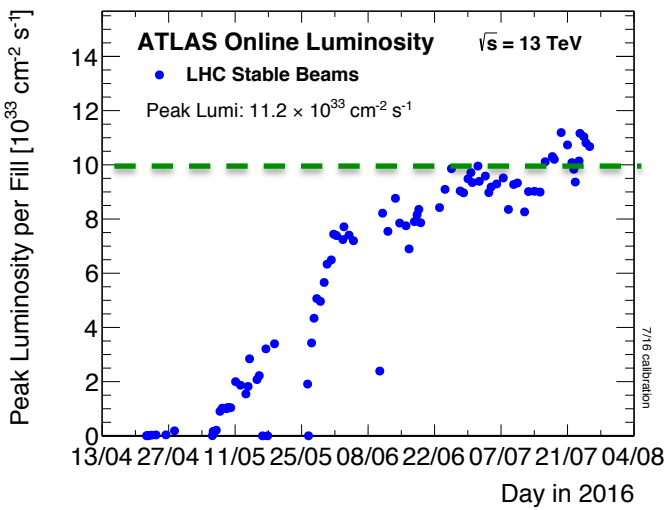


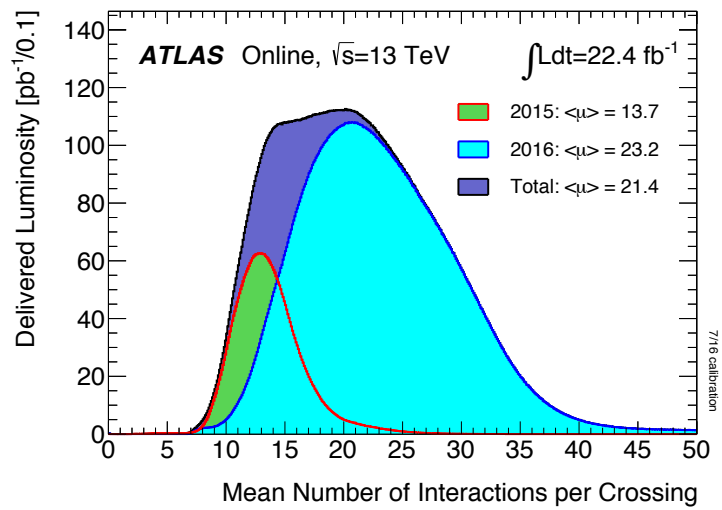
ATLAS TDAQ in Run 2



Run 2 Challenges

- LHC achieved a record instantaneous luminosity of over 10^{34} cm^{-2} s^{-1} (pileup of 30) increasing the event processing time
- Increased Level 1 trigger rate to 100 kHz (75 kHz in 2012)
- An increase in the event size due to a 20% increase in the number of readout channels and the increase in pileup
  - Introducing new detector components: Insertable B-layer (IBL), L1 topological trigger, Fast Tracker (FTK)

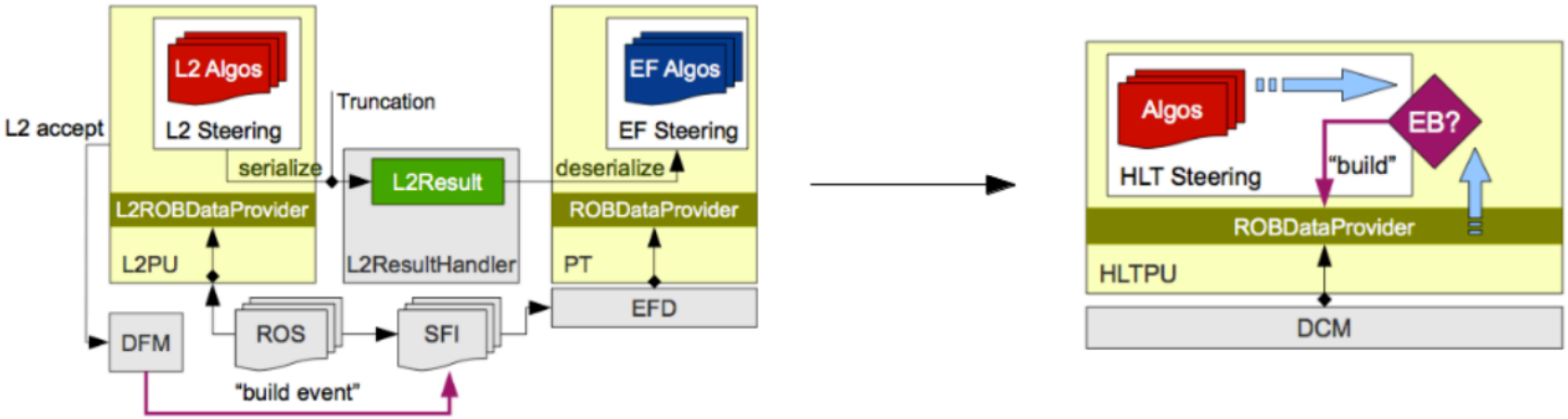




Evolutions in Run 2

To face the new run conditions in Run 2, ATLAS upgraded key parts of the DAQ system by:

- Merging the Level 2 and event filter into a single HLT farm with an increased number of available cores allowing more system flexibility



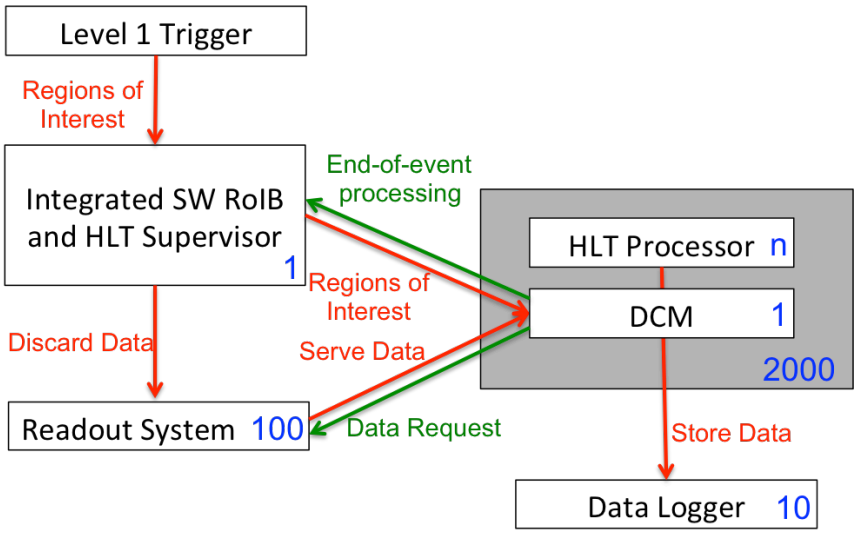
- Upgrading the input and buffer hardware of the readout system to increase the readout rate (also higher density of optical links and larger memory)
- Increasing the network capacity with a simpler and more scalable design
- Using more multi-threaded software design to take advantage of the modern multi-core CPUs
- Data logger system capable of writing at 1 kHz an event size of 1.5 MB

As a result, the DAQ system has more **flexibility, scalability and redundancy**

The ATLAS Dataflow

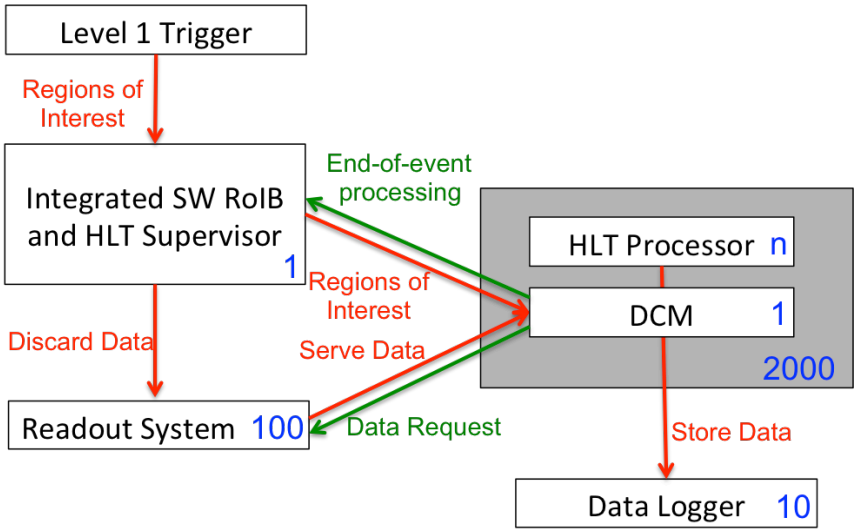
ATLAS operates on the concept of **Regions of Interest** (**RoIs**) where data processing proceeds in stages: Fast algorithms run on RoIs and request more data as needed until the decision to accept or reject the event is made

- Readout System (ROS) buffers front-end data from the detectors and provides a standard interface to the DAQ
- The Region of Interest Builder (RoIB) receives L1 trigger information from the RoIs and combines the information for the HLT supervisor
- The HLT supervisor can handle the input from the RoIB and manage the HLT farm of ~2000 machines at over 100 kHz



The ATLAS Dataflow

- The Data Collection Manager (DCM) handles all I/O on the HLT nodes, including RoI requests from the HLT and full event building
- The HLT processing tasks running the actual HLT algorithms are forked from a single mother process to maximize memory sharing
- The data loggers are responsible for saving accepted events to disk, and sending the files to CERN permanent storage infrastructure (EOS)

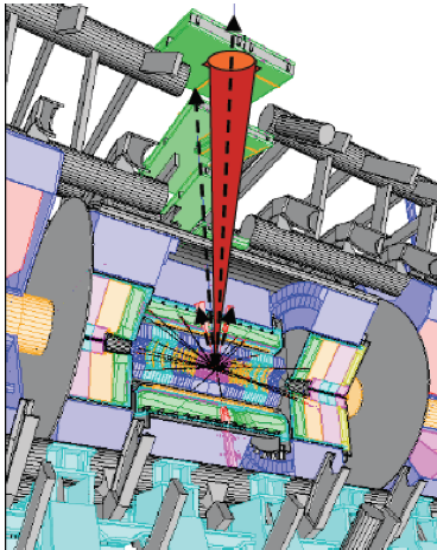


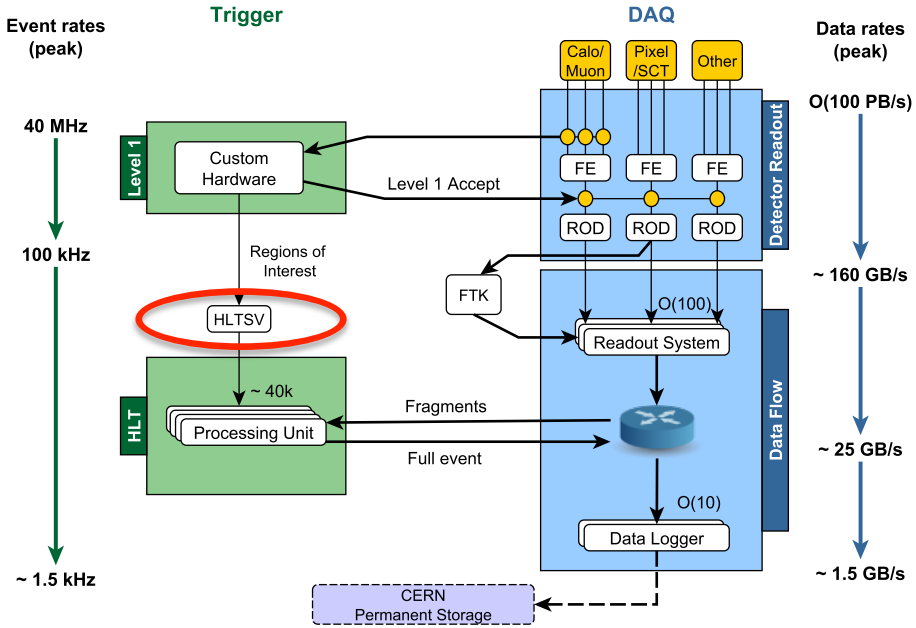
Region of Interest Builder (RoIB)

Guide the ATLAS High Level Trigger (HLT) by

- Taking raw event fragments from the hardware based Level 1 trigger sources
- Assembling all fragments into a single RoI record at 100 KHz over 12 links
- Sending the RoI record to the HLT Supervisor (HLTSV)

The HLTSV uses the RoI record to assign the event to an HLT node

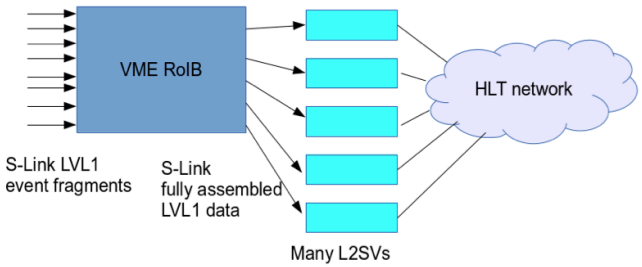




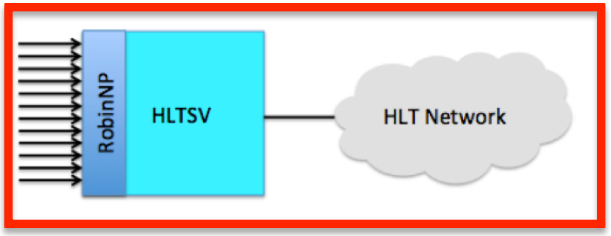
RoIB Evolution

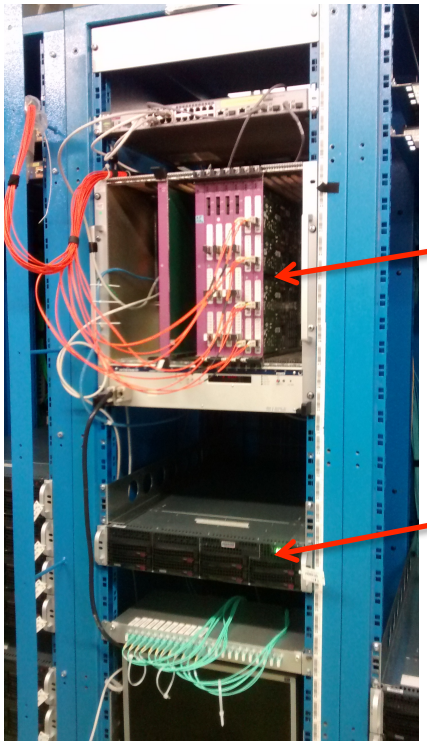
- Migrating the functionality of the custom designed 9U VME RoIB from multi-card VME based into a custom PCI-Express card in a commodity PC
- Allowing the use of more channels and the increase of the RoI record size at higher readout rates

Run 1



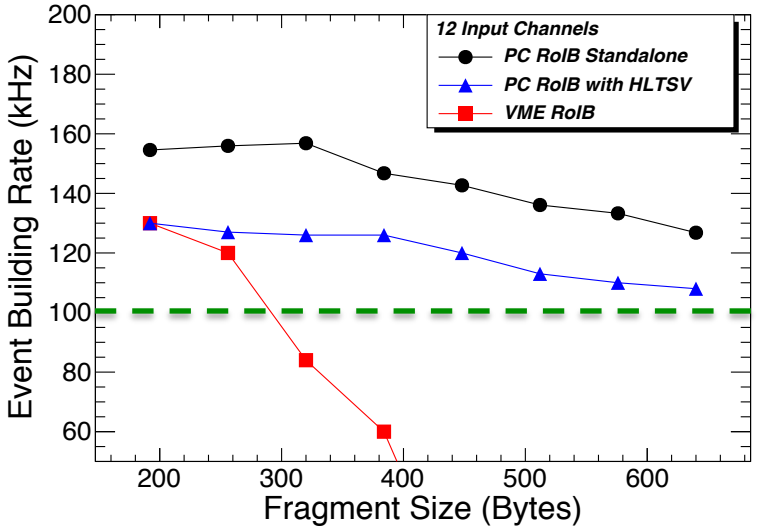
Run2

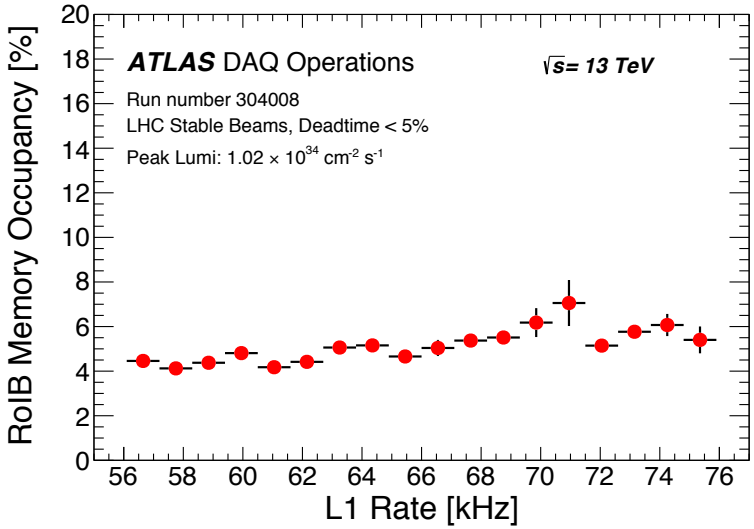


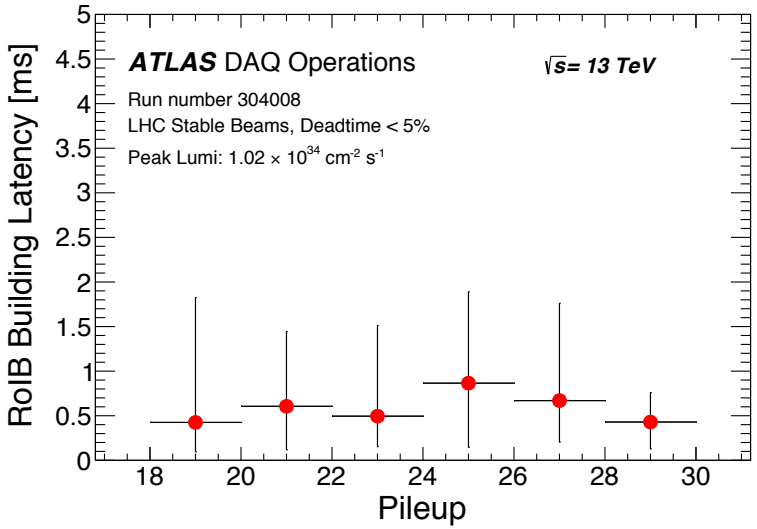


RoIB Performance

- The PC RoIB can sustain a rate of over 100 kHz with realistic ATLAS running conditions
- The HLTSV memory usage of the PC RoIB is at the level 5% level
- The RoIB event assembly does not depend on pileup conditions



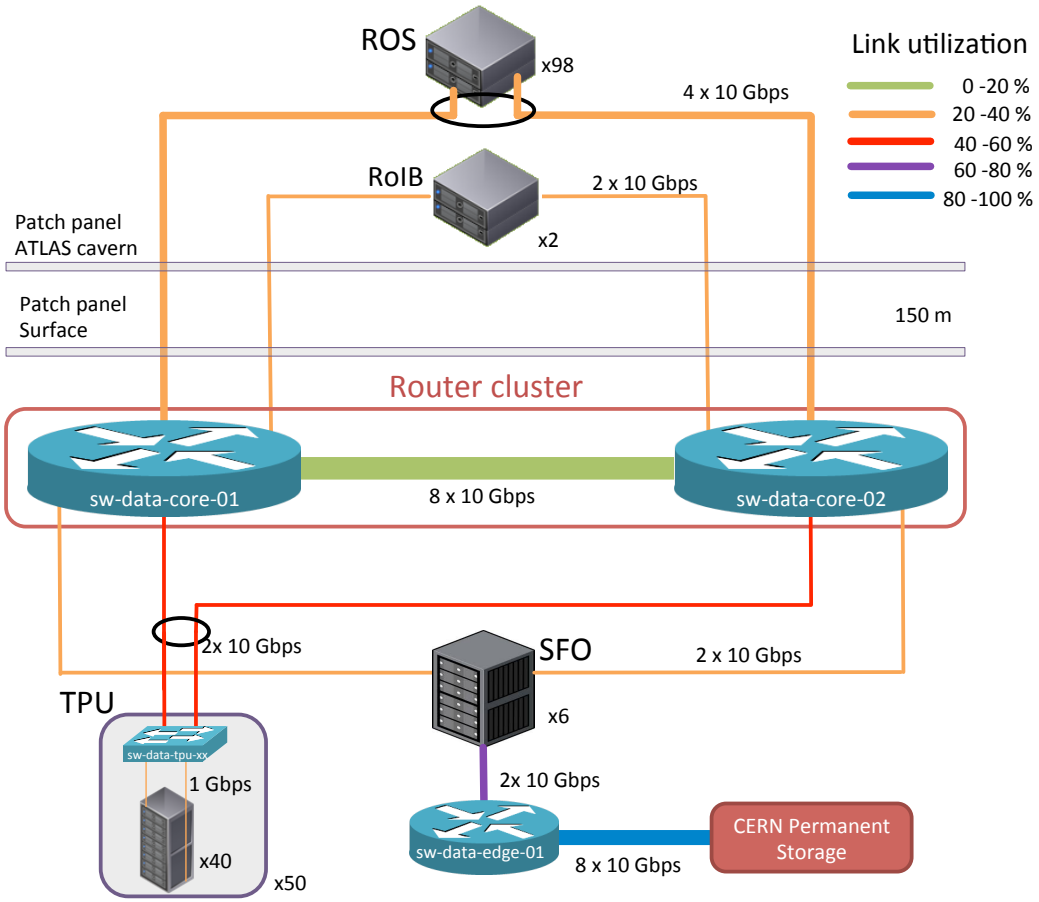




Dataflow Network

- The dataflow network was upgraded and simplified with emphasis on **redundancy**
- The network can handle a larger data volume
- All duplication is for redundancy at every level in case of link/switch failures
- A single network is used for RoI based access, event building, and sending data to the data logger (SFO)

The implemented redundancy allows the network to run without on call expert

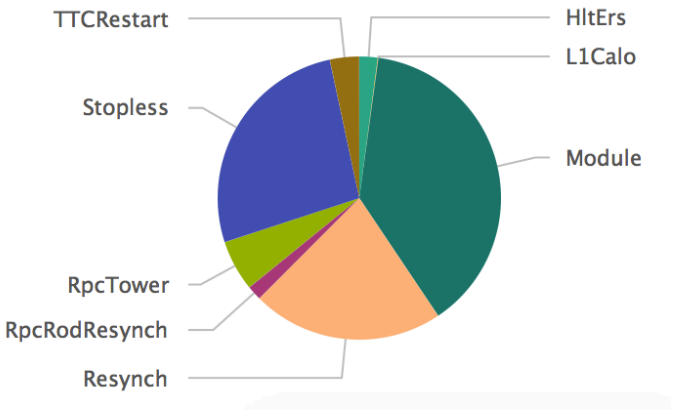


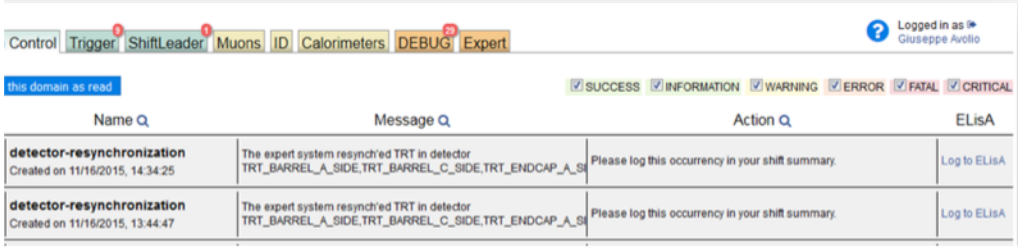
Software

- Switching all software suite to exclusively 64 bit operation
- I/O of the dataflow is based on asynchronous communication using the Boost::ASIO library, while CPU consuming operations are executed on a thread pool that uses Intel Threading Building Block technology for concurrent access to data structures

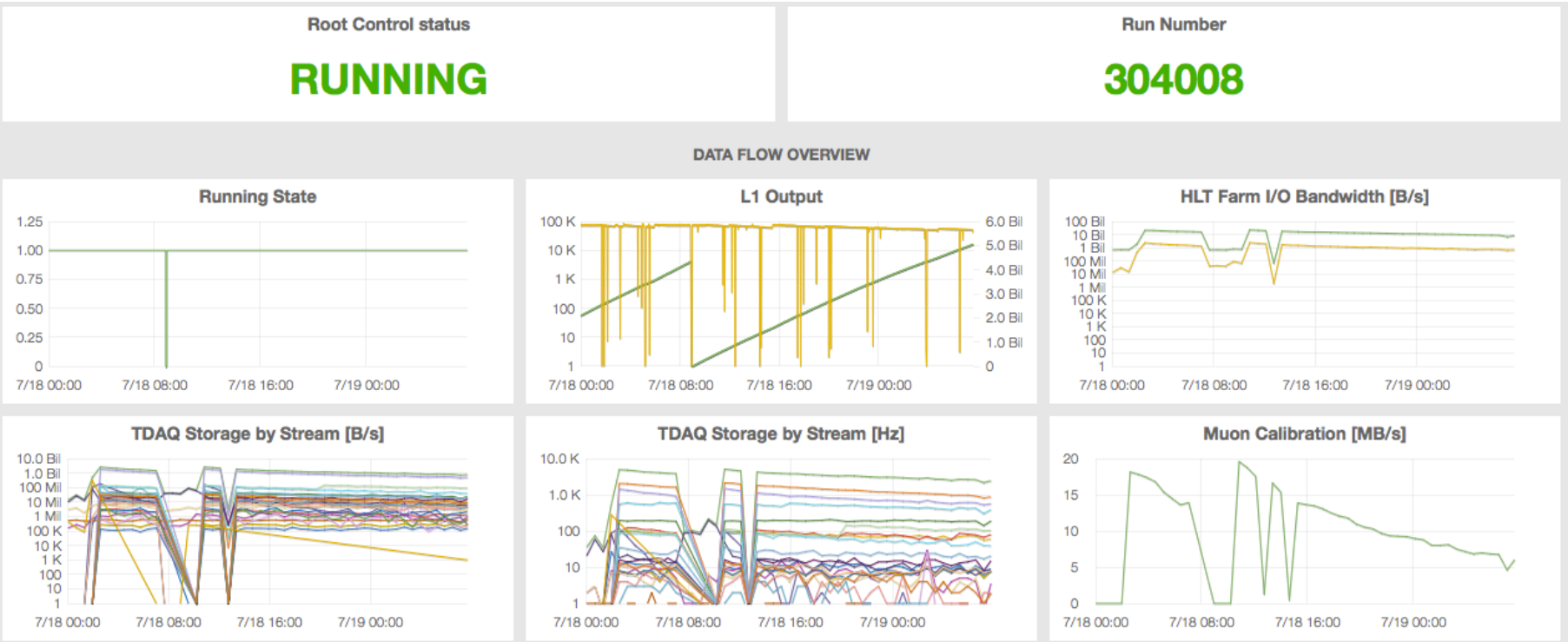
New software tools to support operations

- Central Hint and Information Processor (CHIP): supervises the DAQ system, reacts to detector conditions, and automates procedures
- Shifter Assistant: uses the same engine as CHIP (Complex Event Processing) with a focus on problems requiring human intervention
- P-BEAST: shares operational data produced during the data taking sessions in the form of time series database



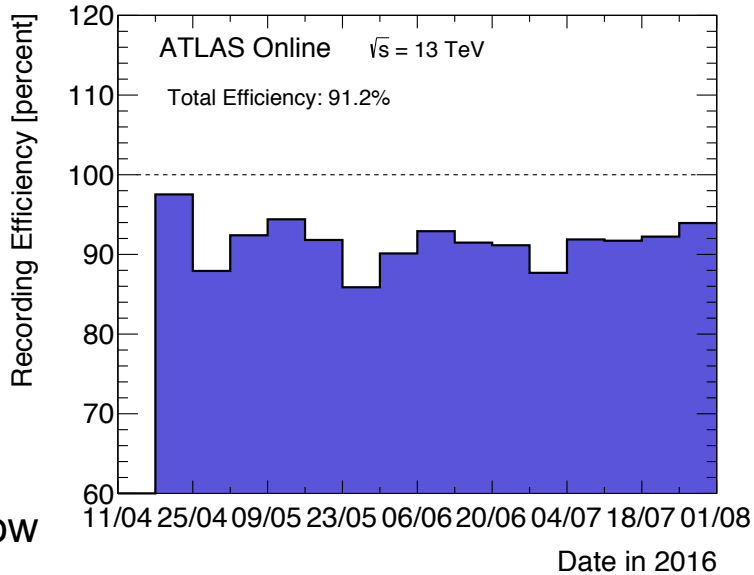


Operational Time Series (P-BEAST)

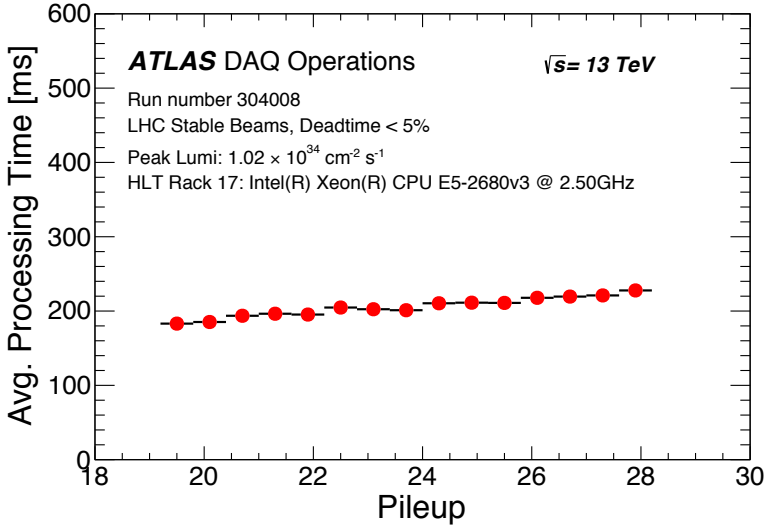


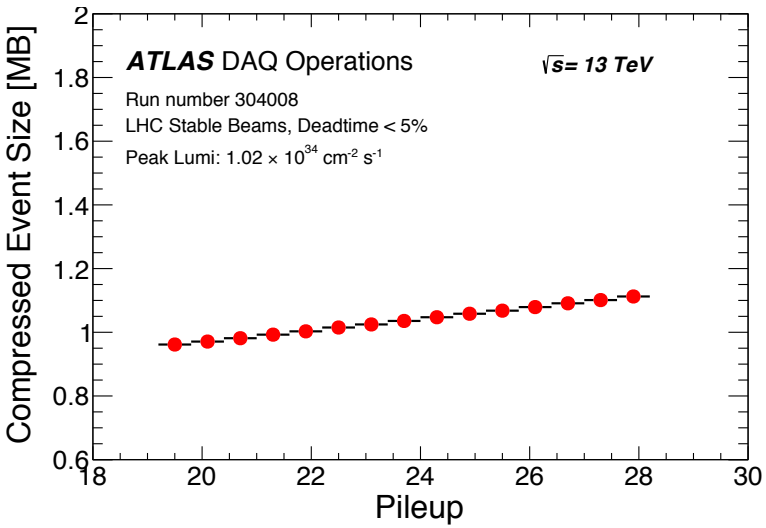
Run 2 Performance

- The 2016 ATLAS data taking efficiency is over 90%
  - Negligible fraction of data loss from the DAQ system
- DAQ system is scaling well with increased instantaneous luminosity
- System capable of handling larger pileup and thus larger event size



The evolution of the average processing time per event and the event size as a function of pileup are shown below





Summary

- Considerable simplification of the ATLAS Dataflow during Run 2 with improved performance and added stability
- Upgraded DAQ system hardware and re-written software to use modern and improved designs
- Enough headroom in performance as the instantaneous luminosity increases (event size, pileup…)
- ATLAS DAQ will continue delivering physics data with high efficiency

Thank you for your attention!

More ATLAS TDAQ talks:

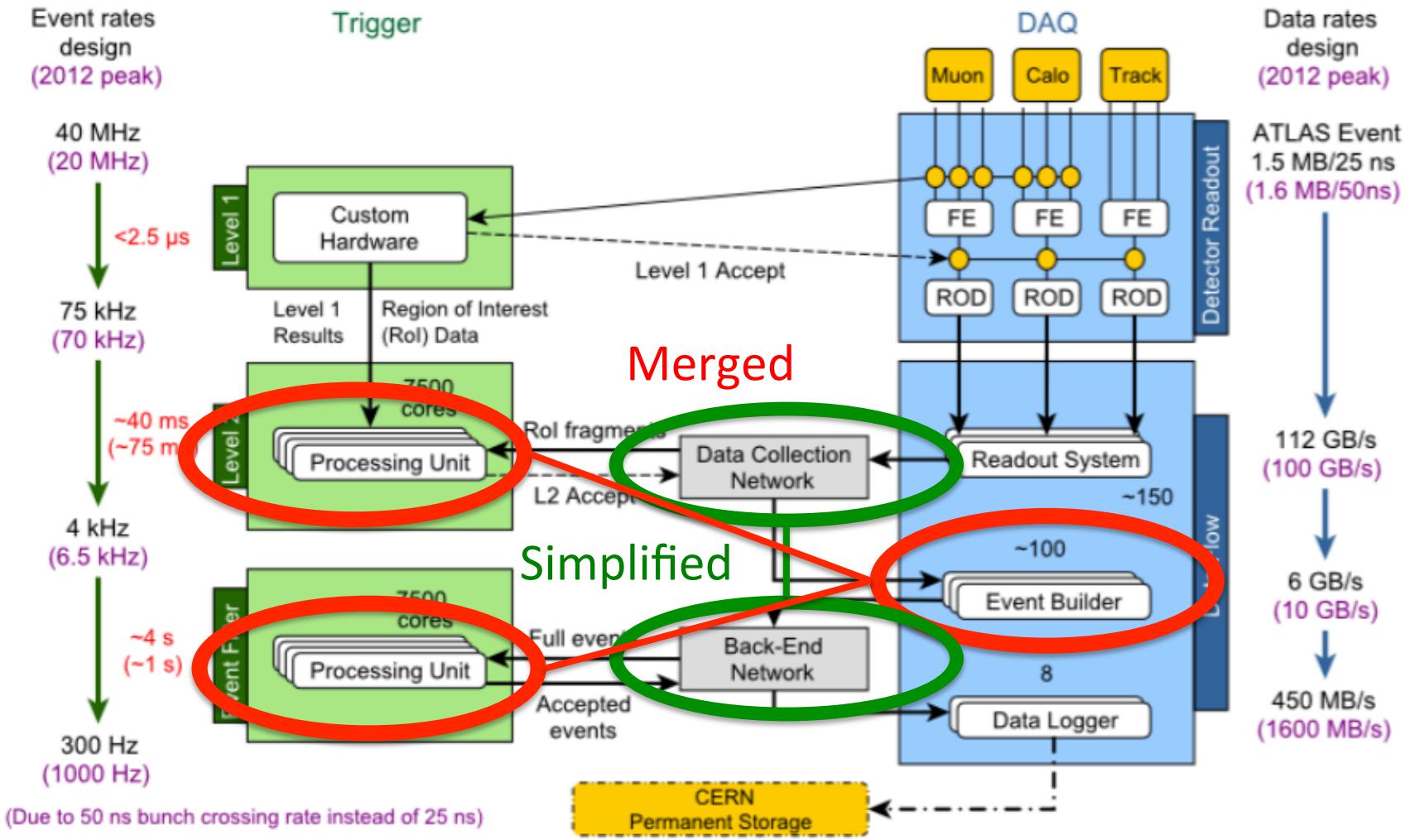
- The ATLAS Run 2 **Trigger**: Design, Menu, Performance and Operational Aspects by J. M. Miguens
- The Upgrade of the ATLAS **Electron and Photon Triggers** and their Performance by F. Monticelli
- Overview of the ATLAS **Fast Tracker** Project by L. Ancu

ATLAS TDAQ posters:

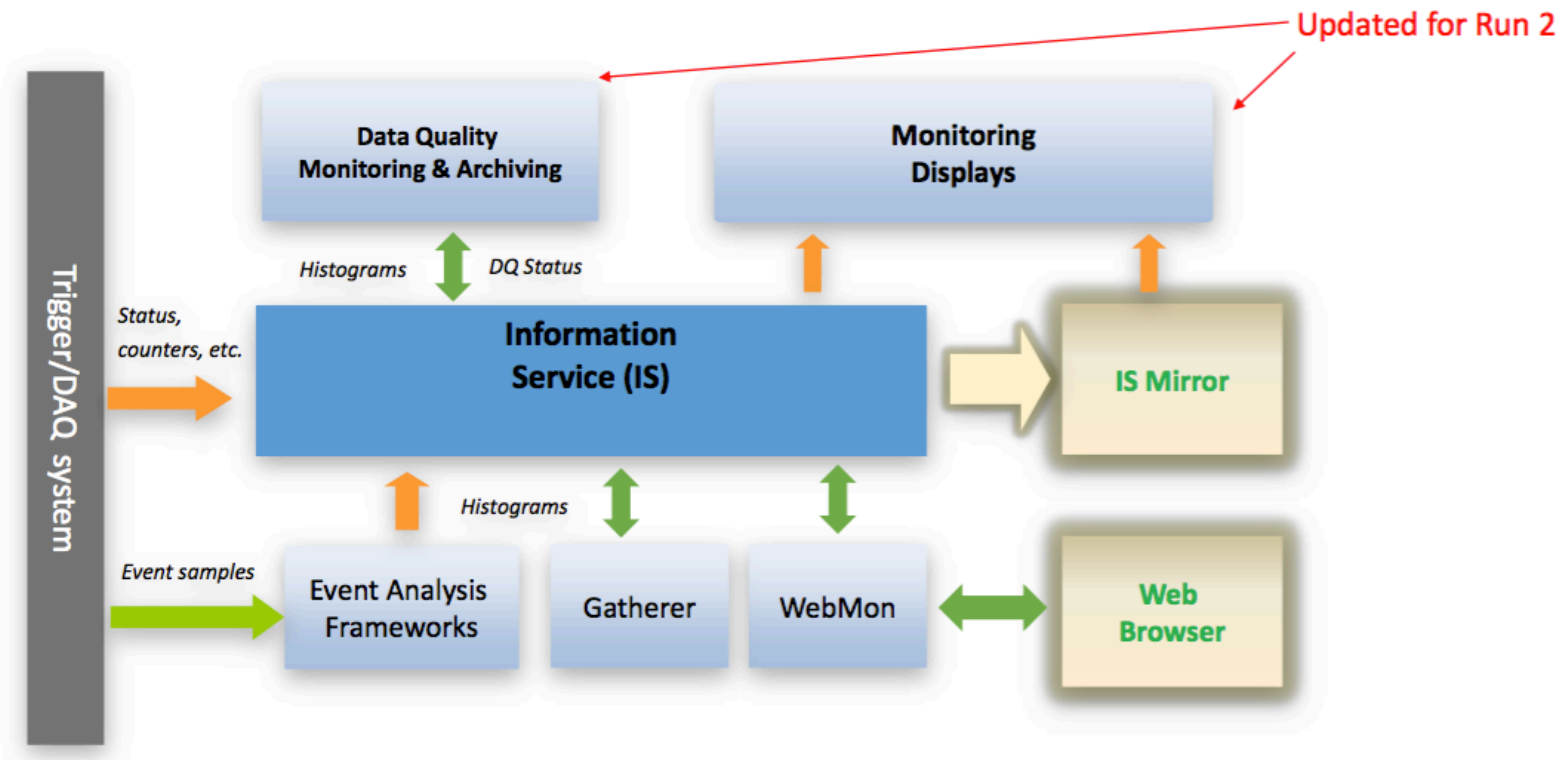
- ATLAS Trigger and Data Acquisition **Upgrades for High Luminosity LHC** by W. K. Balunas
- gFEX, the ATLAS **Calorimeter Global Feature Extractor** for the Phase-I upgrade by D. Miller
- Performance of the ATLAS **Tau Trigger** in Run 2 by G. Besjes
- Real-time **flavour tagging** selection in ATLAS by J. Allison
- ATLAS **jet trigger** performance in 2015 data by C. Herwig
- The design and performance of the ATLAS **Inner Detector trigger** for Run 2 by F. Miano
- The ATLAS **Fast Tracker** Processing Units - track finding and fitting by K. Krizka
- The ATLAS **Fast Tracker** Processing Units - input and output data preparation by A. Bolz

Backup

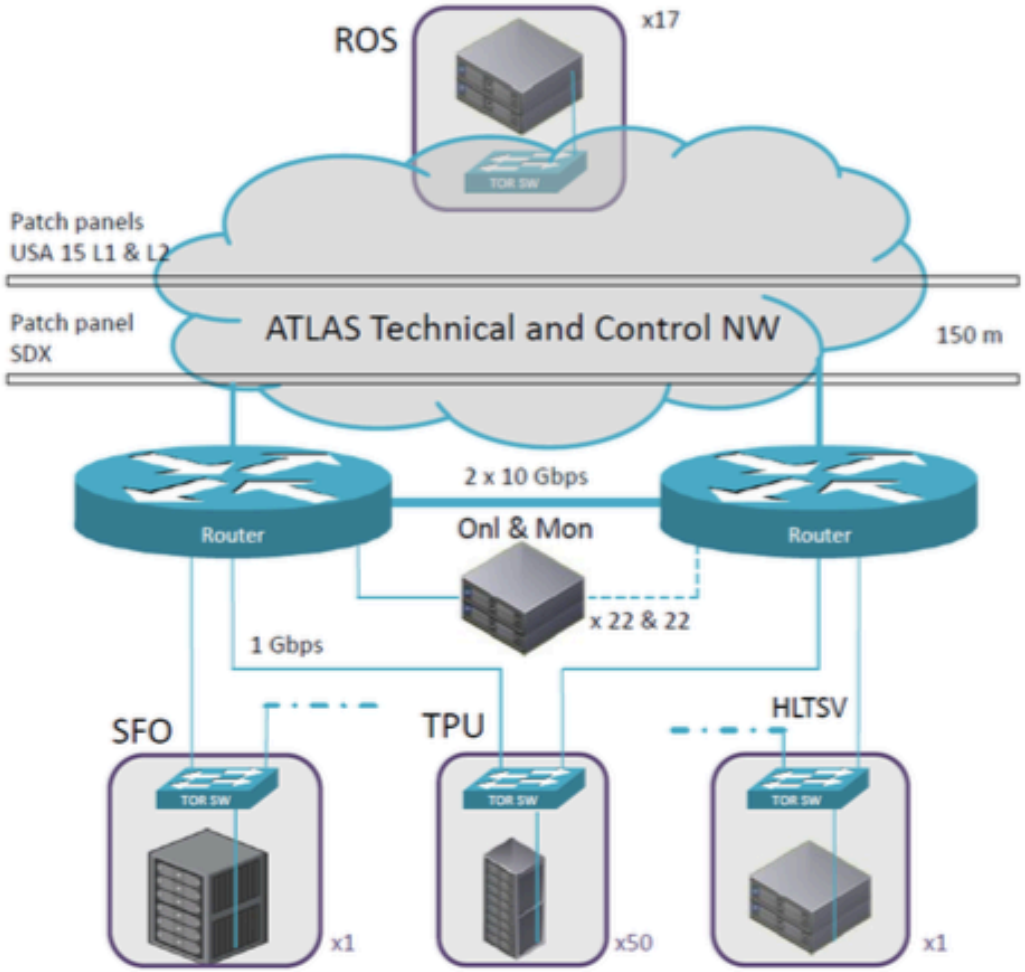
ATLAS TDAQ in Run 1



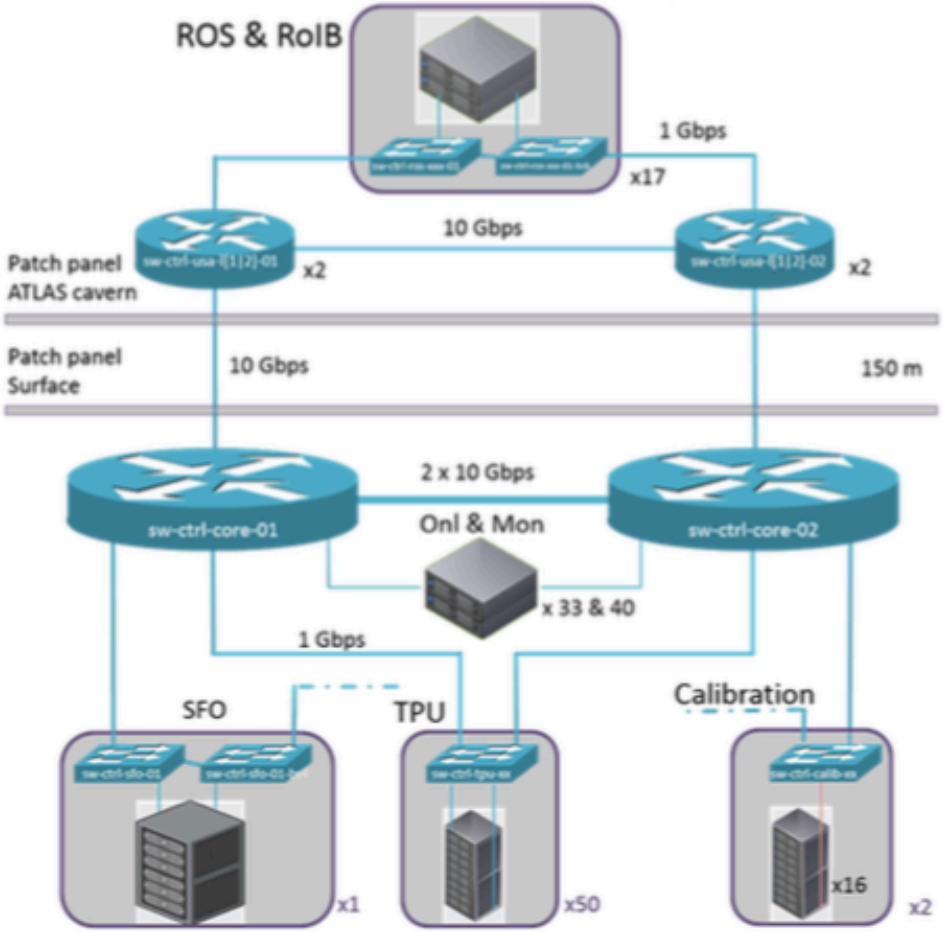
Online Monitoring



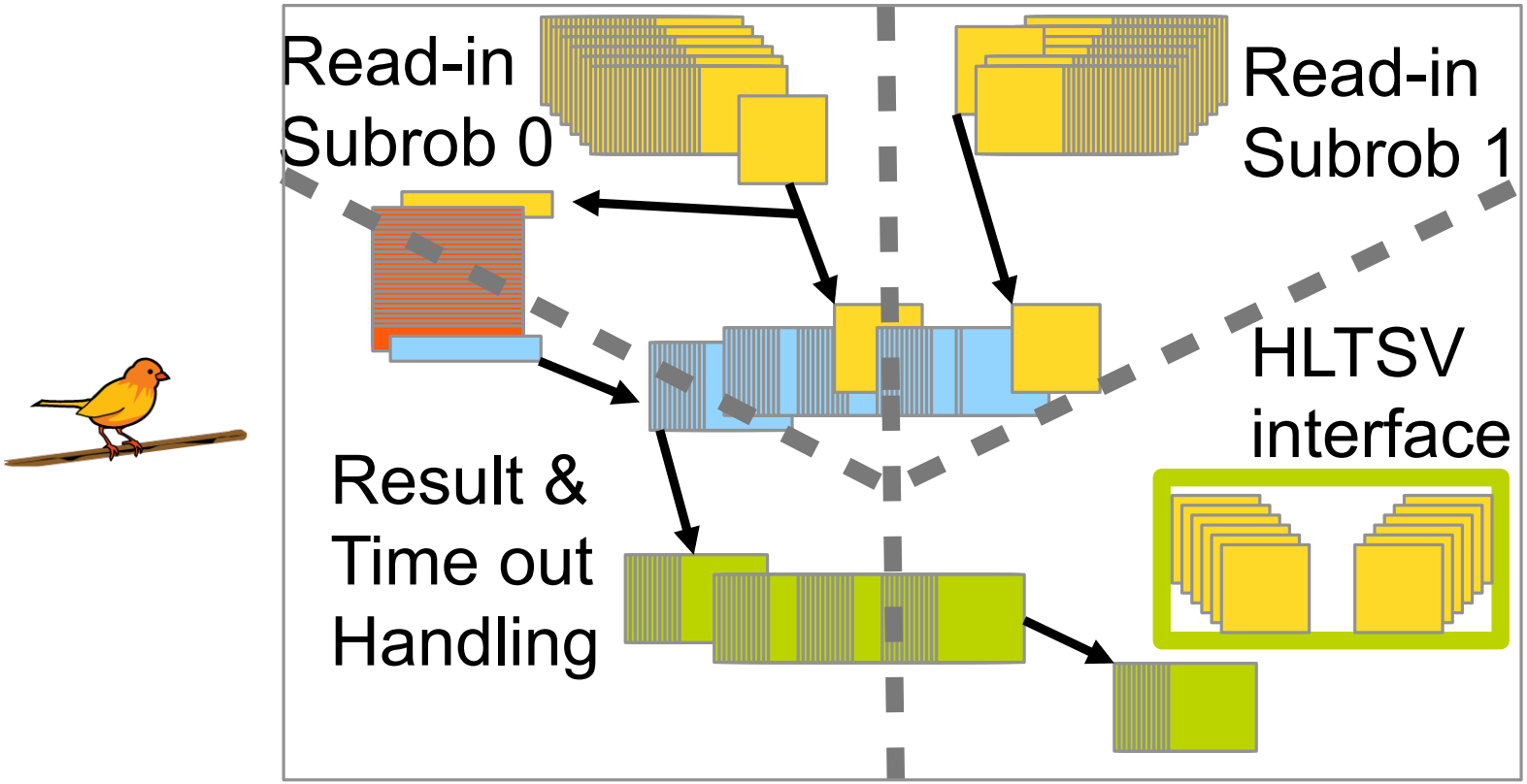
Control Network in Run 1



Control Network in Run 2



PC-RoIB Implementation



- Parallel threads handle data input to the system, event building, event checking, interface to the HLTSV, and publish monitoring information
  - Intel C++ tbb library used for concurrent data objects to minimize contention
- Time out events after ***1 s*** if missing one or more L1 RoI record (oks parameter)
  - 10,000x longer than VME system
- In the case of a timed out L1ID, the PC-RoIB suppresses fragments that arrive late if incomplete

RoIB RobinNP operation