The ATLAS Run 2 DAQ System

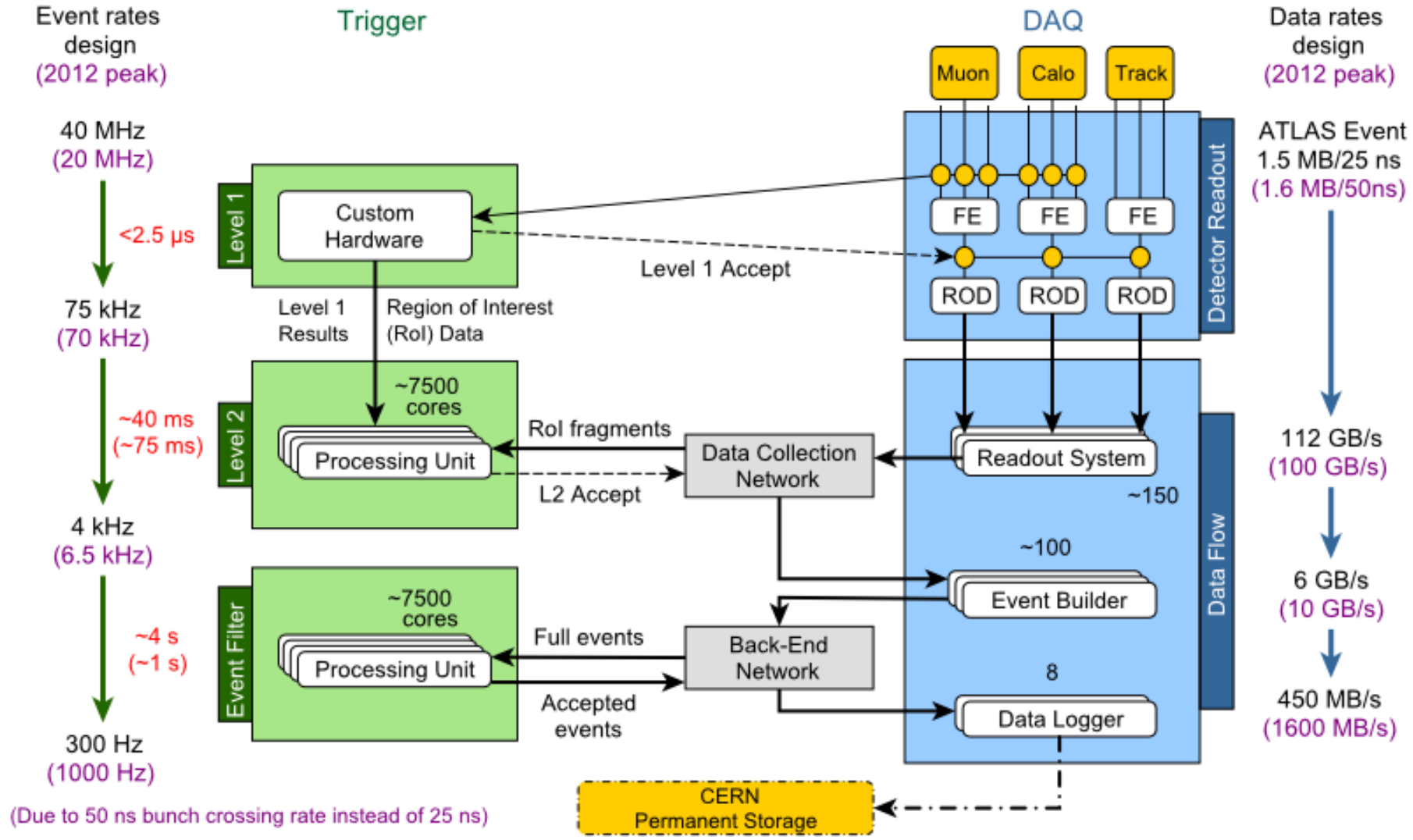
William Panduro Vazquez

2nd DAQ@LHC Workshop, Chateau de Bossey

Outline

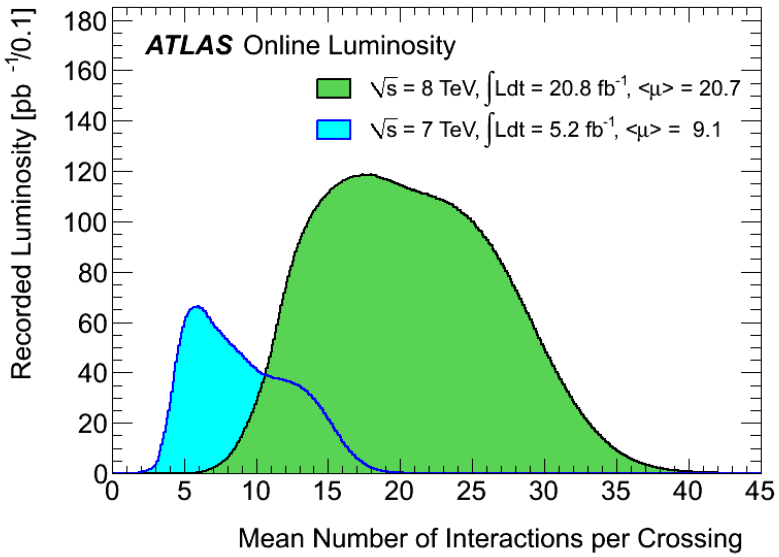
- Run 1 System Recap
- Review of drivers for evolution
- Focus on Individual Run 2 system components
- Run 2 System Summary
- Performance in 2015

ATLAS TDAQ System (Run 1)



Changing Conditions for Run 2

- Run conditions from 2015
  - Pileup *(multiple interactions per bunch crossing)*
  - Peak of 40 in 2012 *(luminosity $7x10^{33}$)*
    - Design: avg pileup of 21 for luminosity 10^{34}
  - For Run 2 estimate max average 55
    - Assuming 25 ns crossings
      - With 50 ns could reach 80
    - Peak value more significant for DAQ system
  - Potentially larger event size (up to 2 MB)
  - Longer HLT processing time
    - Increased buffering capacity required
- ATLAS DAQ requirements changes for Run 2
  - L1 trigger rate 100 kHz (70 kHz in 2012)
  - 20% increase in number of readout channels
    - Higher luminosity and energy requiring more channels to read out existing hardware
    - New detector components and trigger readout
      - Insertable B-layer (IBL) – new pixel detector layer
      - L1 topological trigger
      - In 2016: Fast tracker (FTK)
  - Average rate written to disc ~1 kHz
    - Run 1 reality becomes a requirement
    - Peak rates potentially much higher



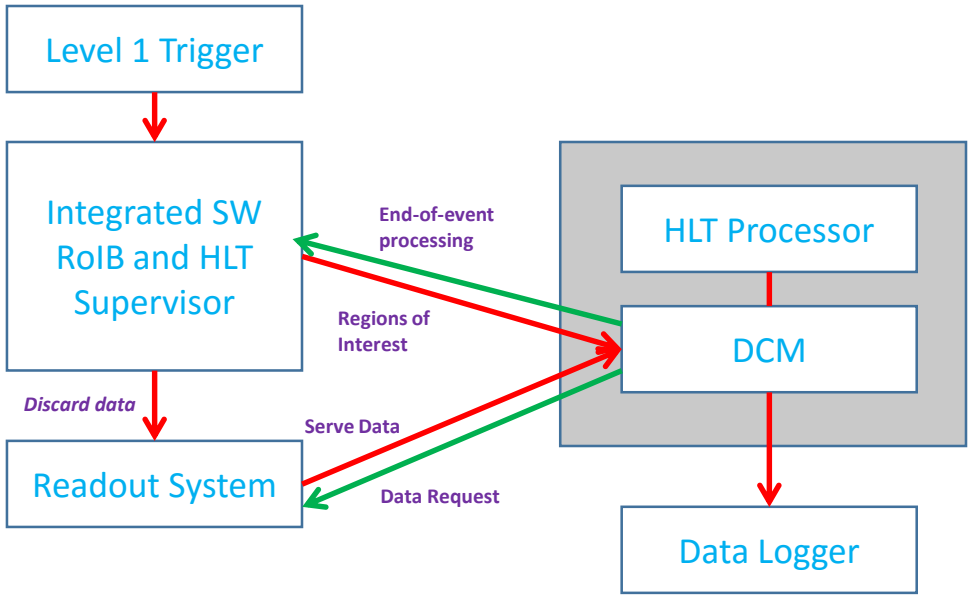
2012 Avg: 20.7

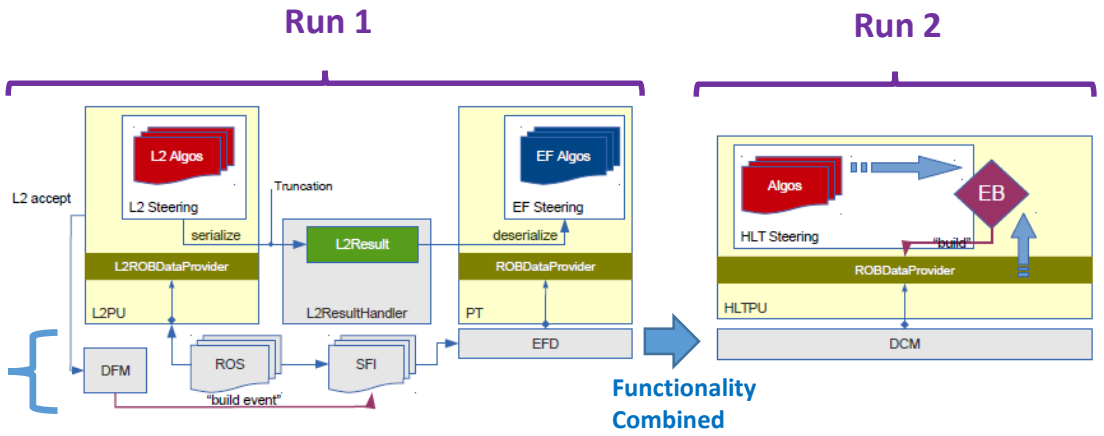
Evolution of Run 2 System Components

- Merger of Level-2 and Event filter processing into single HLT farm
  - New dedicated dataflow components
  - More details in talk by Reiner Hauser in next session
- Readout System (ROS)
  - New hardware and software
  - Significant performance improvement and reduction in system footprint
- Region-of-Interest Builder / HLT Supervisor
  - New software-based RoIB, built with new ROS technology integrated into new single HLT farm supervisor
- Data Logger/SFO
  - Improved data transfer capacity to permanent storage
  - New hardware and re-optimised multithreaded software
- Dataflow network
  - Complete redesign exploiting new technology for significantly improved resilience and redundancy
- Readout Crate Control Computers
  - Replacement of remaining obsolete 32bit machines with new 64bit version
- System Administration
  - New process control and monitoring software, rolling replacement of obsolete nodes
- General overhaul and optimisation of dataflow and operations software
  - More details on dataflow Reiner's talk

High Level Trigger Evolution

- Remove distinction between Level-2 (L2) nodes and Event Filter (EF) nodes
- Previous complicated dataflow and control replaced by Data Collection Manager (DCM)



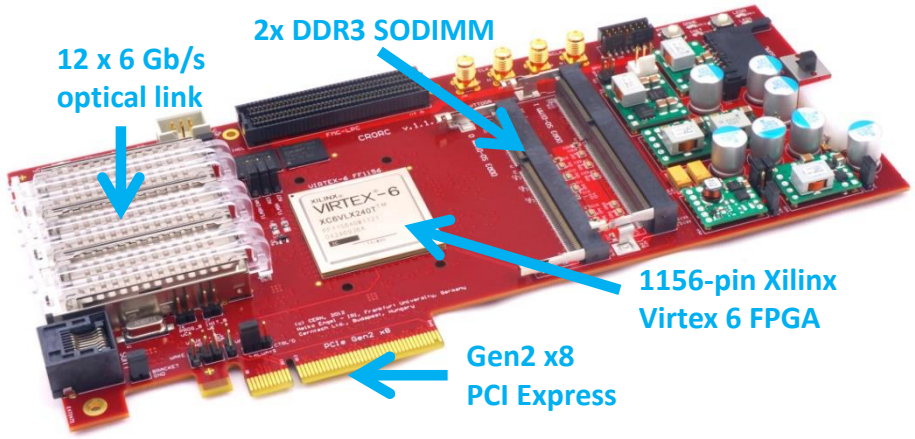


- Common nodes performing complete decision process
  - Optimised and more flexible event building
    - Eliminates previous limit of 7 kHz
- Major saving in bandwidth and processing time
  - No need to pack/unpack and transfer data between nodes
  - Event building no longer needs to duplicate requests for data that the HLT already holds through prior L2 request
- More advanced processing allowing simpler, more efficient dataflow
  - Better HLT node load balancing

Readout System Evolution

- Faster, smaller ROS PCs with higher link density
  - Buffers housed on new PCIe card: **RobinNP**
  - New ROS PCs will be half the height of the old
  - 2 RobinNPs per ROS PC (24 ROLs)
  - Link density 3x greater
    - Significant space, power and cooling saving
    - Much greater scope for later expansion
  - Significant output network capacity improvement (4 x 10 GbE)

ROS Features	Run 1	Run 2
Number of PCs	150	100
Max Number of PCs per Rack	10	15
Average Readout Links per PC	12	24
Max PC Input Bandwidth (MB/s)	1920	3840
Max PC Output Bandwidth (MB/s)	200	4000



- Hardware based on board developed by ALICE (Common Readout Receiver Card or 'C-RORC')
  - Custom ATLAS firmware implementing enhanced 'ROBIN' functionality

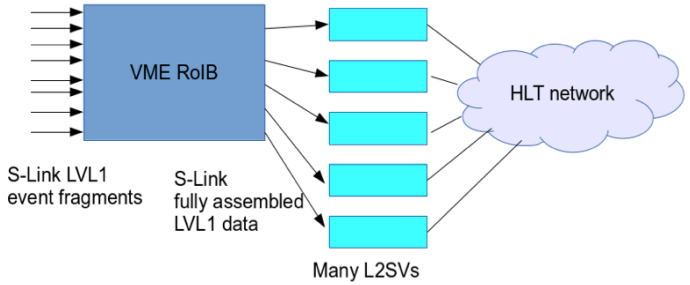
- Data management done by CPU of host PC
  - Original ROBIN tasks managed by on-board Power PC processor
  - Host processing easier to maintain, can upgrade CPU for performance boost

Card	ROBIN	RobinNP
Number of Readout Links	3	12
Total RAM on-board (MB)	192	8192
Memory per link (MB)	64	682
Max Input Bandwidth (MB/s)	480	1920
Output Bandwidth (MB/s)	264	1600 (3200 possible)

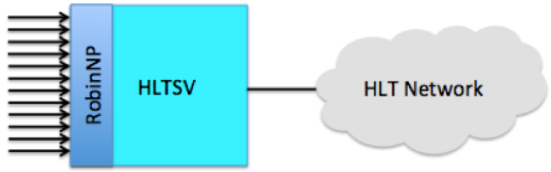
RoIB/HLTSV Evolution

- RoIB assembles Regions of Interest (RoI) from fragments produced by L1 sources (Central Trigger Processor, L1 Calorimeter Trigger, L1 Muon) and sends to HLT farm supervisor process (HLTSV) for checking and assignment to a processing node
  - More in Reiner Hauser's talk later today
  - HLTSV updated independently for 2015 running
- Run1
  - RoIB implemented on custom hardware in VME crate
  - Pre-farm merger a set of Level-2 Supervisor (L2SV) machines sent RoI's to L2 portion of farm for processing
- Drivers for evolution
  - RoIB
    - Ageing hardware - maintenance expertise dwindling
    - Limited bandwidth
    - Difficult to add new features/monitoring
  - HLTSV
    - Merger of L2 and EF farms
    - Opportunity to exploit new hardware and software standards
- New architecture replaces old RoIB and Supervisor machines with one new machine (actually a ROS PC containing RobinNP cards) for 2016
  - Redesigned RoIB software builds RoI's in software after receiving them over s-link via RobinNP
  - Integrated as a plugin to overhauled HLTSV process for sending on to the farm

Run 1



Run 2

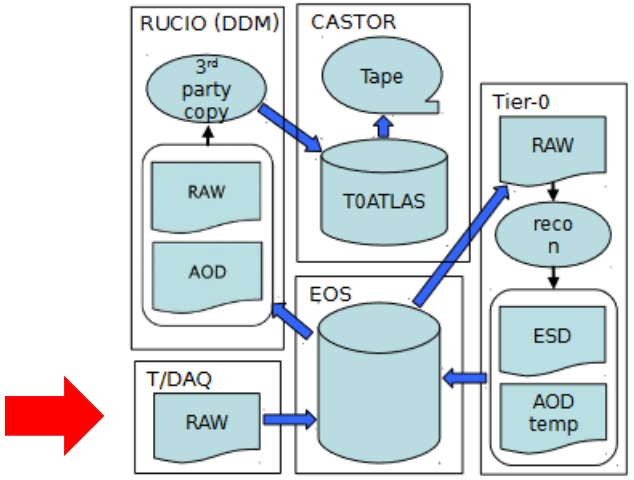


Run 1

Run 2

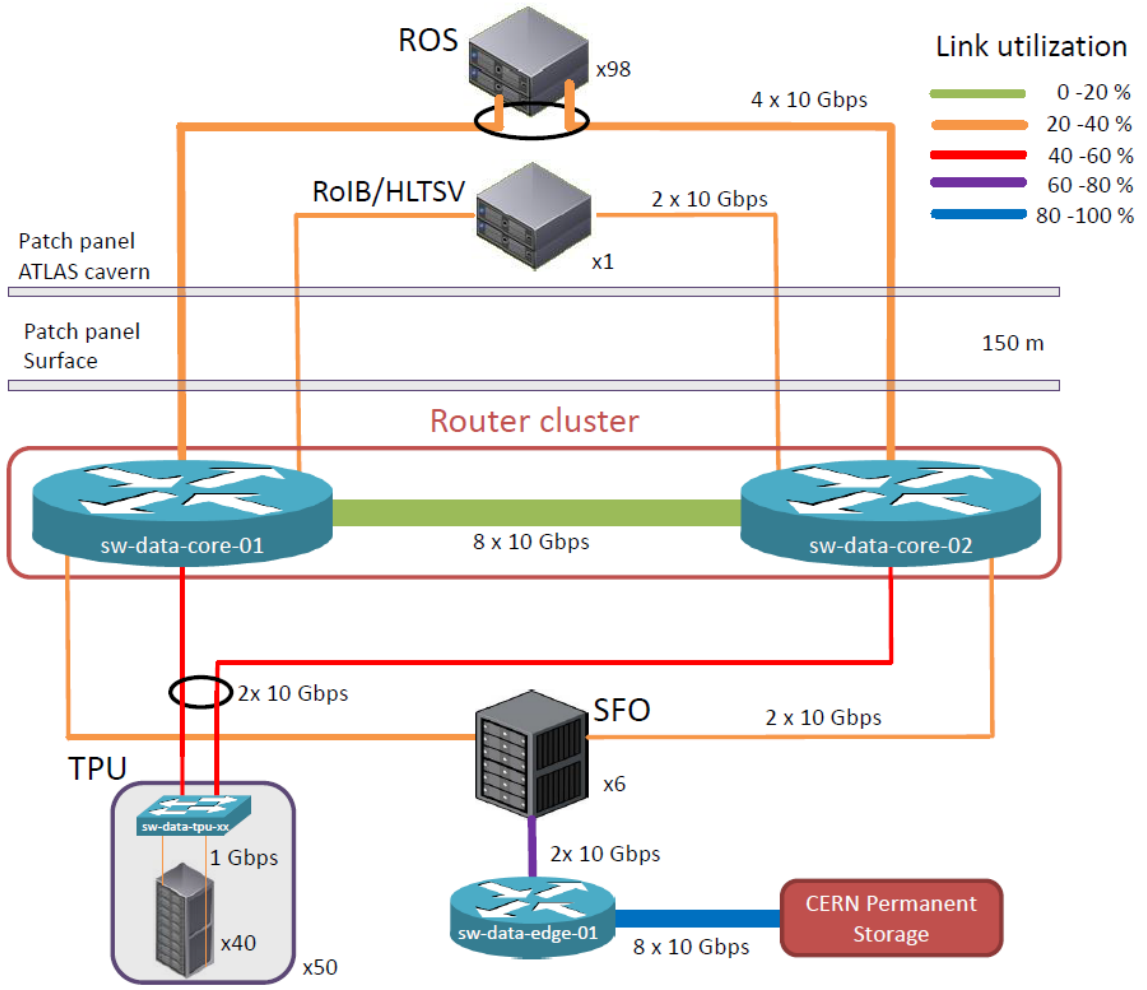
SFO Evolution

- Sub Farm Output (SFO) takes events accepted accepted by HLT farm and writes them to permanent storage
- Existing SFO machines replaced with new hardware during LS1
- Run 1 functionality mostly retained, but software re-written and multi-threading optimised
- SFO storage increased from 150 TB in Run 1 to 340 TB in Run 2
- New system capable of baseline output rate of 1kHz at 1.5 MB event size
- Full details in Reiner Hauser's talk later today



Data Collection Network

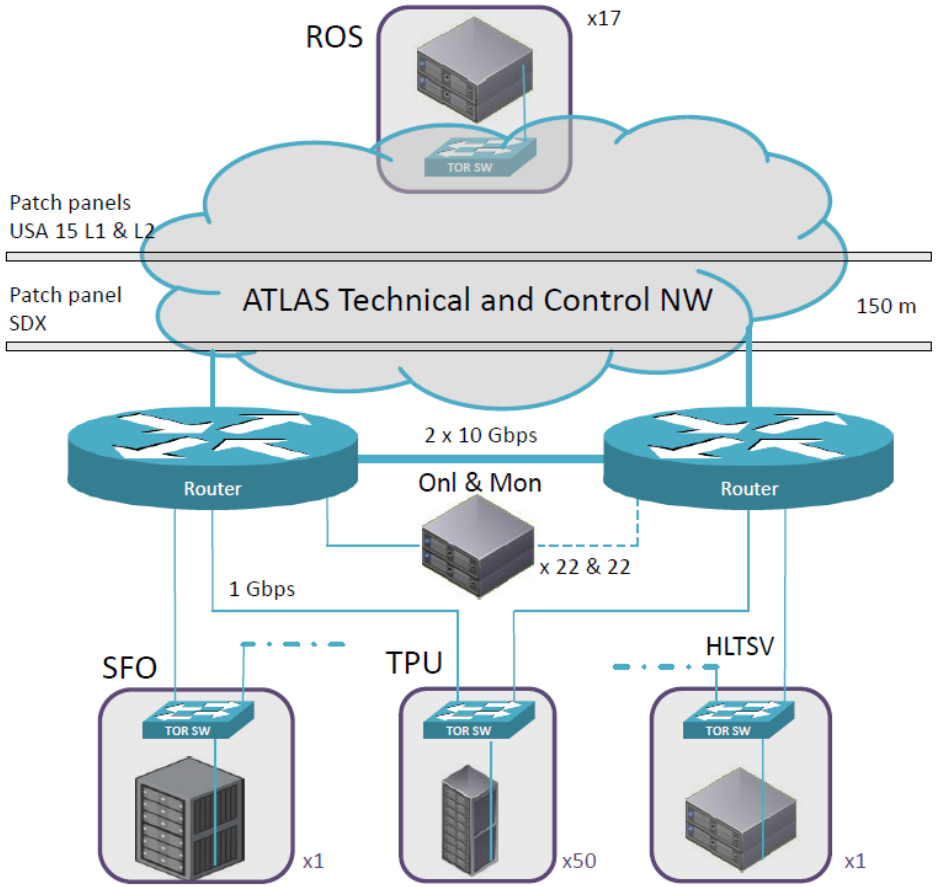
- Simple but performant
- Readout System PC's are directly connected to the routers
  - No need for aggregation
- Multi Chassis Trunking:
  - Brocade protocol for MLXe router clustering
  - Provides Active-Active redundancy
    - By sharing ARP and MAC tables the routers can route independently of each other
    - High capacity link between devices only to cope with HW failures
- Deep buffer Top-of-Rack switches
  - Significant improvement in EB time



Control Network

- Lower BW and performance needed
  - Majority of devices based on HP Procurve family provided by IT
  - For newer installations superseded by Brocade ICX family
- Active-Backup redundancy at all critical levels:
  - Core routers
  - Top-of-Rack switches (not TPUs)
  - Top-of-Rack switch uplinks
  - Infrastructure Host links
- Standard protocols and techniques for redundancy:
  - OSPF, VRRP, Active-backup bonding (Hosts)
- All the redundancy implemented allowed to run the network without on call expert

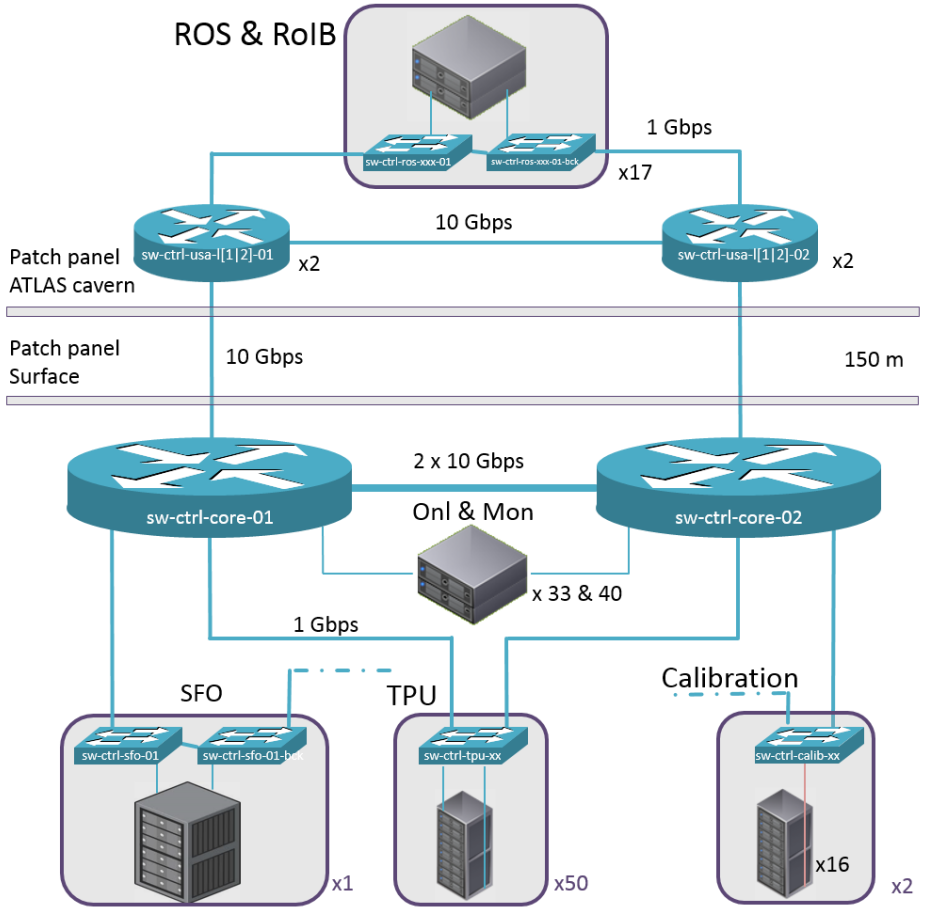
Run 1



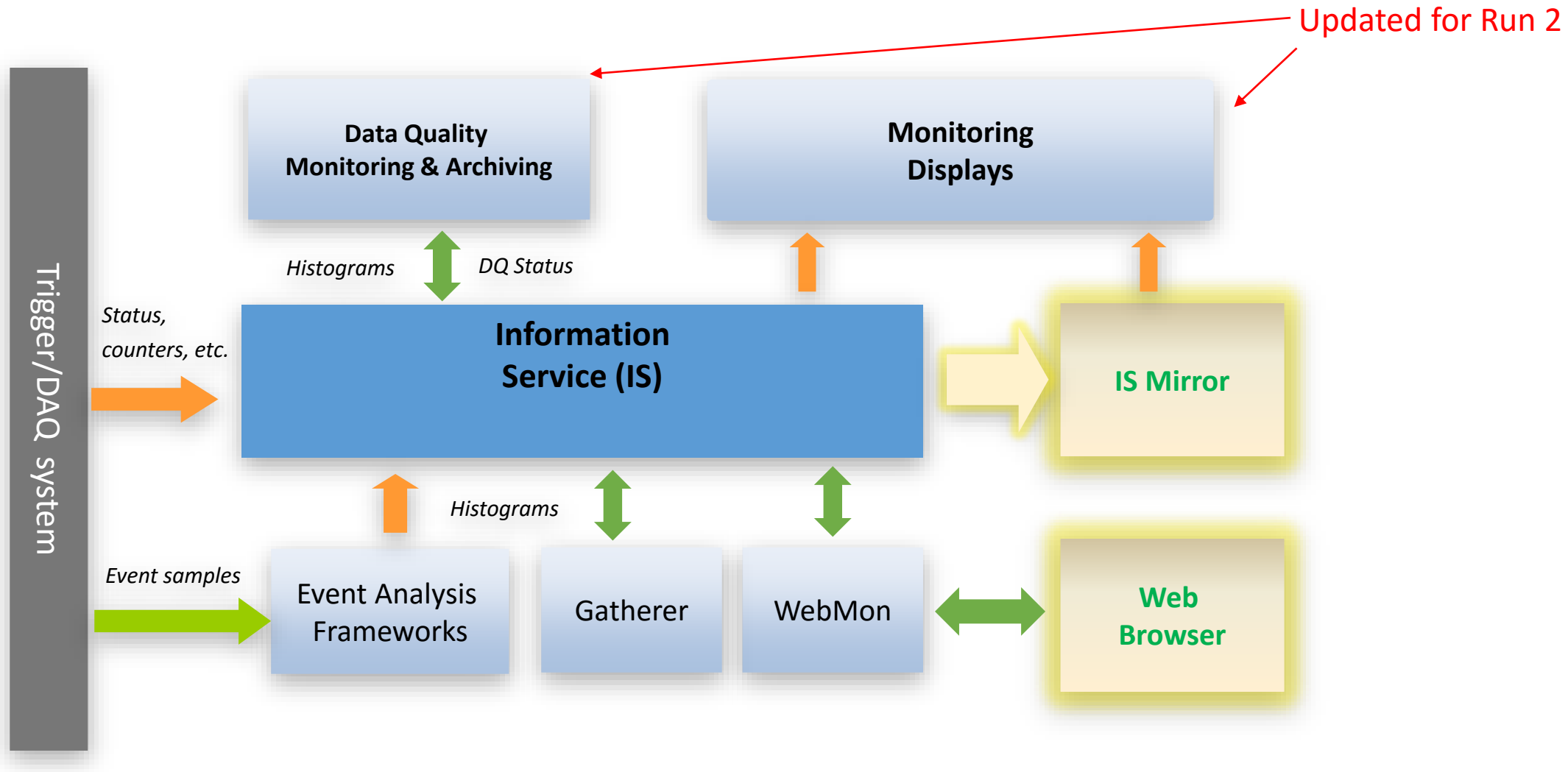
Control Network

- Lower BW and performance needed
  - Majority of devices based on HP Procurve family provided by IT
  - For newer installations superseded by Brocade ICX family
- Active-Backup redundancy at all critical levels:
  - Core routers
  - Top-of-Rack switches (not TPUs)
  - Top-of-Rack switch uplinks
  - Infrastructure Host links
- Standard protocols and techniques for redundancy:
  - OSPF, VRRP, Active-backup bonding (Hosts)
- All the redundancy implemented allowed to run the network without on call expert

Run 2

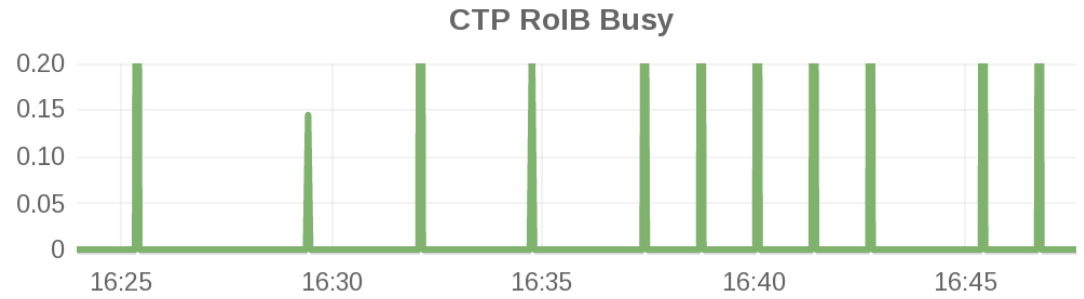


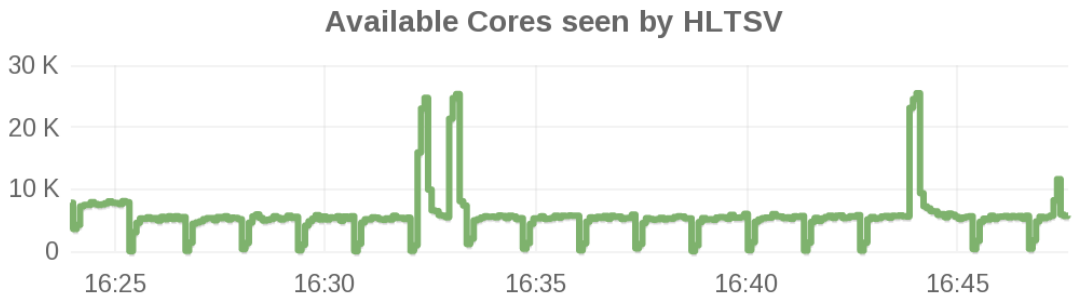
Run 2 Online Monitoring Architecture



Histogram Handling Improvements for 2016

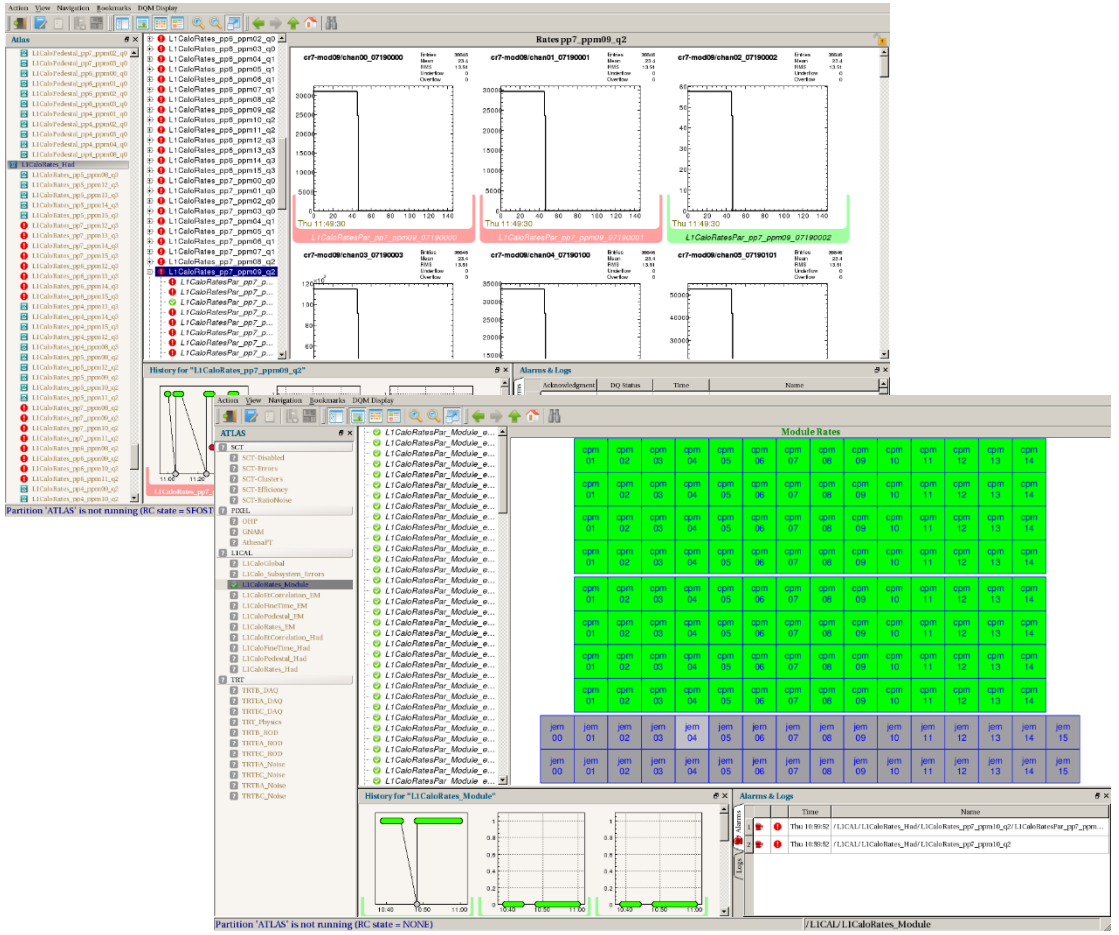
- Each HLT PU produces ~10K histograms
  - Total number of histograms~ 10K * 20K = 200M
- In Run 1 and the start of Run 2 histogram gathering was synchronized across the whole HLT farm
  - Affected HLT performance due to temporary shortage of CPU resources
- Histogram publication algorithms has been changed for 2016
  - Histograms split into groups
  - Groups are published sequentially with some delay between publications





Online Data Quality

- ~200 Monitoring applications analyzing ATLAS events in real-time:
  - Produce ~100K histograms
- Every minute ~ 70K histograms are automatically checked by DQ algorithms:
  - Histograms together with results are presented by a dedicated application



Readout Crate Control Computers

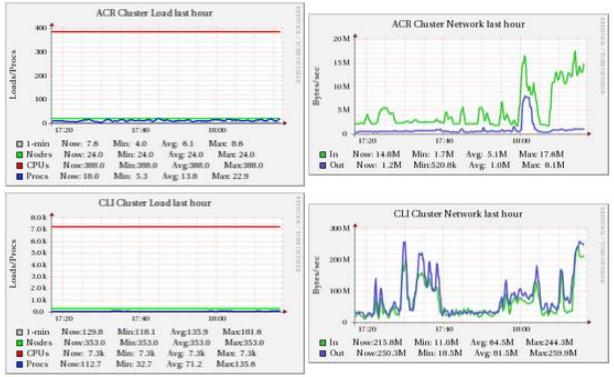
- Majority of detector RODS and associated readout hosted by VMEbus crates controlled by single board computers (SBCs)
- Major parts of system (Liquid Argon and Tile Calorimeters) migrated to 64bit machines during LS1
- Remainder migrated during 2015/2016 winter shutdown
- Had to select new model (Concurrent Technologies VPE-24) due to end-of-life of previous VME interface chip
  - 4-Core Intel Atom E3845 @ 1.91GHz
- ATLAS now running fully 64bit, working through issues related to switch
  - 32bit -> 64bit
  - Single core -> multicore
- Ready for 2016 data taking

Software

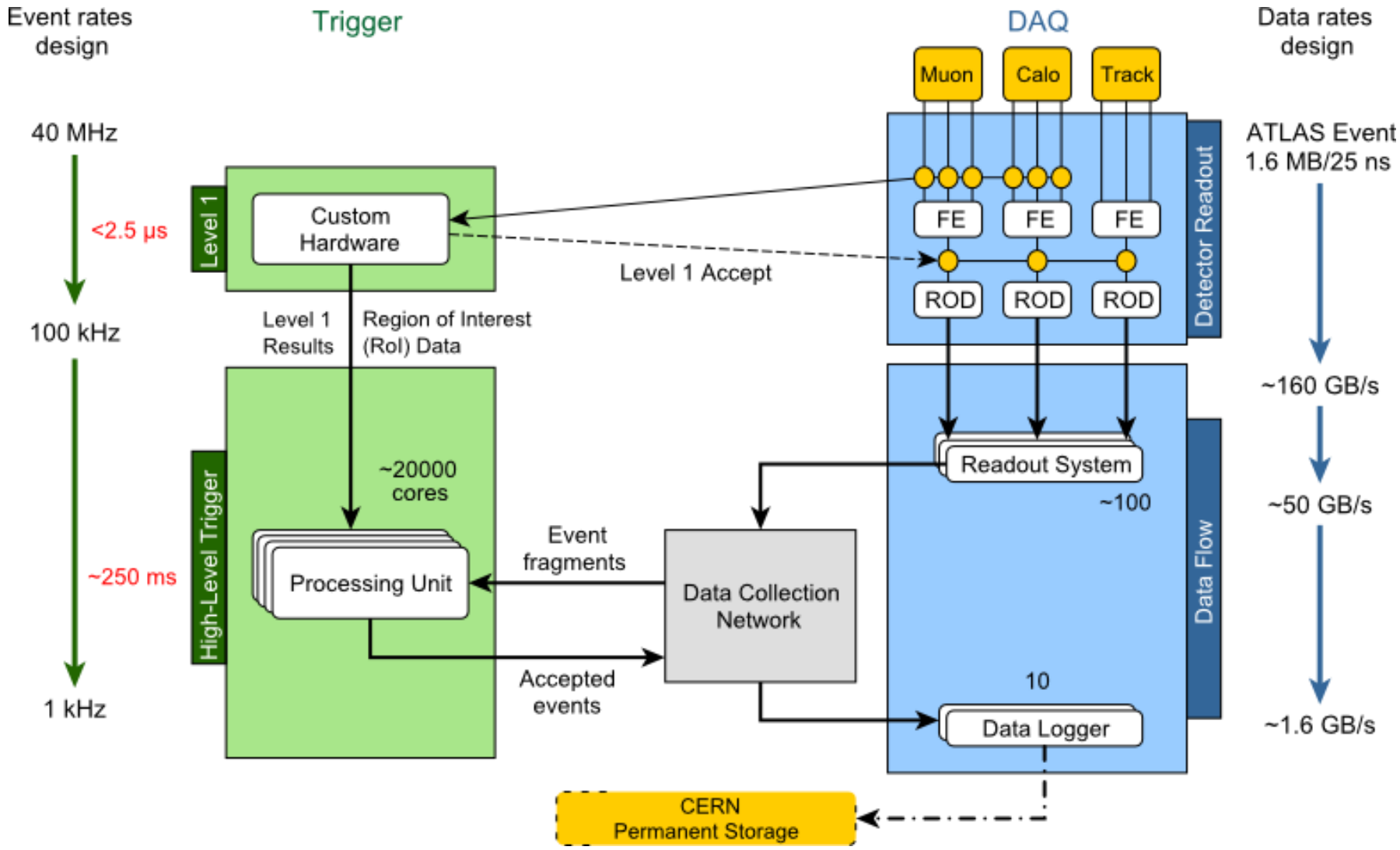
- Migration of complete software suite to exclusively 64bit operation
- Overhauled dataflow software introducing new standards such as boost::asio and improved multithreading (more details in Reiner's talk)
- New software to support operations
  - Central Hint and Information Processor (CHIP)
    - New 'expert system' to monitor and react to detector conditions
      - More details in talk by Giuseppe Avolio
  - P-BEAST operational monitoring archive and dashboard
    - Easy access to operational data time series
      - More details in talk by Igor Soloviev
  - Improved shifter tools
    - Shifter Assistant – provide web interface with advise and instructions to shifters in response to detector and DAQ system conditions (also in Giuseppe's talk)
    - Overhauled DAQ Summary web interface providing easy access to top level run parameters and rates in one page

System Administration

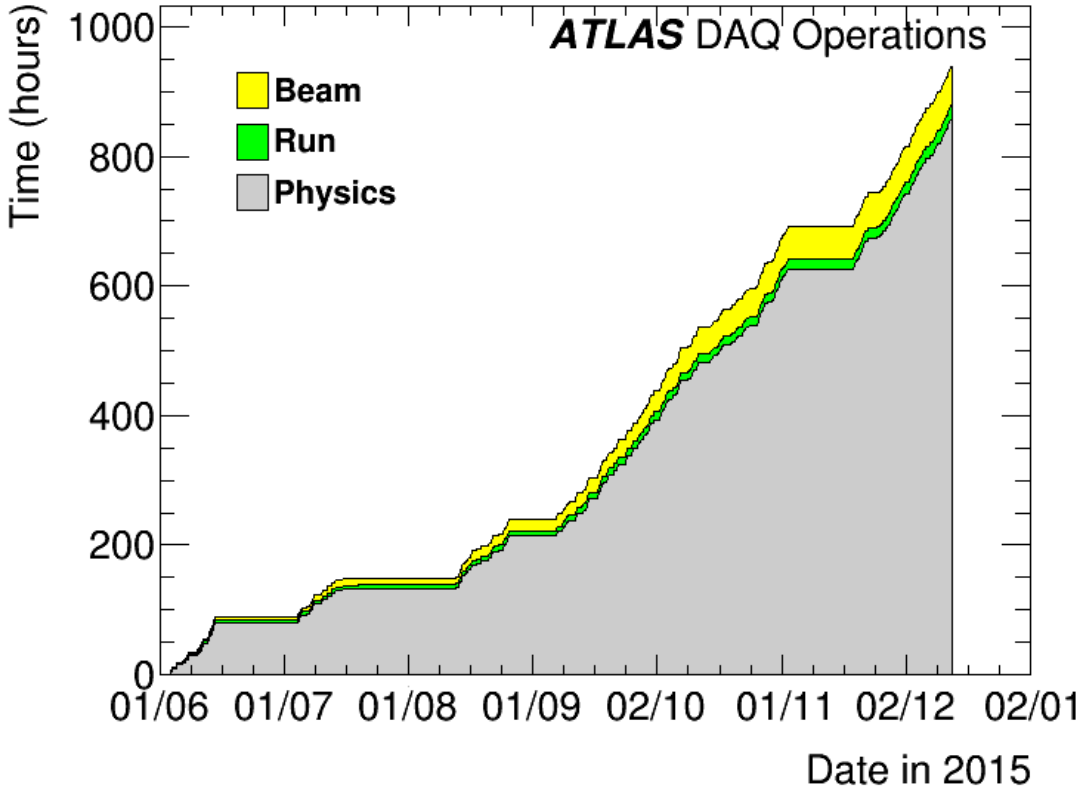
- Hardware
  - Rolling replacement of hardware (HLT nodes, data logger, readout system, crate control single board computers, general service nodes)
  - New ‘Calibration’ Farm set up to support subdetector standalone tests, recycling old event builder hardware
  - Replacement of NetApp filer system
    - SSD cache layer provides better IOPS performance
    - larger disks provide more usable space
    - Both performance and space appear well sufficient for foreseeable needs of Run2
- New Virtual Machines set up for TDAQ applications (e.g. Shifter assistant replay, Splunk)
- Monitoring
  - Migration to Icinga 2 and Ganglia for performance data
  - Native support for distributed and High-Availability operation
  - Much improved coverage of OS, HW, network and service health
- Node management
  - All nodes now fully managed by **Puppet**
    - Automatic prevention of incorrectly configured nodes being used in data taking
  - Migration to PuppetDB for central information storage
    - More flexible and performant than old ‘Dashboard’ interface



ATLAS TDAQ System (Run 2)



2015 Performance



- Overall efficiency during stable beams with detector in physics mode 92.4%
  - Equating to 857 hours of colliding beam data
- System tested successfully up to 100kHz L1 rate
  - Not reached in colliding beam due to Pixel detector protection rate veto
  - Expected to achieve it in 2016 once LHC fully populated with bunches
- Peak output bandwidth from Point 1 to EOS measured at up to 3GB/s (depending on streaming)

Summary and Outlook

- Many significant improvements to performance, stability and maintainability of DAQ system throughout LS1 and during 2015 running
- Good performance of system so far in Run 2, ongoing maintenance program to handle issues and continue to improve
- Many thanks to all who contributed!