ATLAS DAQ from Run 2 to HL-LHC

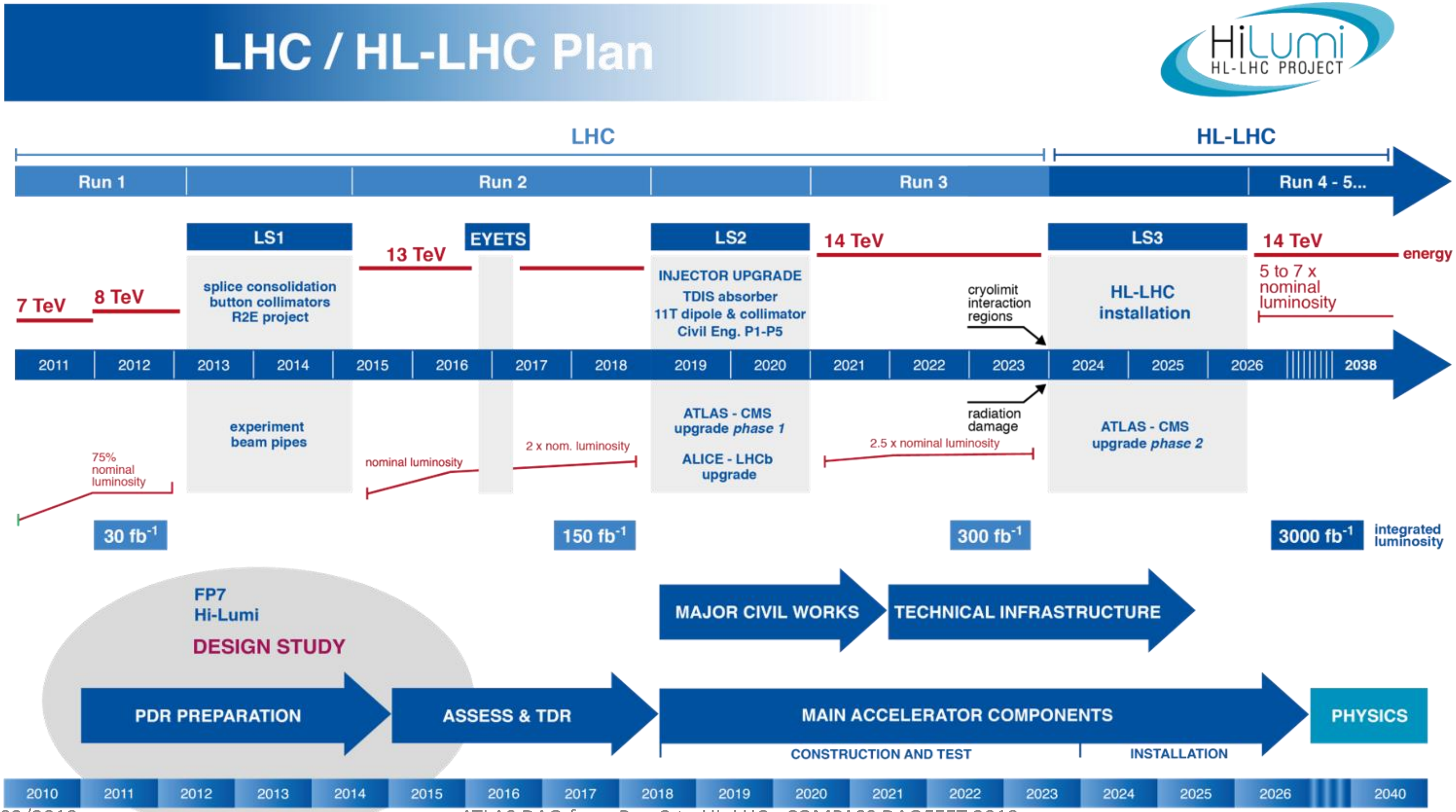
William Panduro Vazquez
On behalf of the ATLAS Collaboration

Overview

- LHC Schedule and Evolution
- ATLAS Trigger Strategy
- DAQ System in Run 2
- Common challenges and evolution
- DAQ System in Run 3
- FELIX and SW ROD
- Towards Run 4 and HL-LHC
- Summary

Talk will touch on Trigger and other ATLAS systems, but focus will be mainly from DAQ/Readout perspective

LHC Evolution and Overall Upgrade Schedule



LHC / HL-LHC Plan

HiLumi HL-LHC PROJECT

LHC

HL-LHC

Run 1

Run 2

Run 3

Run 4 - 5...

LS1

EYETS

LS2

LS3

7 TeV

8 TeV

13 TeV

14 TeV

14 TeV

energy

splice consolidation
button collimators
R2E project

INJECTOR UPGRADE
TDIS absorber
11T dipole & collimator
Civil Eng. P1-P5

cryolimit
interaction
regions

HL-LHC
installation

5 to 7 x
nominal
luminosity

2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2038

radiation
damage

experiment
beam pipes

ATLAS - CMS
upgrade *phase 1*

ALICE - LHCb
upgrade

ATLAS - CMS
upgrade *phase 2*

75%
nominal
luminosity

nominal luminosity

2 x nom. luminosity

2.5 x nominal luminosity

30 fb^{-1}

150 fb^{-1}

300 fb^{-1}

3000 fb^{-1}

integrated
luminosity

FP7
Hi-Lumi

DESIGN STUDY

MAJOR CIVIL WORKS

TECHNICAL INFRASTRUCTURE

PDR PREPARATION

ASSESS & TDR

MAIN ACCELERATOR COMPONENTS

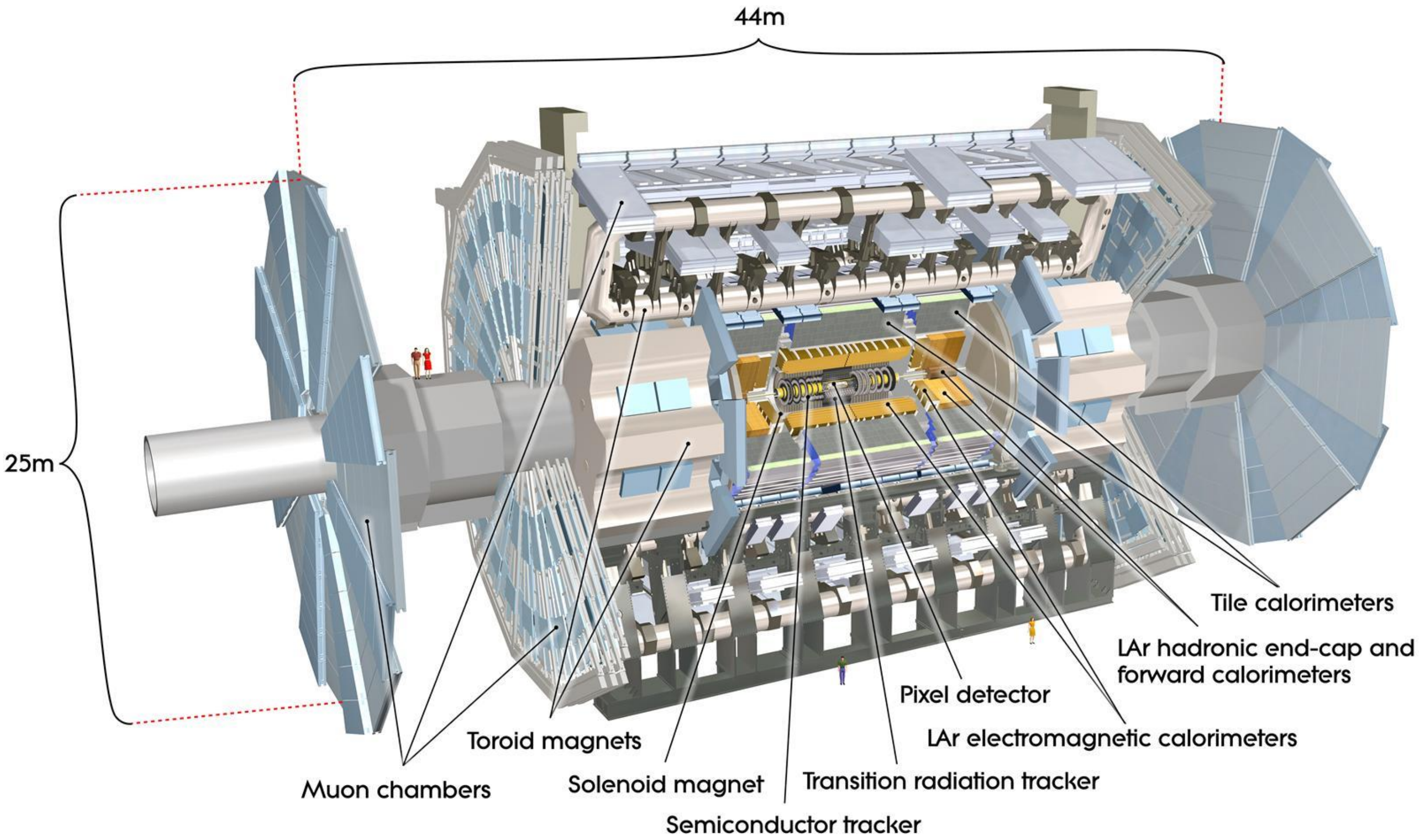
CONSTRUCTION AND TEST

INSTALLATION

PHYSICS

2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2040

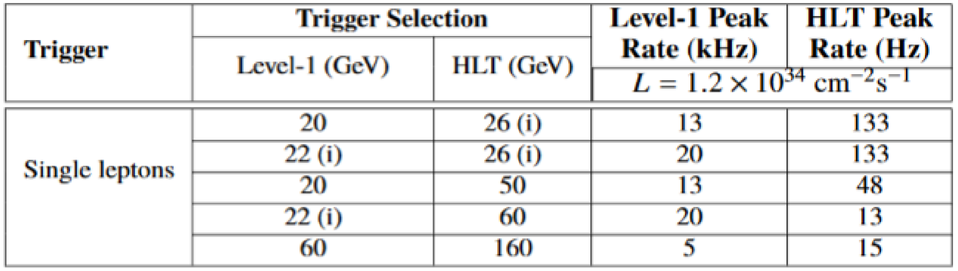
The ATLAS Detector


44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

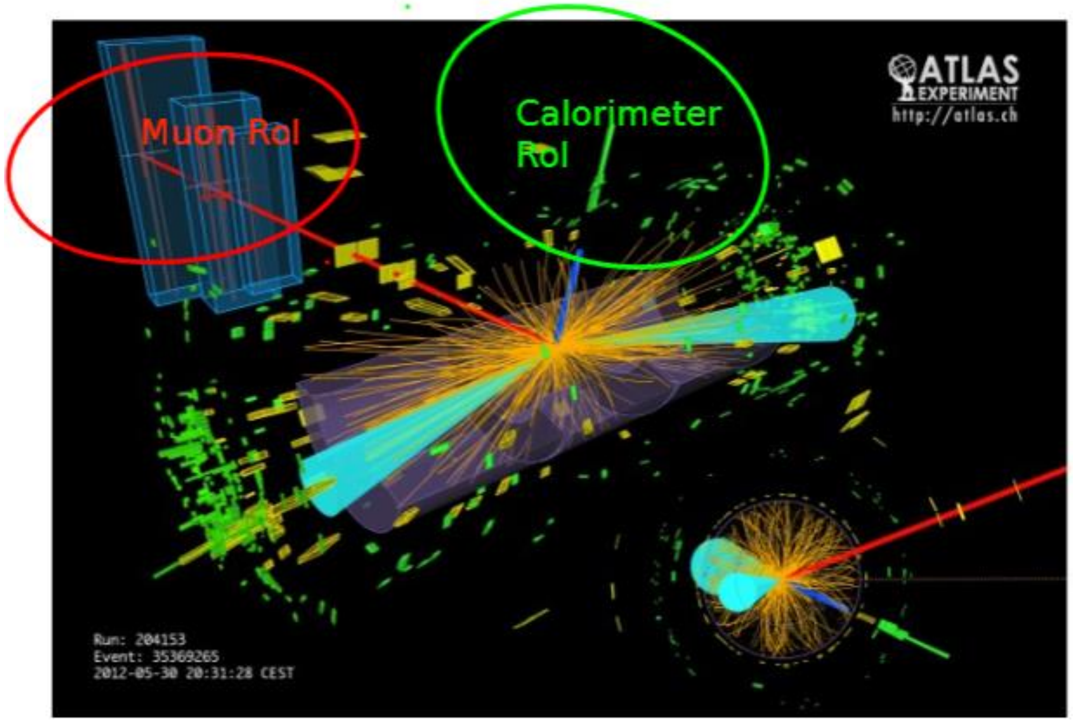
Detector System	Number of Readout Channels
Pixel	80M
SCT	6M
TRT	350k
LAr	185k
Tile	500k
Muon	1.24M

ATLAS Trigger Strategy

- Multi-level trigger system
- Event selection based on Regions of Interest (RoI)
  - Deposits in calorimeters or muon system passing pre-defined criteria, for example transverse energy or isolation
- Trigger Menu defines thresholds at each stage, e.g.:



Trigger	Trigger Selection		Level-1 Peak Rate (kHz)	HLT Peak Rate (Hz)
	Level-1 (GeV)	HLT (GeV)	$L = 1.2 \times 10^{34}$ cm^{-2}s^{-1}	
Single leptons	20	26 (i)	13	133
	22 (i)	26 (i)	20	133
	20	50	13	48
	22 (i)	60	20	13
	60	160	5	15

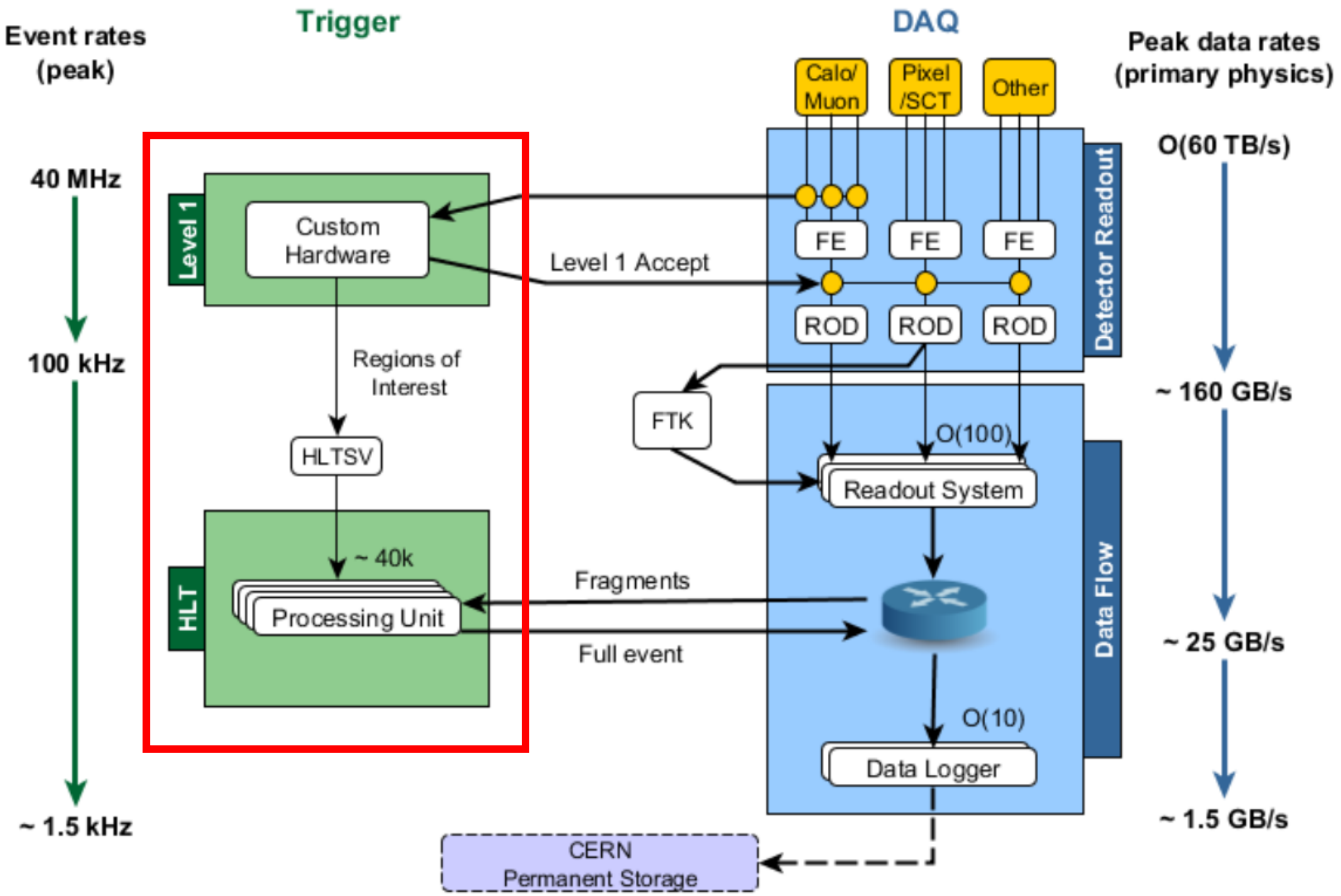


For a better overview of the Trigger itself, consult the Run 2 overview paper

https://arxiv.org/abs/1611.09661

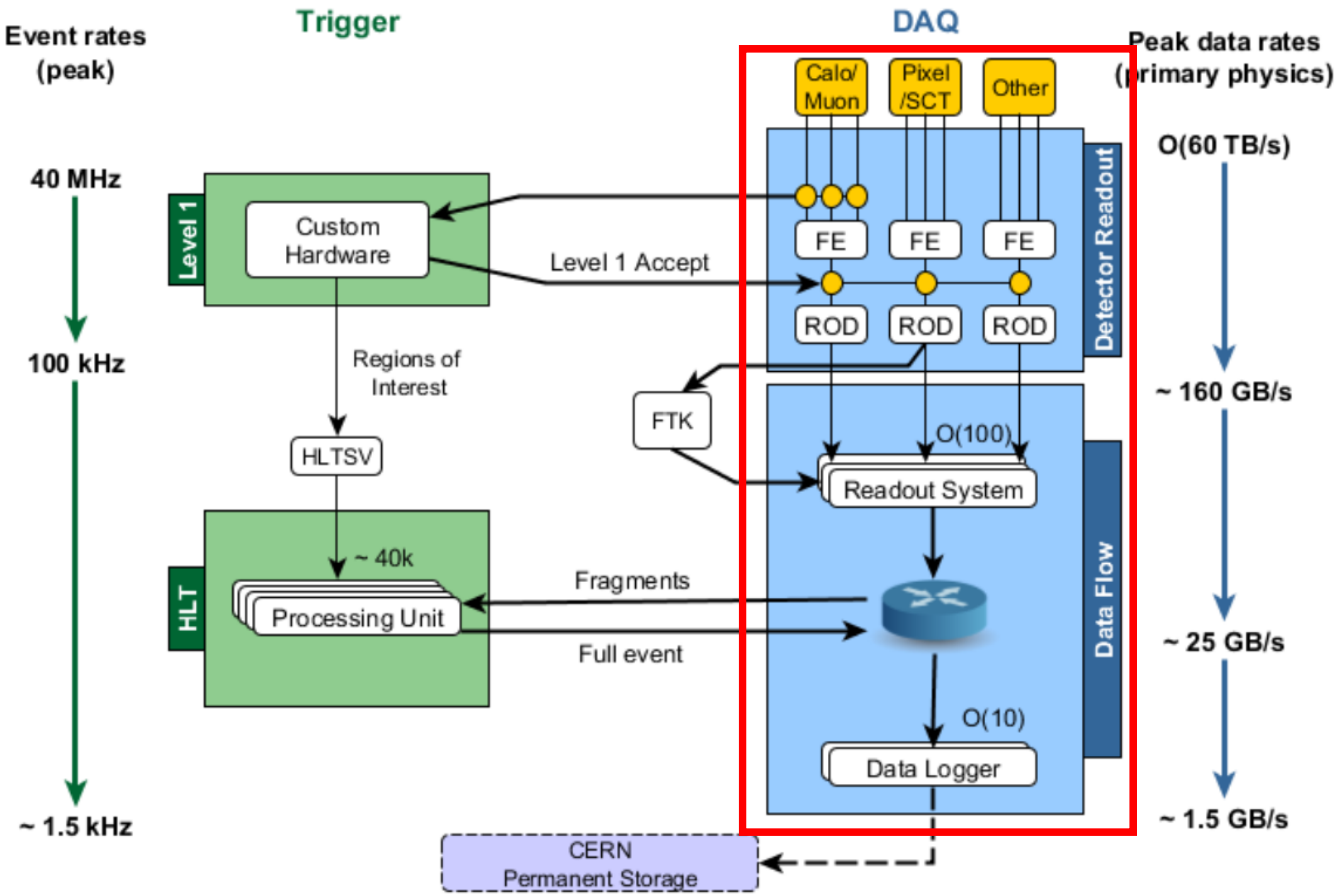
Overall Architecture – Trigger View

- Calorimeter and Muon System information processed by Level-1 trigger hardware – latency **2.5 µs**
- Level 1 accepts events featuring regions of interest passing thresholds defined in trigger menu
  - Tuned to produce up to **100 kHz** of accepts for nominal beam conditions
- Region-of-interest data used to seed processing in software-based High Level Trigger (HLT)
- HLT algorithms perform more complete analysis of event, featuring full event tracking information – latency ~**0.4 s**
  - Algorithms as near as possible to 'offline' reconstruction, but optimised for 'online' use-case
  - Acceptance criteria also based on Trigger menu
- HLT tuned to accept events at approx. **1.5 kHz** for nominal beam conditions



Overall Architecture – DAQ View

- In parallel to HLT seeding, Level 1 Accept also causes front-end detector electronics to read out event data for all other ATLAS detector systems
- Data are sent first to 'Readout Drivers' (RODs)
  - Detector-specific custom hardware (mainly VMEbus)
  - Perform initial data processing and formatting
- After ROD stage, data sent via optical link to Readout System (ROS)
  - First common stage of DAQ system
  - Bank of approx. 100 server PC's featuring custom I/O cards (RobinNP) to receive and buffer data
- ROS serves data to HLT processing node on request over 40GbE network
- Events accepted by HLT sent to data logger system for packaging and transfer to permanent storage offline
  - Typical event size 1.5 MB



Trigger System in Run 2

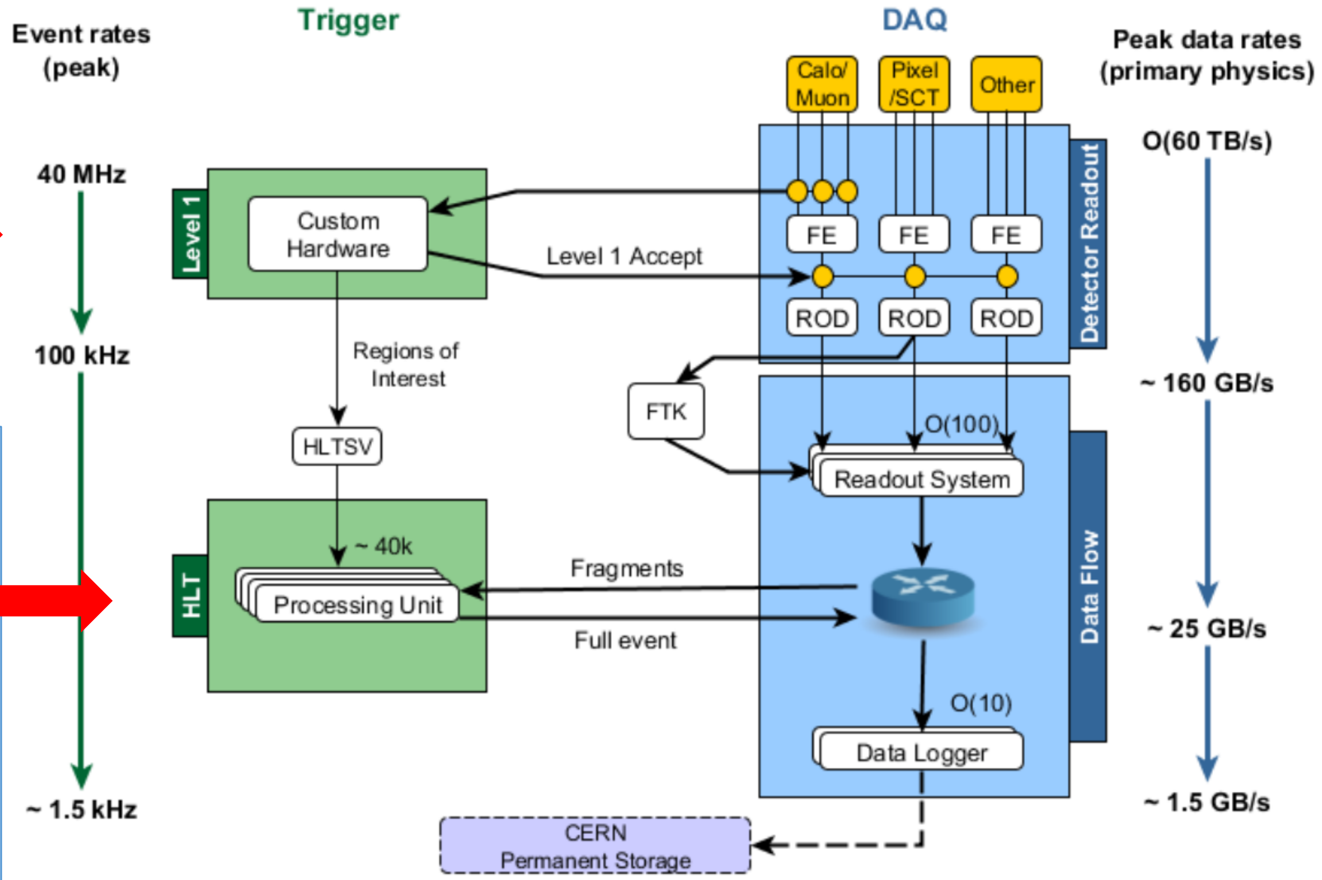
Upgraded Level 1 Trigger System, able to process more Trigger objects in parallel.

New **topological trigger** to use event shape information to improve selection efficiency

HLT processing originally (Run 1) split into two farms. First processed RoI data before triggering full event building. Final selection was then performed on full events in second farm.

In Run 2 the functionality was **merged into a single farm**, building events iteratively while performing selection.

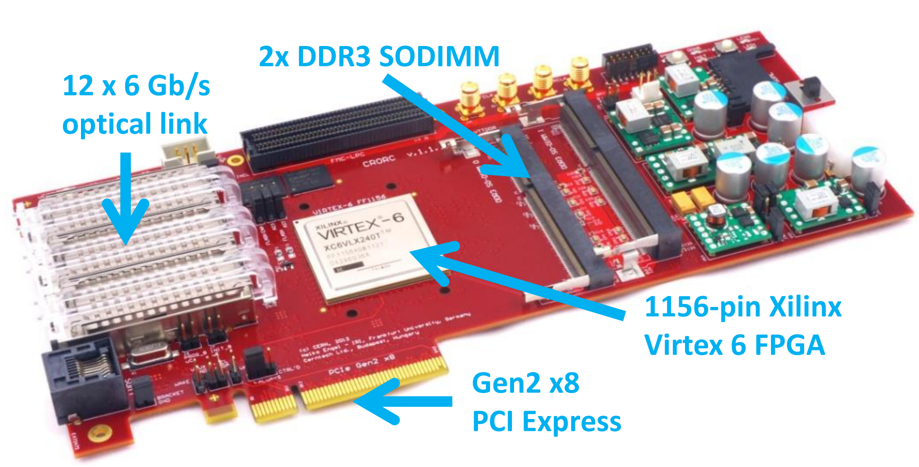
More efficient use of resources, avoids duplication of data, easier to load-balance

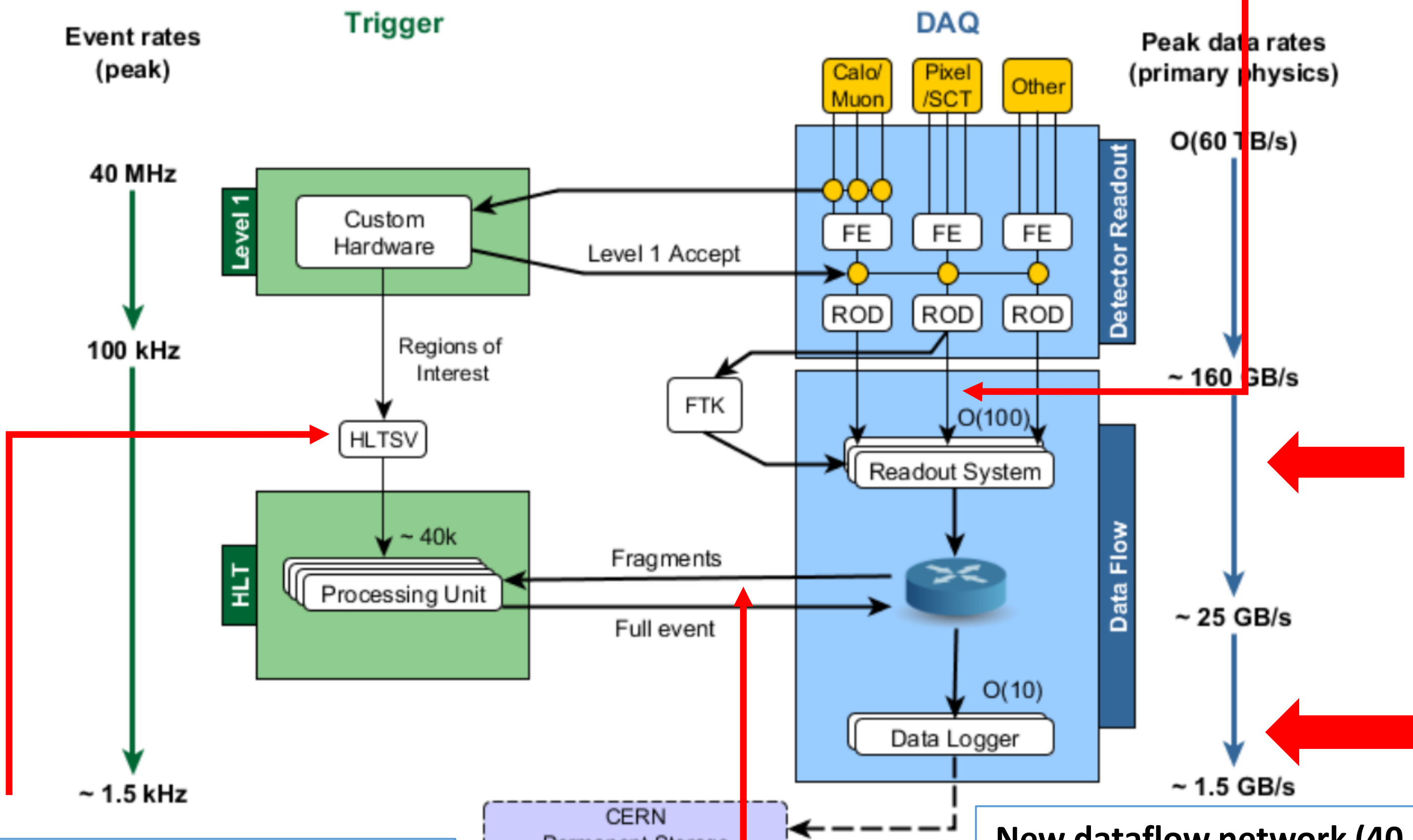


DAQ System in Run 2

Overall data-taking efficiency across all of Run 2: 94.3%

20% more ROD-ROS S-links (2000), due to new and expanded detector readout





Upgraded Readout System, based on RobinNP I/O card.

4x as many optical links as predecessor
6 x output bandwidth
10 x buffer capacity
Previous on-board PPC replaced with processing in host

Hardware developed by ALICE, featuring ATLAS-specific firmware. Typically 2 cards per ROS server.

Updated data logger hardware, with 3 x storage capacity of Run 1 system

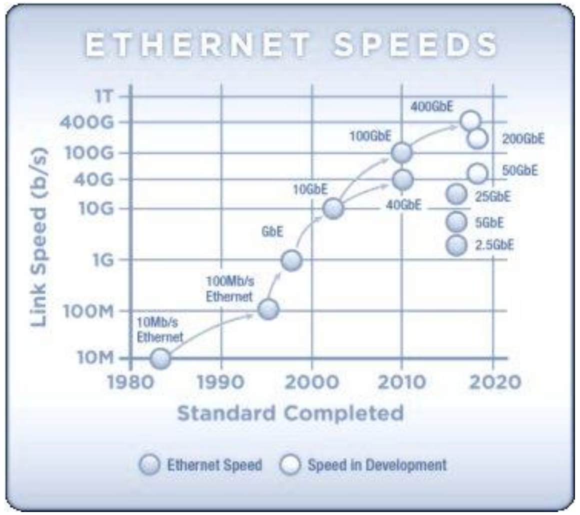
RoI builder also migrated from VMEbus to RobinNP-based system

New dataflow network (40 GbE)
4 x backbone bandwidth
40 x ROS bandwidth

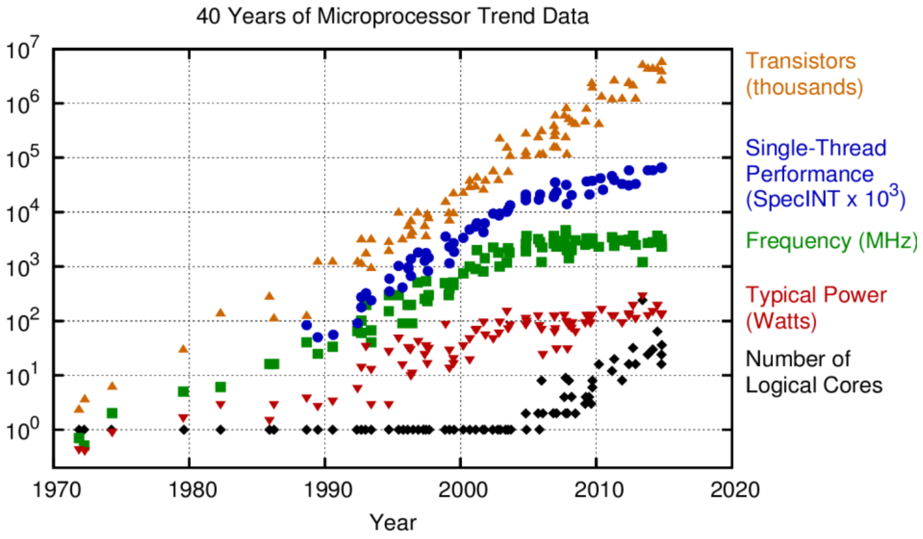
More details can also be found in presentations given to 2016 DAQ@LHC workshop

Common Challenges and Evolution

- ATLAS detector readout electronics ageing
  - Mix of technologies from past 20 years of design
  - Most detectors maintain separate hardware/firmware
    - Maintenance challenge due to technology obsolescence and loss of key personnel
- Technological evolution since system originally designed
  - Server CPU power (both clock speed and core count)
  - Network bandwidth and sophistication
  - Larger, more flexible FPGAs
  - What previously had to be done in hardware may now be done in firmware
  - What was previously done in firmware may now be done in software
- Wider trend towards commoditisation of readout technology
  - ALICE, LHCb, DUNE, many others
- Many more joint standards, meeting common challenges
  - E.g. radiation hard links - GBT/lpGBT project
    - https://ph-dep-ese.web.cern.ch/ph-dep-ese/optical_link/optical_link.html



40 Years of Microprocessor Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

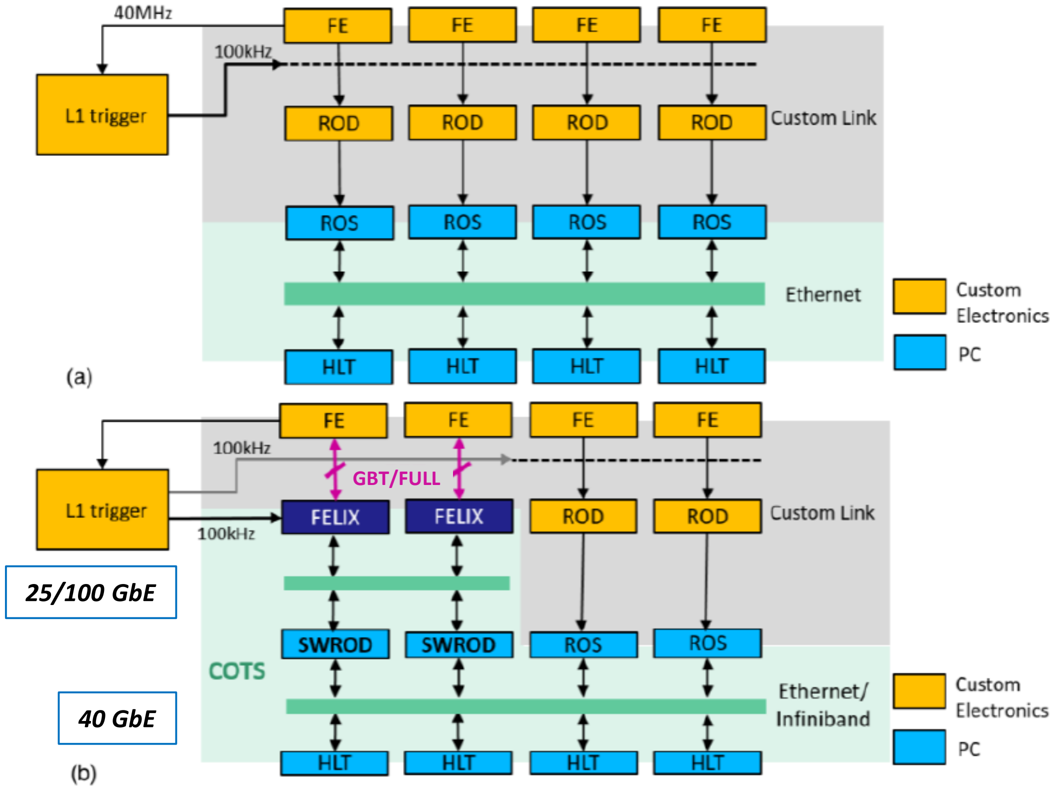
ATLAS in Run 3 – Wider Picture

- LHC luminosity and number of collisions per bunch crossing (pileup) expected to match peak values for Run 2
  - Luminosity 2x10^{34} cm^{-2}s^{-1} at pileup ~55 (design values 1x10^{34} cm^{-2}s^{-1} at pileup 27)
  - History has shown us this may evolve to larger values throughout the run
  - Means larger, more complex events to process while maintaining physics and DAQ performance
- New detector components
  - New Muon Small Wheels, new calorimeter and calorimeter trigger electronics (FEX), new RPC electronics for some sectors
- Further improvements to muon trigger electronics at Level 1
- Move to further align online and offline processing in HLT, further exploiting multithreading
- Expect to complete commissioning of Fast TracKer (FTK) system, providing hardware-accelerated tracking at Level 1 rate to HLT
- Changes for Run 2 and Run3 described in Phase-I TDR: https://cds.cern.ch/record/1602235

ATLAS Readout in Run 3

- Given wider trends and operational experience, ATLAS chose to develop new readout platform. Moving common hardware nearer to detector. Exploit commodity electronics where possible.

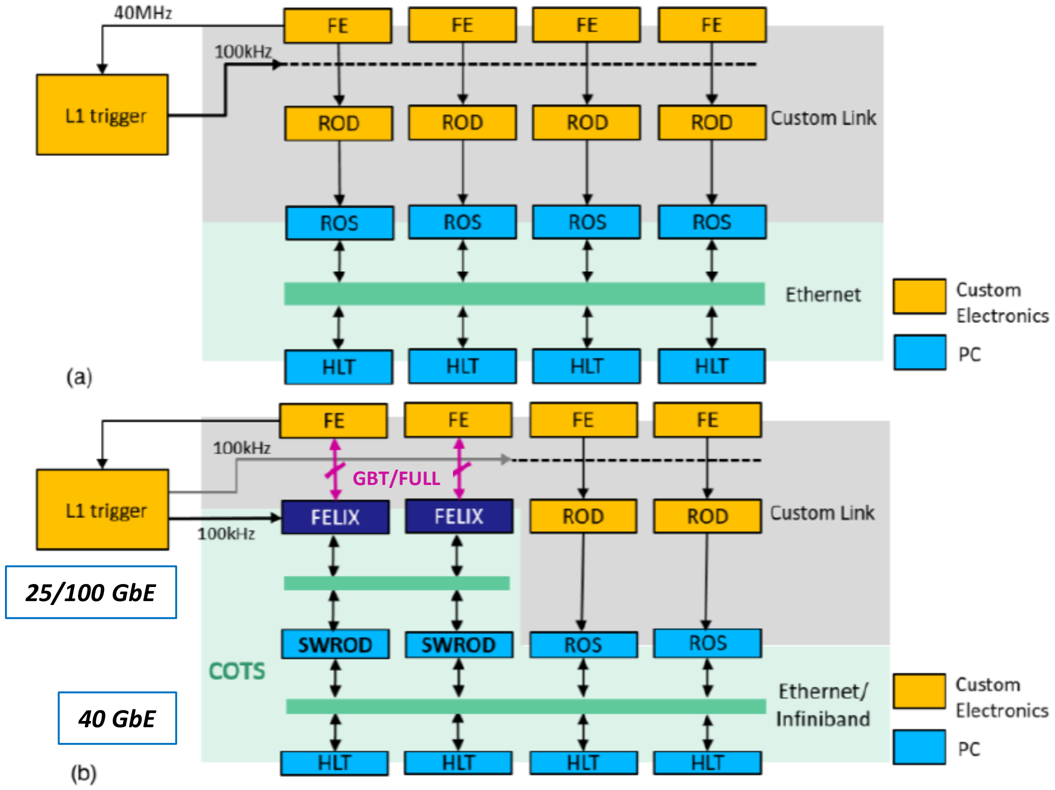
- Replace detector specific HW+FW RODs with new components
  - FELIX
    - Front-End Link eXchange
    - Connect directly to detector front-end electronics (or trigger hardware)
    - Receive and configurably route data from detector directly to client applications over high performance network
    - Route Level 1 Trigger clock and control signals to detector electronics
    - Able to interface both with GBT protocol (4.8 Gb/s raw) and directly to remote FPGA via high bandwidth ‘FULL mode’ protocol (9.6 Gb/s raw)



ATLAS Readout in Run 3

- Given wider trends and operational experience, ATLAS chose to develop new readout platform. Moving common hardware nearer to detector. Exploit commodity electronics where possible.

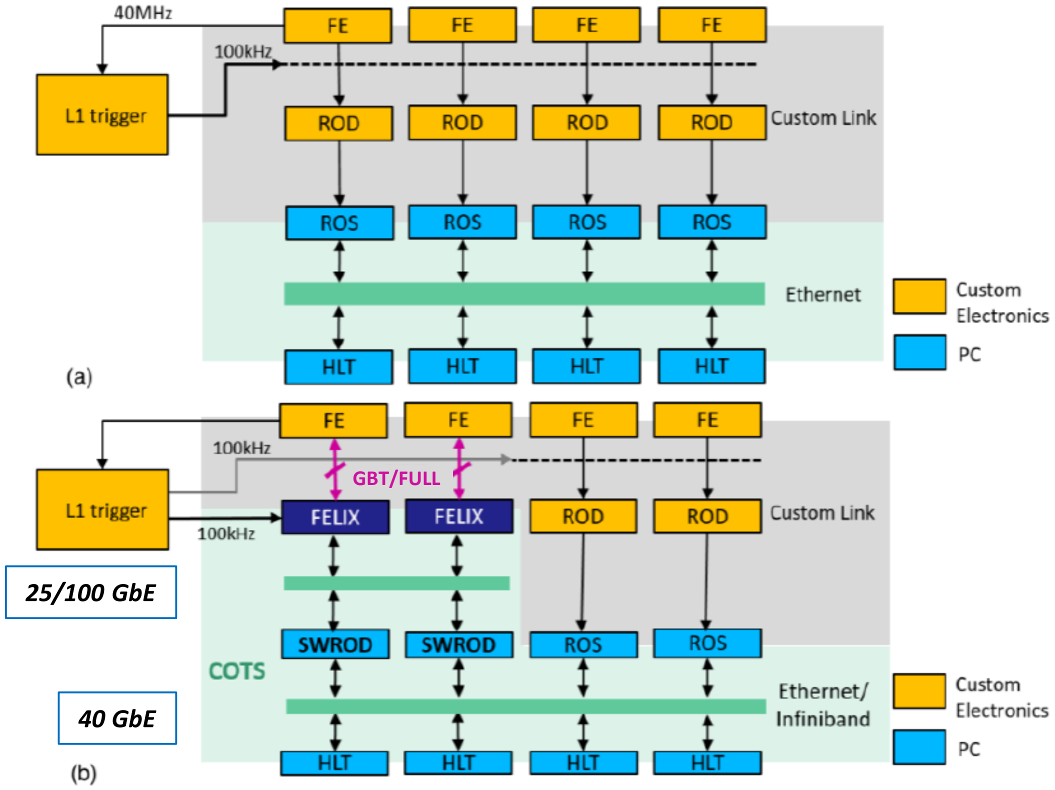
- Replace detector specific HW+FW RODs with new components
  - SW ROD
    - Software processes running on servers connected to FELIX via high bandwidth network
    - Common platform for data aggregation and processing – enabling detectors to insert their own processing software into data path
      - Previously performed in ROD hardware
    - Buffers data and serves it upon request to HLT
      - Interface indistinguishable from ROS



ATLAS Readout in Run 3

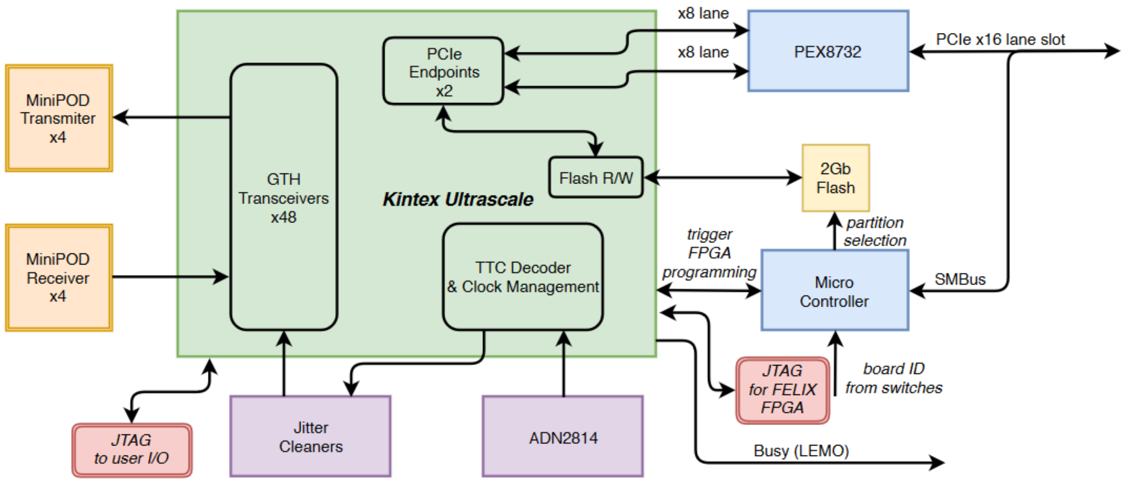
- Given wider trends and operational experience, ATLAS chose to develop new readout platform. Moving common hardware nearer to detector. Exploit commodity electronics where possible.

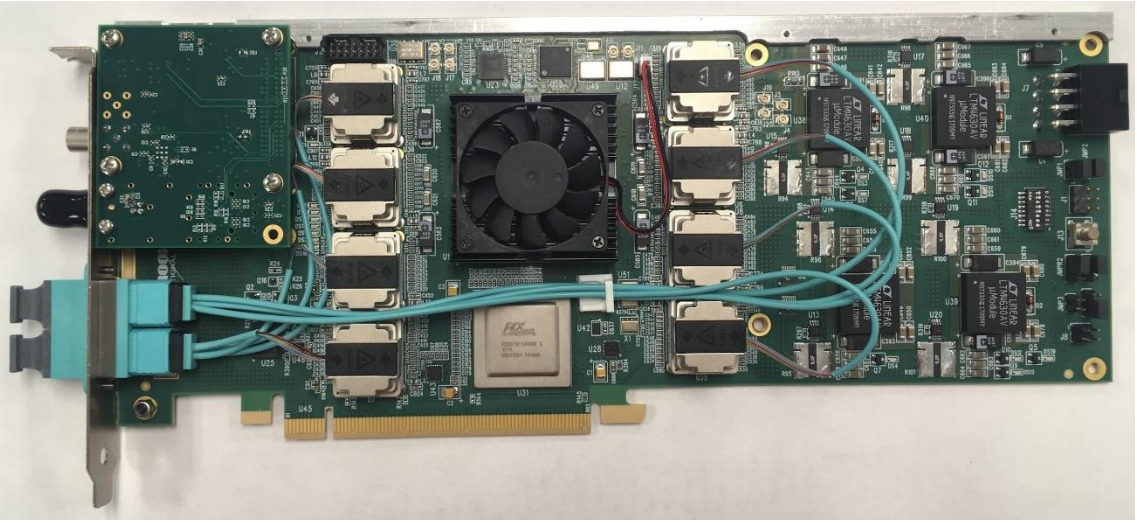
- Control and monitoring applications also now distributed among servers connected to data network
- In Run 3, FELIX and SW ROD will be deployed for newly installed detector systems (Muon New Small Wheels, new calorimeter readout and trigger etc)
- Legacy ROS system will remain for rest of system until Run 4



Inside FELIX

- FELIX consists of one or two PCIe I/O cards hosted by a commodity server
  - I/O card itself is custom built, but common across all subsystems
    - Final design developed by team at Brookhaven National Laboratory
    - Xilinx Kintex Ultrascale FPGA (XCKU115-FLVF-1924)
  - Connected to ATLAS Timing, Trigger and Control System (TTC) via customisable mezzanine
    - Also exist for TTC-PON and White Rabbit protocols
    - Includes interface to BUSY system
  - Interface via MTP24/48 connector, fanned out to MiniPODs
    - Firmware supports 24 optical links for dataflow purposes
  - PCIe Gen3 x 16 for communication with host server
    - 74 Gb/s observed in lab benchmarking, expect to be able to get nearer to 100 Gb/s with further FW/SW optimisation
- Dual 25 GbE or 100 GbE output network from host (depending on use case)

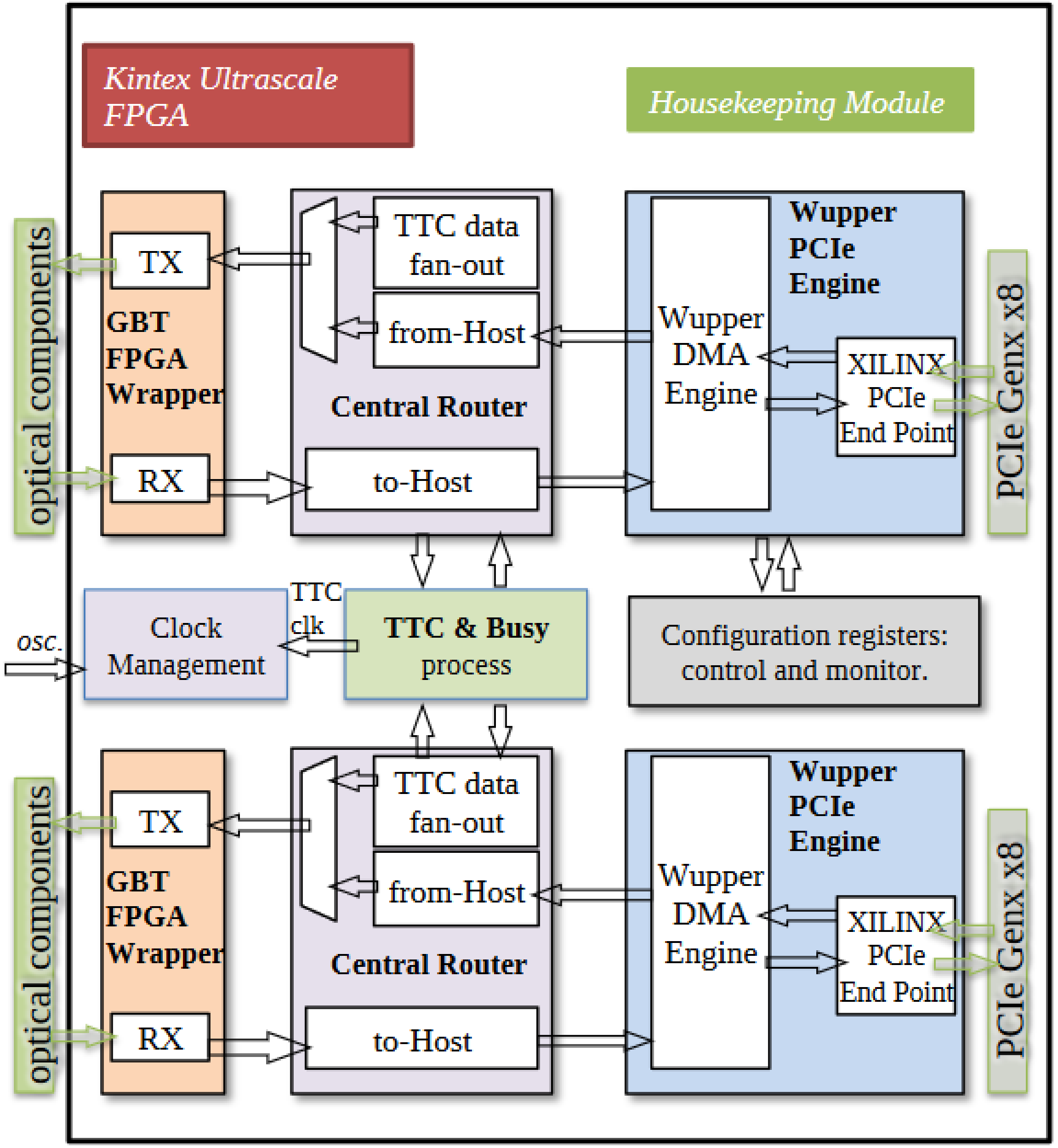




FLX-712: ATLAS Production Board for Run 3

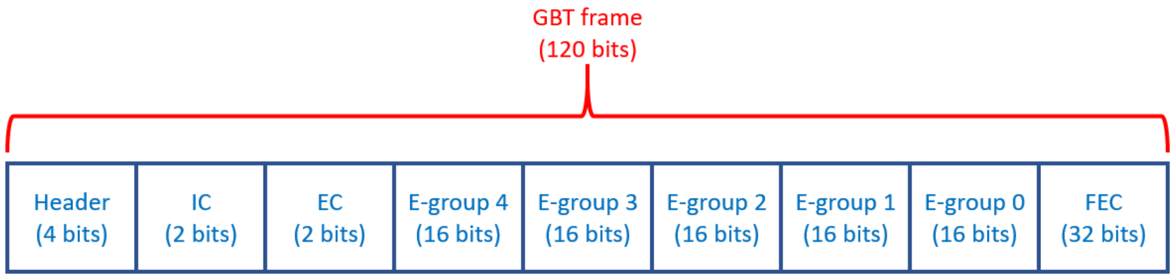
Inside FELIX - Firmware

- Two identical sets of blocks, each attached to separate PCIe Gen3 x 8 end point
- Bi-directional communication paths to and from front-end
- Link wrapper (GBT or FULL mode)
- TTC and BUSY interface wrapper
- Central Router
  - Core of FELIX functionality
  - Decodes and decomposes incoming data packets from front-end (currently 8b10b and HDLC available) into logical blocks for transfer to host server
  - Encodes data from host server for sending to front-end
- Wupper PCIe Engine
  - Manages PCIe bus and DMA communication with host



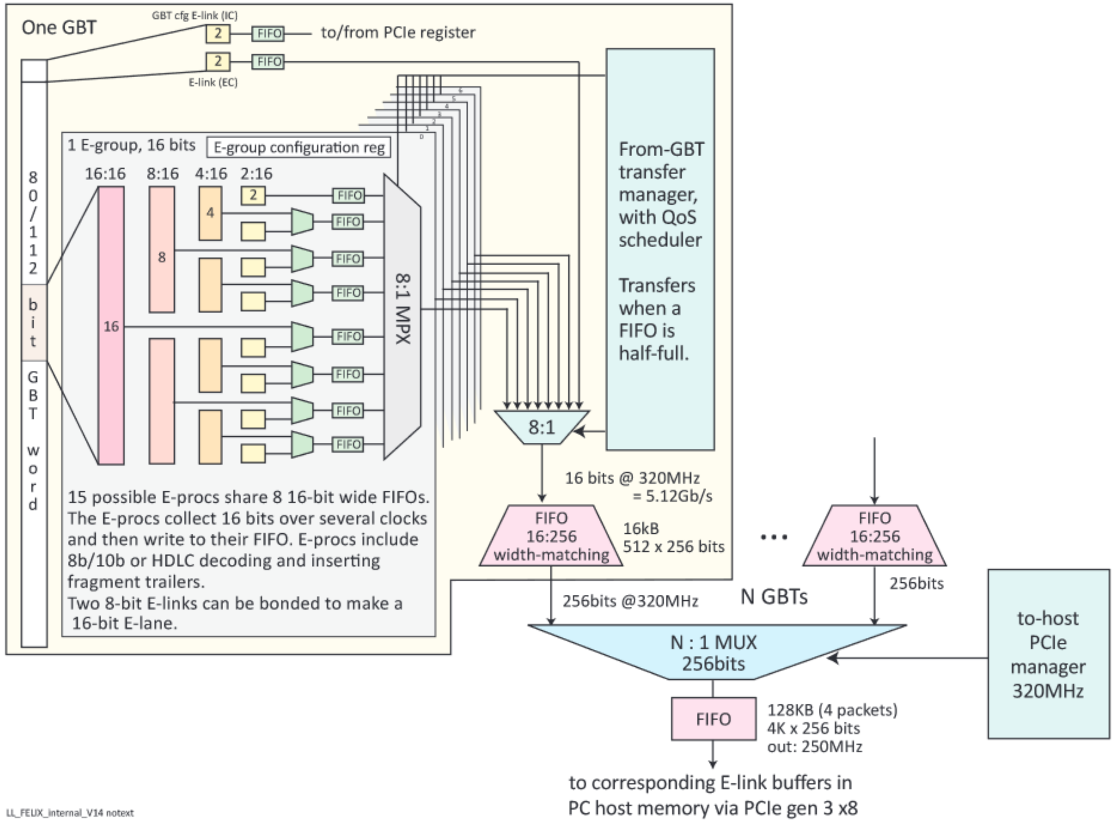
Inside FELIX - GBT

- Developed as part of radiation-hard Versatile Link project
- Implemented in front-ends through dedicated ASIC
    - FPGA version available for development
- 3.2 Gb/s user payload (before decoding)
- 24 links serviced per FELIX I/O card at full bandwidth
- Each GBT frame received contains multiple logical ‘E-links’
    - In ATLAS allows lower bandwidth electrical signals from front-end chips to be aggregated for transfer over one higher bandwidth pipe
    - E-links can be 2, 4, 8 or 16 bits wide
        - An E-group contains multiple E-links depending on their width
    - Dedicate channels for control data
    - Forward error correction built into protocol (radiation hardness)
- FELIX Central Router extracts/packages E-link data according to configuration
- E-link specific packets can then be transferred to/from network end-point via host server
- Significant FPGA resource load (LUTs) to manage routing configurably – mitigated using in-built FIFOs in silicon



GBT frame (120 bits)

Header (4 bits)	IC (2 bits)	EC (2 bits)	E-group 4 (16 bits)	E-group 3 (16 bits)	E-group 2 (16 bits)	E-group 1 (16 bits)	E-group 0 (16 bits)	FEC (32 bits)



From-GBT transfer manager, with QoS scheduler

Transfers when a FIFO is half-full.

15 possible E-procs share 8 16-bit wide FIFOs. The E-procs collect 16 bits over several clocks and then write to their FIFO. E-procs include 8b/10b or HDLC decoding and inserting fragment trailers.
Two 8-bit E-links can be bonded to make a 16-bit E-lane.

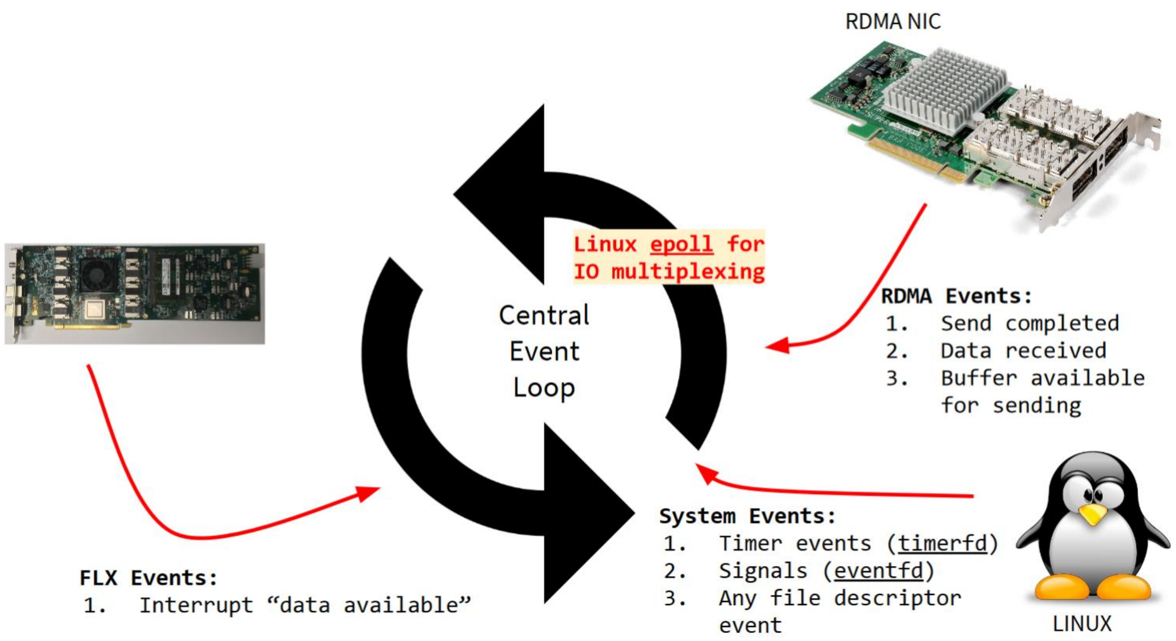
to corresponding E-link buffers in PC host memory via PCIe gen 3 x8

Inside FELIX – FULL mode

- Much simpler protocol for communication with remote FPGA without need for radiation hardness
- Only for communication from front-end to FELIX
  - Communication in other direction via GBT protocol
- Each link has no formal payload substructure
  - Central Router significantly less complicated, far lower FPGA resource consumption
- Single 32-bit wide frame with 8b10b encoding
  - Built-in checksum
  - Control signals (e.g. for BUSY can be inserted into data stream by detector)
- FELIX can assert flow control ‘XOFF’ signal to front-end via GBT link
- 7.68 Gb/s user payload to FELIX (after decoding)
- 24 links serviced per FELIX I/O card (12 at full bandwidth)

Inside FELIX - Software

- Primary dataflow and control through **FelixCore** application running as a daemon on host server
- FELIX firmware transfers data to host ring buffer via continuous DMA
  - Data split into fixed size ‘blocks’ for transfer
- Event driven software architecture
  - Incoming DMA triggers packet processing and transfer to NIC
    - Re-composes blocks back into complete packets
    - Designed to eliminate need to make copies of data to maximise processing speed
  - Handle signals from FELIX to front-end with same approach
- Network transfers make use of RDMA to maximise throughput and efficiency
- Comprehensive suite of test applications available for commissioning and development



RDMA NIC

Linux epoll for IO multiplexing

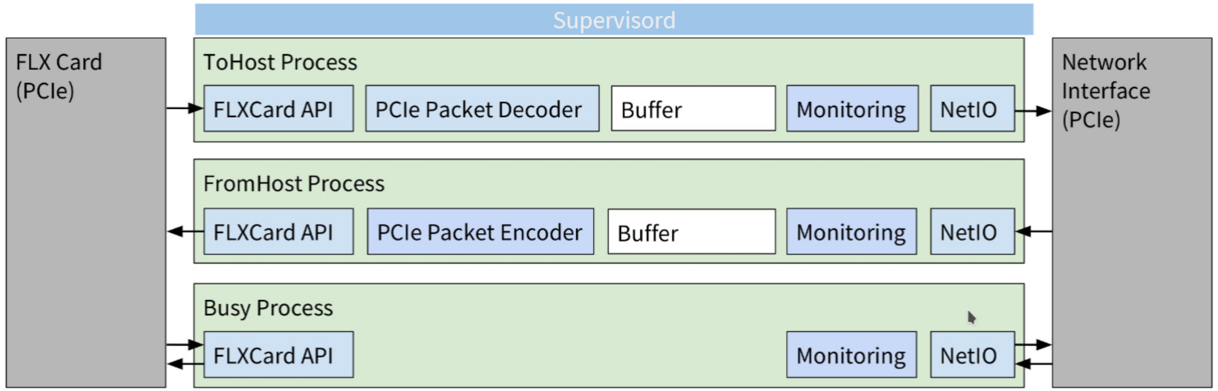
Central Event Loop

RDMA Events:
1. Send completed
2. Data received
3. Buffer available for sending

FLX Events:
1. Interrupt “data available”

System Events:
1. Timer events (timerfd)
2. Signals (eventfd)
3. Any file descriptor event

LINUX



Supervisord

FLX Card (PCIe)

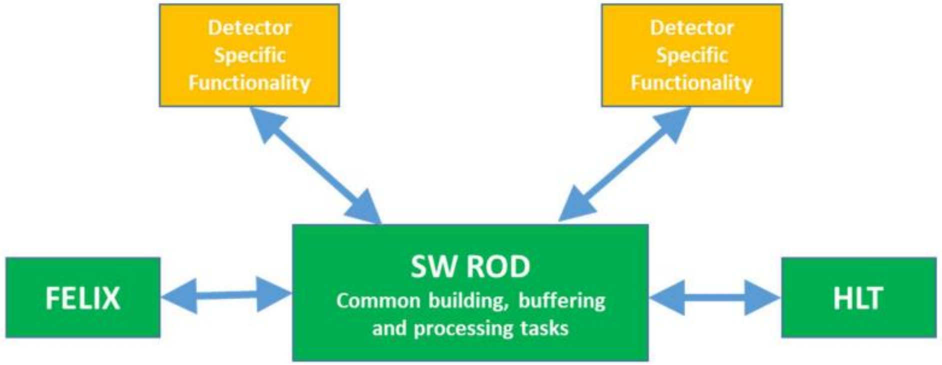
ToHost Process: FLXCard API | PCIe Packet Decoder | Buffer | Monitoring | NetIO

FromHost Process: FLXCard API | PCIe Packet Encoder | Buffer | Monitoring | NetIO

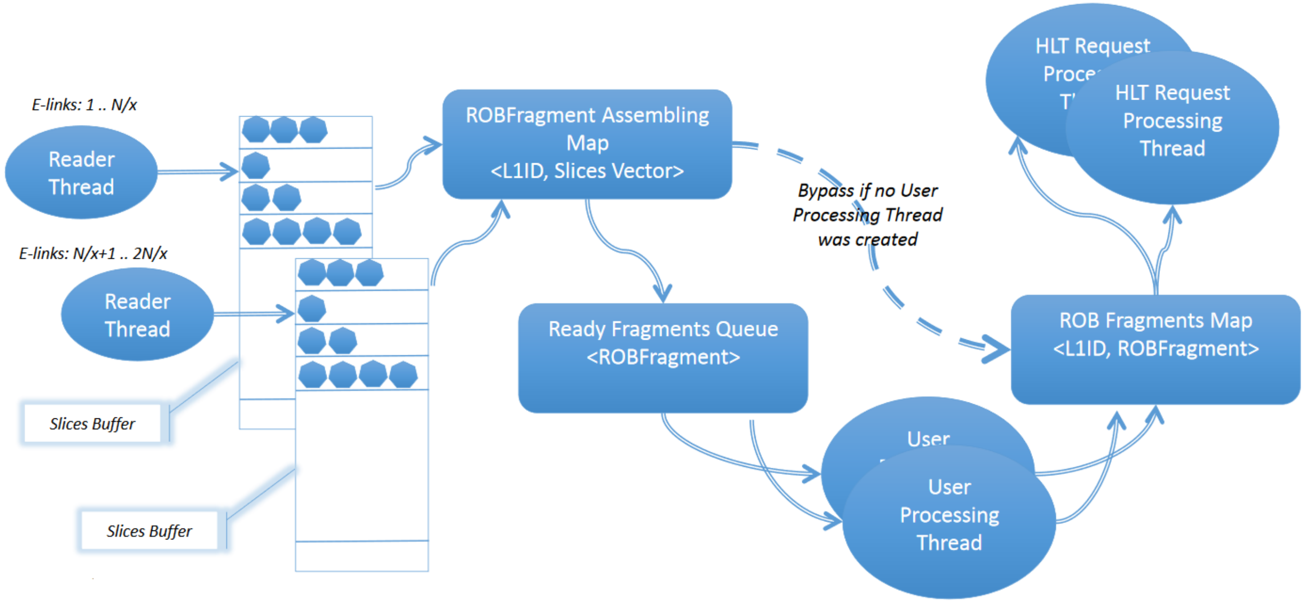
Busy Process: FLXCard API | Monitoring | NetIO

Network Interface (PCIe)

SW ROD


Detector Specific Functionality
Detector Specific Functionality
FELIX
SW ROD
Common building, buffering and processing tasks
HLT

- The SW ROD is FELIX’s logical counterpart
  - Software processes running on banks of server PC’s connected to FELIX over high bandwidth network
  - Subscribes to and receives event data from FELIX and facilitate sub-detector specific processing
    - Where the data handling actions in original hardware RODs now reside
    - Data coming in on multiple links can be configurable aggregated into larger packets for transfer to HLT
      - Data from multiple FELIX servers handled by a single SW ROD
    - Possible to implement monitoring feature either in SW ROD, or in separate process sampling data from it over the network

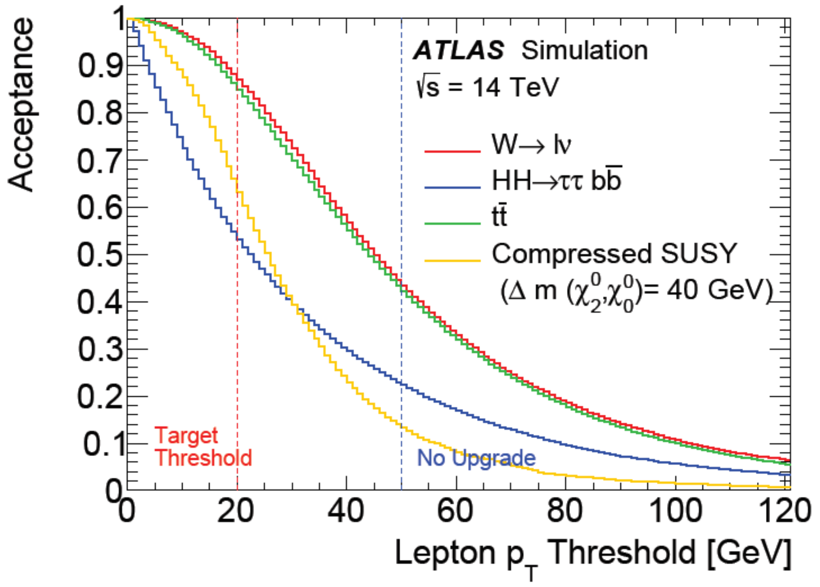

E-links: 1 .. N/x
Reader Thread
E-links: N/x+1 .. 2N/x
Reader Thread
Slices Buffer
Slices Buffer
ROBFragment Assembling Map <L1ID, Slices Vector>
Ready Fragments Queue <ROBFragment>
Bypass if no User Processing Thread was created
HLT Request Proce T
HLT Request Processing Thread
ROB Fragments Map <L1ID, ROBFragment>
User
User Processing Thread

FELIX & SW ROD Status

- Steady development over past few years, now reaching production maturity
  - Final hardware design complete, firmware and software nearing feature completeness
  - Performance testing confirms ability to satisfy ATLAS Phase-I rate requirements
  - Sample systems provided to subdetector test stands to provide early validation and integration opportunities
- Industrial tender process for manufacture of FELIX I/O cards complete, production in progress.
- Currently validating candidate servers for FELIX and SW ROD
- Total system size – approx. 100 I/O cards, 60 FELIX servers, 35 SW RODs
  - Installed alongside legacy ROS system (~100 servers, ~200 RobinNP I/O cards)
- Aiming to install full system at ATLAS P1 in summer 2019

Towards Run 4 and HL-LHC

- Run 4 will see significantly more challenging collision environment
- Luminosity range from 5x10^{34} cm^{-2}s^{-1} at pileup ~140 to 7.5x10^{34} cm^{-2}s^{-1} at pileup ~200
- Without technology upgrades will need to increase trigger thresholds in order to maintain rates
- ATLAS proposing a programme of upgrades to enable us to keep thresholds low
  - All new silicon based inner tracker (ITk)
  - Further improvements to calorimeter and muon electronics
  - Proposed new High Granularity Timing Detector (HGTD)
  - Comprehensive redesign of Trigger and DAQ system, including upg trigger electronics and all new DAQ components
  - Implement ability to add GPU or FPGA co-processors to HLT
- For Trigger and DAQ, more can be found in the TDR
  - https://cds.cern.ch/record/2285584



DAQ System in Phase-II

- Approx factor of 10 increase in rate through all areas of system – event size 5.2 MB (c.f. 1.5 MB now)
- FELIX readout for all subdetector links
- Data Handler replaces ‘SW ROD’ from Phase-I
- All buffering moved into central ‘Storage Handler’
  - Total storage ~36 PB
- Proposed distributed file system covering Data/Storage handlers and Event Filter nodes
- Hardware tracking (HTT) available as co-processor to Event Filter in regions of interest or full scan mode

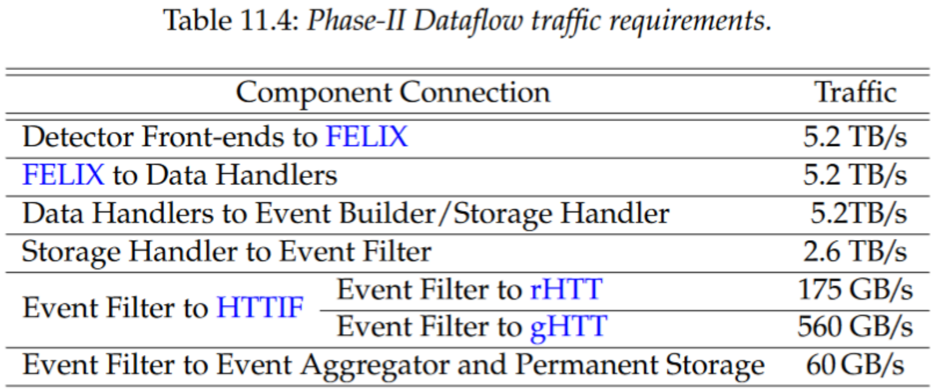
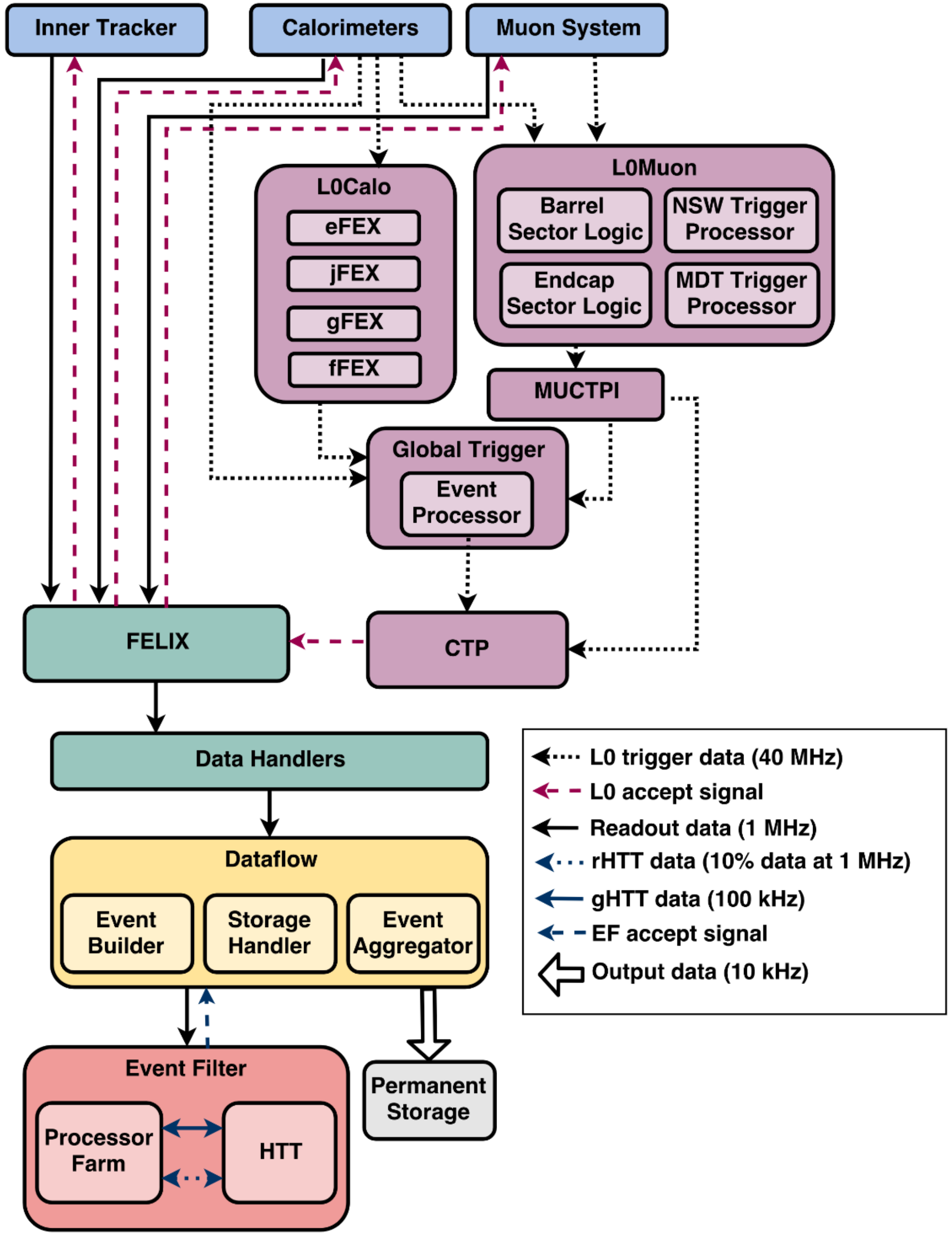


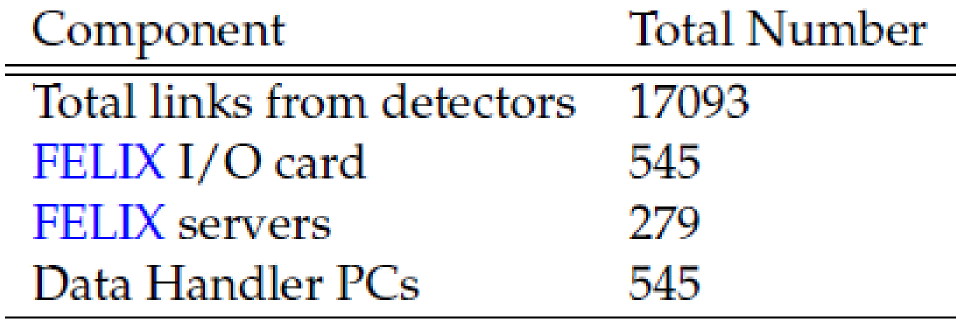
Table 11.4: *Phase-II Dataflow traffic requirements.*

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

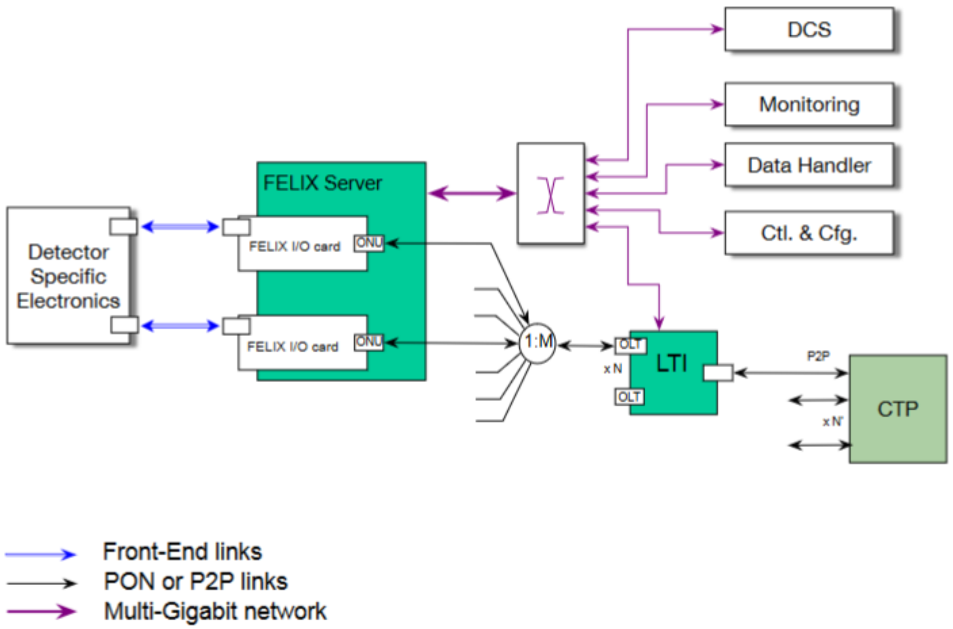


FELIX in Phase-II

- Learn from Phase-I design and operations experience
- In Phase-II, must support new interfaces
  - New TTC distribution network
  - Link Protocols
    - lpGBT
    - Aurora
- Increase link density to potential 48 per I/O card (protocol dependent)
- Support for time critical subdetector processing functionality within firmware
- Exploit technological advancement
  - Bus technology (PCIe Gen 4 and beyond)
  - Processing power
  - Network bandwidth up to 400 GbE
- Currently in early prototyping and requirements capture phase
- Full programme of R&D, leading to construction ready for installation in 2025



Component	Total Number
Total links from detectors	17093
FELIX I/O card	545
FELIX servers	279
Data Handler PCs	545



Summary and Outlook

- ATLAS DAQ system underpinned successful data taking in Run 1 and Run 2
  - Programme of upgrades to meet new challenges, starting before Run 2, continuing through Runs 3 and 4
- Run 3 will see beginning to move towards common readout platform (FELIX) across all ATLAS subsystems (completed in Run 4)
  - Run 3 developments nearing completion ahead of installation this summer
  - Already ramping up on Run 4 R&D
- Several other experiments have adopted FELIX as the basis for their DAQ systems and more are exploring its viability
  - Exploring possibility of making FELIX firmware and software available via open source distribution for wider benefit
  - Dedicated CERN team offering DAQ support for non-LHC experiments and actively promoting the use of FELIX based solutions where appropriate.
- Thanks for your time!