The ATLAS FELIX System

William Panduro Vazquez
On behalf of the ATLAS Collaboration
Overview

• LHC Schedule and Evolution
• ATLAS TDAQ System Overall Design
• 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
The ATLAS Detector

- **Detector System** | **Number of Readout Channels**
- Pixel | 80M
- SCT | 6M
- TRT | 350k
- LAr | 185k
- Tile | 500k
- Muon | 1.24M
Overall Architecture – Trigger View

- Calorimeter and Muon System information processed by Level-1 trigger hardware
- L1 accepts events featuring regions of interest passing thresholds defined in trigger menu
  - Tuned to produce up to 100kHz 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
  - 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, L1 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
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 $2 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ at pileup $\sim 55$ (design values $1 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ 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 feature extractor electronics (FEX), new RPC electronics for some sectors
• Further improvements to muon trigger electronics at L1
• Move to further align online and offline processing in HLT, further exploiting multithreading, with flexibility to add GPU or FPGA co-processors
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/igGBT project
    - [https://ph-dep-ese.web.cern.ch/ph-dep-ese/optical_link/optical_link.html](https://ph-dep-ese.web.cern.ch/ph-dep-ese/optical_link/optical_link.html)
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 RODs with new components
  • FELIX
    • Front-End Link eXchange
      • PCIe cards hosted in a server
    • Connect directly to detector front-end electronics (or trigger hardware)
    • Receive and configurable route data from detector directly to client applications over high performance network
    • Route L1 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 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 (Phase-I)

• 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
  • Dual 25 GbE or 100 GbE output network from host (depending on use case)
    • Software also supports Infiniband
Inside FELIX (Phase-I)

• 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
- 3.36 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 host

### GBT frame (120 bits)

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Header</td>
<td>4 bits</td>
</tr>
<tr>
<td>IC</td>
<td>2 bits</td>
</tr>
<tr>
<td>EC</td>
<td>2 bits</td>
</tr>
<tr>
<td>E-group 0</td>
<td>16 bits</td>
</tr>
<tr>
<td>E-group 1</td>
<td>16 bits</td>
</tr>
<tr>
<td>E-group 2</td>
<td>16 bits</td>
</tr>
<tr>
<td>E-group 3</td>
<td>16 bits</td>
</tr>
<tr>
<td>E-group 4</td>
<td>16 bits</td>
</tr>
<tr>
<td>FEC</td>
<td>32 bits</td>
</tr>
</tbody>
</table>

- 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 8/16/10b or HDLC decoding and inserting fragment trailers.
- Two 8-bit E-links can be bonded to make a 16-bit E-lane.
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
- 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 (Phase-I) FELIX I/O card (12 at full bandwidth)
Inside FELIX (Phase-I)

• 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
SW ROD

- 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

14/04/2019

The ATLAS FELIX System - Workshop on CEPC
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
• 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 upwards of $5 \times 10^{34}$ cm$^{-2}$s$^{-1}$ at pileup $\sim 200$
- 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 TDAQ system, including upgraded trigger electronics and all new DAQ components
- For TDAQ, more can be found in the TDR
  - [https://cds.cern.ch/record/2285584](https://cds.cern.ch/record/2285584)
TDAQ 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.

<table>
<thead>
<tr>
<th>Component Connection</th>
<th>Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detector Front-ends to FELIX</td>
<td>5.2 TB/s</td>
</tr>
<tr>
<td>FELIX to Data Handlers</td>
<td>5.2 TB/s</td>
</tr>
<tr>
<td>Data Handlers to Event Builder/Storage Handler</td>
<td>5.2 TB/s</td>
</tr>
<tr>
<td>Storage Handler to Event Filter</td>
<td>2.6 TB/s</td>
</tr>
<tr>
<td>Event Filter to HTTIF</td>
<td>Event Filter to rHTT 175 GB/s</td>
</tr>
<tr>
<td>Event Filter to gHTT</td>
<td>Event Filter to gHTT 560 GB/s</td>
</tr>
<tr>
<td>Event Filter to Event Aggregator and Permanent Storage</td>
<td>60 GB/s</td>
</tr>
</tbody>
</table>
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
    - IpGBT
    - Aurora
    - Potential 25 GbE protocol based on FULL mode (t.b.c.)
- 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 through to installation in 2025
FELIX Outside of ATLAS

• Multiple ‘proof of concept’ studies, potentially leading to full scale deployment
  • sPHENIX, BELLE-II, protoDUNE etc...

• FELIX can operate in both triggered and triggerless environments
  • Trigger interface in Phase-I FELIX adaptable via separate mezzanine card
  • Expect similar flexibility in future evolutions

• Firmware flexible enough to facilitate introduction of experiment-specific processing
  • Already under exploration in DUNE context

• Overall effort supported by CERN detector technologies group
  • Aiming to make technology developed in an LHC context available and useful for other experiments

• Working on licensing matters required to make entire FELIX platform (hw/fw/sw) open source
  • PCIe core (Wupper) already available on Open Cores
Summary and Outlook

• ATLAS TDAQ 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

• FELIX (or equivalent technologies) under investigation by many experiments in the field
  • Exploring possibility of making FELIX firmware and software available via open source distribution for wider benefit

• Thanks for your time!