FELIX and SW ROD

commissioning the new ATLAS readout system

2020 Virtual IEEE NSS

Carlo Alberto Gottardo
NIKHEF, Radboud U. Nijmegen
On behalf of the ATLAS TDAQ Collaboration

ATLAS Trigger and Data Acquisition System

- The Large Hadron Collider collides protons at a 40 MHz rate
- ATLAS detects the collision products
- The maximum event data rate for permanent storage is (currently) 1 kHz
- The trigger systems select interesting events

DAQ architecture in Run 2 (2015-18)

Readout Driver (ROD)

- several types of links from Front End (FE) to ROD
- **VME boards**, of about a dozen flavours developed and maintained by detectors
- Connected via point-to-point optical link to **ROBinNP** cards

Readout System (ROS)

- commodity computers hosting the ROBinNPs
- transfers data to the High-Level Trigger farm over the network

Upgrades for Run 3

In Run 3:

- similar conditions to Run 2, higher energy expected
- new detector and trigger systems to improve background rejection
- new readout system: FELIX and SW ROD

Detector upgrades:
New Small Wheels
Muon Barrel (BIS7/8)
LAr Calorimeter readout

[image: <https://cds.cern.ch/record/1095924/>]

DAQ architecture in Run 3

New readout chain for upgraded systems

Front-End Link EXchange (FELIX)

- custom PCIe card hosted on commercial computer
- the interaction with FE includes readout, configuration, trigger, clock distribution, Slow Control monitoring, BUSY
- ~100 cards, 60 host PCs

Software Readout Driver (SW ROD)

- software running on commercial computer
- builds and aggregates events, detector-specific data processing
- ~30 servers

FELIX + SW ROD

- reduced hardware/firmware effort
- data transport decoupled from data processing
- industry standard data networks earlier in the readout chain
- aggregation of many links into a single higher speed network link

A peek to the Run 4+ architecture

- ×10 trigger rate (1 MHz)
- ×3 interactions per bunch-crossing (200)
- ×20 data readout rate (5.2 TB/s)

FELIX

- for all detector systems
- new implementation under development

Data Handlers

- successor of SW ROD, same functionality
- interfaces to the Event Filter, successor of the HLT

The FELIX system for Run 3

The FLX-712 card

- FPGA: Xilinx Kintex UltraScale XCKU115
- 8 MiniPODs to support up to 48 bidirectional optical links
- 16-lane PCIe Gen3 (two 8-lane Endpoints with a switch)
- Flash and Micro-controller to support firmware update
- interfacing to various Timing Trigger and Control (TTC) systems
- BUSY output

hosted on FELIX PC
Intel Xeon E5-1660 v4 @ 3.2GHz
32 GB DDR4 2667 MHz memory
Mellanox Connect-X 25/100 GbE

Working modes & firmware

- FULL**
 - 24 links, 9.6 Gb/s each
 - FPGA-to-FPGA connection
 - 12 fully saturate the PCIe bandwidth
- GBT**
 - 24 links, 4.8 Gb/s each
 - conforming to GigaBit Transceiver protocol used by the radiation tolerant GBTx ASIC [1]
 - each link aggregates multiple *e-links* with configurable bandwidth (80/160/320 Mbps)
 - heavy utilisation of FPGA resources for configurable logic

Most intense use of FELIX expected in ATLAS

Mode	Chunk size	Rate per link	Links per card	Total chunk rate per card	Total data rate per card
FULL	4800 bytes	100 kHz	12	1.2 MHz	46 Gbps
GBT	40 bytes	100 kHz	192	19.2 MHz	7.5 Gbps

[1] doi: 10.5170/CERN-2009-006.342

The felix-star software

- interrupt-driven event loop architecture
- asynchronous non-blocking operations
- uses RDMA technology for low overhead transfers (netio-next library, based on libfabric)
- single thread, two processes per card
- allows coalescing of small messages to decrease I/O overhead

Performance

GBT-mode

2 cards, 392 links, 40 byte chunks, 25 GbE
stable operation up to 150 kHz

FULL-mode

1 cards, 12 links, 5 kB chunks, 100 GbE
maximum data rate ~200 kHz limited by network
back-pressure triggers FULL-mode traffic control

For more results see <https://cds.cern.ch/record/2704279>

Performance

Data acquisition at 1 MHz trigger rate

On FELIX testbed at CERN

- 32 links, 260 Mbps each
- 1 MHz random trigger rate
- Stable transfer rate, no errors at $\times 10$ design trigger rate!

Testing & Integration

Control and monitoring system

- FE control and configuration data paths under test
- GBTx and GBT Slow Control Adapters integrated in FELIX testbed
- Detector Slow Control software update to netio-next libraries

SW ROD

- several SW ROD tests run with FELIX emulator
- integration with actual hardware in progress

The SW ROD for Run 3

SW ROD requirements

Functional requirements

- Aggregate data according to detector-specific algorithms and input data format
- Support for GBT and FULL mode readout via FELIX
- Support for multiple data handling procedures (transfer to HLT, write to disk, calibration...)

Performance requirements (worst case scenario)

Mode	FELIX cards per SW ROD	Total chunk rate	Total data rate
FULL	1	1.2 MHz	46 Gbps/s
GBT	6	115 MHz	4.6 GB/s (40-byte chunks)

SW ROD PC

Dual Intel Xeon Gold 5218 CPU @ 2.3 GHz
16x2 physical cores
96 GB DDR4 2667 MHz memory
Mellanox ConnectX-5 100 Gb to FELIX
Mellanox ConnectX-4 40 Gb to HLT

to be aggregated into bigger fragments according to their trigger IDs!

With 20 cores × 3 GHz → ~200 CPU operations per input chunk!

GBT fragment building algorithm

Buffer slices not present in FULL mode

GBT-mode performance

Setup

- software data generator: equivalent to 6 FELIX cards
- $6 \times 192 = 1152$ e-links, 6 aggregated output channels (AOC), 40-byte chunks
- Same netio-next network library used by felix-star

Results

- Performance limited by network bandwidth
- ~30% overhead for 40-byte chunks
- Good scaling with number of threads

SW ROD scalability

Setup

- 48 e-links, 40-byte chunks
- software data generator: equivalent to 1 FELIX card
- Same netio-next network library used by felix-star

Results

- 1 MHz rate reached only a limited number of *e-links*
- CPU-limited rates
- memory management brings significant overhead
 - `tbb::concurrent_queue` used in place of `new/delete` brings ~40% performance improvement
- Further optimisation under study

Full-mode performance

Setup

- software data generator: equivalent to 1 FELIX card
- In: 24 links, 5 kB chunks, 6 reading threads
- Out: variable number of AOCs
- Same netio-next network library used by felix-star

Results

- Performance limited by network bandwidth except in the case with 1 AOC (investigating)
- Small protocol overhead for large data chunks

Summary

- In Run 3 FELIX + SW ROD replace (part of) ROD + ROS system
 - more flexibility, reduced custom design
 - architecture for Run 4 and beyond
- Hardware commissioning and deployment on-going
- Software development ongoing
- Performance exceeding the Run 3 requirements

BACKUP

FELIX users in ATLAS Run 3

Muon Spectrometer [GBT mode]

- New Small Wheels (NSW)
 - sTGC (Small-strip Thin Gap Chamber)
 - MicroMegas (Micro Mesh Gaseous Structure)
- BIS78 (Barrel Inner Small MDT sector 7/8)

L1 calorimeter trigger [FULL mode]

- gFEX (Global Feature Extractor)
- jFEX (Jet Feature Extractor)
- TREX (Tile Rear Extension)
- ROD, Hub for eFEX (Electron Feature Extractor)

Liquid Argon Calorimeter [48-ch GBT / FULL mode]

- LTDB (LAr Trigger Digitizer Board, custom GBT)
- LDPB (LAr Digital Processing Board, FULL)

Tile Calorimeter test system [FULL mode]

FELIX hardware block diagram

FELIX firmware architecture