FELIX: the new detector interface for ATLAS

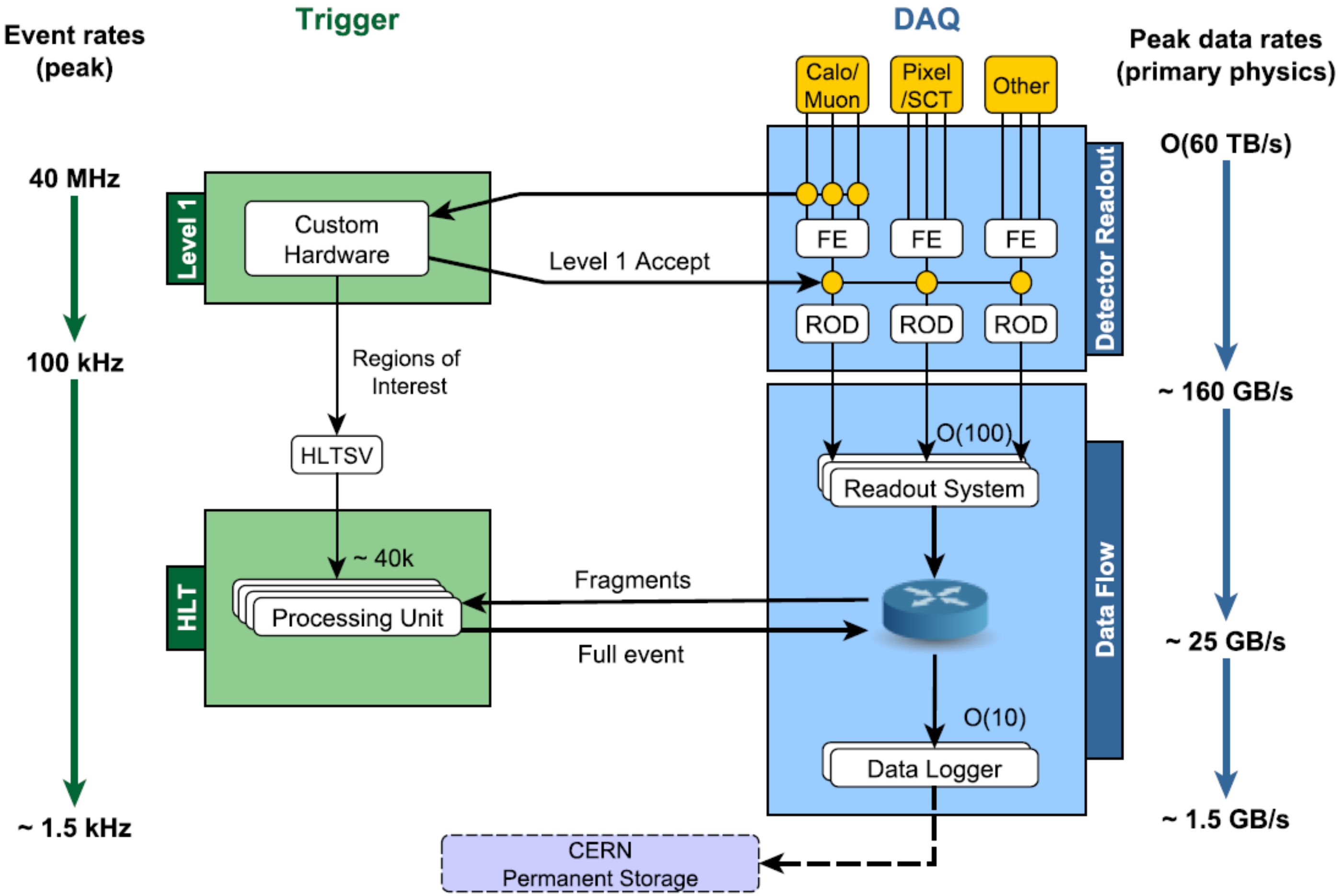
William Panduro Vazquez
On behalf of the ATLAS TDAQ Collaboration

Overview

- ATLAS TDAQ System Overall Design (Run 2)
- Common challenges and evolution
- DAQ System in Run 3
- FELIX and SW ROD
- Hardware Production
- Performance, Integration and Commissioning
- Summary

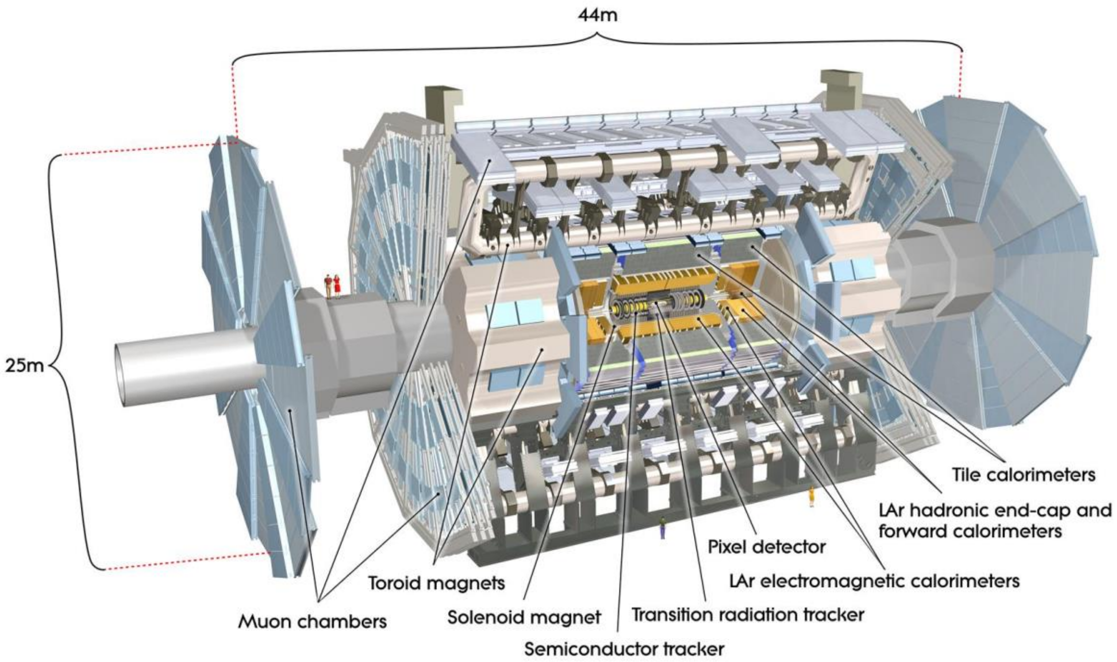
Overall TDAQ Architecture (Run 2)

- L1 hardware trigger based on Calorimeter and Muon information in Regions-of-Interest (RoI)
- ROI data used to seed processing in software-based High Level Trigger (HLT)
  - Performs more complete analysis of event
  - Uses full event tracking information
- In parallel to HLT seeding, L1 Accept also causes front-end detector electronics to transfer relevant data 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
  - Data buffered in custom PCIe I/O card (RobinNP)
- 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



ATLAS in Run 3 – Wider Picture

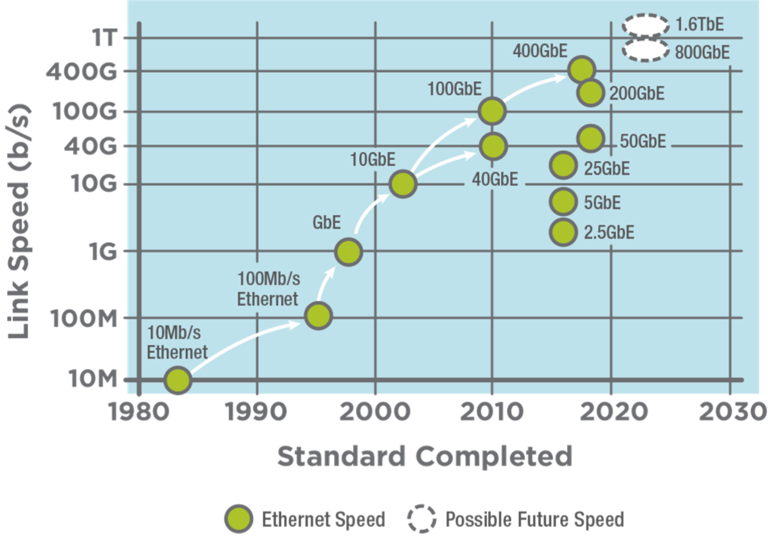
- LHC luminosity and number of collisions per bunch crossing (pileup) expected to match peak values for Run 2
  - Luminosity $2\mathrm{x}10^{34}$ $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ at pileup ~55 (design values $1\mathrm{x}10^{34}$ $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ at pileup 27)
  - May evolve to larger values throughout the run
  - 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


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

Common Challenges and Evolution

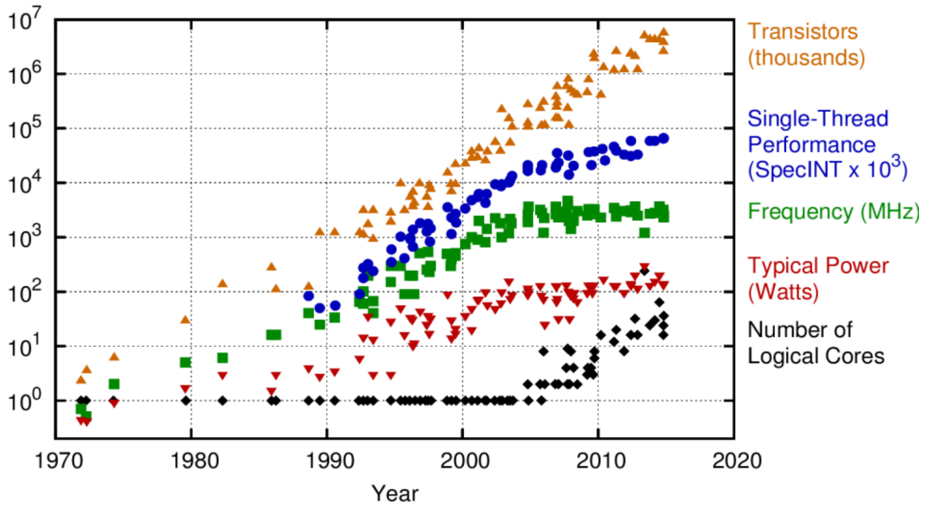
- ATLAS detector readout electronics ageing
  - Mix of technologies from past 20 years of design
  - Most detectors maintain separate hardware/firmware
  - Maintenance challenge due to technology obsolescence and loss of key personnel
- Technological evolution since system originally designed
  - Server 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://espace.cern.ch/GBT-Project/default.aspx

ETHERNET SPEEDS



NEXT ETHERNET ERA
ethernet alliance

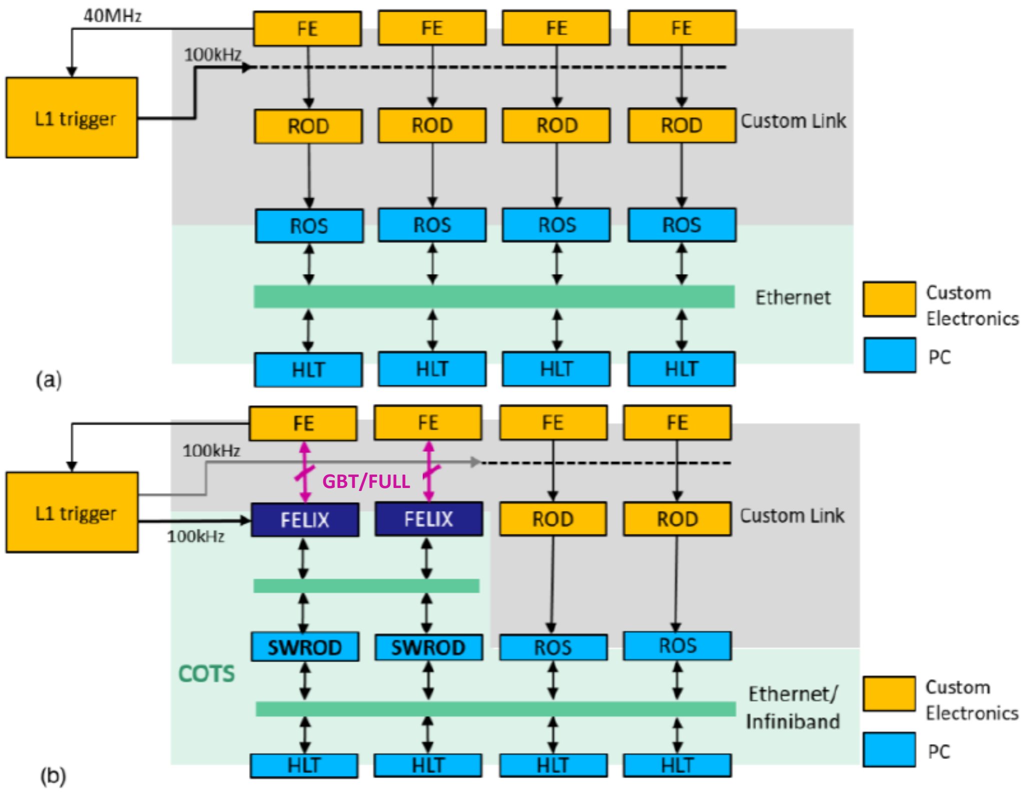
40 Years of Microprocessor Trend Data



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

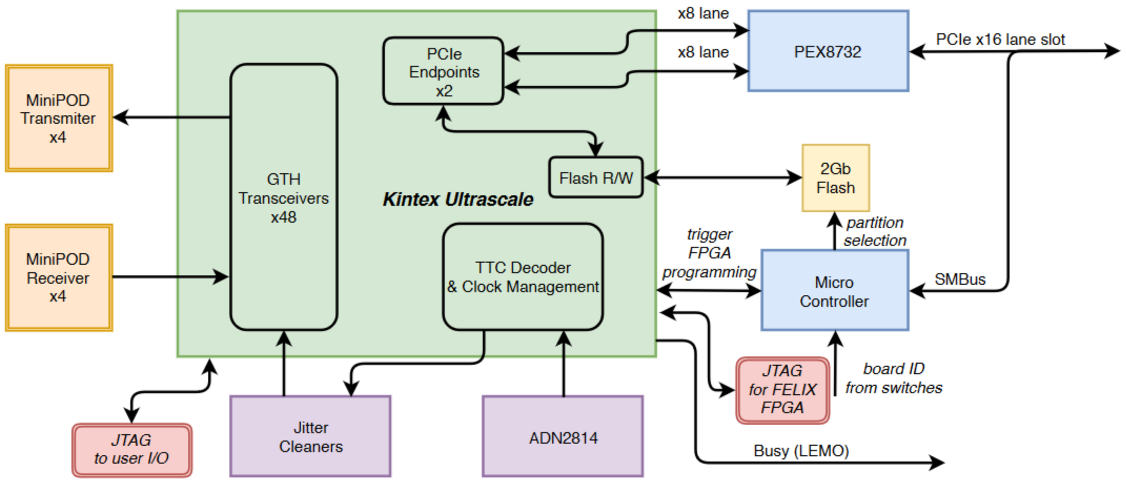
ATLAS Readout in Run 3

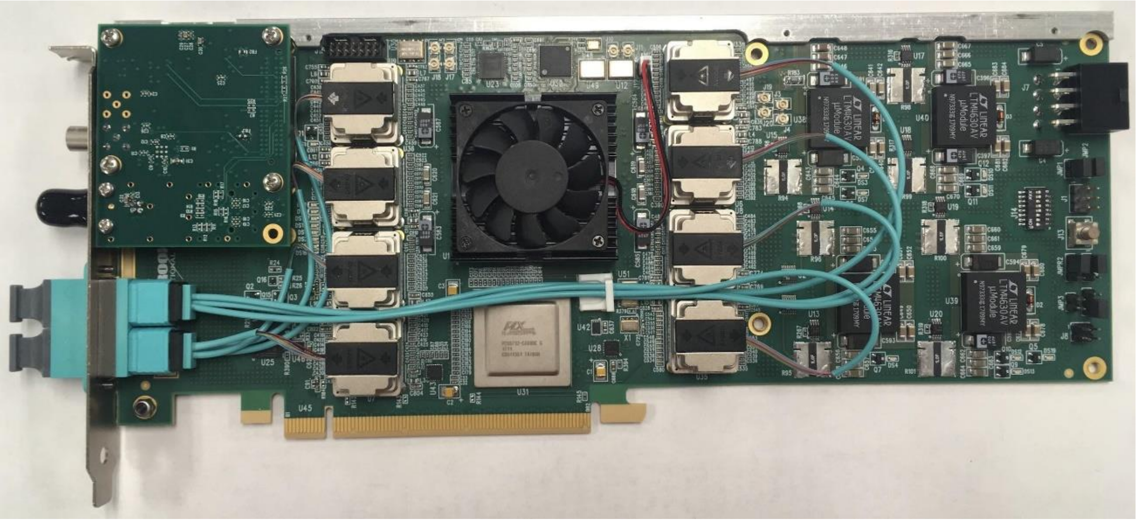
- Given wider trends and operational experience, ATLAS chose to develop new readout platform. Moving common hardware nearer to detector. Exploit commodity electronics where possible.
- FELIX
  - Front-End Link eXchange
    - PCIe cards hosted in a server
  - Connect directly to detector front-end electronics (or trigger hardware)
  - Receive and route data from detector directly to clients over high bandwidth network
  - Route L1 Trigger clock and control signals to detector electronics
  - Able to interface both with GBT protocol and directly to remote FPGA via high bandwidth 'FULL mode' protocol
- SW ROD
  - Software process 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
  - Buffer data and serves it upon request to HLT
    - Interface indistinguishable from legacy readout (ROS)
- Control and monitoring applications also now distributed among servers connected to data network



Inside FELIX

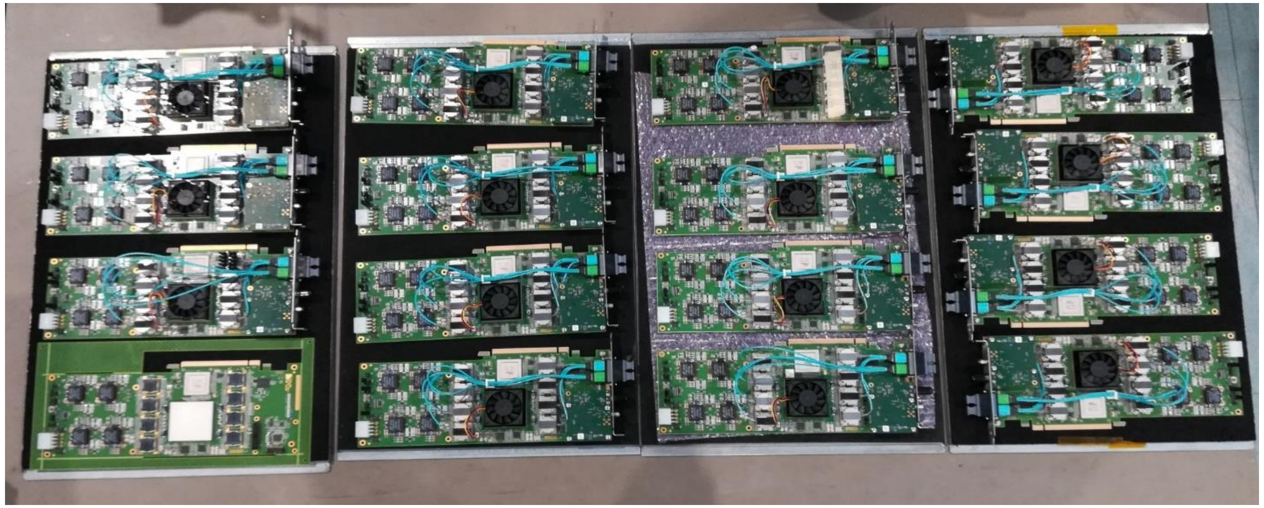
- Each FELIX system consists of one or two PCIe I/O cards hosted by a commodity server
    - I/O card itself is custom built, but common across all subsystems
        - Xilinx Kintex Ultrascale FPGA (XCKU115-FLVF-1924)
    - Connected to ATLAS Timing, Trigger and Control System (TTC) via customisable 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





FLX-712: ATLAS Production Board for Run 3

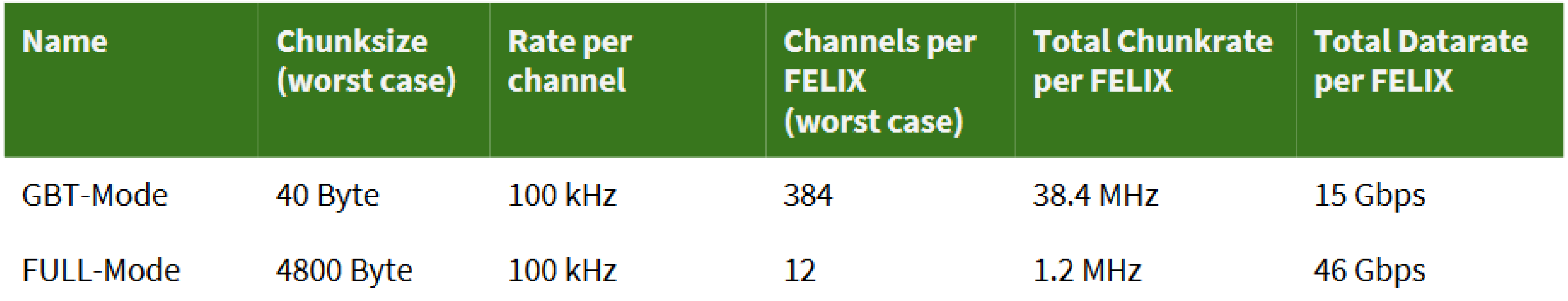
Hardware Status



- Currently in series production phase
  - Scheduled delivery of full complement of FLX-712 cards, plus FELIX and SW ROD servers at end of 2019
- 23 pre-production cards delivered through spring/summer 2019 and subjected to robust series of tests
  - Low level: Eye Scans/BER tests, impedance, thermal cycling, X-ray & X-ray fluorescence
  - High Level: Functional tests – dataflow, BUSY system integration, long term stability
  - All cards fully validated, majority now circulated to ATLAS sub-detector teams to enable surface commissioning of detector components and trigger electronics
- Total system size – approx. 95 I/O cards, 60 FELIX servers, 30 SW RODs
- Aiming to install in ATLAS electronics cavern in early 2020

Performance, Integration and Commissioning

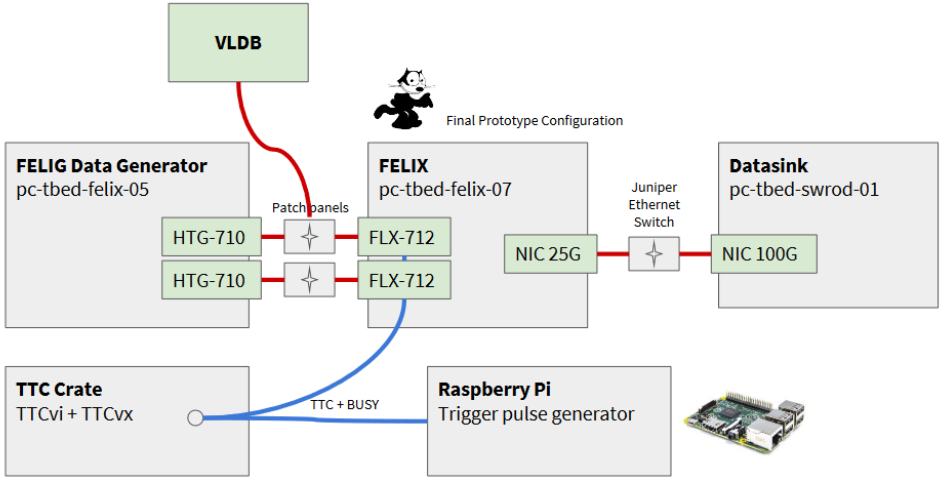
- FLX-712 pre-production cards and servers undergoing thorough series of tests in realistic use cases
- Run at realistic L1 trigger rate with random packet sizes and trigger patterns
  - Use worst case packet sizes below as average value for distribution
  - L1 Accepts can arrive close together in groups
- Test interface with Detector Control System (DCS)
  - Control data moving in to-detector direction in parallel to bulk dataflow

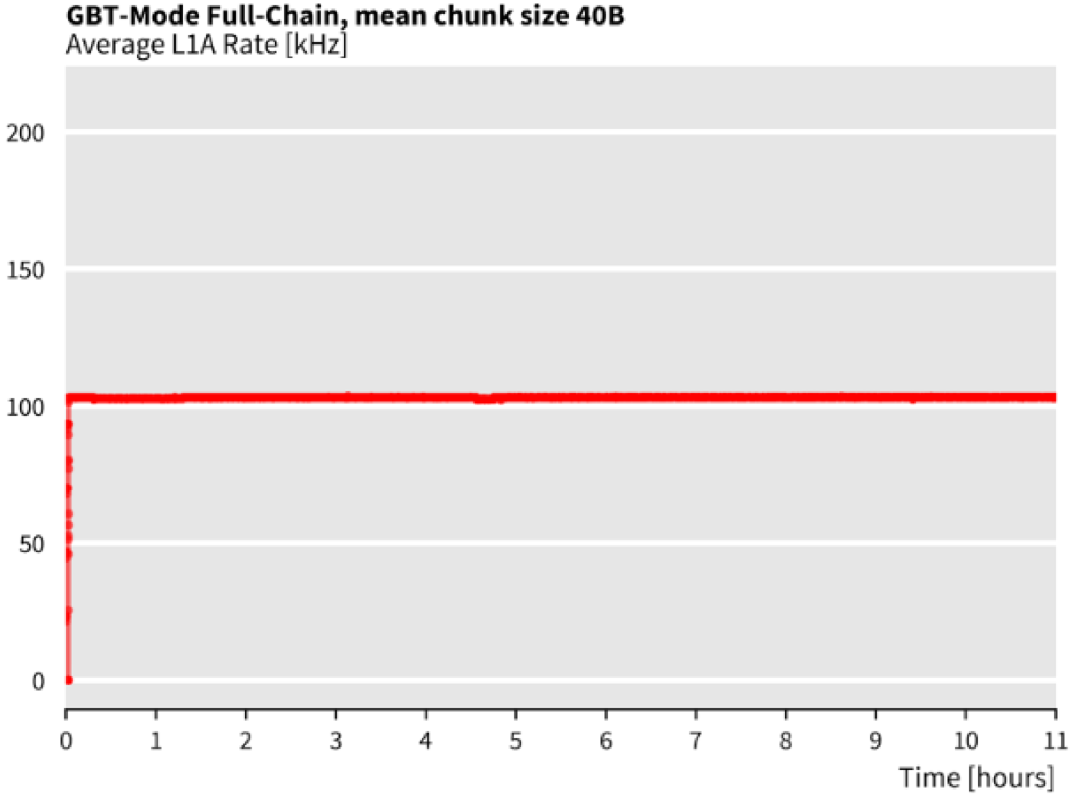


Name	Chunksize (worst case)	Rate per channel	Channels per FELIX (worst case)	Total Chunkrate per FELIX	Total Datarate per FELIX
GBT-Mode	40 Byte	100 kHz	384	38.4 MHz	15 Gbps
FULL-Mode	4800 Byte	100 kHz	12	1.2 MHz	46 Gbps

Performance, Integration and Commissioning

- GBT Mode
  - Stable multi-hour operation mimicking duration of longer than average LHC fill
  - Verified reliable parallel communication with VLDB (CERN test board featuring DCS components)
    - More complete DCS test in development with full software stack

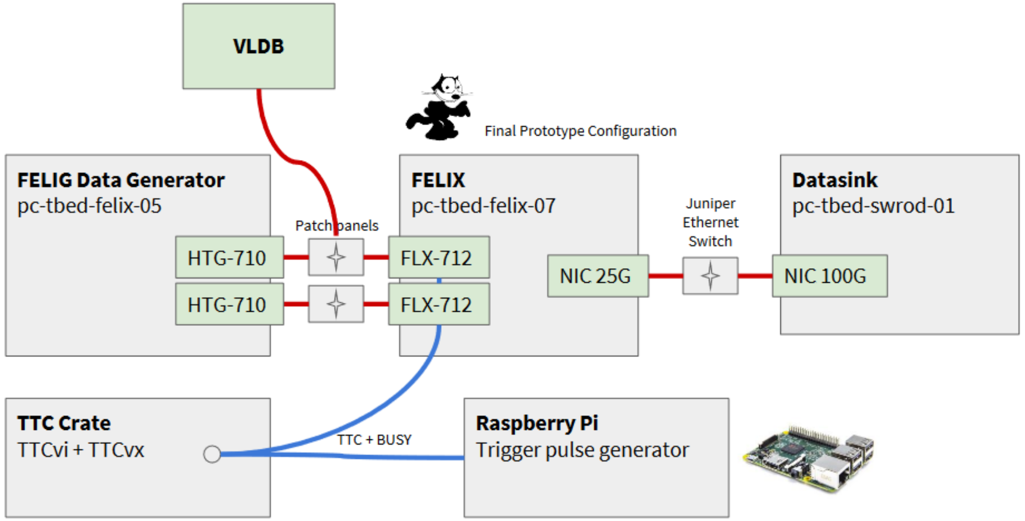




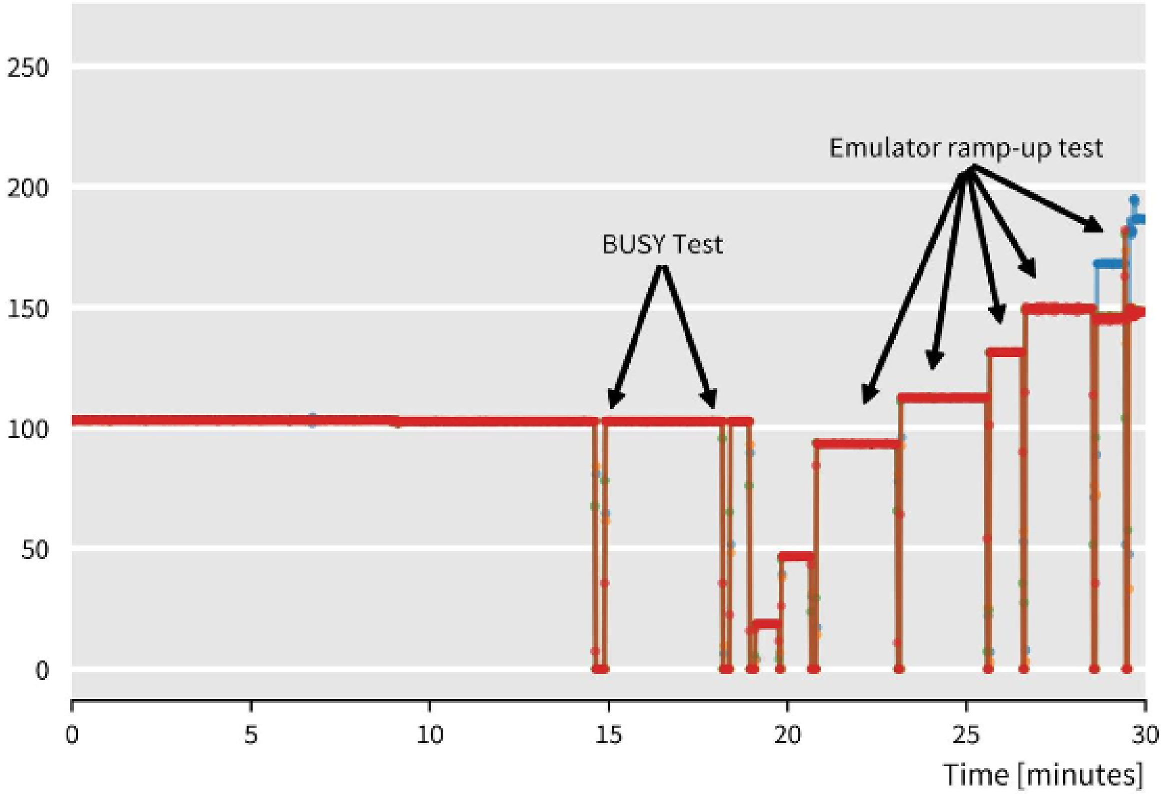
Approx 12.5 Gbps Network Throughput

Performance, Integration and Commissioning

- GBT Mode
  - Test correct propagation of BUSY signal – leading to dataflow halt
    - Manually triggered
  - Finally ramp emulators to 150 kHz, demonstrating rates 50% above expectation

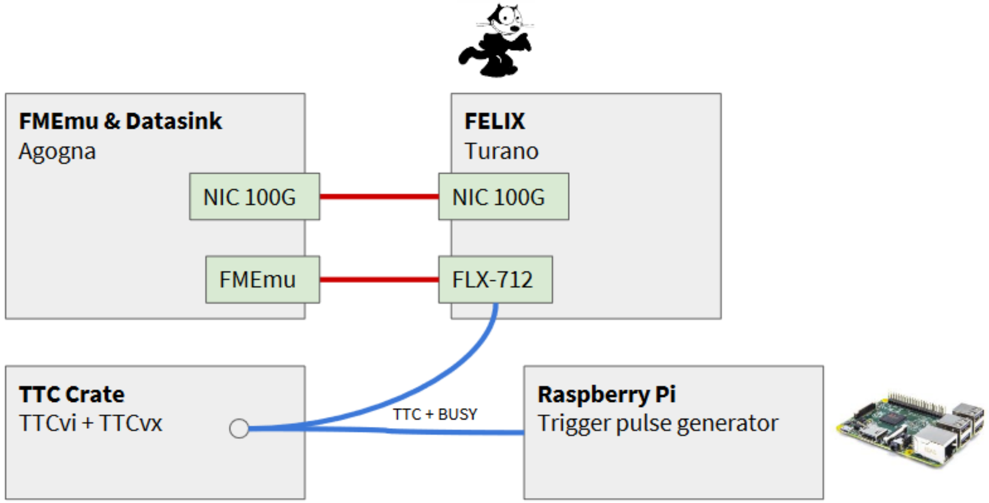


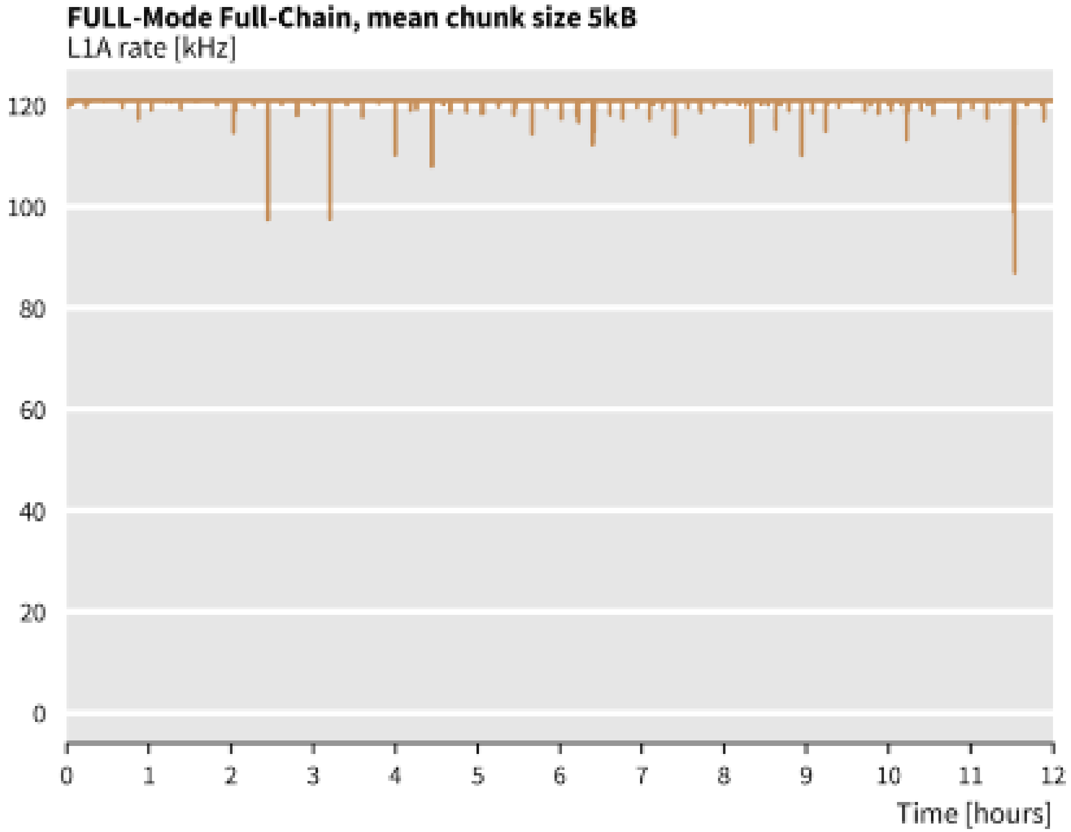
GBT-Mode Full-Chain, mean chunk size 40B
Message rate per channel at client [kHz]



Performance, Integration and Commissioning

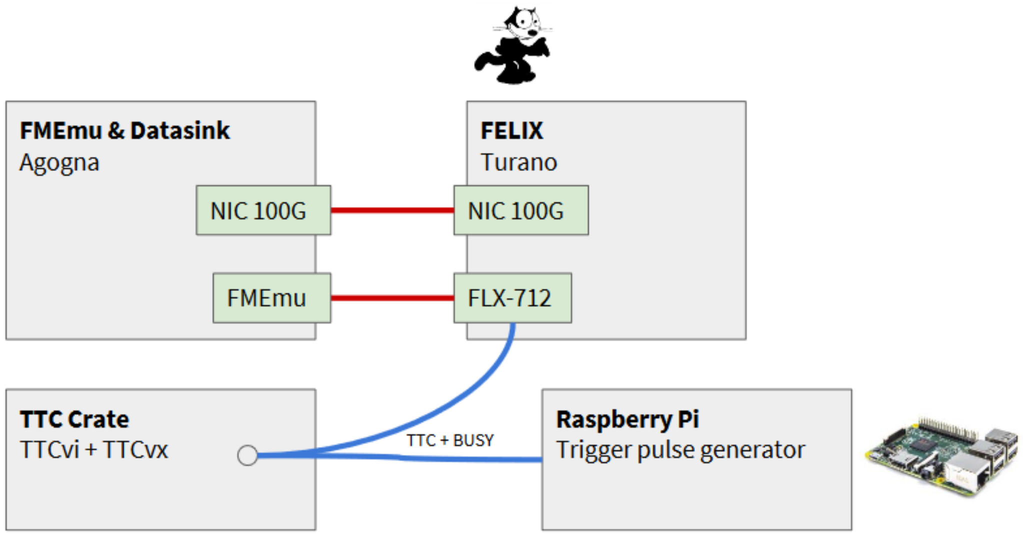
- FULL Mode
  - Stable multi-hour operation mimicking duration of longer than average LHC fill





Performance, Integration and Commissioning

- Stress test – increase L1 rate until backpressure mechanism kicks in to cap dataflow (XOFF)
    - Achieved 300 kHz



FMEmu & Datasink
Agogna

NIC 100G

FMEmu

FELIX
Turano

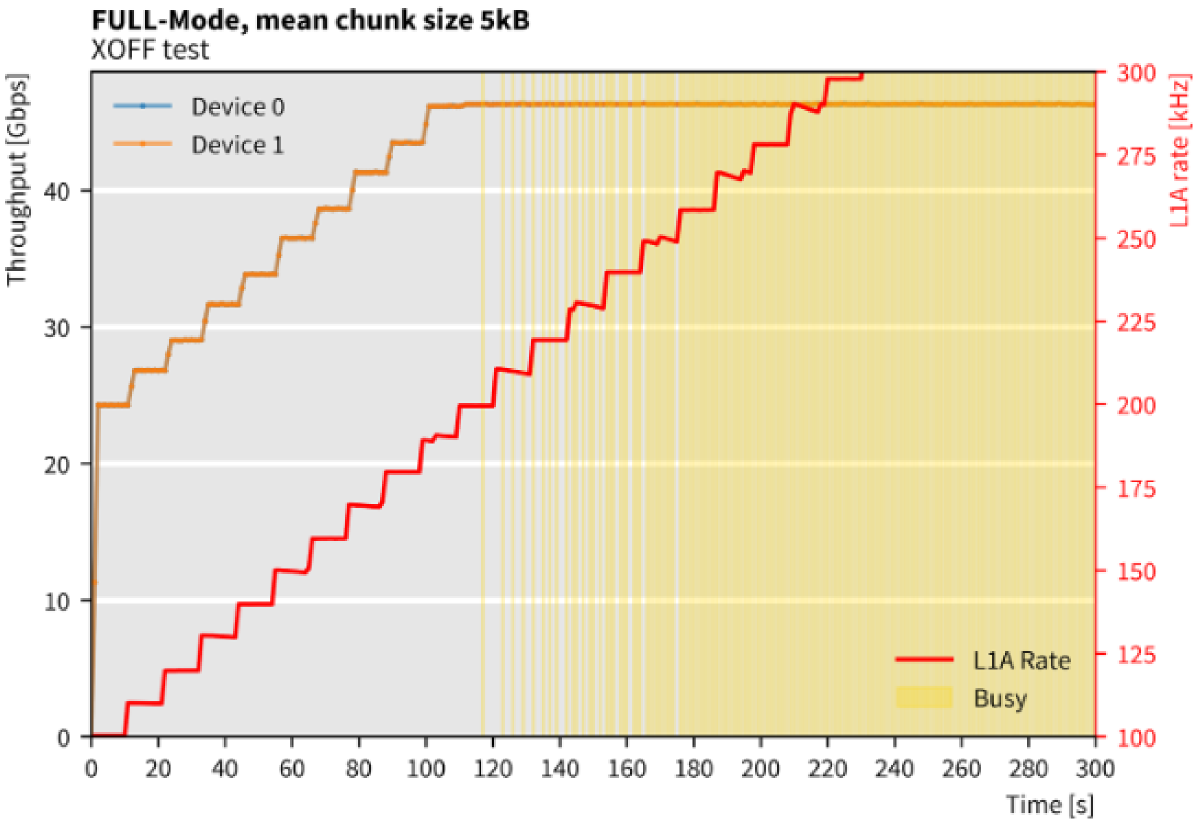
NIC 100G

FLX-712

TTC Crate
TTCvi + TTCvx

TTC + BUSY

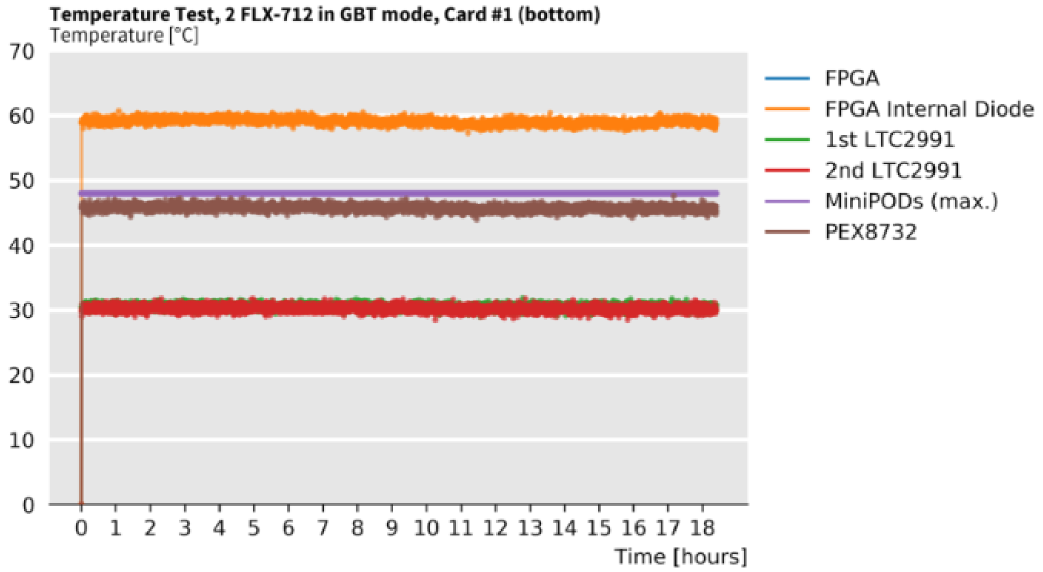
Raspberry Pi
Trigger pulse generator

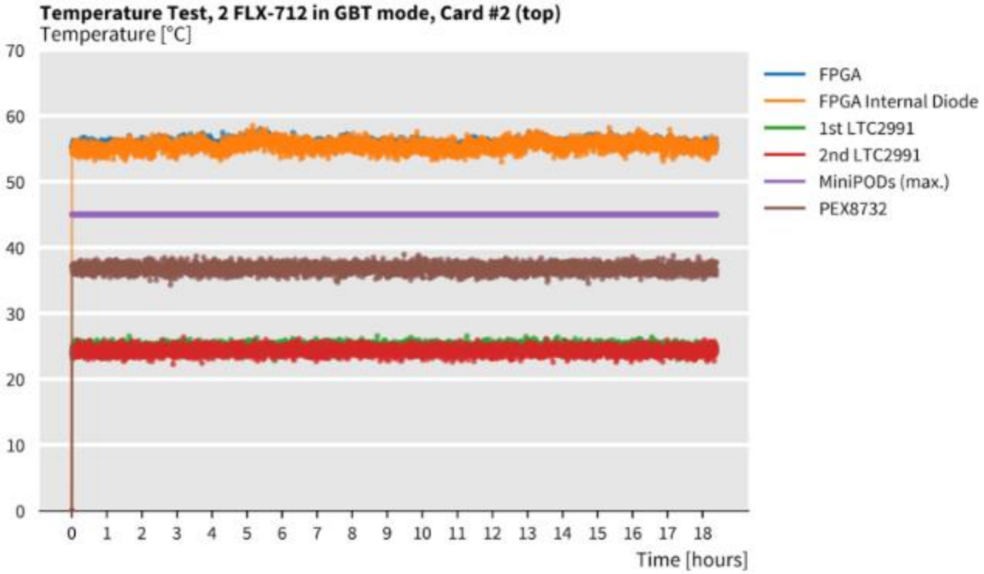


FULL-Mode, mean chunk size 5kB
XOFF test

Performance, Integration and Commissioning

- Monitored temperatures of FLX-712 card components in all integration tests
- Example here from most difficult scenario – two cards in 2U server with 48 MiniPODs active
  - Rack not optimally cooled
- Stable at acceptable levels over 18 hours





Summary and Outlook

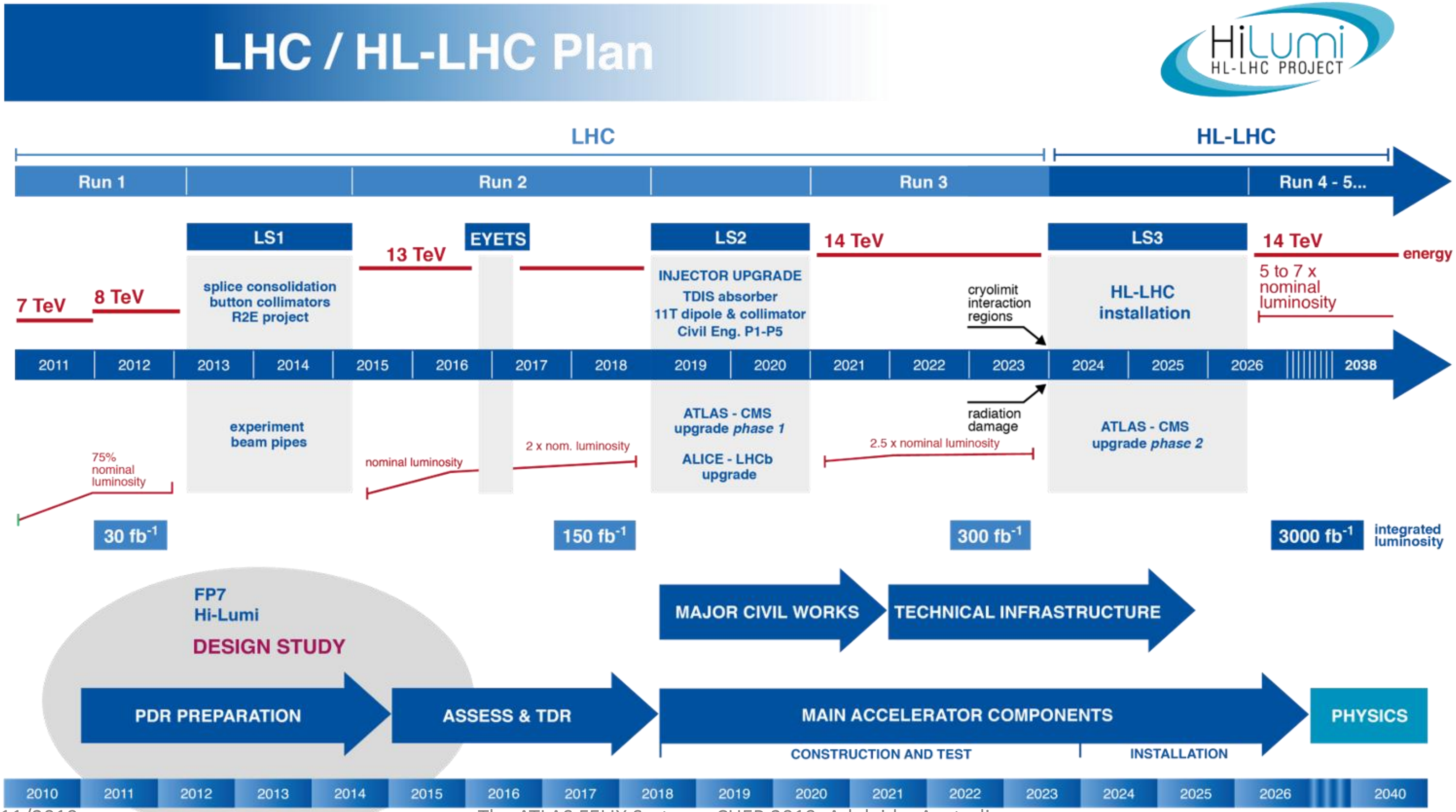
- ATLAS TDAQ system underpinned successful data taking in Run 1 and Run 2
- Run 3 will see first part of move towards new common readout platform (FELIX and SW ROD)
  - See talk by Joern Schumacher on Thursday afternoon (Track 5) for more on FELIX software
  - See talk by Revital Kopeliansky on Thursday afternoon (Track 1) for more on Run 4
- Final FELIX and SW ROD hardware in final production
  - Performance of pre-production samples well exceed Run 3 requirements
    - In use by ATLAS detector systems to facilitate integration of front-end electronics
- 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!

Extra Slides

References

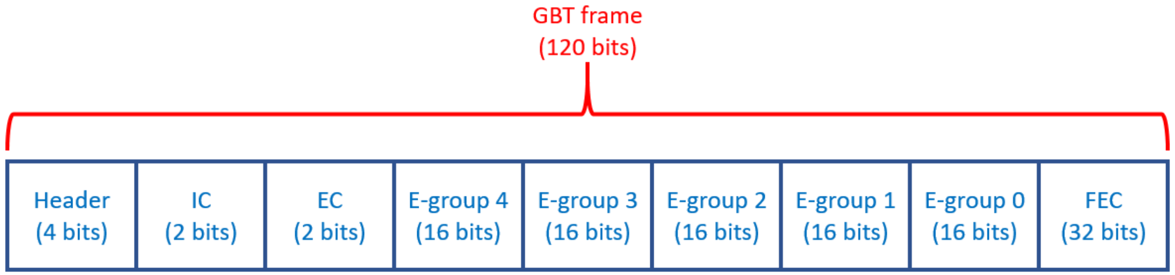
- ATLAS TDAQ Phase-I Technical Design Report
  - https://cds.cern.ch/record/1602235
- FELIX Project Webpage
  - https://atlas-project-felix.web.cern.ch/atlas-project-felix/

LHC Evolution and Overall Upgrade Schedule



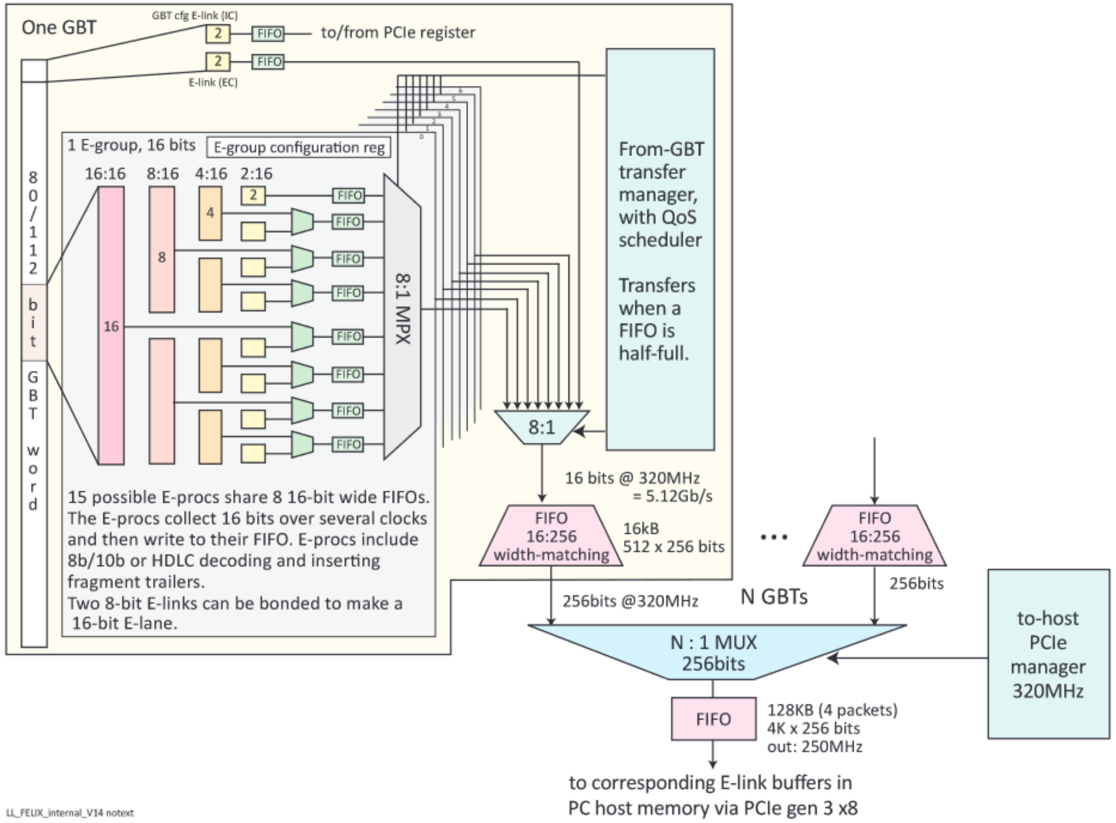
Inside FELIX - GBT

- 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)

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



One GBT
GBT cfg E-link (IC)
2
FIFO
to/from PCIe register
2
FIFO
E-link (EC)
80/112 bit GBT word
1 E-group, 16 bits
E-group configuration reg
16:16
8:16
4:16
2:16
16
8
4
2
FIFO
8:1 MPX
From-GBT transfer manager, with QoS scheduler
Transfers when a FIFO is half-full.
8:1
16 bits @ 320MHz = 5.12Gb/s
FIFO 16:256 width-matching
16kB 512 x 256 bits
...
FIFO 16:256 width-matching
256bits
256bits @320MHz
N GBTs
to-host PCIe manager 320MHz
N : 1 MUX 256bits
FIFO
128KB (4 packets) 4K x 256 bits out: 250MHz
to corresponding E-link buffers in PC host memory via PCIe gen 3 x8
LL_FELIX_internal_V14 notext

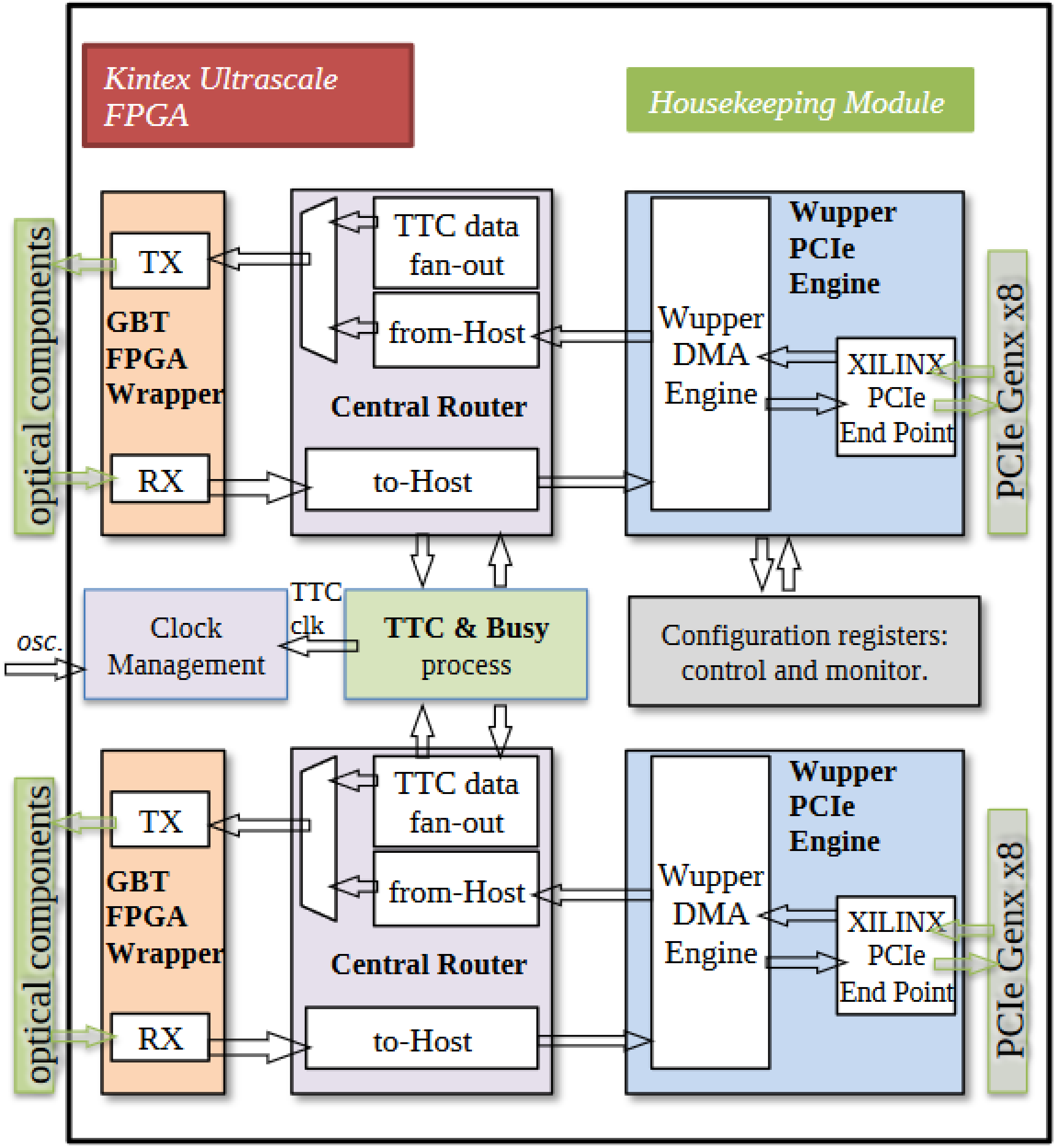
15 possible E-procs share 8 16-bit wide FIFOs. The E-procs collect 16 bits over several clocks and then write to their FIFO. E-procs include 8b/10b or HDLC decoding and inserting fragment trailers.
Two 8-bit E-links can be bonded to make a 16-bit E-lane.

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

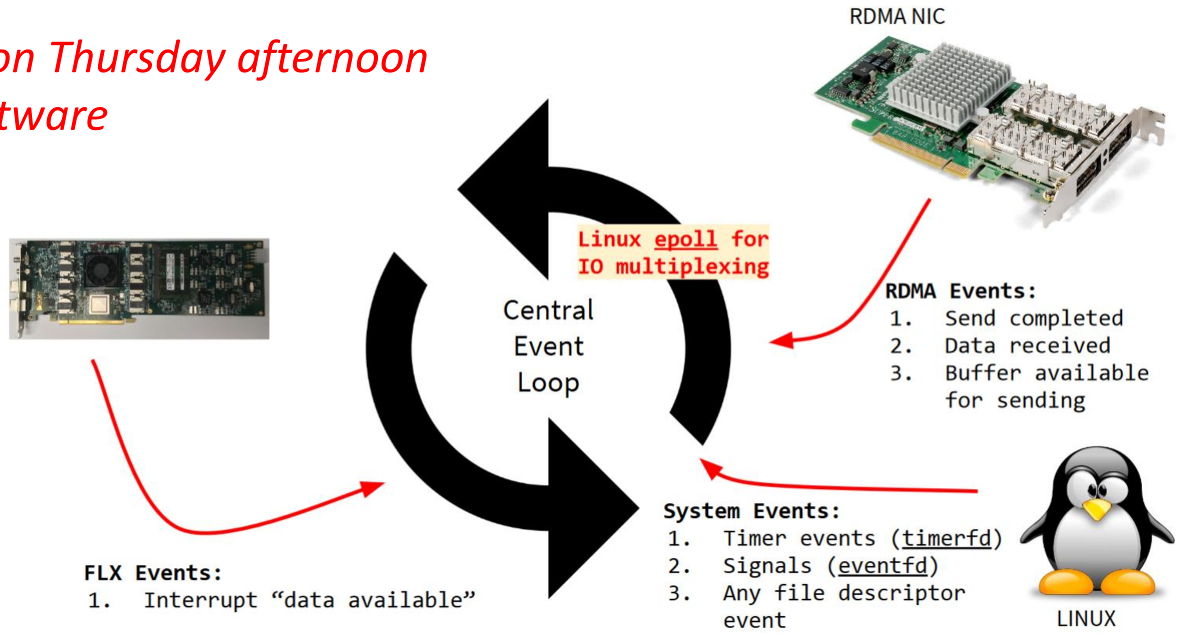
- FELIX firmware
  - Two identical sets of blocks, each attached to separate PCIe Gen3 x 8 end point
  - Bi-directional communication paths to and from front-end
  - Link wrapper (GBT or FULL mode)
  - TTC and BUSY interface wrapper
  - Central Router
    - Core of FELIX functionality
    - Decodes and 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

See talk by Joern Schumacher on Thursday afternoon (Track 5) for more on FELIX software

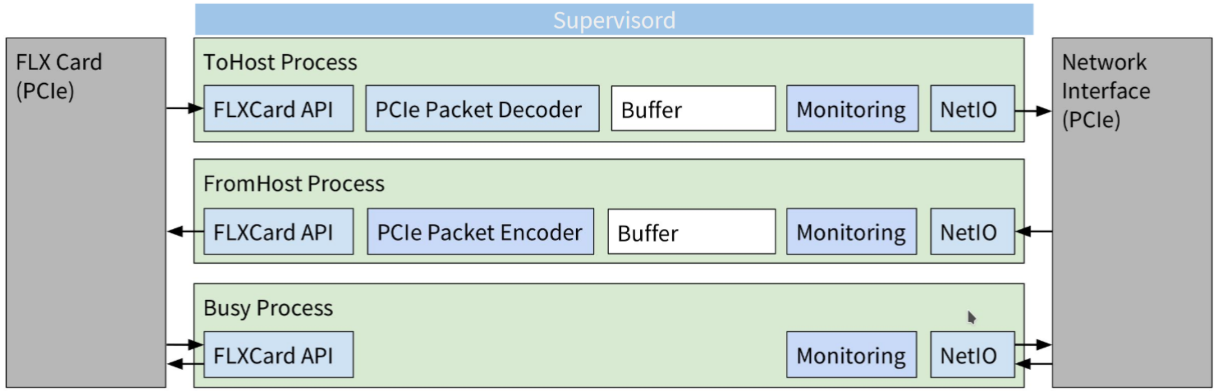
- FELIX Software
  - Primary dataflow and control through **FelixStar** application
    - Running as a daemon on host server
  - FELIX firmware transfers data to host ring buffer via continuous DMA
    - Data split into fixed size 'blocks' for transfer
  - Event driven software architecture
    - Incoming DMA triggers packet processing and transfer to NIC
      - Re-composes blocks back into complete packets
      - Eliminates need to make copies of data
      - Maximises 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



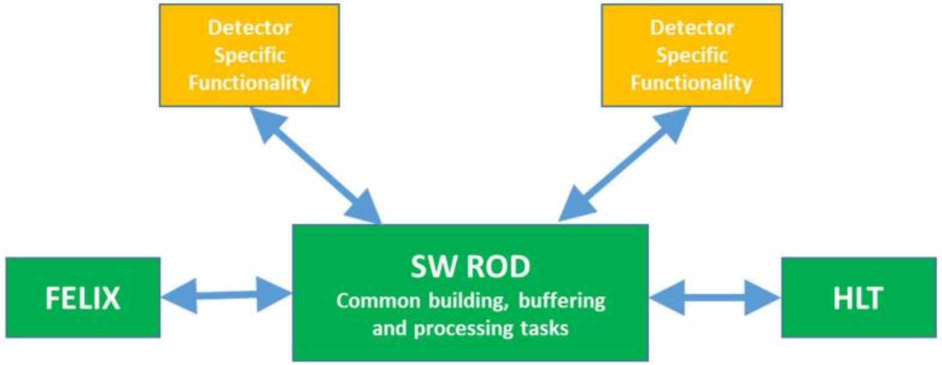
RDMA Events:
1. Send completed
2. Data received
3. Buffer available for sending

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

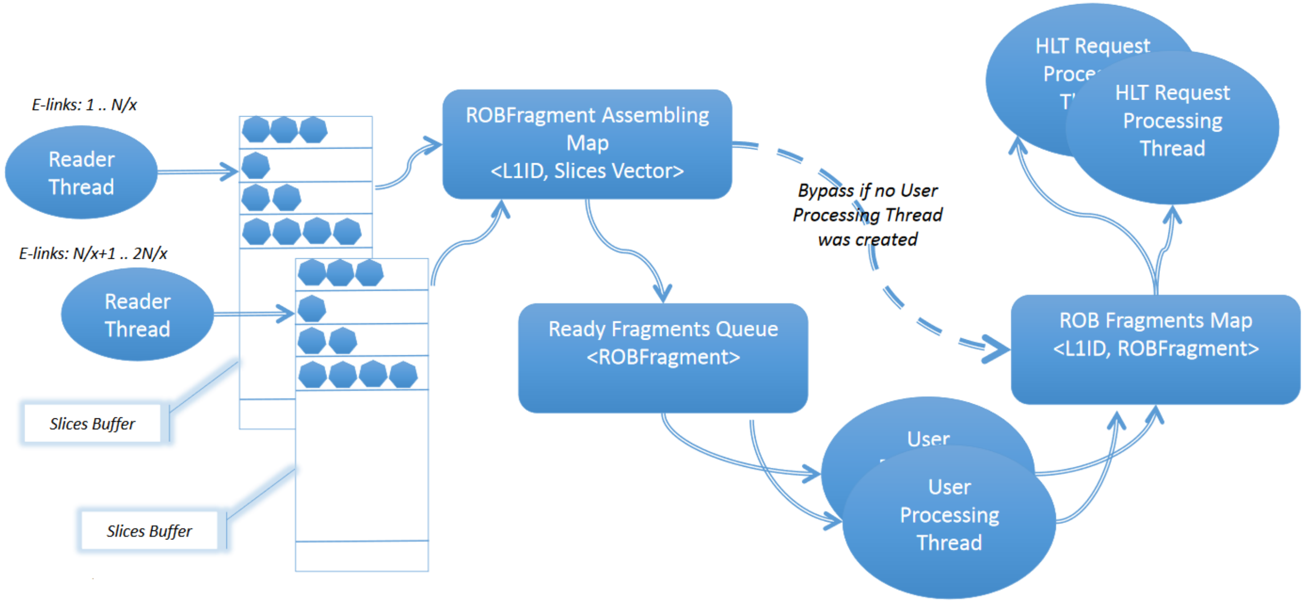
FLX Events:
1. Interrupt "data available"



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
  - Subscribes to and receives event data from FELIX and facilitates sub-detector specific processing
    - Data handling actions in original hardware RODs now reside here
    - Data coming in on multiple links can be aggregated into larger packets for transfer to HLT
      - Data from multiple FELIX servers handled by a single SW ROD
    - Custom and common monitoring tools supported
    - Possible to sample SW ROD event data from other process on 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