FELIX and the SW ROD

Commissioning the new detector interface for the ATLAS trigger and readout system

William Panduro Vazquez

On behalf of the ATLAS TDAQ Collaboration

ATLAS Upgrades for LHC Run 3

2027

Run 4+ (HL-LHC)
√s = 14 TeV
max pileup 140-200

2024

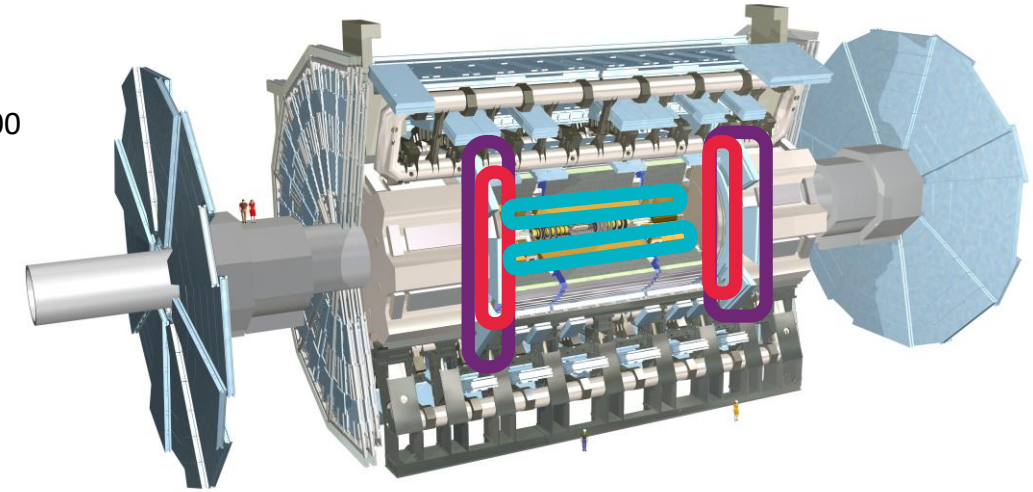
2022

Run 3
√s = 13-14 TeV
max pileup 60

2018

2015

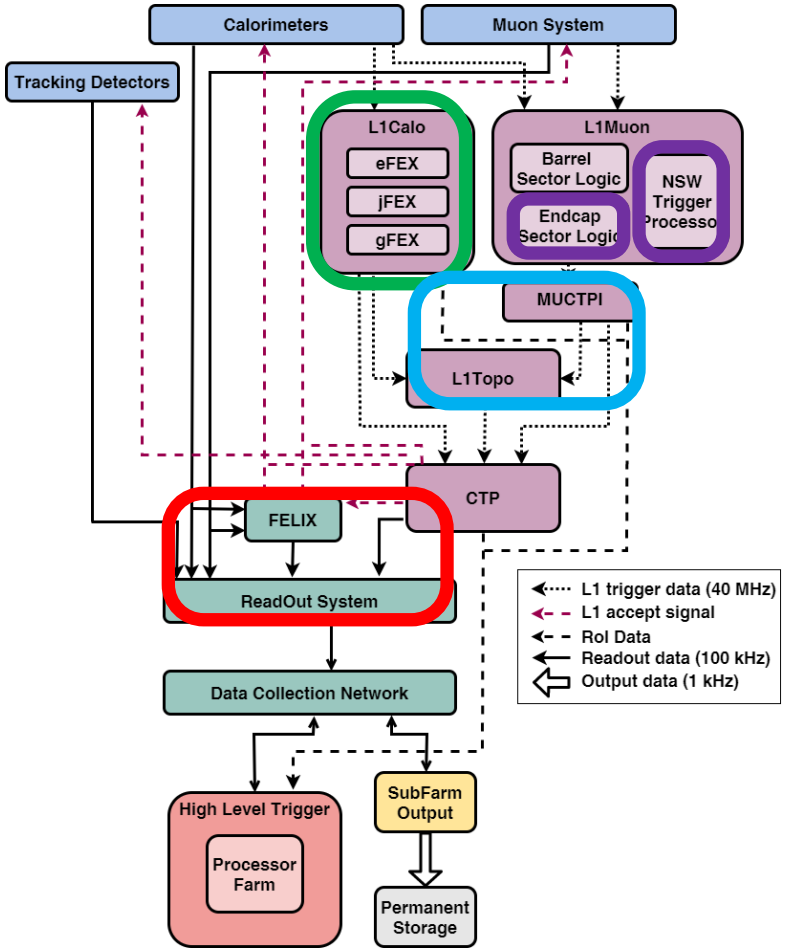
Run 2
√s = 13 TeV
max pileup 60



- Upgraded detector systems
  - Muon System
    - New Small Wheels
    - Inner Barrel RPCs (BIS 7/8)
  - Calorimeters
    - Liquid Argon (LAr) digital readout
    - Tile Run 4 demonstrator

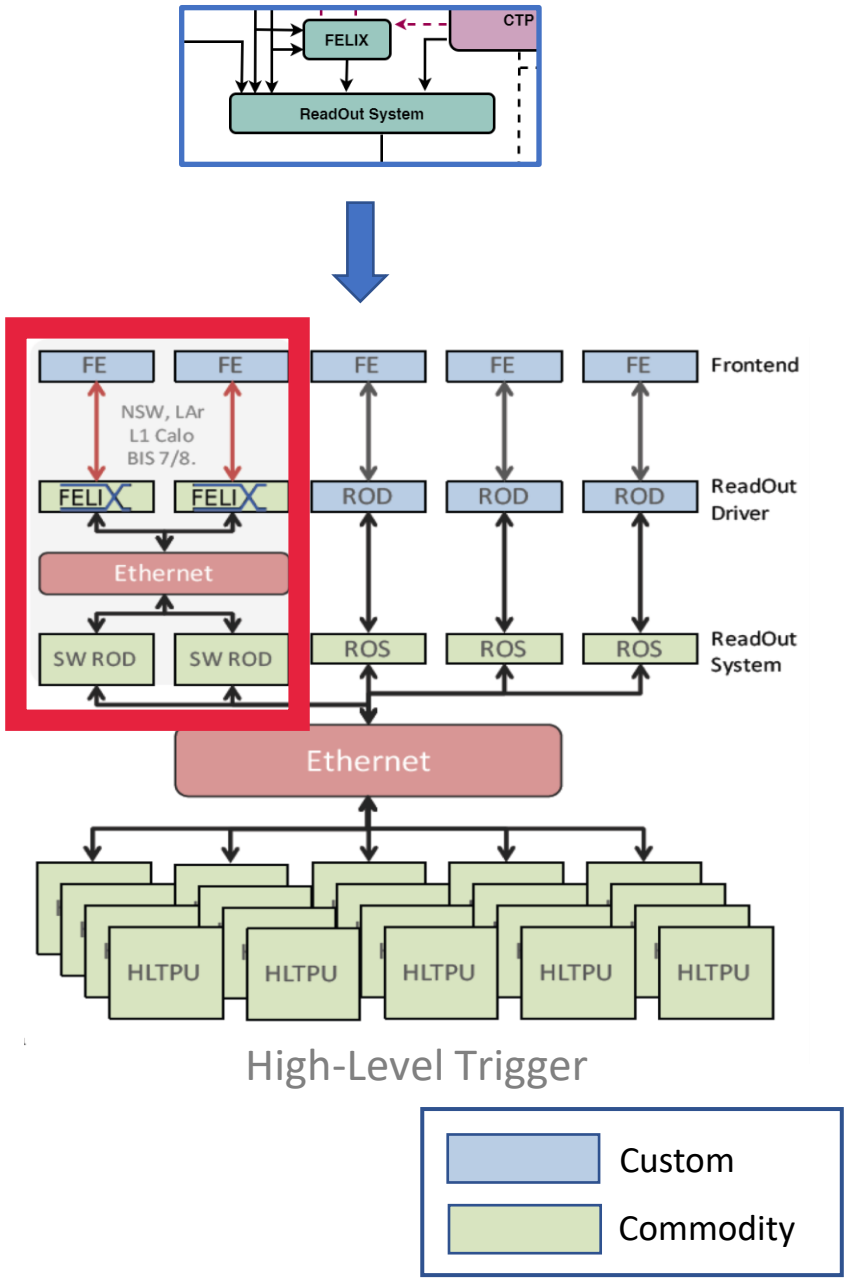
Pileup = number of interactions per LHC bunch crossing

- Trigger & DAQ Upgrades
  - L1 Calorimeter Trigger
  - L1 Muon Endcap Sector Logic Electronic & NSW Trigger Processor
  - Upgraded MuCTPi and Topological Trigger
  - FELIX & Software Readout Driver (SW ROD)



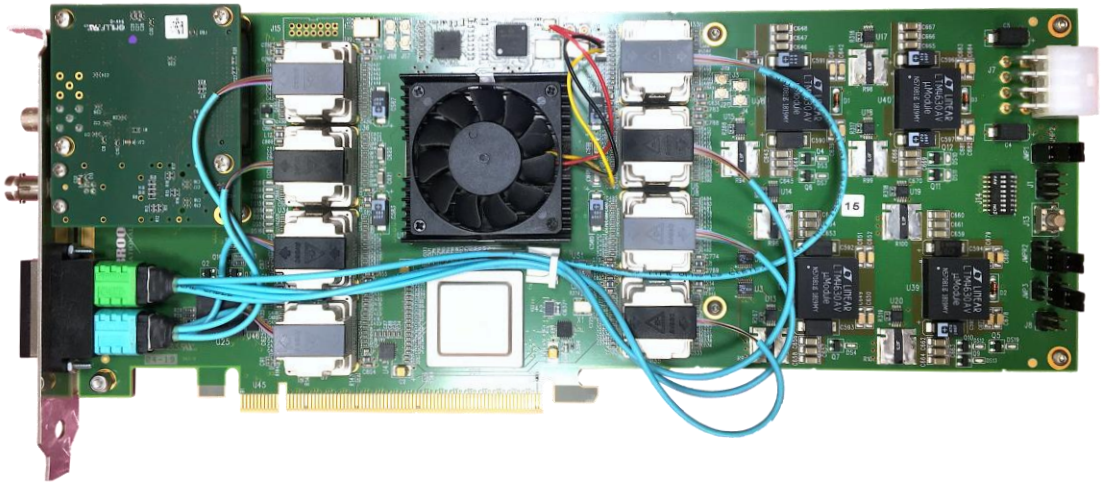
DAQ architecture in Run 3

- Overall goal to reduce amount of custom hardware in readout path and provide common platform for all ATLAS subsystems
  - Reduced hardware/firmware effort due to common and commodity components
  - Make the most of 20 years of technological advancement since original ATLAS designs
  - Data transferred to industry standard networks as early as possible in readout path
- Introducing FELIX and SW ROD components alongside legacy Readout System (ROS)
  - Only for the new trigger and detector systems mentioned in previous slide
  - Extend to rest of ATLAS in Run 4
- Front-End Link eXchange (FELIX)
  - Custom PCIe card hosted in commercial server
  - FE interface for readout, configuration, trigger, clock distribution, Slow Control monitoring and BUSY
- Software ROD (SW ROD)
  - Software running on commercial servers
  - Subscribes to FELIX data streams for primary readout
  - Able to build event fragments from multiple FELIX streams and facilitate detector-specific data processing
  - Buffers data during HLT decision, exposing identical interface to legacy ROS
- System size in Run 3
  - FELIX: Approx 100 cards, 60 host PCs
  - SW ROD: Approx 30



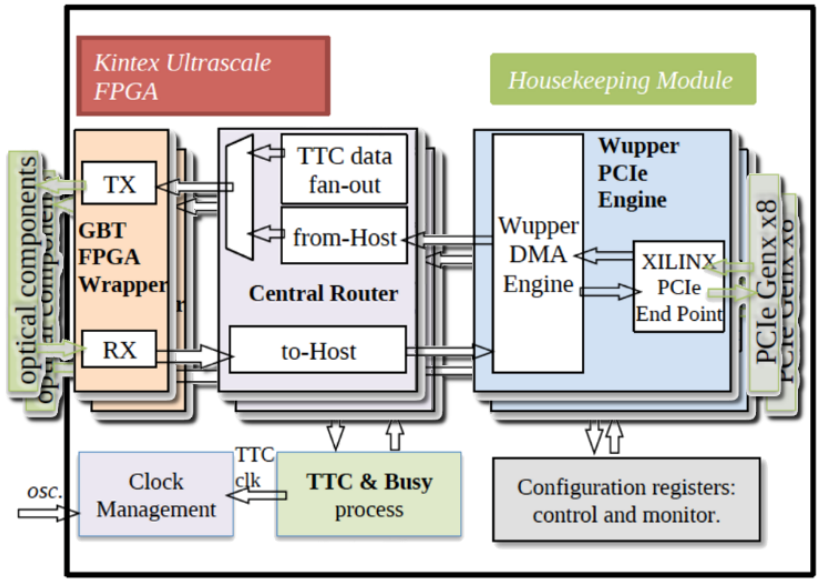
The FELIX Card (FLX-712)

- Common hardware platform for all detectors using FELIX readout
  - FPGA: Xilinx Kintex UltraScale XCKU115
  - Can host either 4 or 8 MiniPODs to support 24 or 48 bidirectional optical links
  - 16-lane PCIe Gen3 (two 8-lane Endpoints with a switch)
  - Flash and Micro-controller to support firmware update
  - Interface to the Timing Trigger and Control (TTC) system via swappable mezzanine
  - BUSY output over LEMO cable
- Host Server
  - Intel Xeon E5-1660 v4 @ 3.2GHz
  - 32 GB DDR4 2667 MHz memory
  - Mellanox Connect-X 5 NIC: 25/100 GbE



Firmware and Flavours

- GBT Mode
  - 24 links, 4.8 Gb/s each conforming to GBT protocol used by the radiation tolerant GBTx ASIC
  - Each link aggregates up to 42 *E-Links* with configurable bandwidth (80/160/320 Mbps)
  - Each E-Link carries a signal with either 8b10b or HDLC encoding, or TTC distribution to the front end electronics
  - Heavy utilisation of FPGA resources for configurable logic
- FULL Mode
  - 24 links, 9.6 Gb/s each, 8b10b encoded
  - Designed for FPGA-to-FPGA connection outside of rad hard environments
  - 12 links saturate the PCIe bandwidth if bandwidth is fully occupied
  - To front end: 4.8 Gb/s GBT with E-Links for TTC distribution and configuration

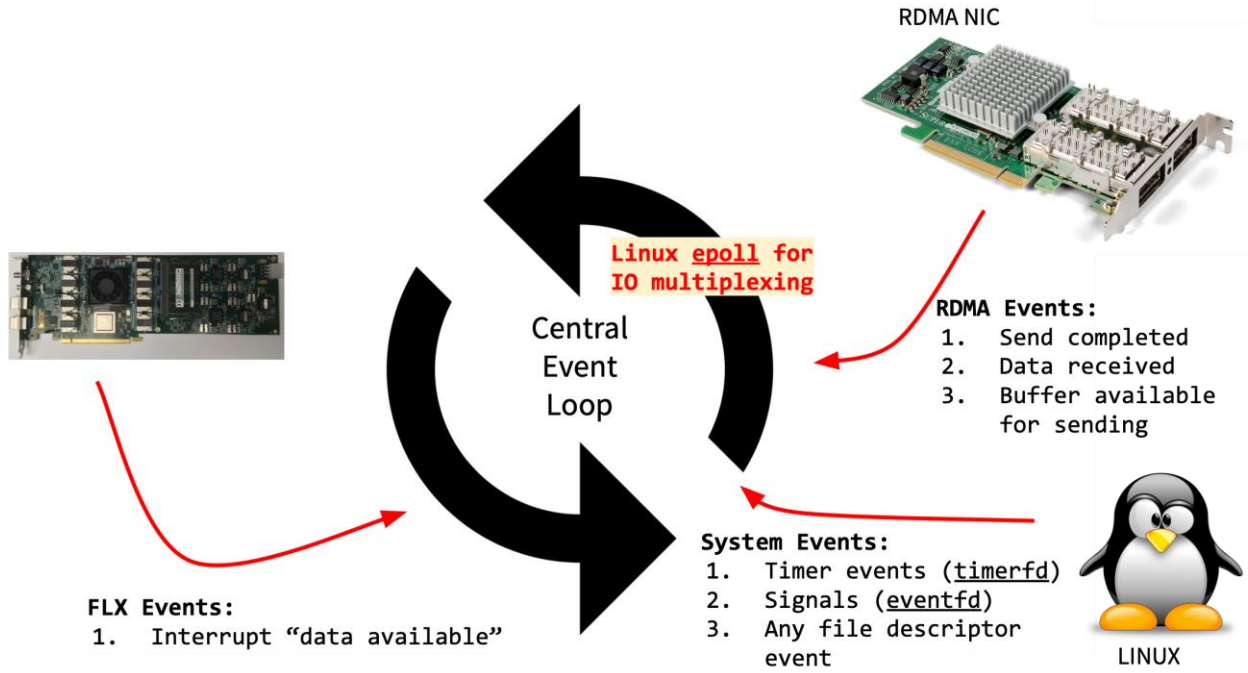


Mode	Packet size	Rate per link	Links per card	Total packet rate per card	Total data rate per card
FULL	4800 bytes	100 kHz	12	1.2 MHz	46 Gbps
GBT	40 bytes	100 kHz	192	19.2 MHz	7.5 Gbps

Heaviest FELIX use cases expected in ATLAS

Software

- Higher level routing platform (felix-star) running as daemon on FELIX host
  - Routes data to and from card from network peers
  - 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
  - Network protocol (netio-next) coalesces small messages to decrease I/O overhead

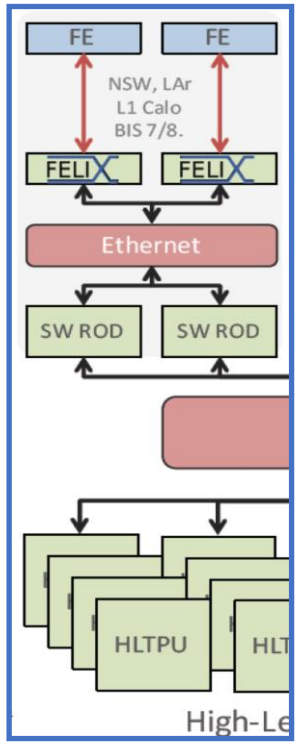


- Performance well in excess of Run 3 requirements in lab tests (see extra slides for details)
- Device driver and suite of low level management tools provided to aid commissioning and integration

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 (buffering, transfer to HLT, writing to disk, calibration processing…)
- Performance Requirements

Mode	FELIX cards per SW ROD	Total packet rate	Total data rate
FULL	1	1.2 (2.4) MHz	46 Gb/s (5-kB packets)
GBT	6	**115 MHz**	37 Gb/s (40-byte packets)



- SW ROD Server Choice
  - Dual Intel Xeon Gold 5218 CPU @ 2.3 GHz
  - 16 x 2 physical cores
  - 96 GB DDR4 2667 MHz memory
  - Mellanox ConnectX-5 100 Gb to FELIX
  - Mellanox ConnectX-4 40 Gb to HLT

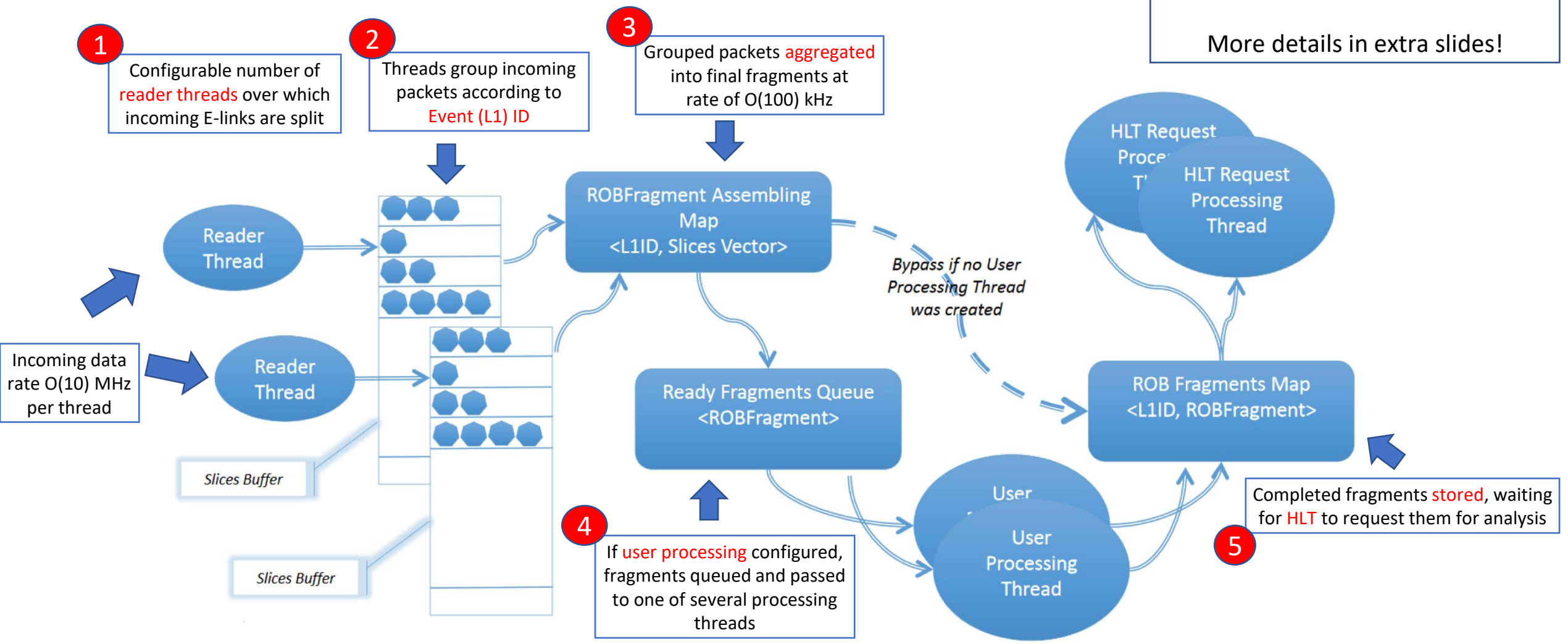
To be aggregated into bigger fragments according to their trigger IDs!

With ~7.4×10^{10} cores × Hz → max 640 CPU operations per input packet, including dataflow overhead!

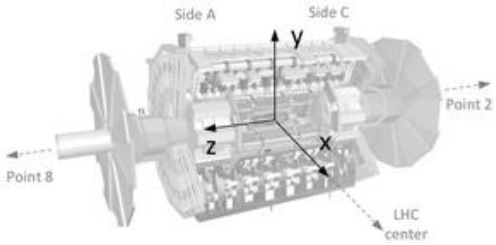
GBT Fragment Building Algorithm

Performance demonstrated to be well above Run 3 requirements

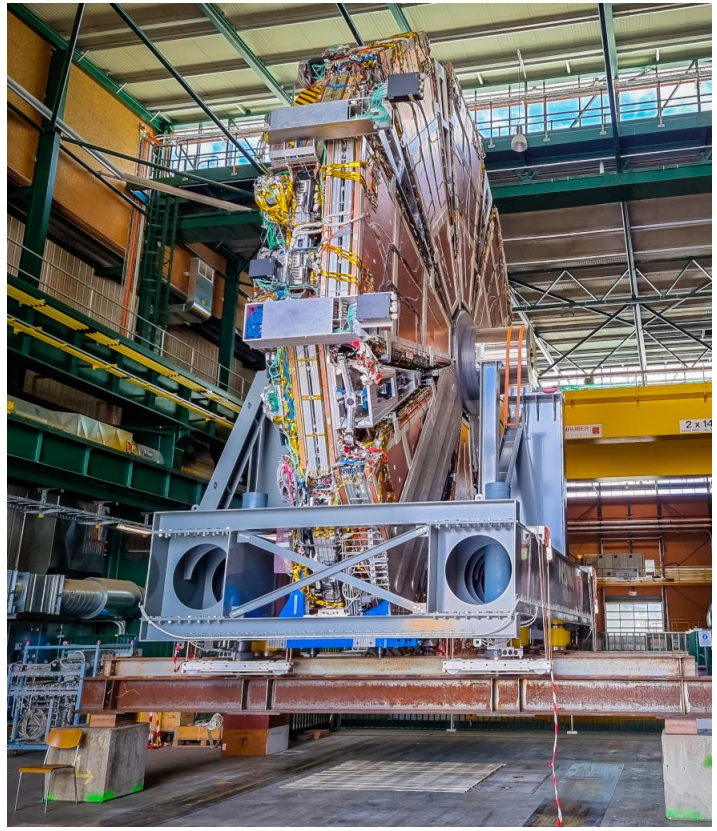
More details in extra slides!



Highlights from detector commissioning

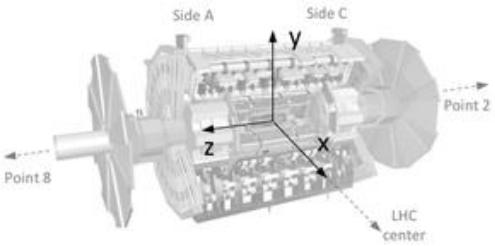


- New Small Wheels
  - Successful surface commissioning of first of two wheels (A-side)
    - FELIX platform operating stably in surface tests for both detector control and readout
    - Readout rates in excess of Run 3 requirements (100 kHz) demonstrated
    - Detector control interface operating stably with high uptime
  - Installation of A-side in ATLAS cavern took place on July 12th
    - All FELIX and SW ROD servers already in installed in electronics cavern
  - C-side surface commissioning now in full swing ahead of proposed installation in late 2021
- BIS 7/8
  - Updated readout electronics installed in ATLAS cavern and connected to FELIX and SW ROD
  - Successful low level communication tested, now moving to full software stack

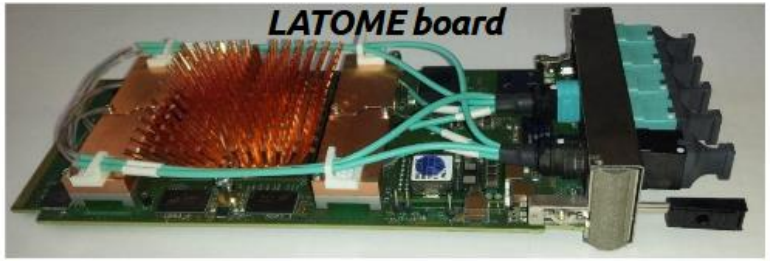


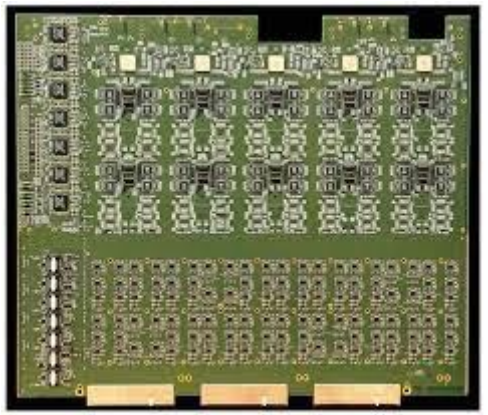
NSW-A on surface, ready for installation

Highlights from detector commissioning

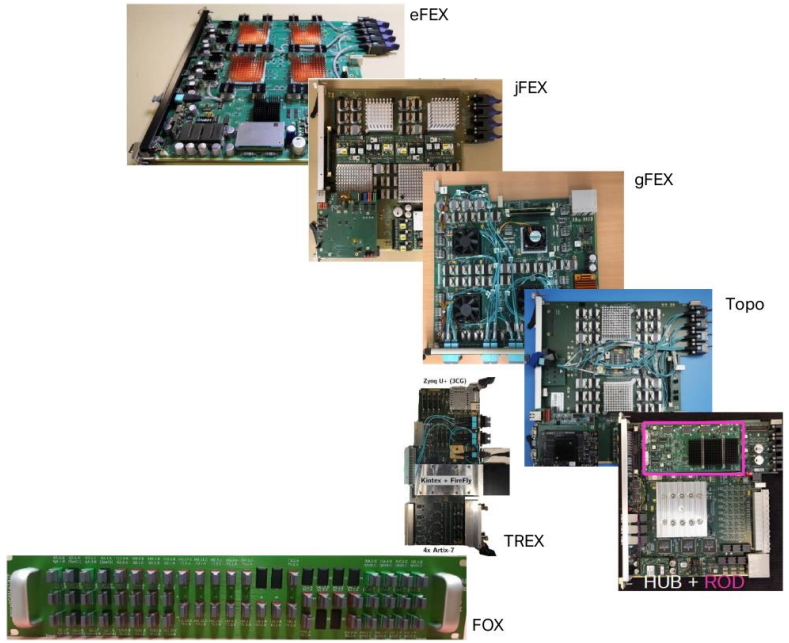


- LAr Calorimeter
  - FELIX and SW ROD installation complete
  - Integration with LAr electronics ongoing
  - Successful low rate tests of full A-side readout
  - A-side detector control interface operating stably with high uptime
  - C-side integration ramping up





LTDB board



- L1 Calorimeter Trigger
  - Successful surface tests of full FELIX and SW ROD readout path for majority of systems
  - In the process of installing and commissioning modules in ATLAS electronics cavern
    - FELIX and SW ROD cavern installation complete

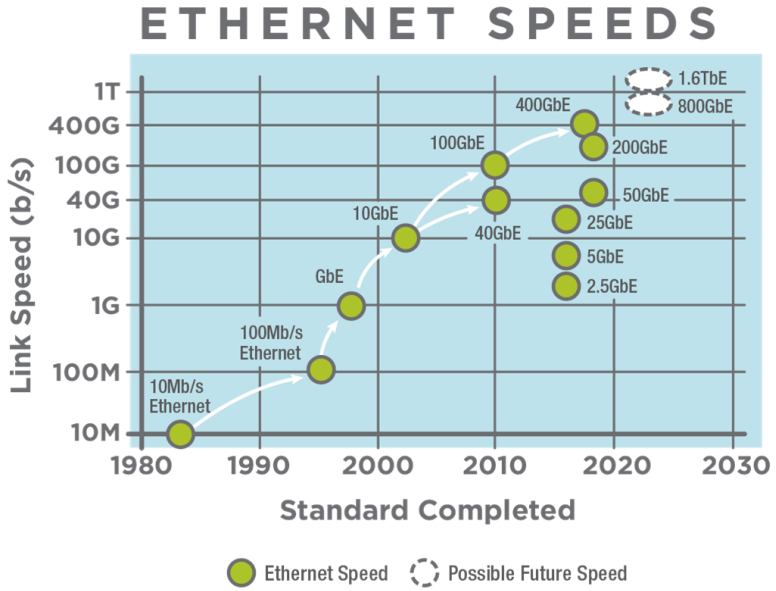
Summary and Outlook

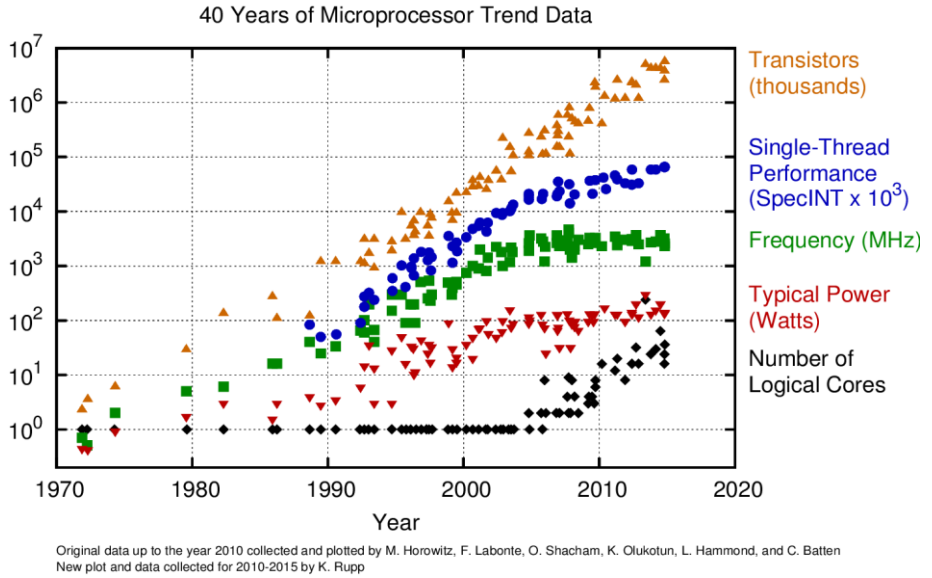
- In Run 3 ATLAS introducing FELIX and the SW ROD as the new readout interface
  - Run alongside legacy readout system, taking care of data from new trigger and detector systems installed for Run 3
  - Will become sole readout system in Run 4
  - Design makes the most of recent technological advancements for increased flexibility and reduced use of custom hardware
- FELIX and SW ROD hardware installation largely complete
  - Subdetector connection and commissioning ongoing throughout 2021
- Ongoing programme of firmware and software development in support of operations
- Performance exceeding the Run 3 requirements

Extra Slides

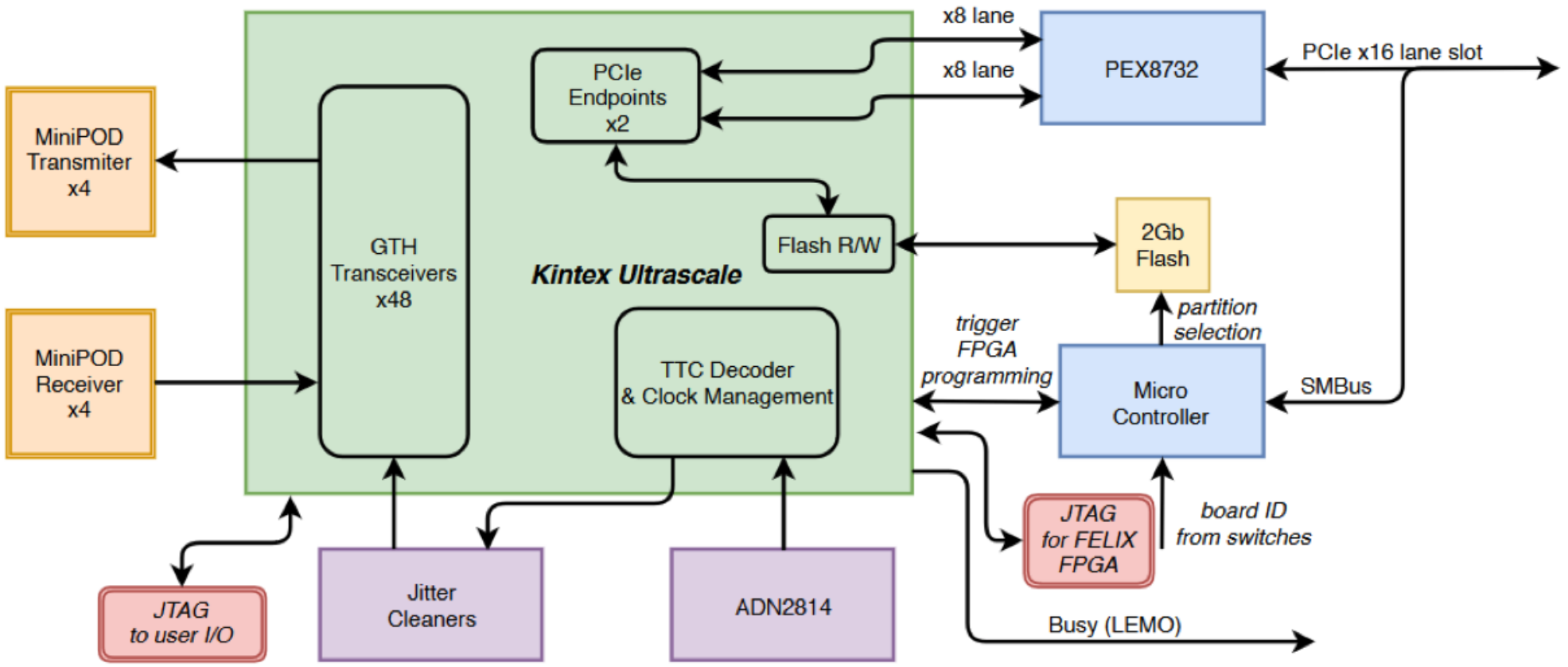
Common Challenges and Evolution

- 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

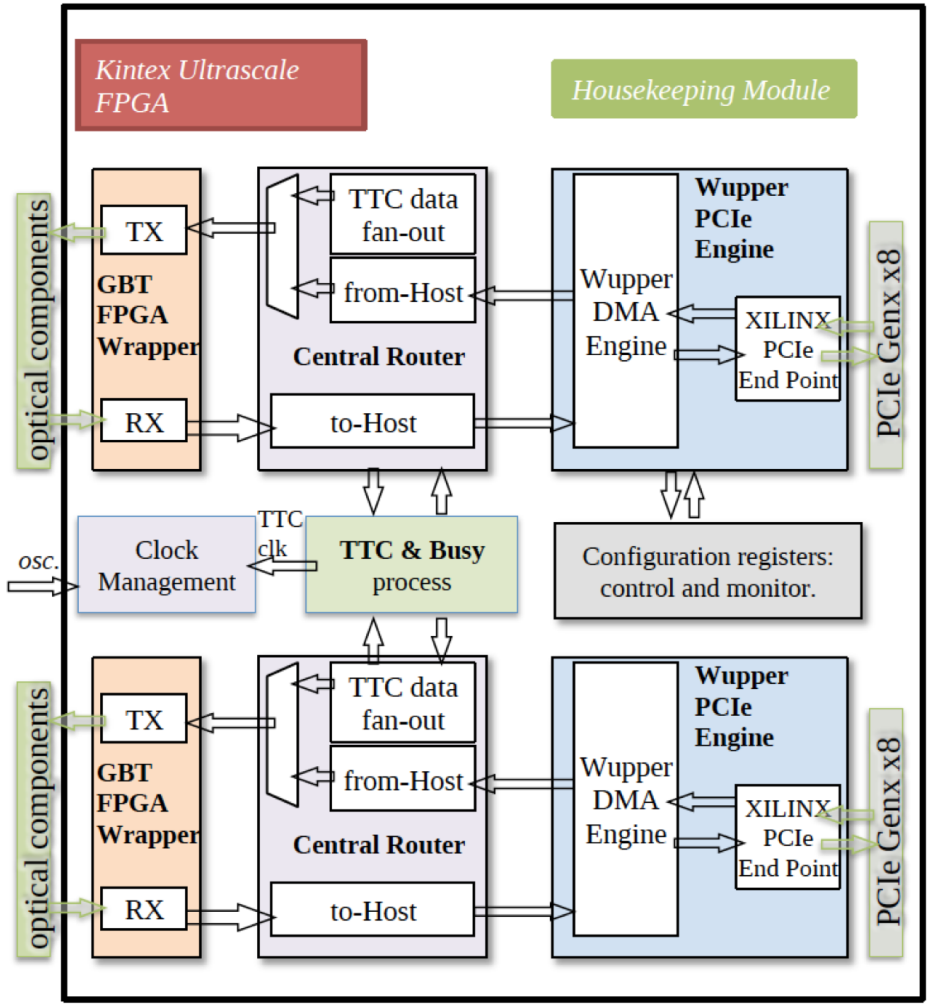




FELIX Hardware Block Diagram

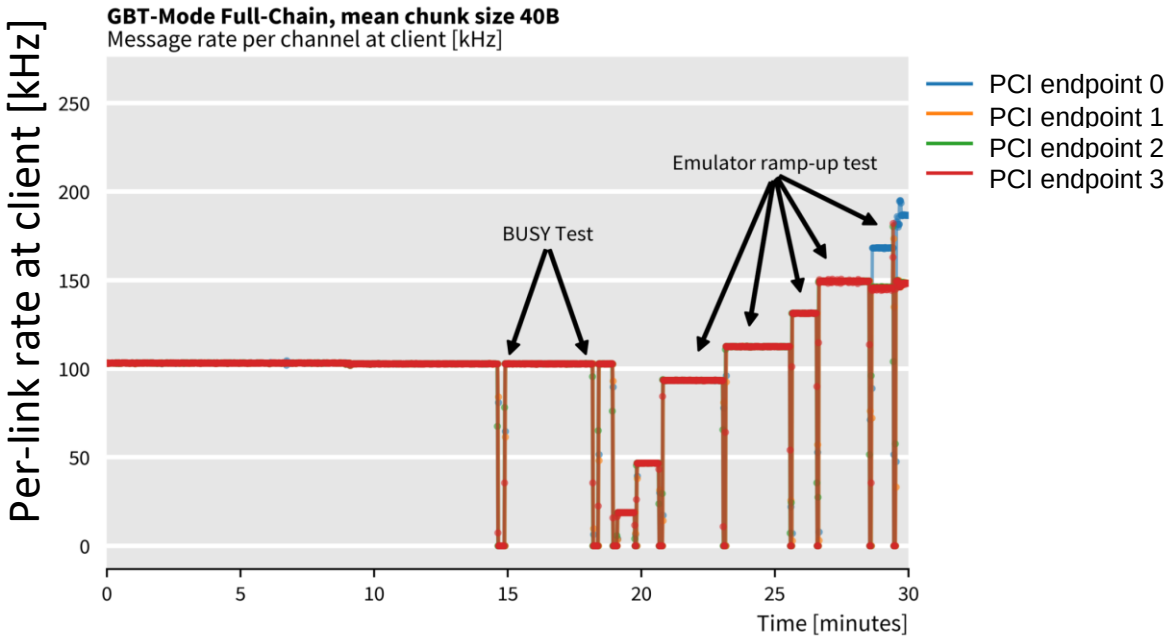


FELIX Firmware Architecture



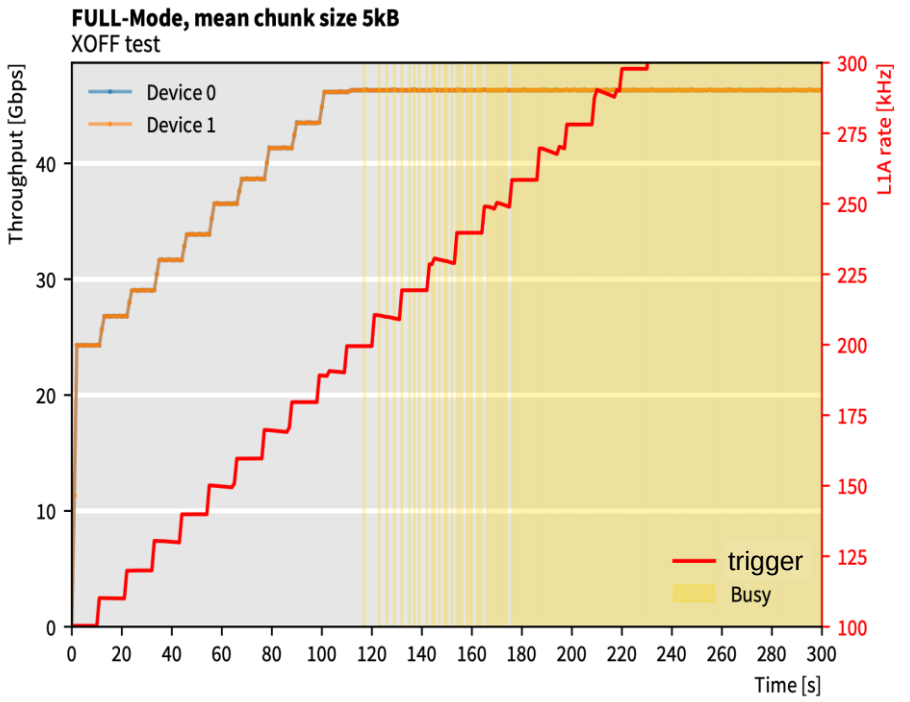
FELIX Performance (in the lab)

- GBT Mode
  - 2 cards, 392 links, 40 byte packets, 25 GbE
  - Stable operation up to 150 kHz

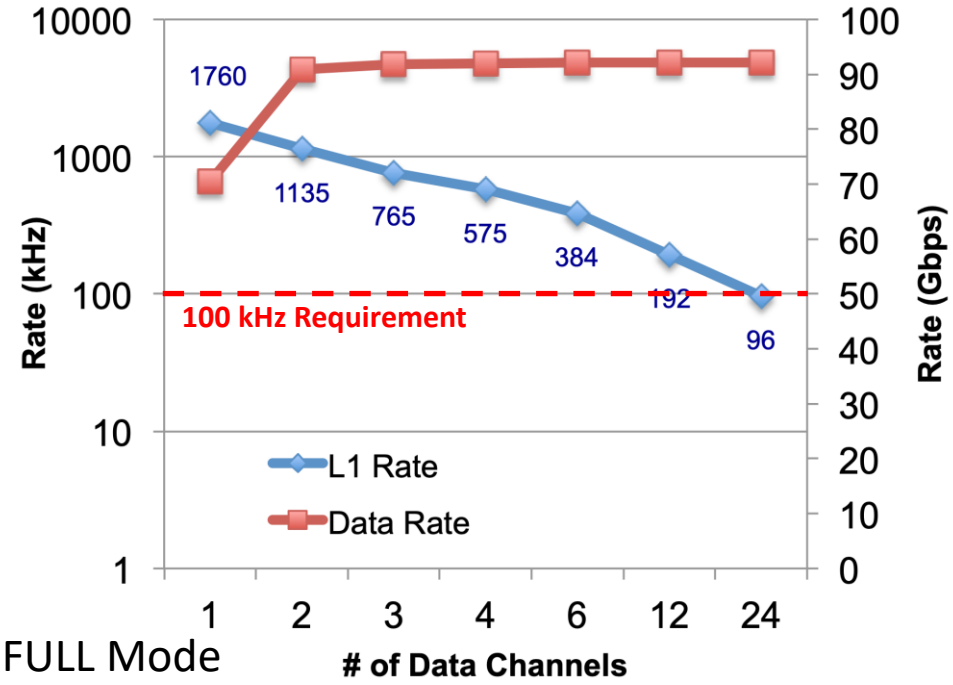


For more results see https://cds.cern.ch/record/2704279

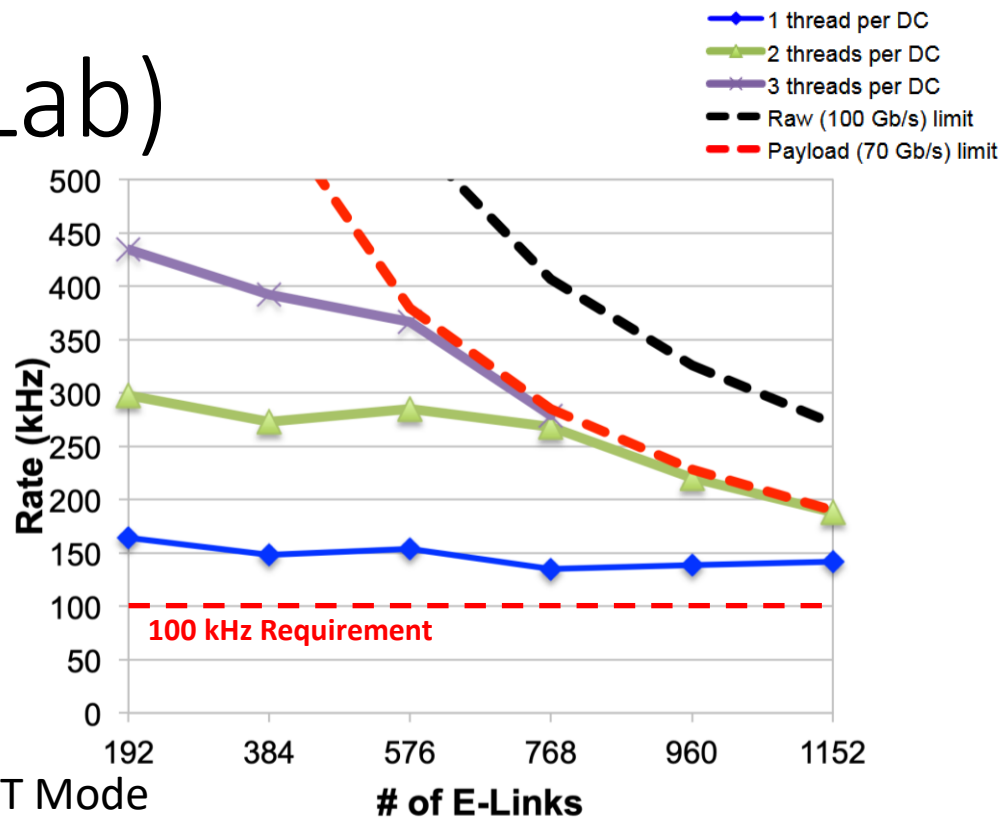
- FULL Mode
  - 1 card, 12 links, 5 kB packets, 100 GbE
  - Maximum data rate ~200 kHz limited by network
  - Backpressure triggers FULL-mode traffic control



SW ROD Performance (in the Lab)



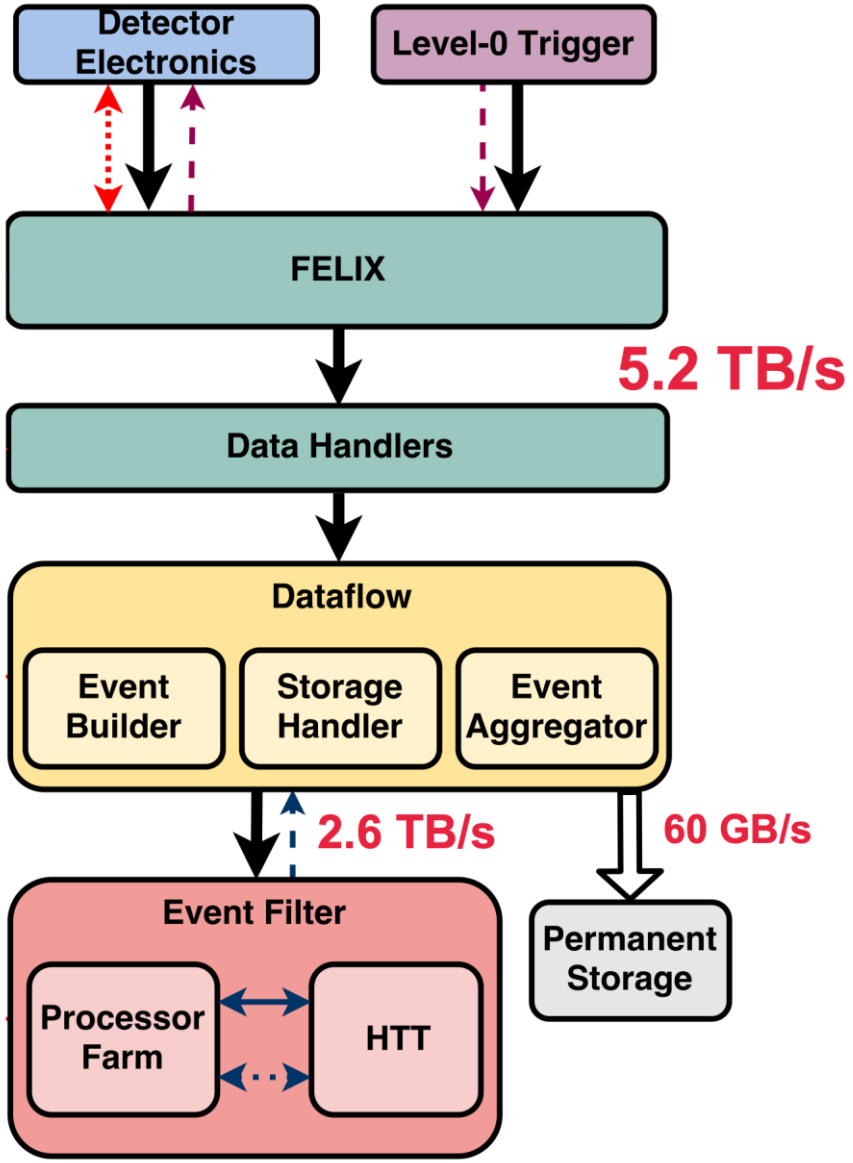
- FULL Mode
  - Software data generator: equivalent to 1 FELIX card
  - In: 24 links, 5 kB packets, 6 reading threads
  - Out: variable number of aggregated output channels (AOCs)
  - Performance limited by network bandwidth except in the case with 1 AOC (further optimisation in progress)
  - Small protocol overhead for large data packets



- GBT Mode
  - Software data generator: equivalent to 6 FELIX cards, each servicing 192 E-links with 40-byte packets
    - 1152 E-links in total
    - 6 AOCs
  - Performance limited by network bandwidth
  - ~30% overhead for 40-byte packets (optimisation in progress)
  - Good scaling with number of threads

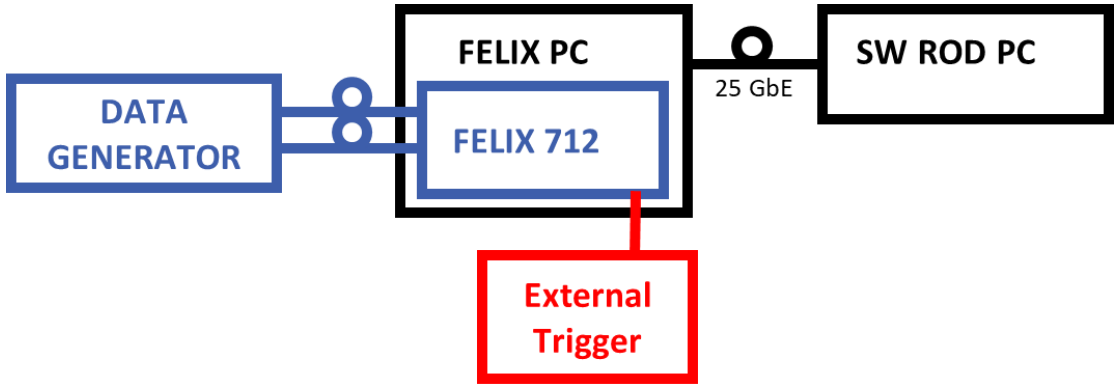
A look ahead to Run 4

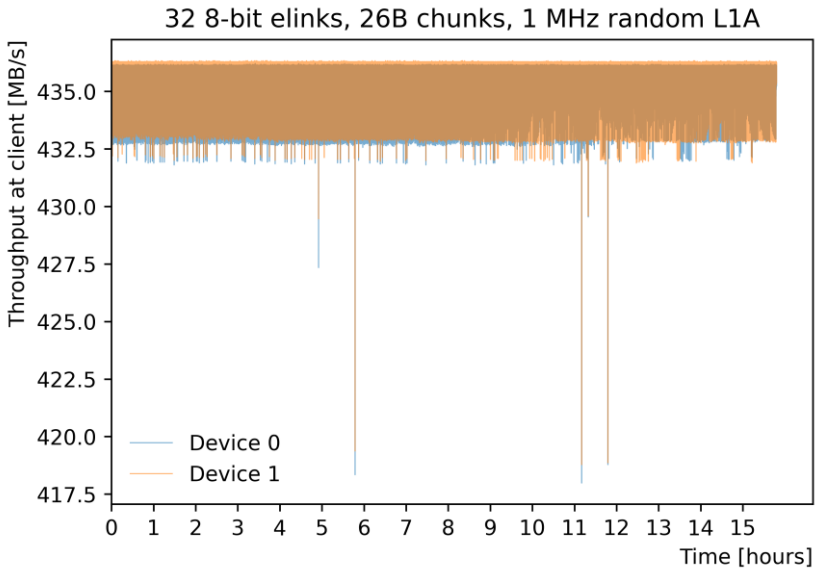
- Comparison with Run 3
  - 10 x trigger rate (1 MHz)
- 3 x pileup (200)
- 20 x readout rate (5.2 TB/s)
- FELIX deployed for all detector systems
  - New implementation under development
- Data Handlers
  - Successor of SW ROD, same functionality
  - Interfaces to new dataflow system (Storage Handler)

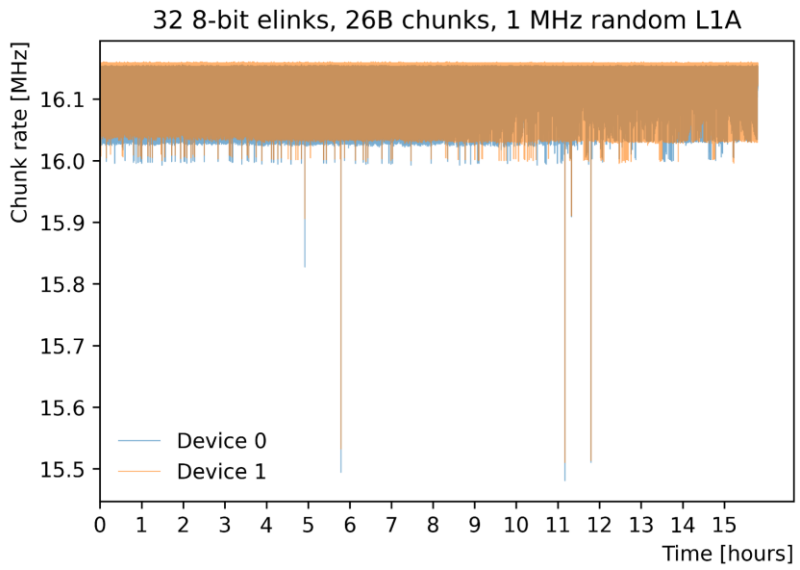


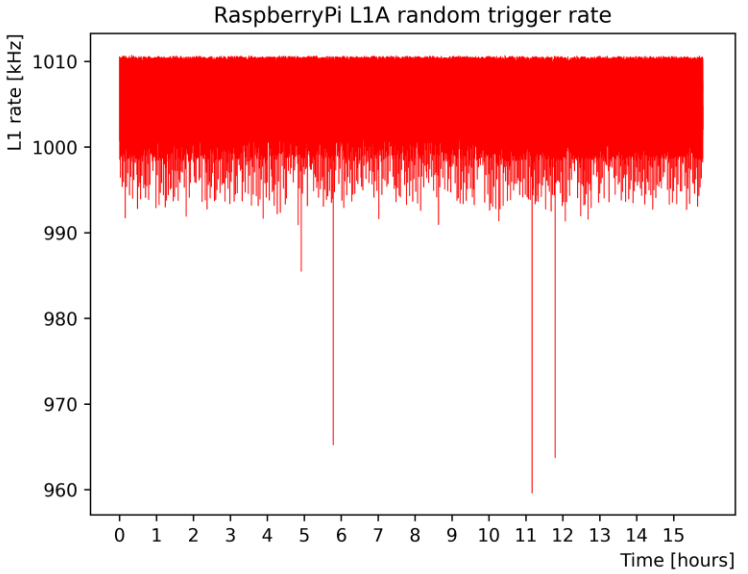
FELIX Performance (in the lab)

- Run 4 scalability testing
  - Use Run 3 FELIX for data acquisition at 1 MHz trigger rate
  - 32 links, 260 Mbps each
  - 1 MHz random trigger rate
  - Stable transfer rate, no errors at 10 x design trigger rate!









PCI endpoint 0

PCI endpoint 0