FELIX
the new ATLAS readout system from Run 3 to High Luminosity LHC

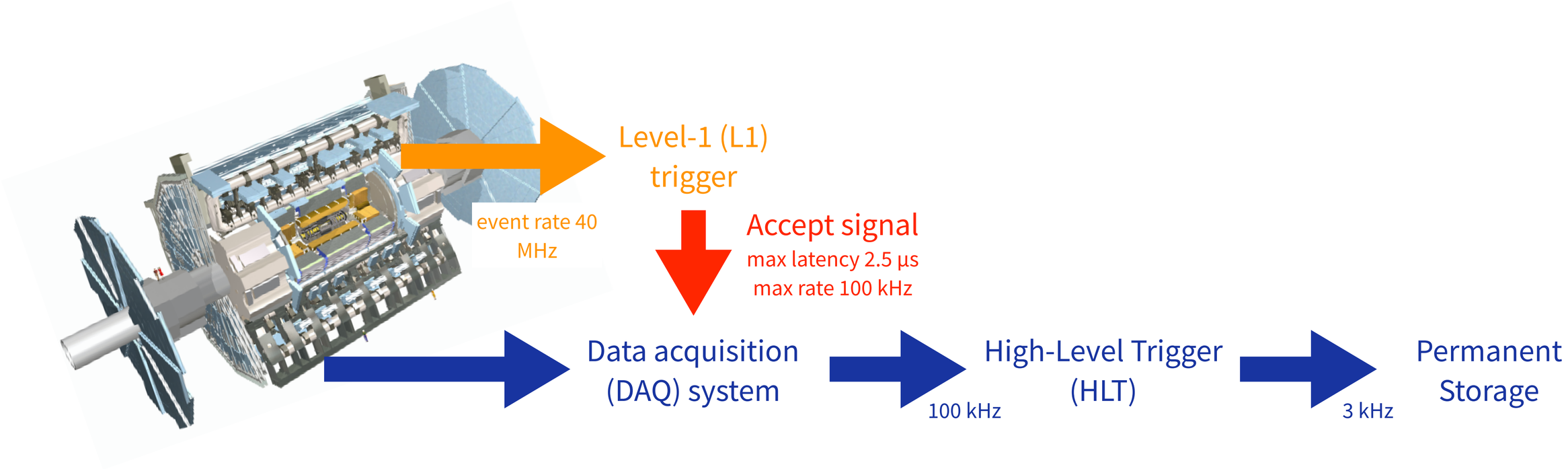
ICHEP 2022 - 09/07/2022

Carlo A. Gottardo

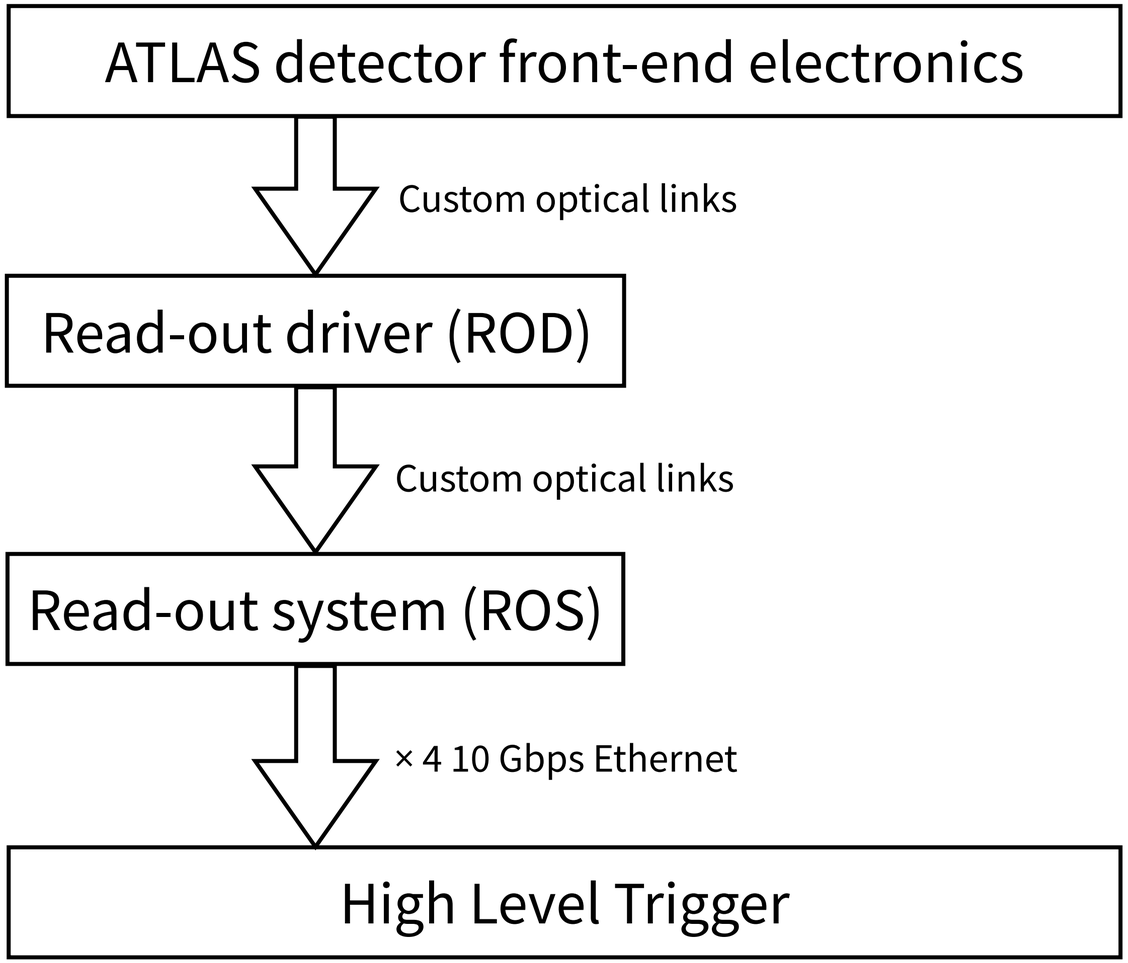
on behalf of the ATLAS TDAQ Collaboration

ATLAS trigger and data acquisition system (DAQ)

- the Large Hadron Collider collides protons at a 40 MHz rate
- the maximum event rate for physics to permanent storage is 3 kHz
- two-level trigger to select events



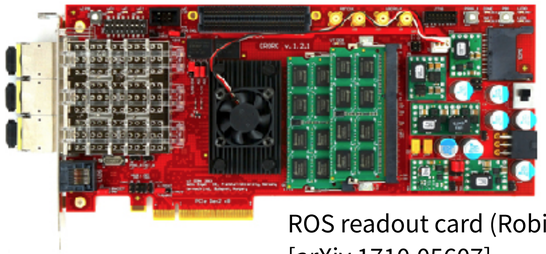
Run 2 DAQ architecture



- sub-detector specific ROD boards running data processing

ROD for muon drift tubes
P. Jansweijer, NIKHEF

- ROS: set of custom readout cards hosted by commodity computers
- data buffering during HLT event processing

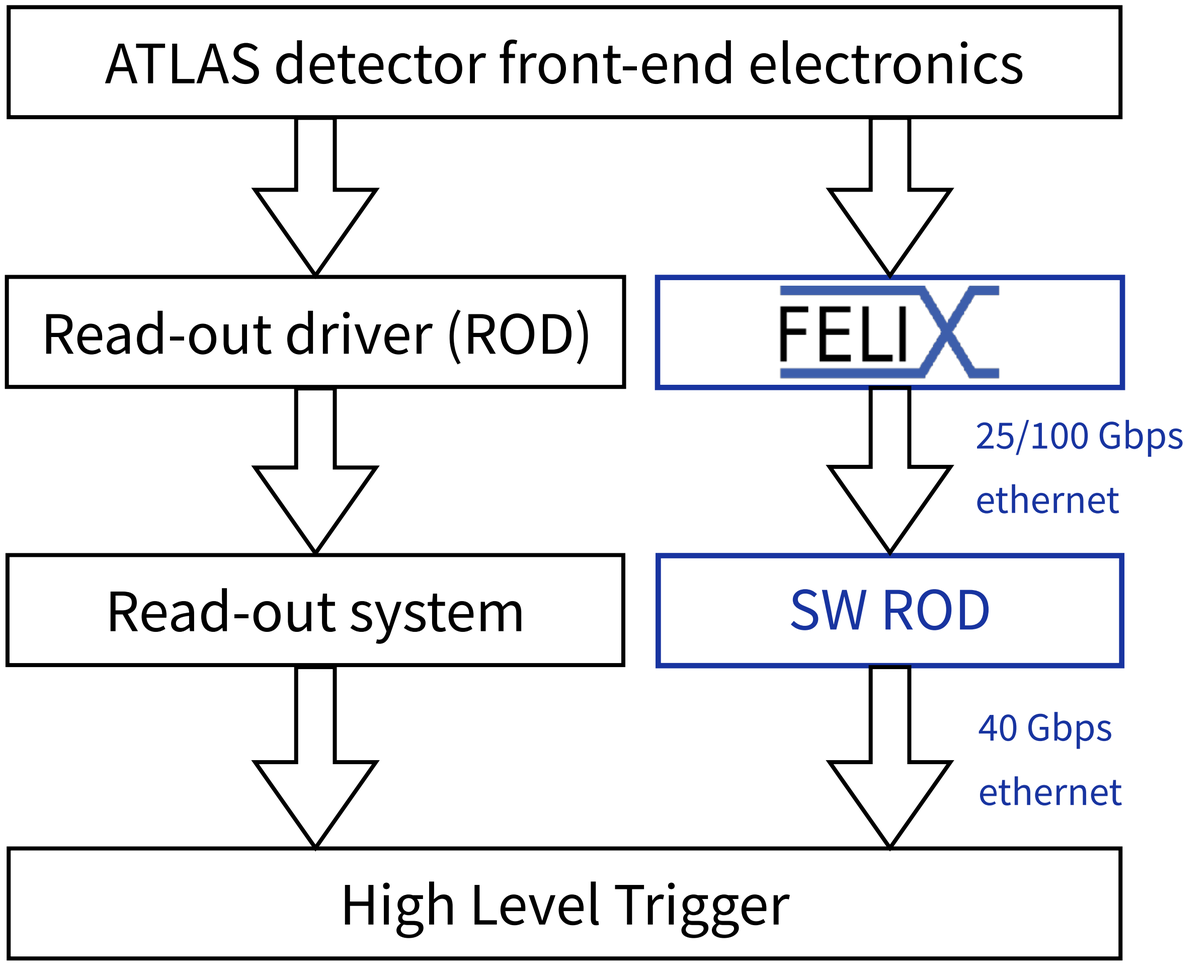


ROS readout card (RobinNP)
[arXiv 1710.05607]

Run 2 architecture remains in place for most sub-detectors

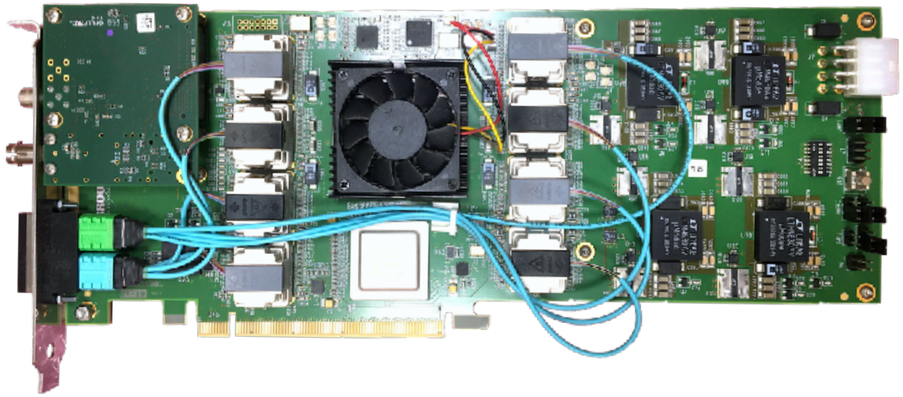
Run 3 DAQ architecture

New FELIX-based DAQ architecture



FELIX

- PCIe cards hosted by commodity computers
- only custom component in new architecture
- data routing, no processing

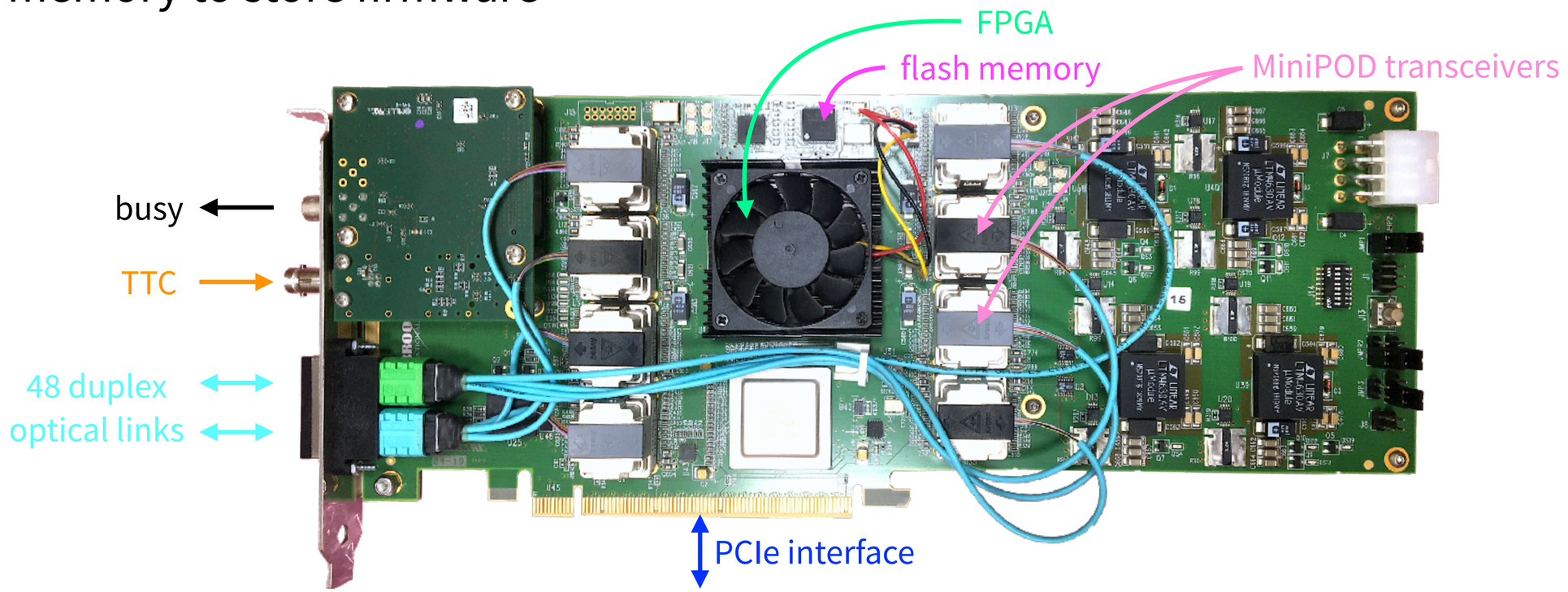


SW ROD

- software in charge of data processing and aggregation, monitoring
- hosted by commodity computers

The FLX-712 FELIX card

- 8 MiniPODs to support up to 48 bidirectional optical links (4 MiniPODs/24 links in most common version)
- FPGA Xilinx Kintex UltraScale XCKU115, 16-lane PCIe Gen3
- interface to Timing, Trigger and Control (TTC) systems
- flash memory to store firmware

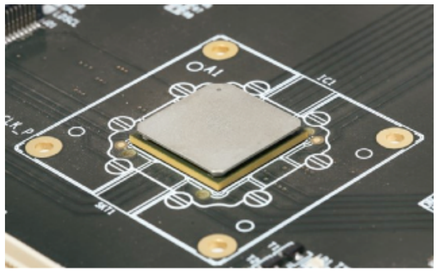


- Around 300 boards produced, distributed in ATLAS, ProtoDUNE-SP, NA62, LUXE and others

Firmware

Two main flavours:

- **FULL**: 24 links per card, 9.6 Gb/s each in the to-host direction
- **GBT**: to interface to GBTX
  - GBTX is a radiation-hard ASIC developed at CERN [1]
  - used as on-detector data stream aggregator
  - GBT firmware supports 24 4.8 Gb/s bi-directional GBT links
  - Each GBT link carries multiple data streams (*e-links*) of configurable bandwidth



GBTX, by CERN EP-ESE

ATLAS benchmarks

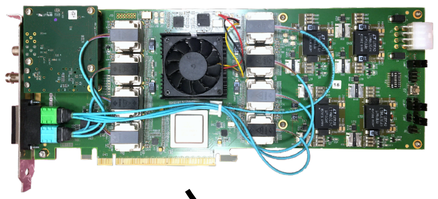
Mode	Message size	Rate per link	(e)links per card	Total message rate per card	Total data rate per card	Use case
FULL	4800 bytes	100 kHz	12	1.2 MHz	46 Gbps	LAr calo, LAr calo trigger
GBT	40 bytes	100 kHz	192	19.2 MHz	7.5 Gbps	New Small Wheels

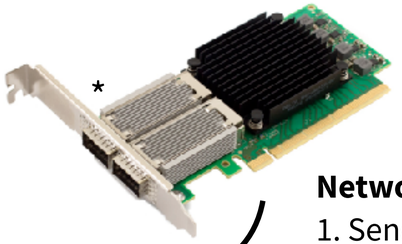
[1] doi: 10.5170/CERN-2009-006.342

Software

Readout application

- interrupt-driven event-loop architecture
- asynchronous non-blocking operations
- custom network library built on top of libfabric [1]
- remote direct memory access (RDMA) technology for low overhead transfers





*Nvidia/Mellanox ConnectX-5
[image by storagereview.com]

FLX-712 events:
1. data available
2. busy state

Network events:
1. Send completed
2. Data received
3. Buffer available for sending

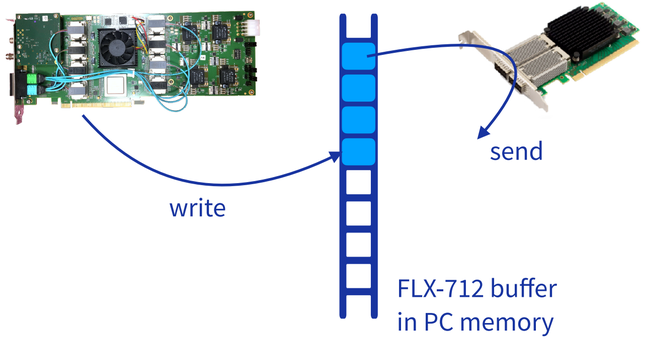
System events:
1. Timer events (`timerfd`)
2. Signals (`eventfd`)
3. Any file descriptor event

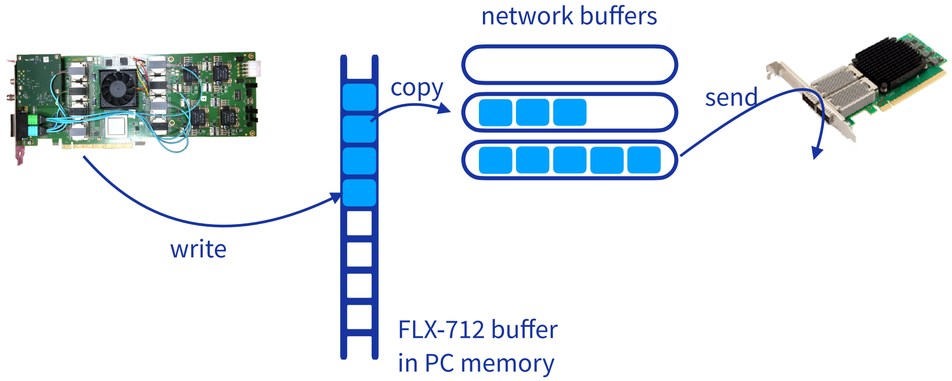
[1] https://ofiwg.github.io/libfabric/

Software

Readout application

- server publishes links/e-links, clients subscribe
- two data transfer approaches: zero-copy, data coalescence





- user-friendly API hides the complexity of network library for client applications

API functions

```
subscribe(elink_number)
unsubscribe(elink_number)
send_data(elink_number)
```

API callback hooks

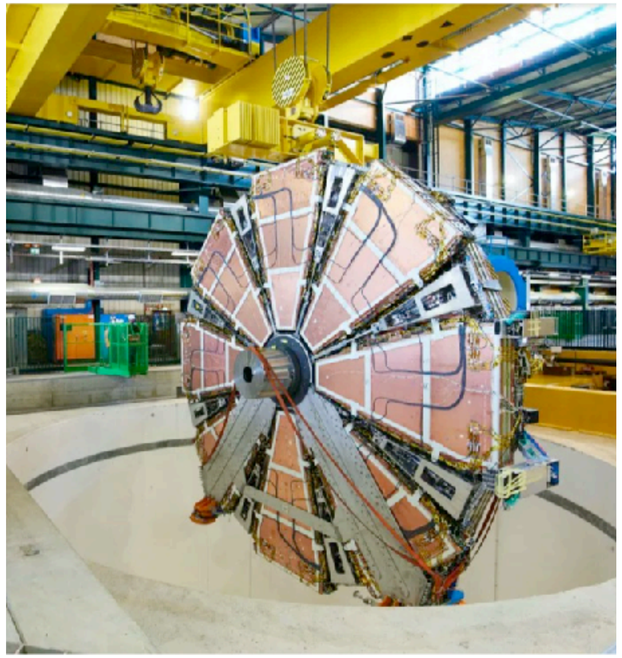
```
on_message_received(elink_number)
on_connection_established(elink_number)
on_disconnection(elink_number)
```

FELIX in ATLAS

Installation and function

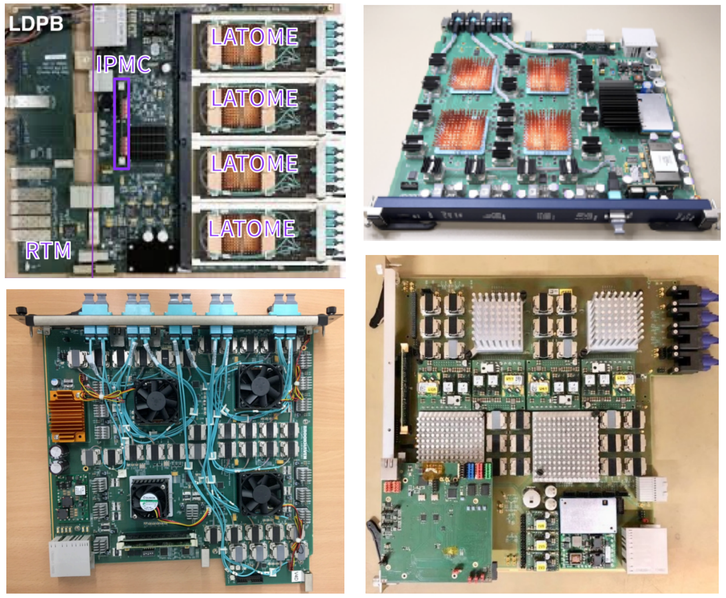
- 64 FELIX PCs, 105 FLX-712 cards installed in ATLAS counting room
- FELIX used for readout, control/monitoring and clock & trigger routing for

New Small Wheels



[CERN tweet]

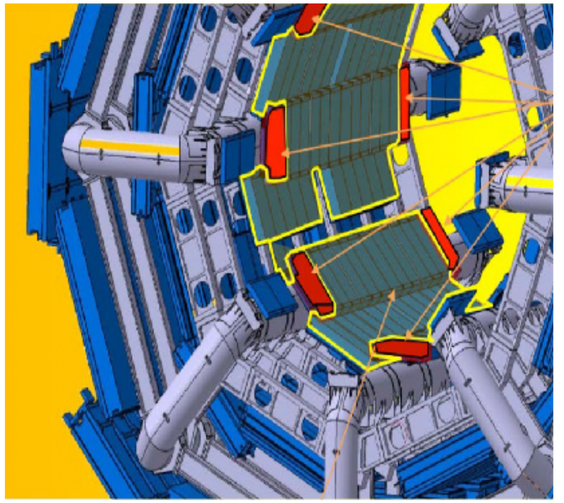
LAr digital readout and Calo trigger



[CDS records 2777152, 2778165, 2789104]

More in H. Cai, N. Themistokleous, C. Tosciri talks

Barrel RPC upgrade



https://cds.cern.ch/record/754973

FELIX PC

Intel Xeon E5-1660 v4 @ 3.2GHz
8 cores, 32 GB DDR4 RAM
nvidia Connect-X5 25/100 GbE

FELIX in ATLAS

Control and monitoring

Application control and monitoring based on Supervisor [1]

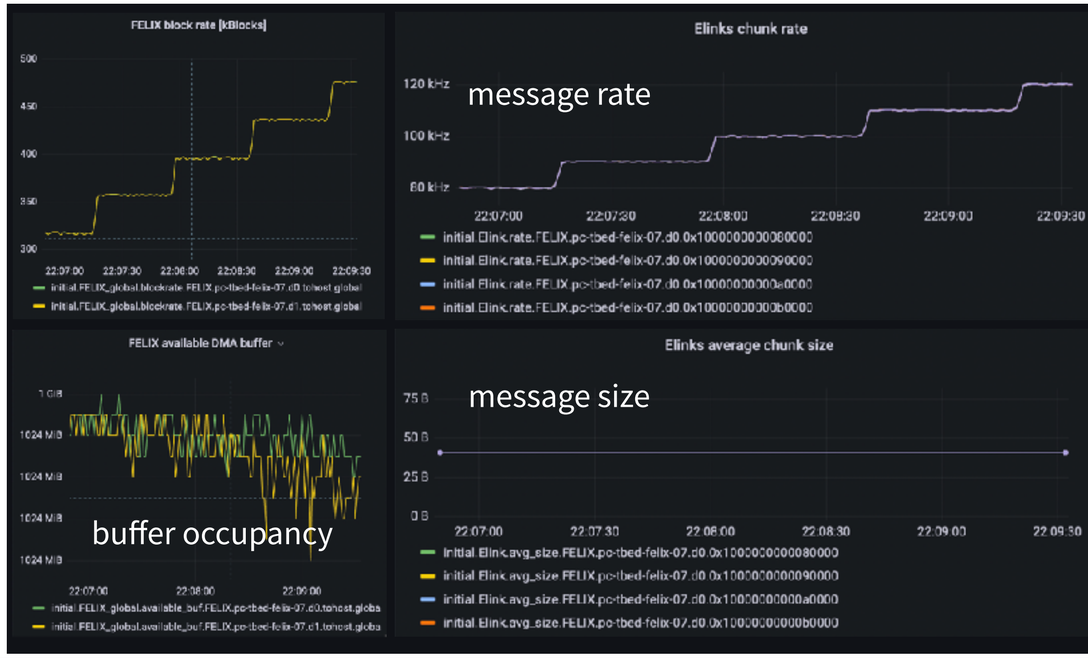
- automatic start of all DAQ applications at boot time
- automatic restart in case of crash
- status monitoring & control via web interface

Search:

Host	flx-init	feconf		felix-tohost		register	felix2atlas
ATCN-only links	c0	d0	d1	d0	d1		
pc-tdq-flx-l1c-efex-00	DONE	DONE	DONE	RUNNING	RUNNING	RUNNING	RUNNING
pc-tdq-flx-l1c-gfex-00	DONE	DONE	DONE	RUNNING	RUNNING	RUNNING	RUNNING
pc-tdq-flx-l1c-jfex-00	DONE	DONE	DONE	RUNNING	RUNNING	RUNNING	RUNNING
pc-tdq-flx-l1c-trex-00	DONE	DONE	DONE	RUNNING	RUNNING	RUNNING	RUNNING
pc-tdq-flx-l1c-trex-01	DONE	DONE	DONE	RUNNING	RUNNING	RUNNING	RUNNING

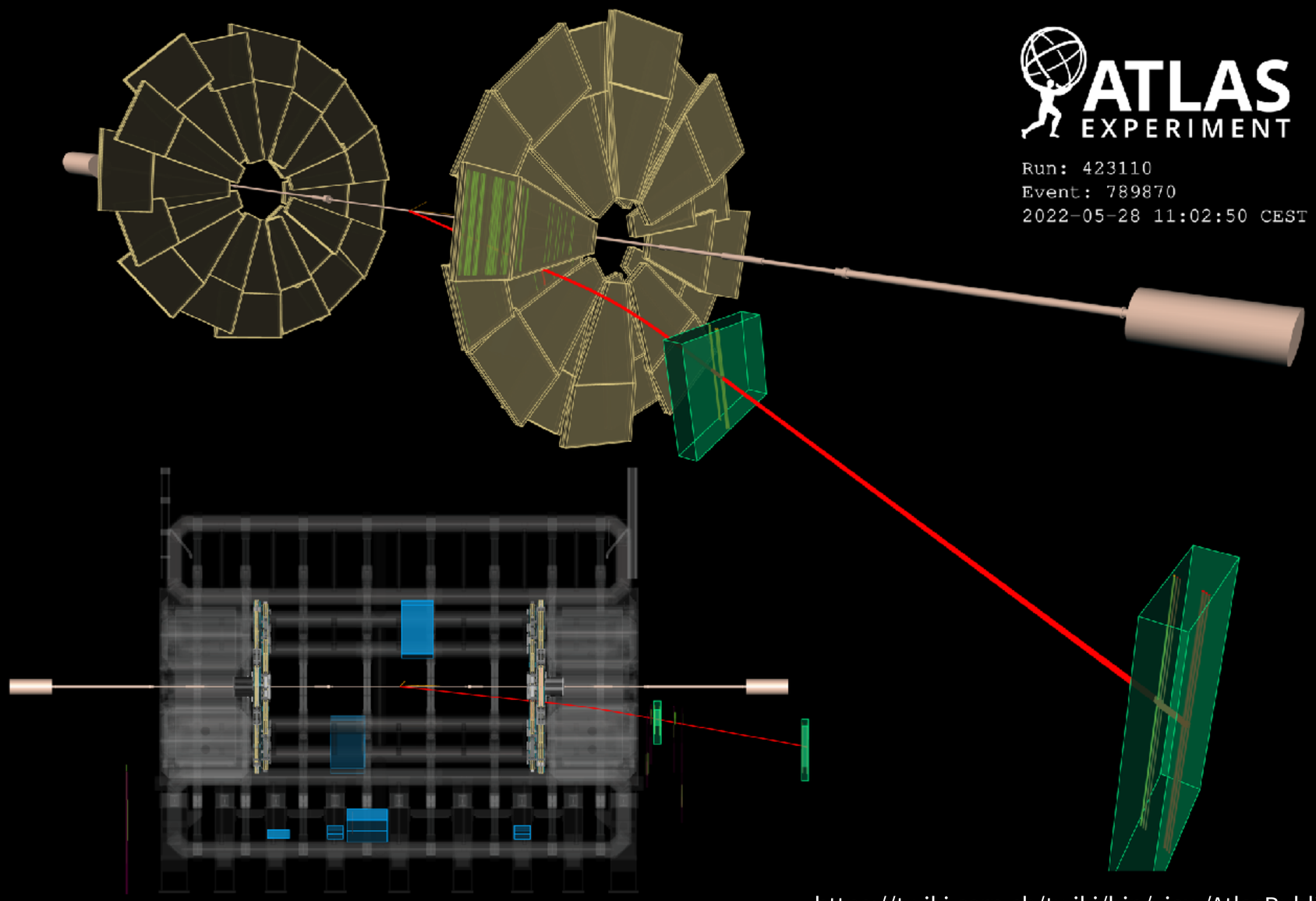
Monitoring integrated in ATLAS infrastructure

- operational monitoring [2] with Grafana [3] dashboard
- log messages in Error Reporting System [4]
- conditions of PC, FLX-712 and network recorded by computer cluster monitoring



[1] http://supervisord.org/
[2] doi: 10.1051/epjconf/202024501020
[3] https://go2.grafana.com
[4] doi: 10.1088/1742-6596/608/1/012004

FELIX in ATLAS



https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EventDisplayRun3Collisions

FELIX for High-Luminosity LHC

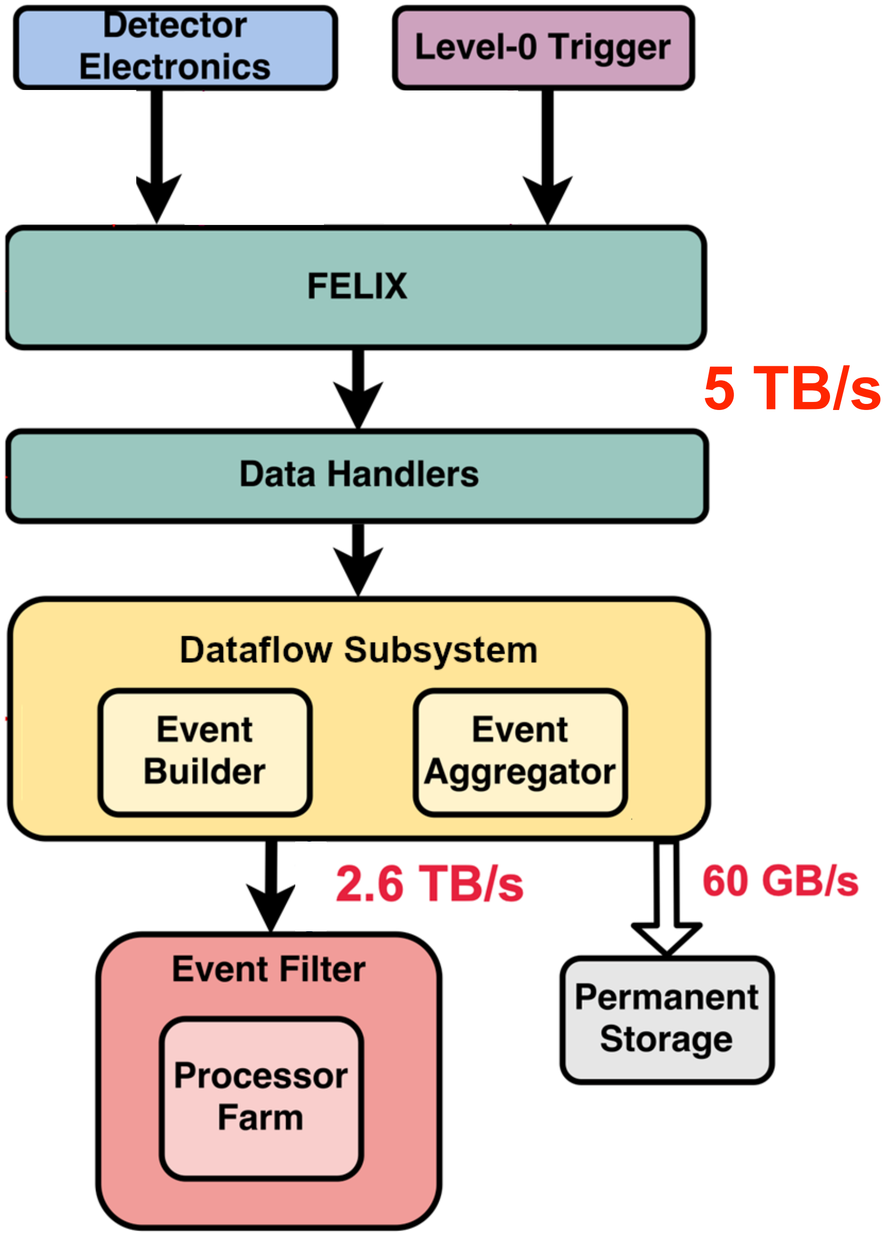
ATLAS DAQ in Run 4

Run 4 conditions

- 1 MHz L1 trigger rate (×10 Run 3)
- up to 200 interactions per bunch-crossing (×3 Run 3)
- 5.2 TB/s data throughput (×20-30 Run 3)

FELIX requirements

- readout of all sub-detectors
- ~14000 optical links with bandwidth [2.5, 25] Gb/s
- support for new detector-specific functionalities
  - e.g. continuous “trickle” reconfiguration of new tracker front-end electronics (pixel, strips)



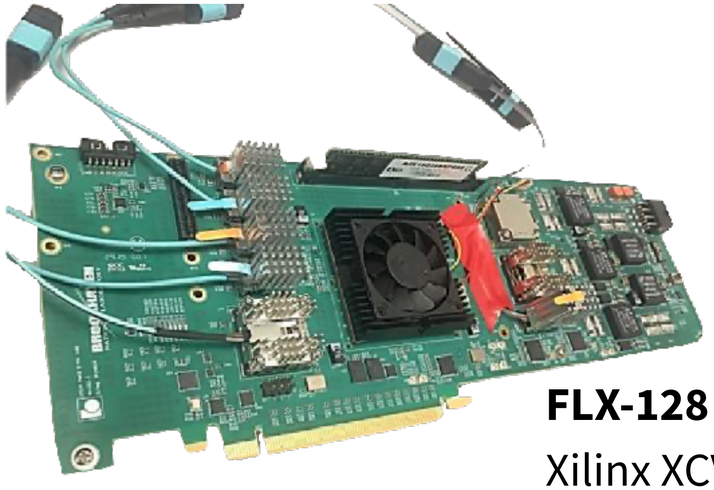
FELIX hardware upgrade

A new FELIX card is necessary to support

- increased maximum link bandwidth (10 → 25 Gbps)
- new timing/trigger interface

Prototypes

- FLX-128 and FLX-181 prototypes built in 2019 and 2021
- new FPGAs, fourth/fifth generation PCIe, new optical transceivers (FireFly™)



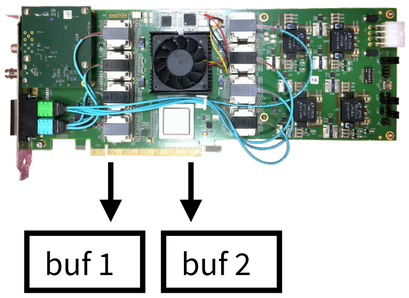
FLX-128
Xilinx XCVU9P FPGA

FLX-181
Xilinx VM1802 FPGA
up to 24 links 25 Gb/s

FELIX firmware upgrade

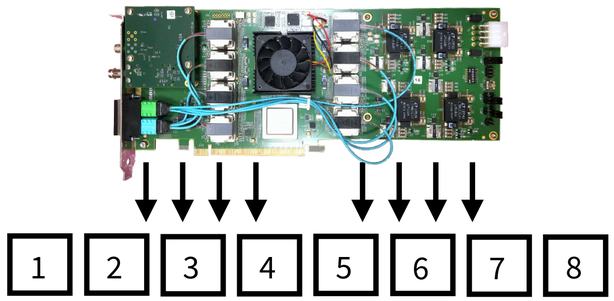
New firmware to support

- additional data encodings (adopted by detectors and TTC system)
- higher speed links and PCIe interface
- larger number of buffers in computer memory



data of first set of 12 (or 24) links

data of second set of 12 (or 24) links



8 buffers. Configurable elink destination.

New flavours in addition to GBT and FULL:

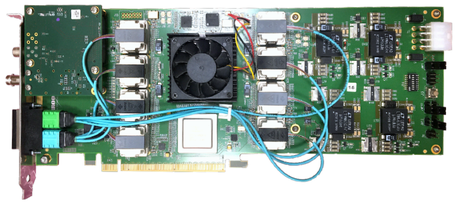
LPGBT (evolution of GBTX), PIXEL & STRIP (custom LPGBT), Interlaken (64b/67b encoding)

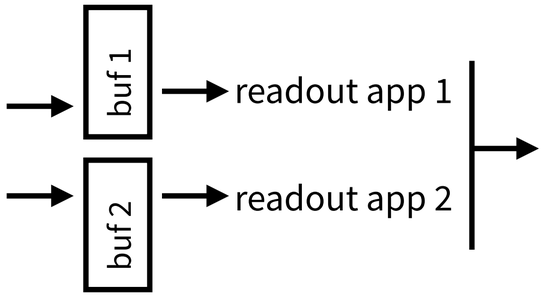
Run 4 firmware regularly built and available for different FLX models

FELIX software upgrade

Same software architecture as in Run 3 but different deployment scheme

Run 3: only two readout applications per card

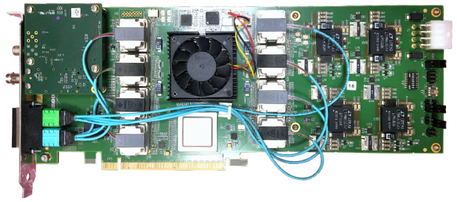


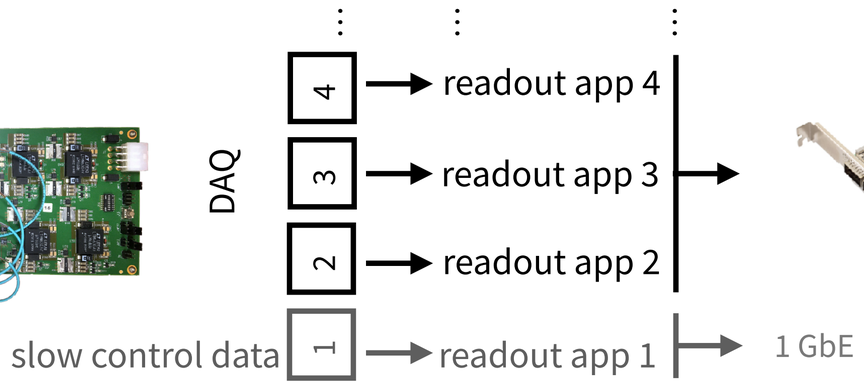


25 / 100 GbE

CPU Intel Xeon E5-1660 v4
8 cores, 3.2 /3.8 GHz

Run 4: up to eight readout applications.Different data types can de decoupled.





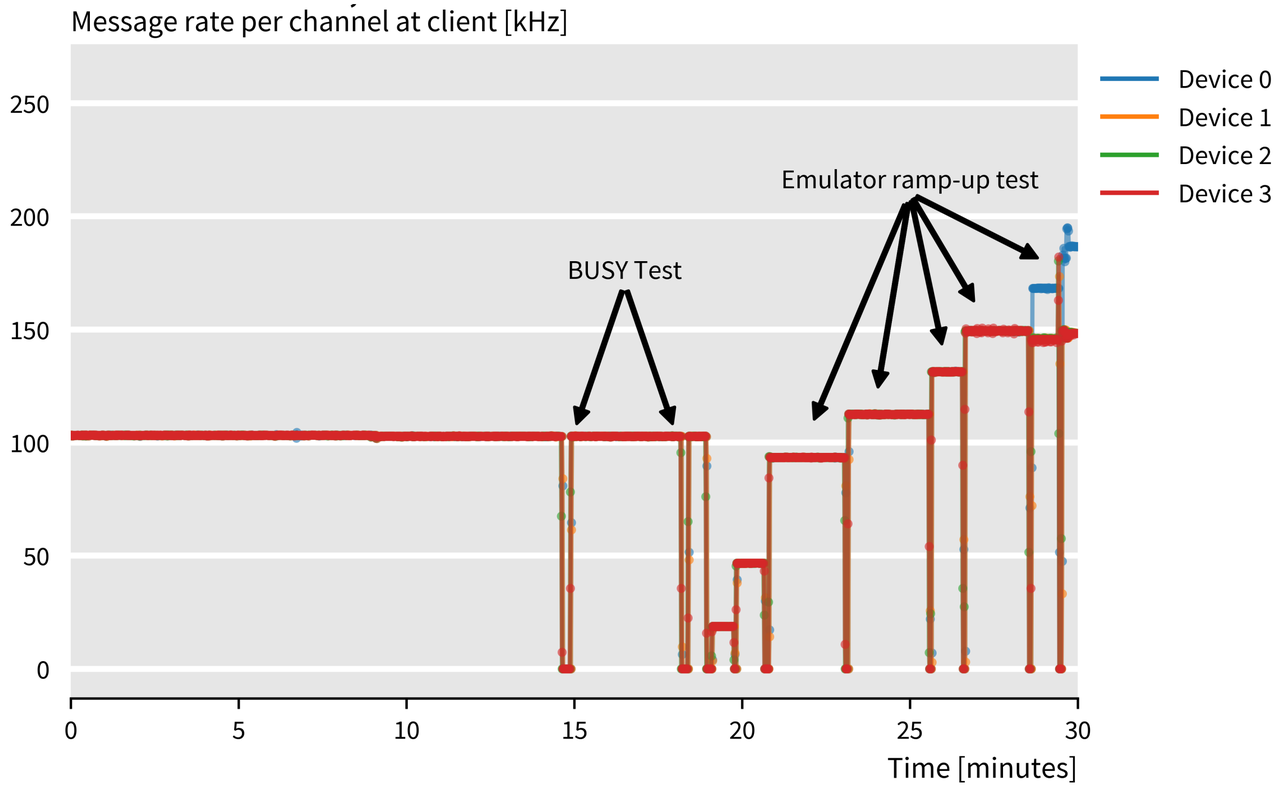
200 / 400 GbE

CPU > AMD Epyc Milan 7313P
16 cores, 3.0 /3.7 GHz

FELIX performance from Run 3 to Run 4

In the Run 3 benchmark scenarios FELIX can handle

- up to 150 kHz trigger rate with data coalescence (96 e-links per DMA buffer, 40 byte messages)
- up to 120 kHz in zero-copy mode (6+6 links, 4.8 kB messages)



Theoretically with ×4 buffers

150 kHz → 600 kHz trigger rate

120 kHz → 480 kHz trigger rate

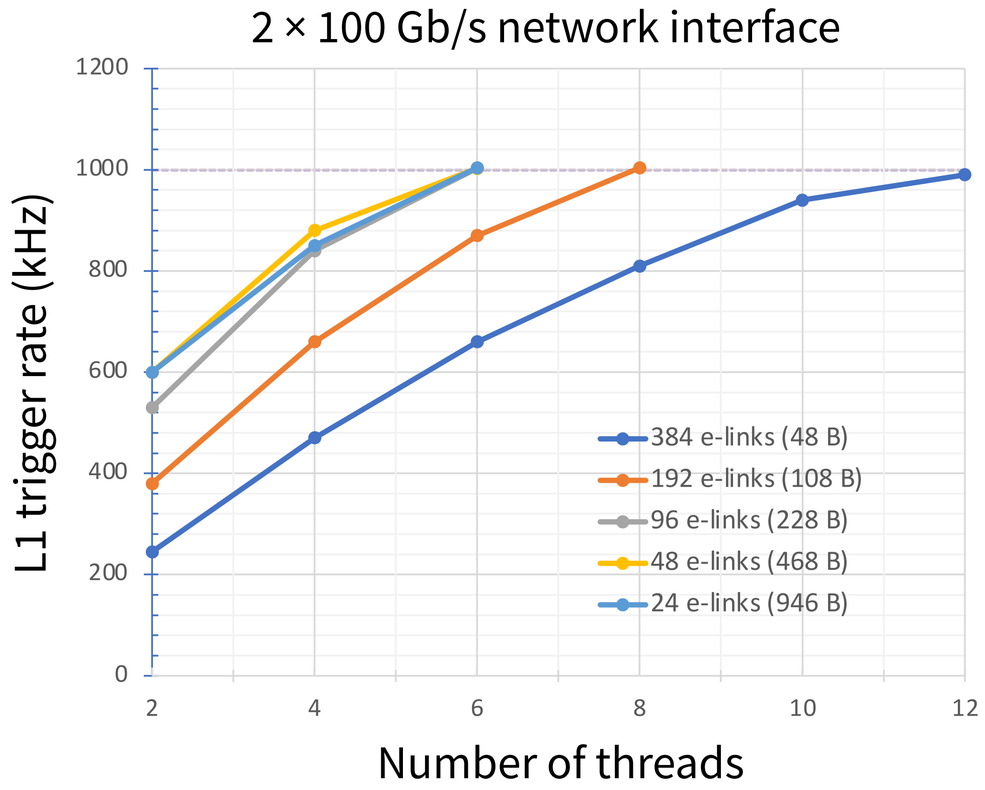
1 MHz already reachable e.g. 24 FULL links with 192 byte messages

Ongoing tests to identify bottlenecks with newer PCs exceed predictions.

Bottleneck identification

Insight on network library

- also used in SW ROD performance test with simulated FELIX input
- test on Run 3 SW ROD PC equipped with two 100 Gb/s network interfaces
- Run 3 SW ROD can handle Phase-II data aggregation rates using 12 CPU cores



Run 3 SW ROD PC

Dual Intel® Xeon® Gold 5218 CPU

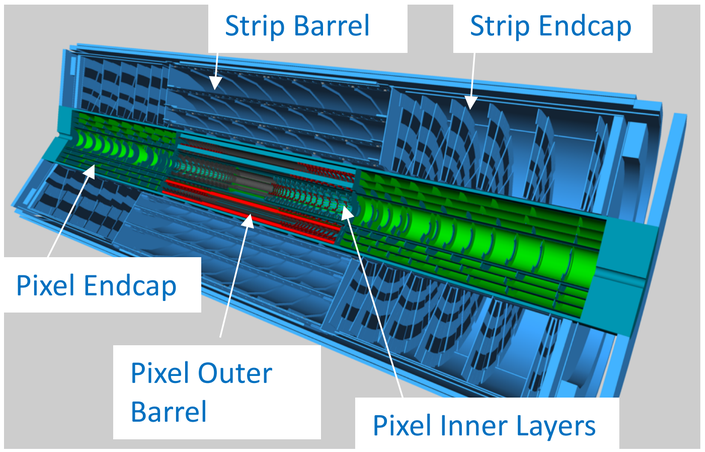
16x2 cores 2.30 - 3.90 GHz, 96 GB RAM

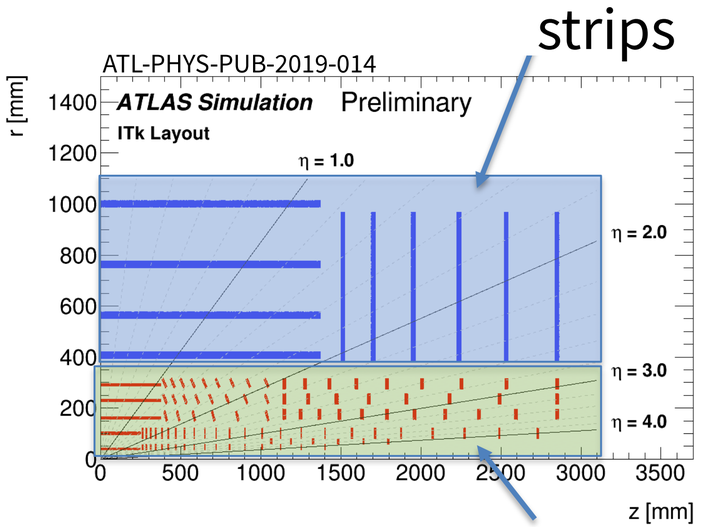
Message size defined to saturate effective network bandwidth (184 Gbps) at 1 MHz trigger rate

Integration with new systems

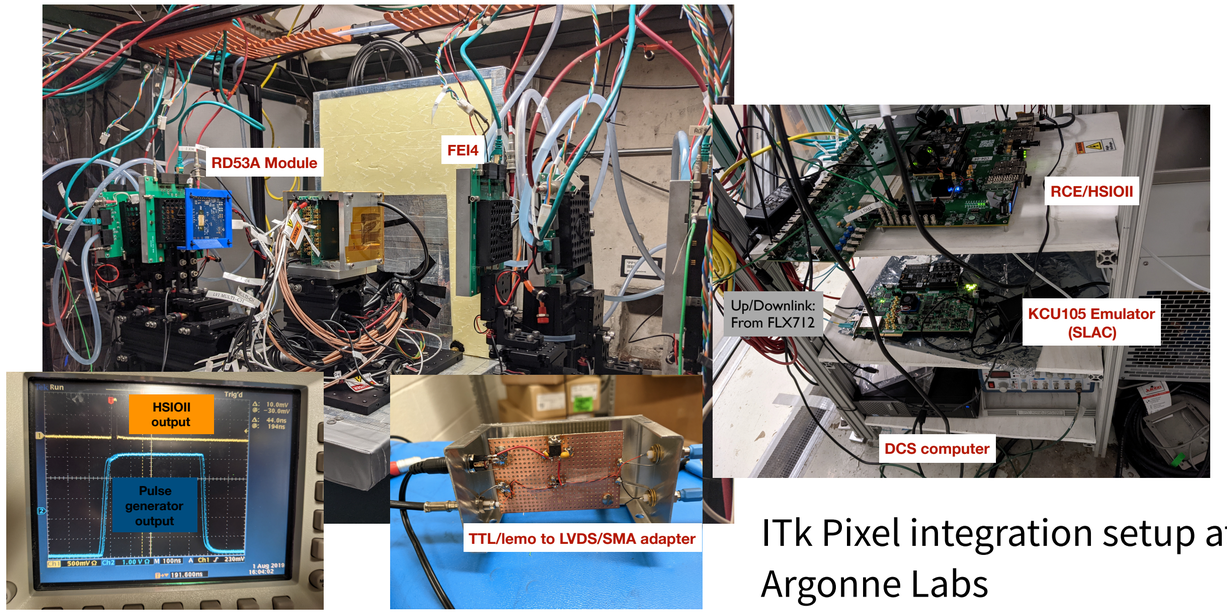
Overview of upgrades in O. Jinnouchi's talk on Friday

Inner tracker (ITk)



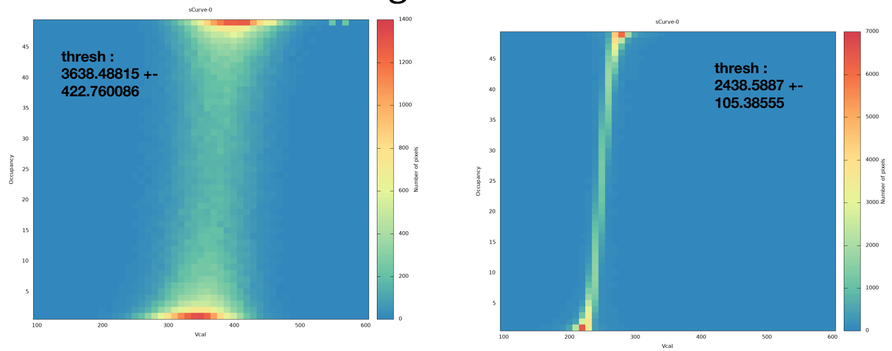


ITk pixel



ITk Pixel integration setup at Argonne Labs

ITk Pixel thresholds tuning via FELIX

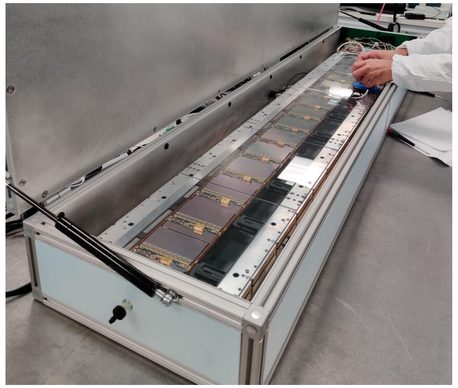


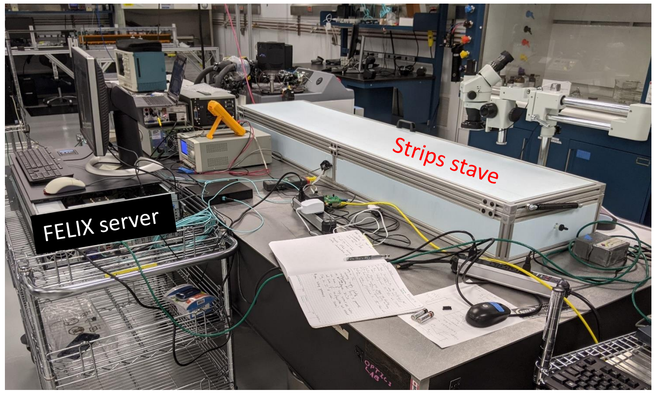
ITk pixel in S. D'Auria talk on Thursday

[images by Marco Trovato]

Integration with new systems

ITk strips

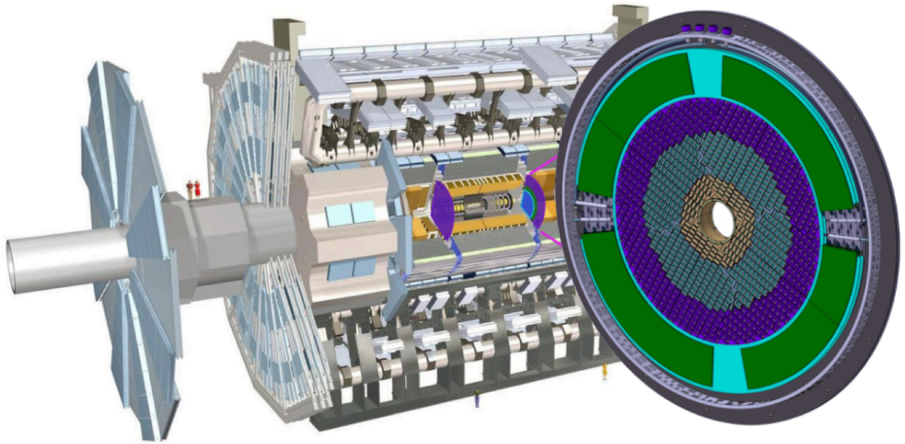


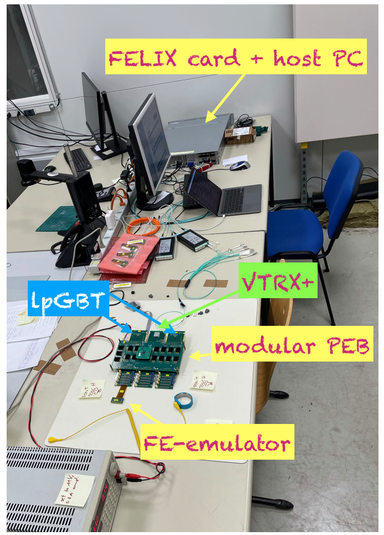


ITk Strips stave at CERN

- FELIX STRIPS firmware functional
- configuration and readout via FELIX
- cross-check with small scale alternative readout hardware

High Granularity Timing Detector





- integration ongoing at CERN
- tests with lpGBT firmware
- more on Z. Liang's talk on Thursday

Summary

Run 3

- FELIX integrated in the ATLAS DAQ infrastructure
- data-taking started, gaining experience in final operational environment and conditions

Run 4 and beyond

- new hardware prototypes under development
- firmware that allows for a scalable software approach
- tests planned with most recent CPUs and network interfaces