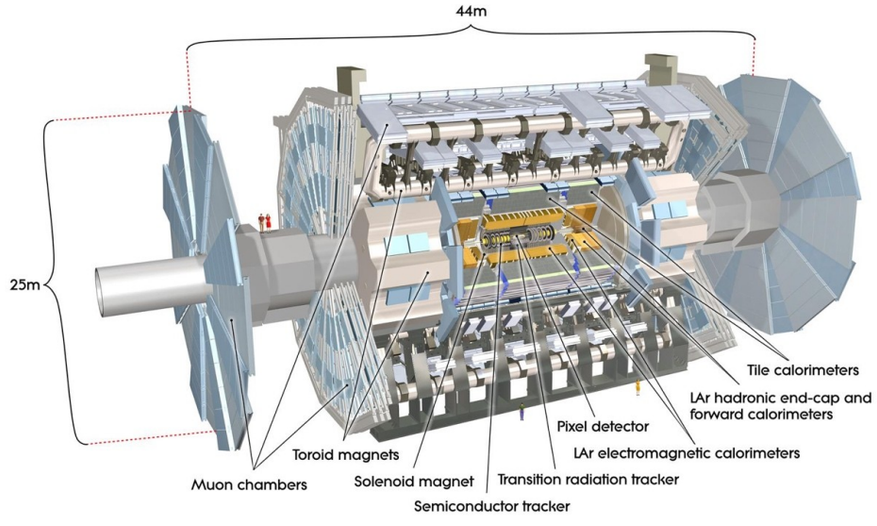
FELIX: commissioning the new detector interface for the ATLAS trigger and readout system

Roberto Ferrari
INFN Pavia

On behalf of the ATLAS TDAQ Collaboration

22nd Virtual IEEE Real Time Conference
12 October 2020

Overview



ATLAS TDAQ evolution for Run 3

FELIX* readout system

Performance & commissioning

Conclusions

*FELIX : Front-End LInk eXchange → (custom) PCIe cards hosted in COTS servers

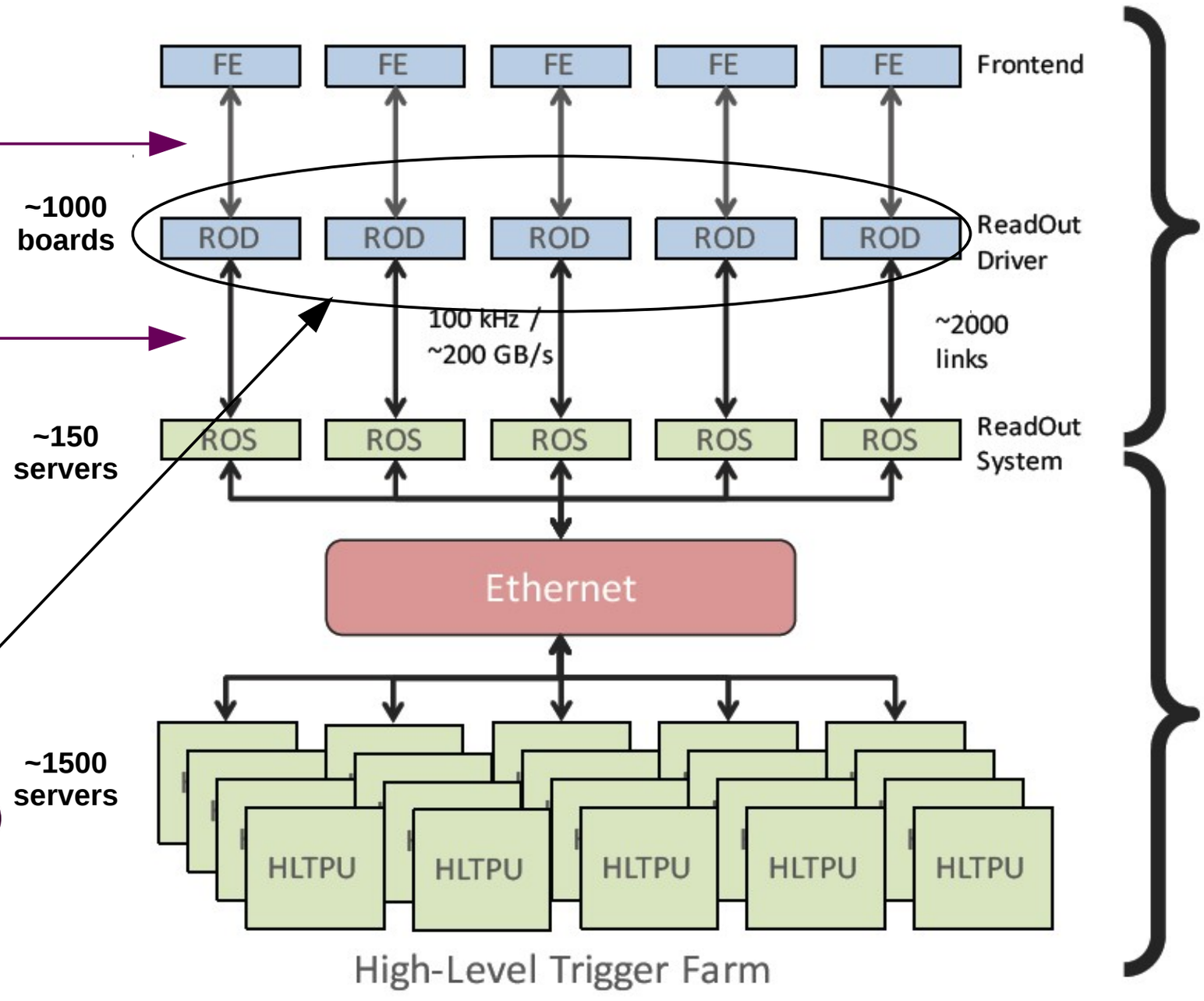
ATLAS TDAQ in Run 2

~ 2 MB events, ~ 50 GB/s network bandwidth,
~ 1.5 GB/s recording throughput

Custom point-to-point links

Point-to-point S-LINKs*

VME (detector-specific)



*S-LINK: CERN Simple Link

ATLAS TDAQ in Run 2

•Custom HW/protocols for Front-End (FE) readout

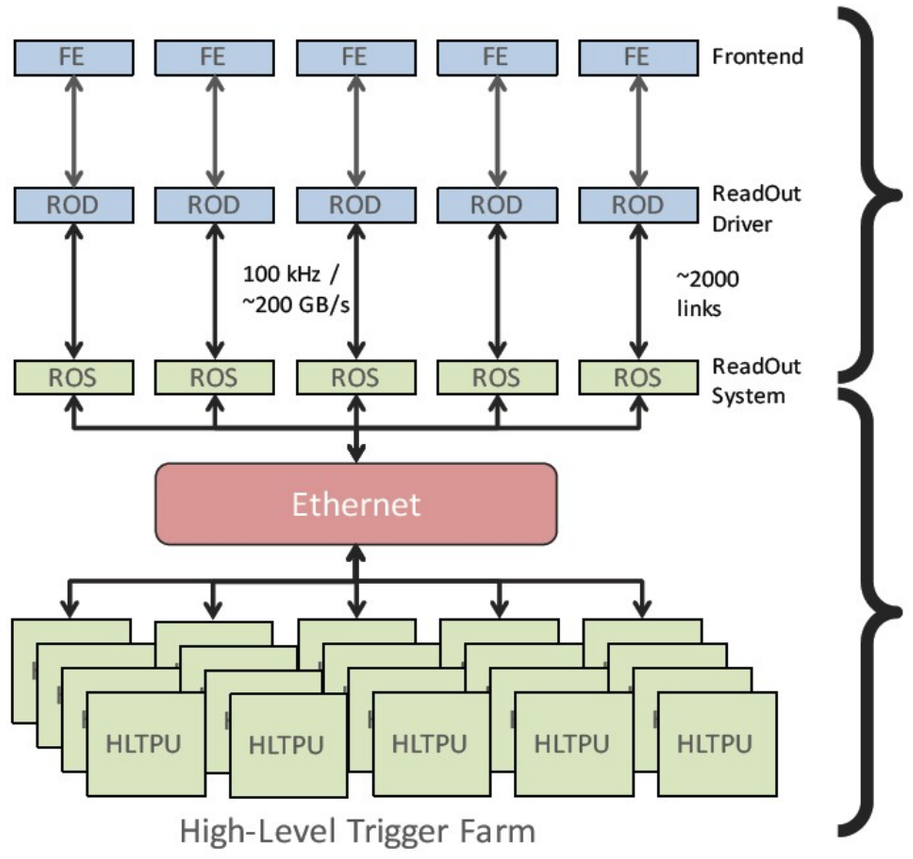
•Data buffered in FE elx waiting for L1 trigger (max latency ~2.5 μs)

•Trigger and LHC clock sent to both FE elx and (detector specific) ReadOut Drivers (RODs)

•RODs send data to ReadOut System (ROS) which buffers them for High-Level Trigger (HLT) requests

•HLT finalises event selection in two steps

Readout system:
~1 k ROD boards
~150 ROS servers



Upgrade for Run 3

LHC Phase-I Upgrade → ATLAS trigger & detector upgrade

1) new muon detectors in both forward and transition regions

→ additional readout channels more than full present muon spectrometer

2) new trigger elx in calorimeter system

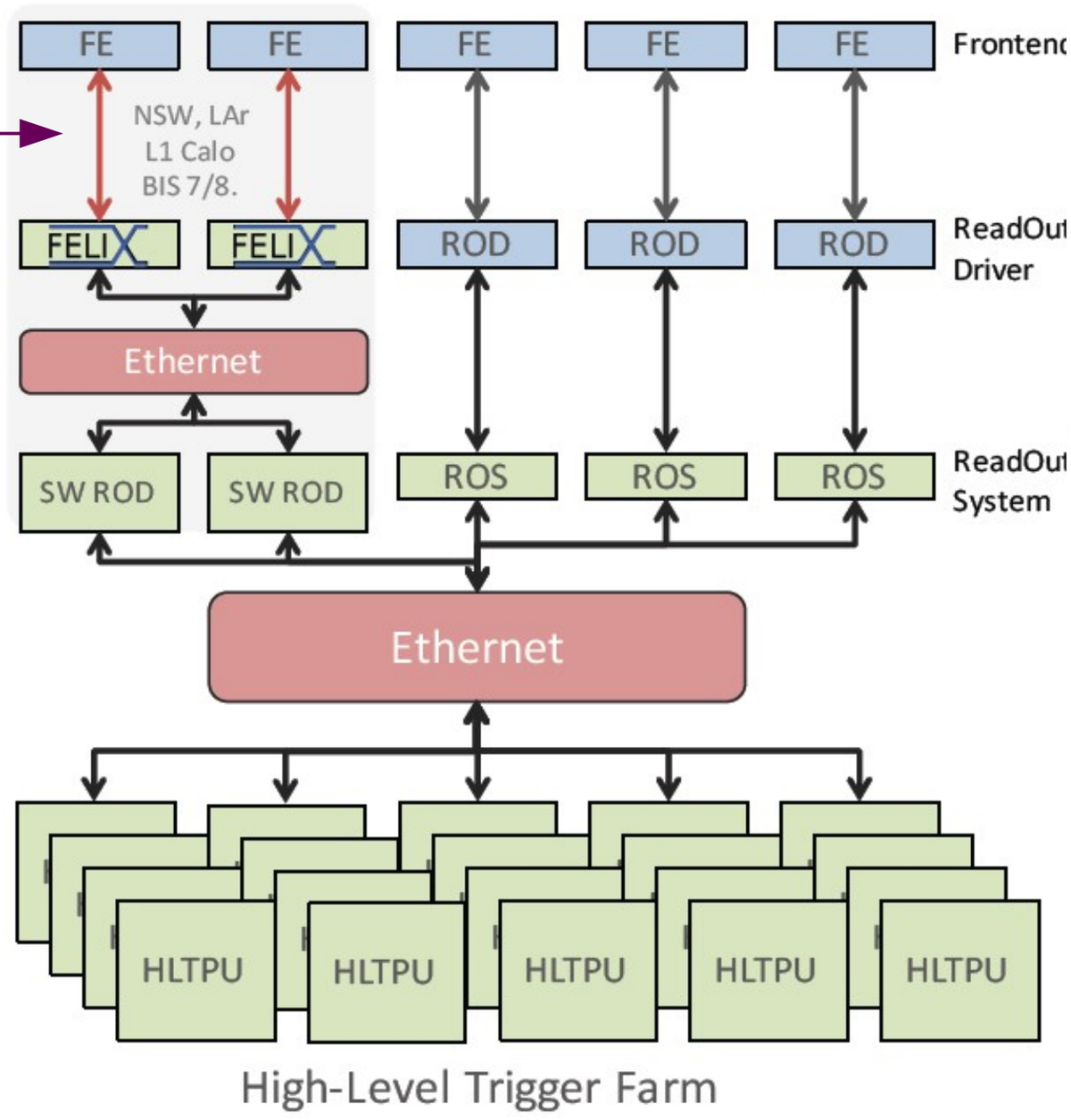
→ upgraded readout architecture

Upgrade for Run 3

Same requirements as Run 2

GBT* or FULL mode links

25-100 Gb/s Ethernet



Custom elx component including FELIX cards

PCs (COTS)

*GBT: GigaBit Transceiver with Versatile Link

Upgrade for Run 3

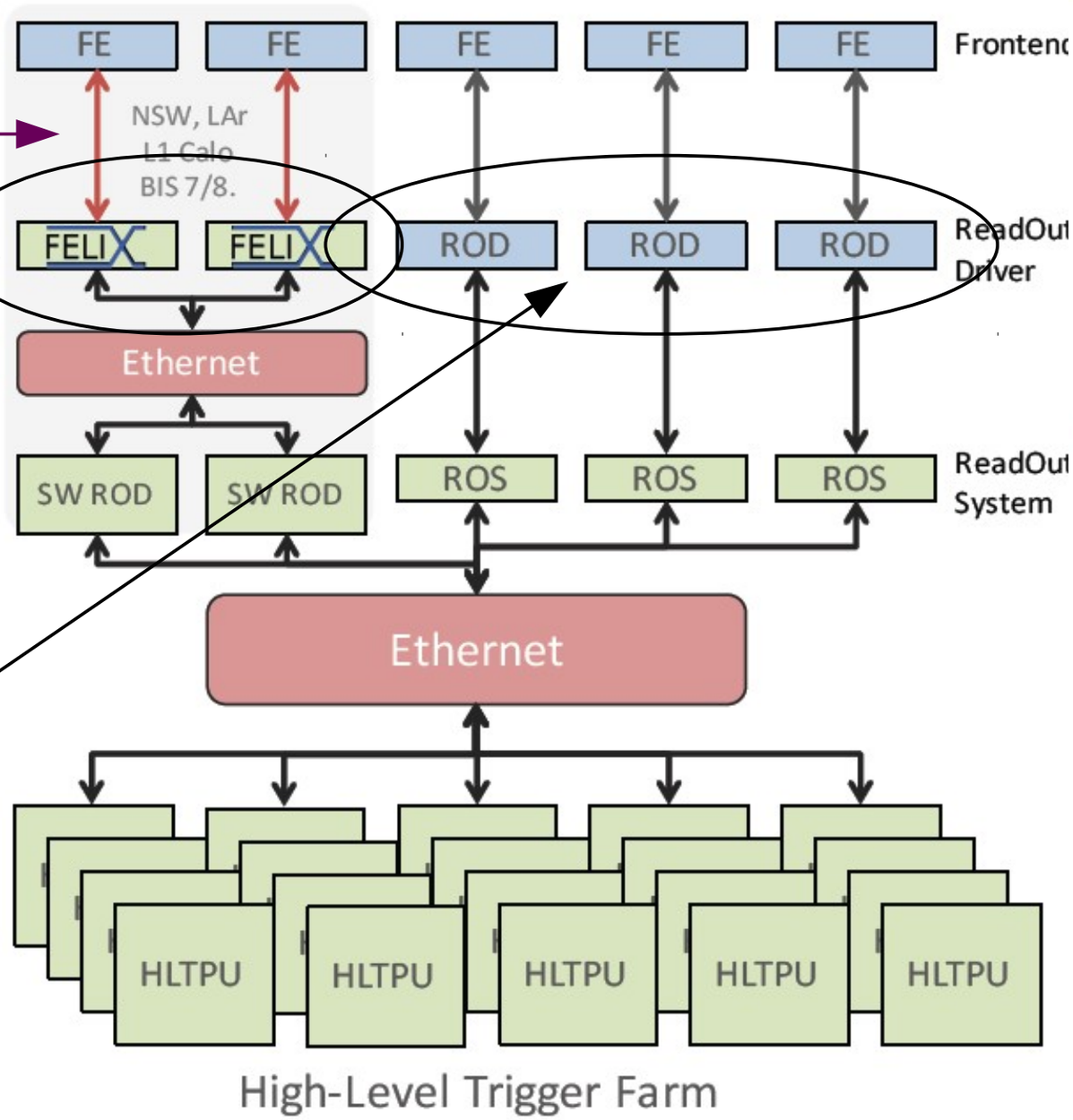
Same requirements as Run 2 but reduced custom components

GBT* or FULL mode links

25-100 Gb/s Ethernet

PCIe Gen3 (TDAQ specific)

VME (detector-specific)

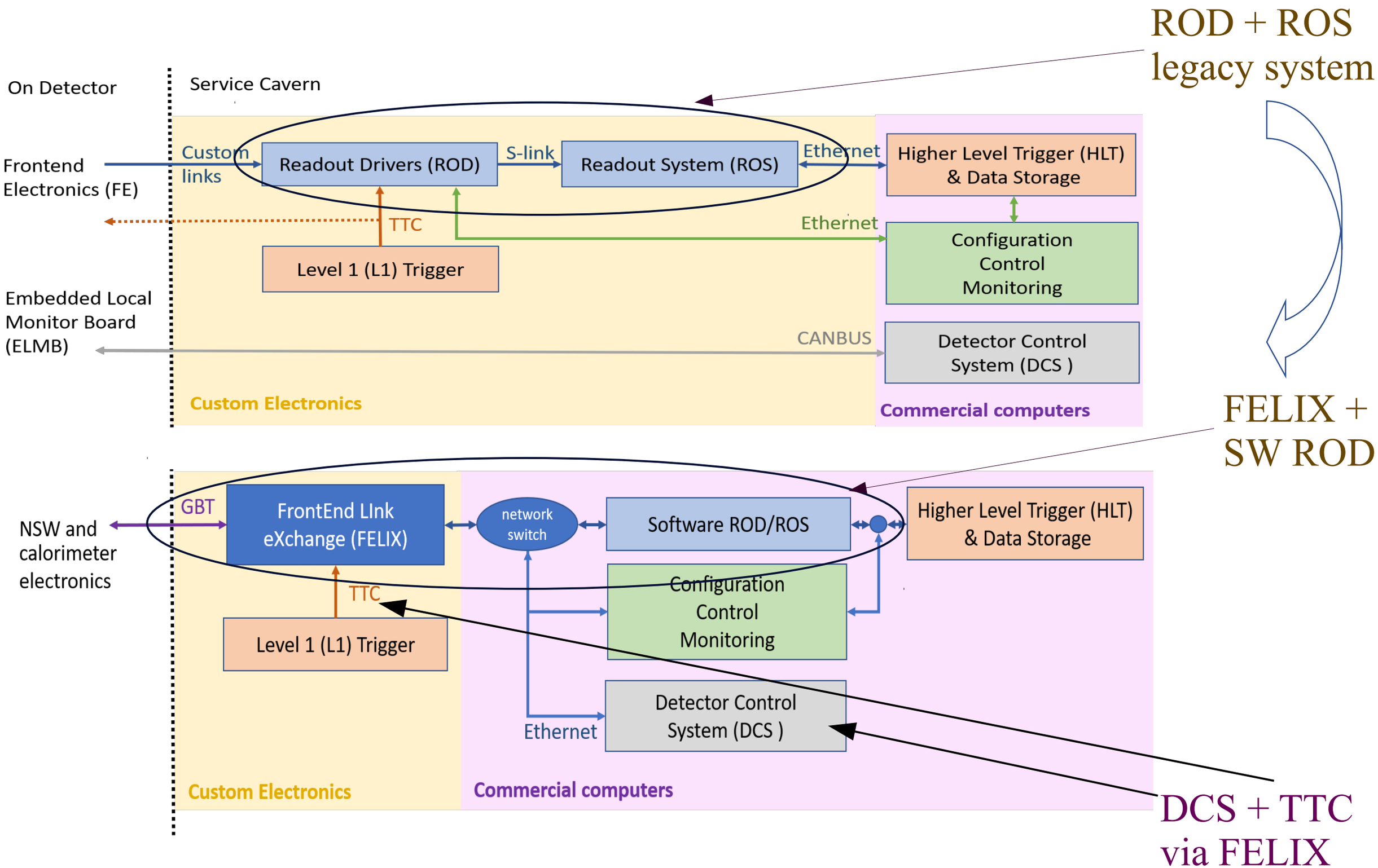


Custom elx component including FELIX cards

PCs (COTS)

*GBT: GigaBit Transceiver with Versatile Link

New Readout Architecture



New Readout Architecture

FELIX :

data/signal/message routing from/to FE elx
detector state agnostic
pushes detector fragments to SW ROD servers

SW ROD :

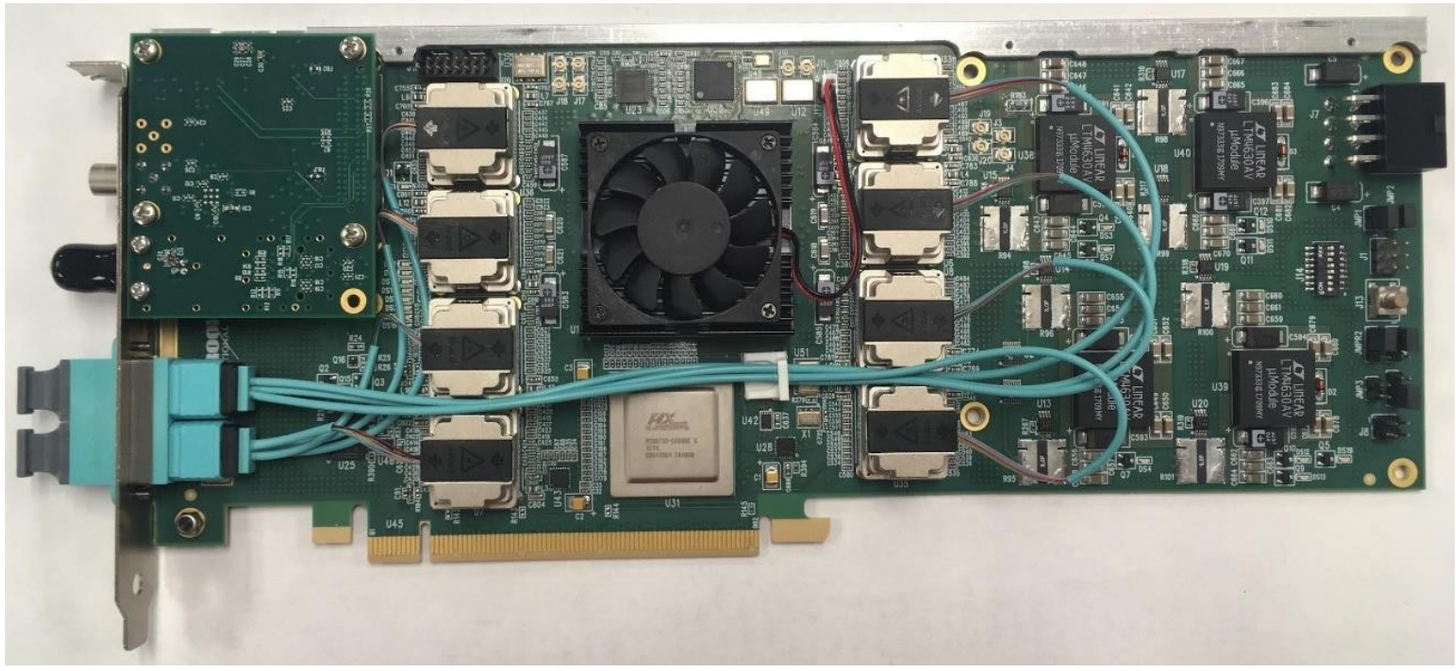
data collecting and processing
supporting configuration, calibration, control, and monitoring
interface to HLT

Run 3 FELIX system:

~100 FLX boards / 60 servers

~30 SW ROD servers

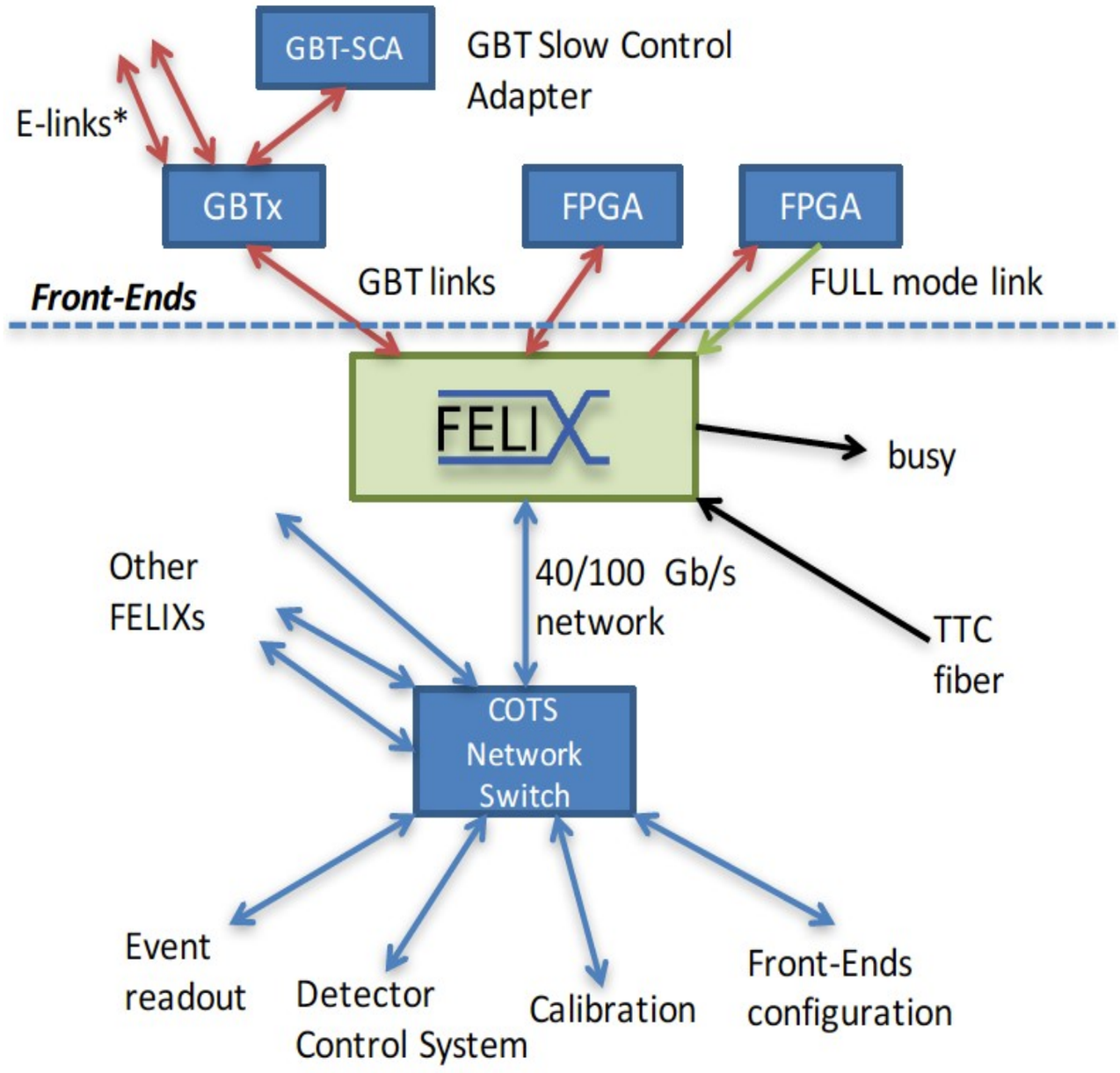
FELIX system



FLX-712 : ATLAS production board for Run 3

FELIX functionality

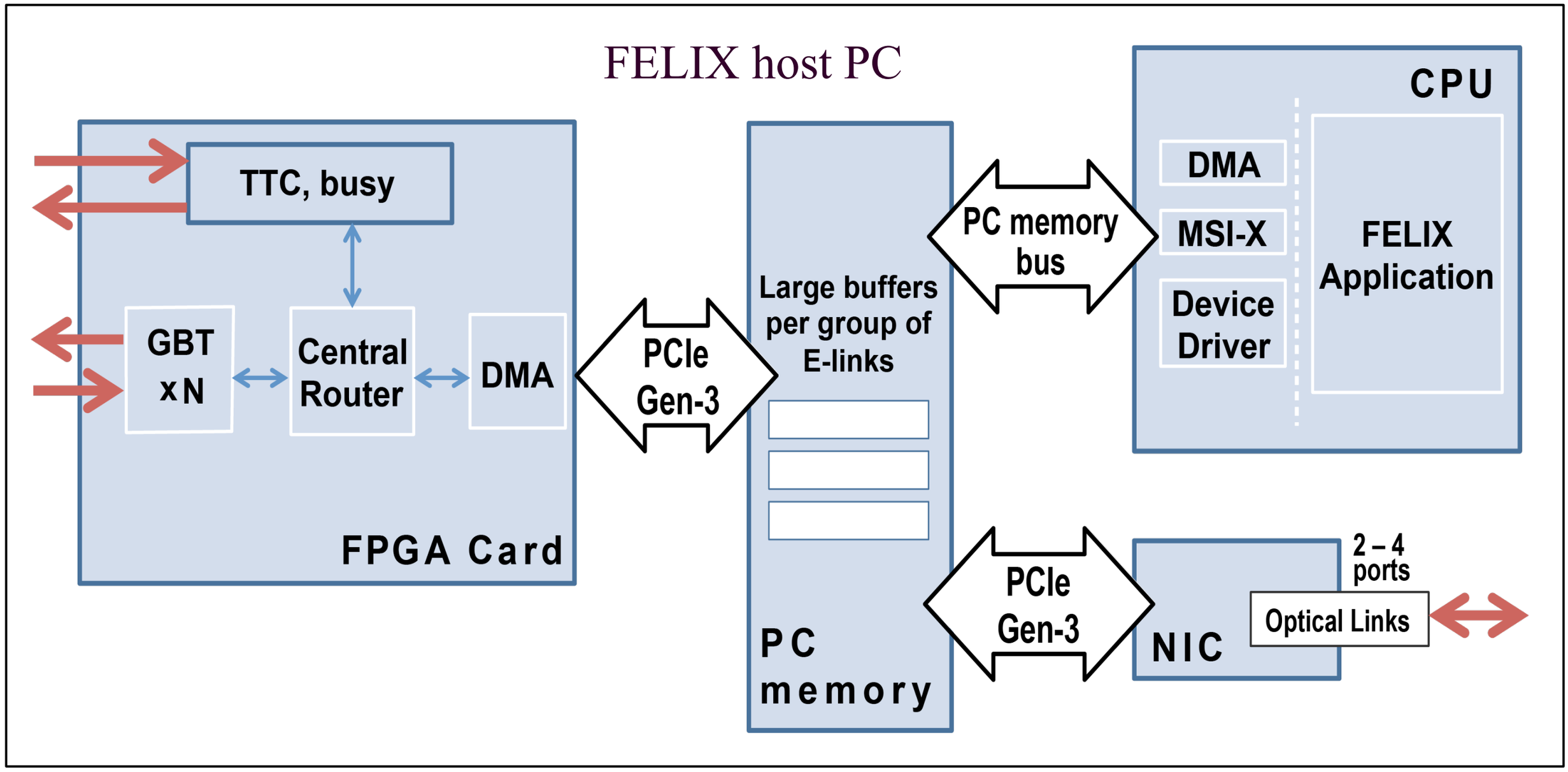
- Router between FE serial links and commercial network
- Data transport decoupled from data processing
- Get and distribute TTC (Timing, Trigger and Control) signals
- GBT-mode configurable e-links*
- Detector independent



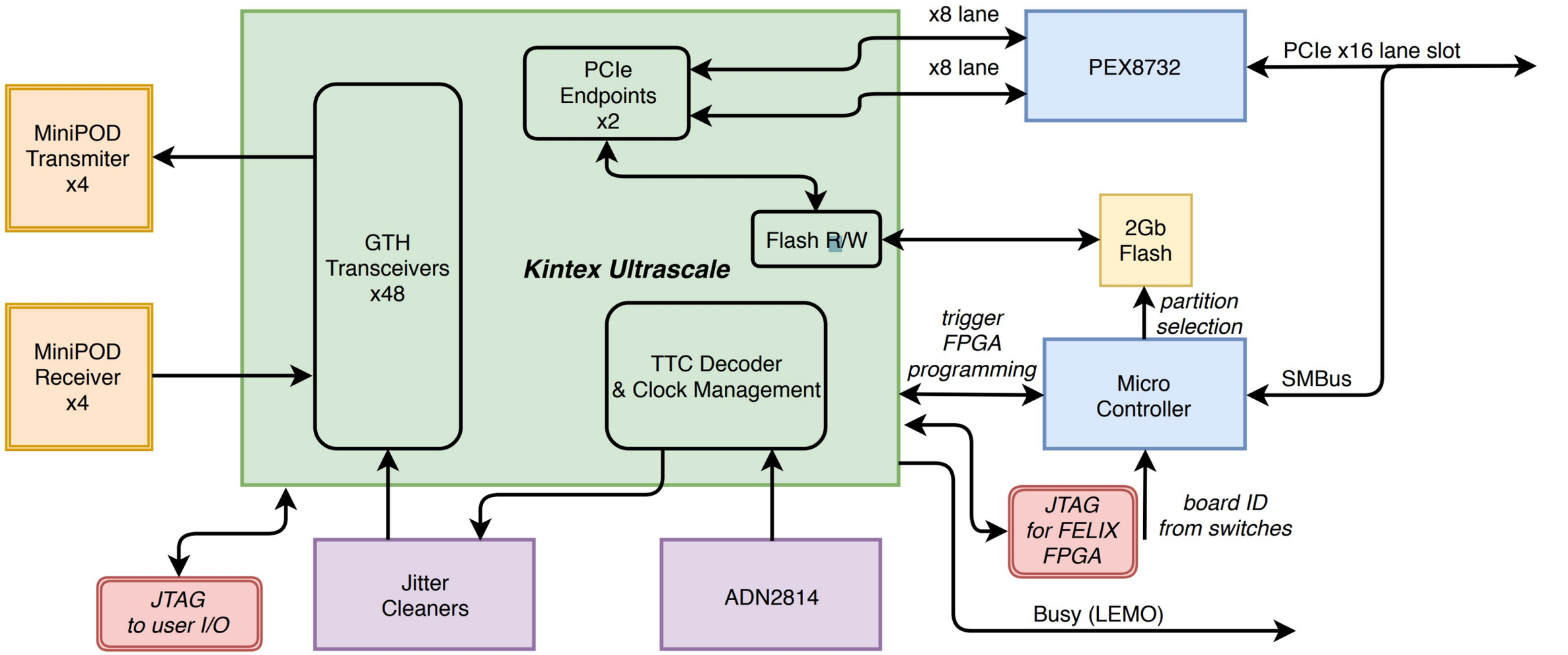
*e-link: data mux/demux protocol (more physical electrical links packed over one single GBT link)

FELIX block diagram

PC hosting up to two PCIe FELIX cards + network card



FLX-712 card features



- Kintex UltraScale FPGA
- 8 MiniPODs
- 16-lane PCIe Gen3
- Flash and µ-controller for FW update
- On-board jitter cleaner
- Timing mezzanine to interface TTC system

- 24 x 4.8 Gb/s links @ PCIe limit
- 12 x 9.6 Gb/s links @ PCIe limit
- 48 links as TTC dispatcher

Rates

Run 3 parameters for FELIX readout (worst cases)

Name	<chunksize>	Rate per channel	Channels per FELIX	Chunkrate per FELIX	Datarate per FELIX
	Bytes	kHz		MHz	Gb/s
GBT Mode	40	100	384	38.4	12
FULL Mode	4800	100	12	1.2	46

GBT Mode → FPGA-resource limited

FULL Mode → PCIe-bandwidth limited

Firmware flavours

	FLX-712 # chans
GBT dynamic - all combinations of e-links (2,4,8) and modes (8b/10b, HDLC)	4+4
GBT semi-static - static & configurable links	12+12
FULL - 6+6 channel matches max PCIe bw - 12+12 channel @ lower bw	12+12
LTDB* mode - only clock distribution, trigger, slow control and monitork	24+24 (LTDB)

*LTDB: Liquid Argon Digitizer Board

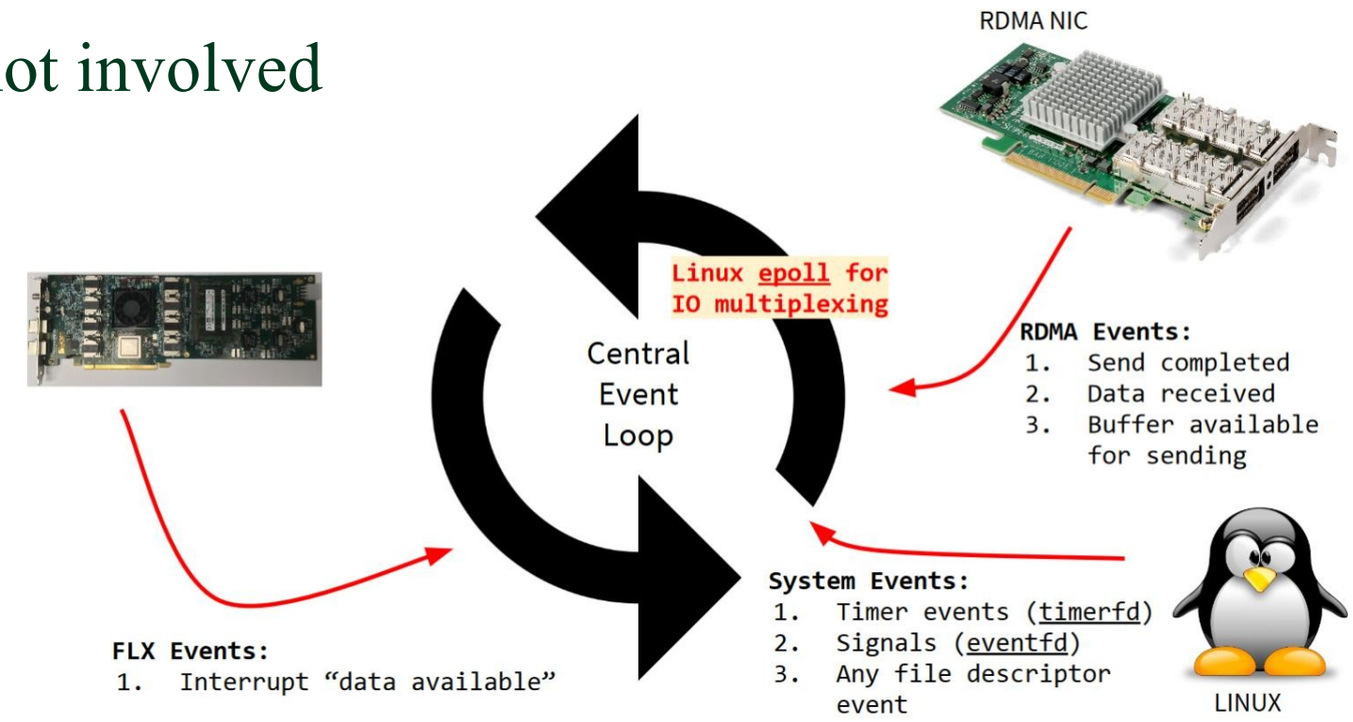
Software

low level sw → basic configuration/monitoring (e-link conf., felix monitoring)

higher level sw → data rate and channel monitoring

felix-star

- single-threaded event loop, any operation is one event
- networking based on new communication library: NetIO-next
- data transfer uses RDMA i.e. kernel not involved
- higher performance

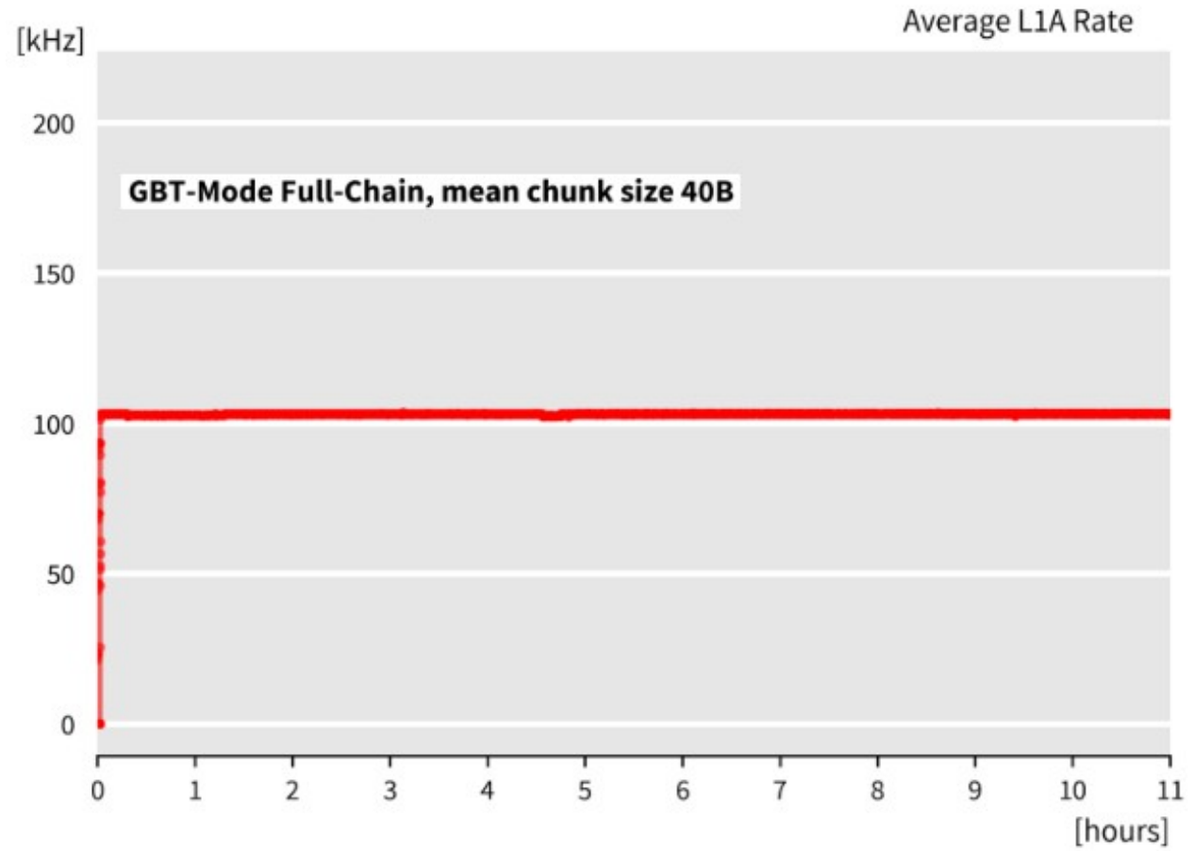


Performance, integration and commissioning

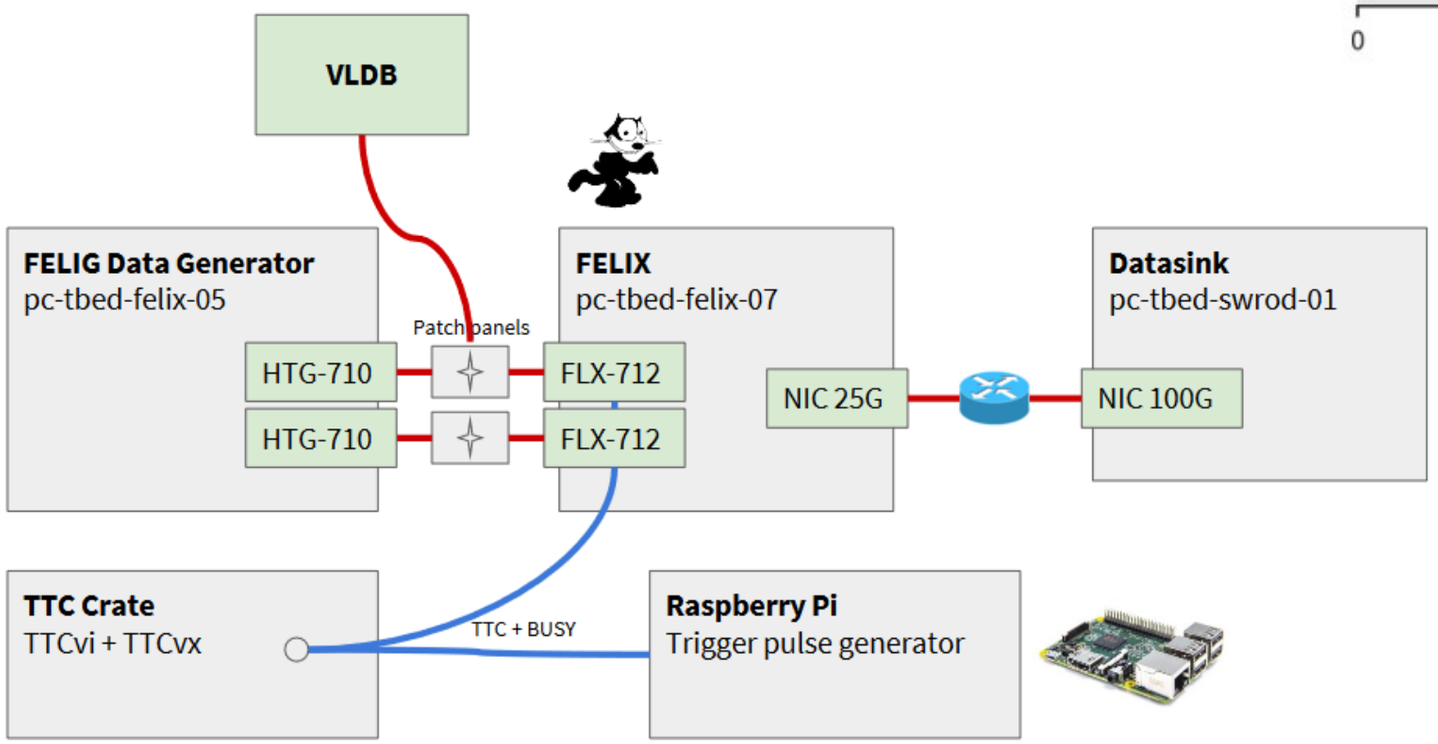
GBT mode

- stable multi-hour operation (longer than average LHC fill)

- reliable parallel communication with test board featuring DCS components



~ 12.5 Gbps network throughput

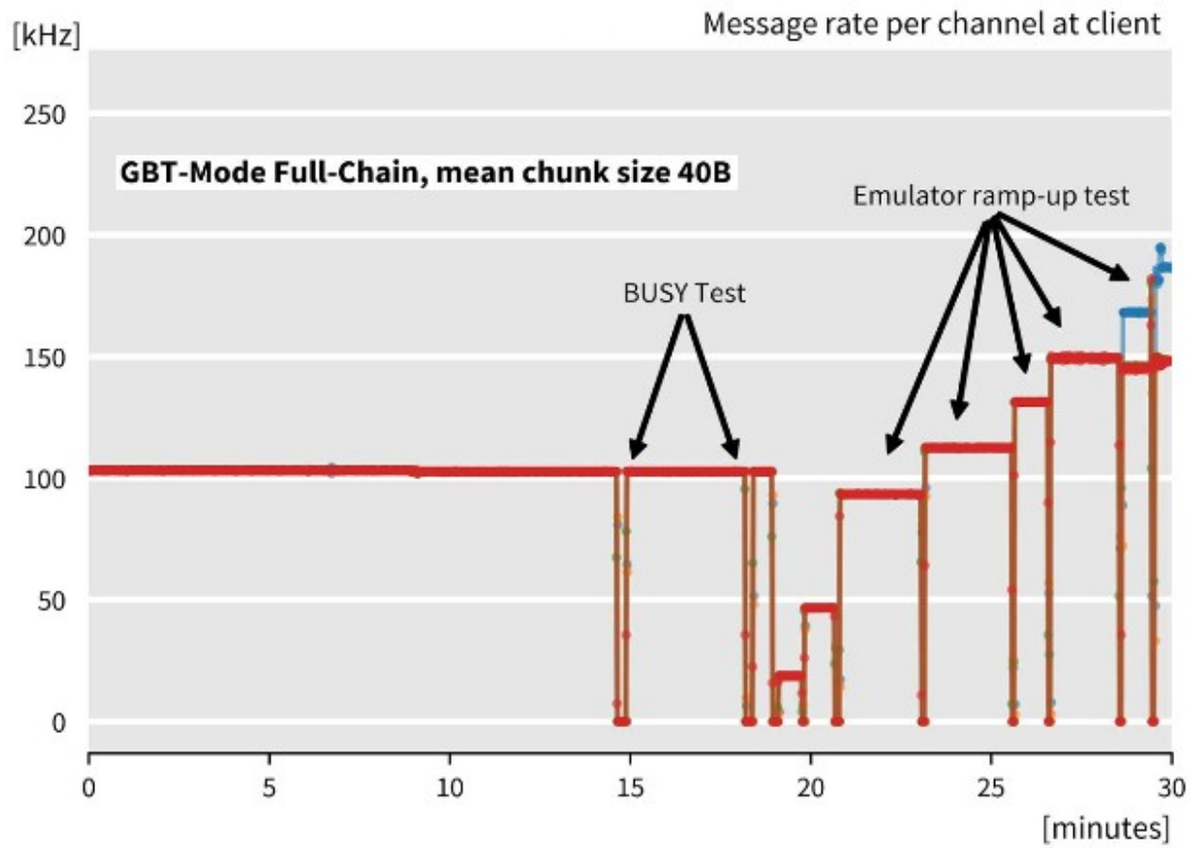


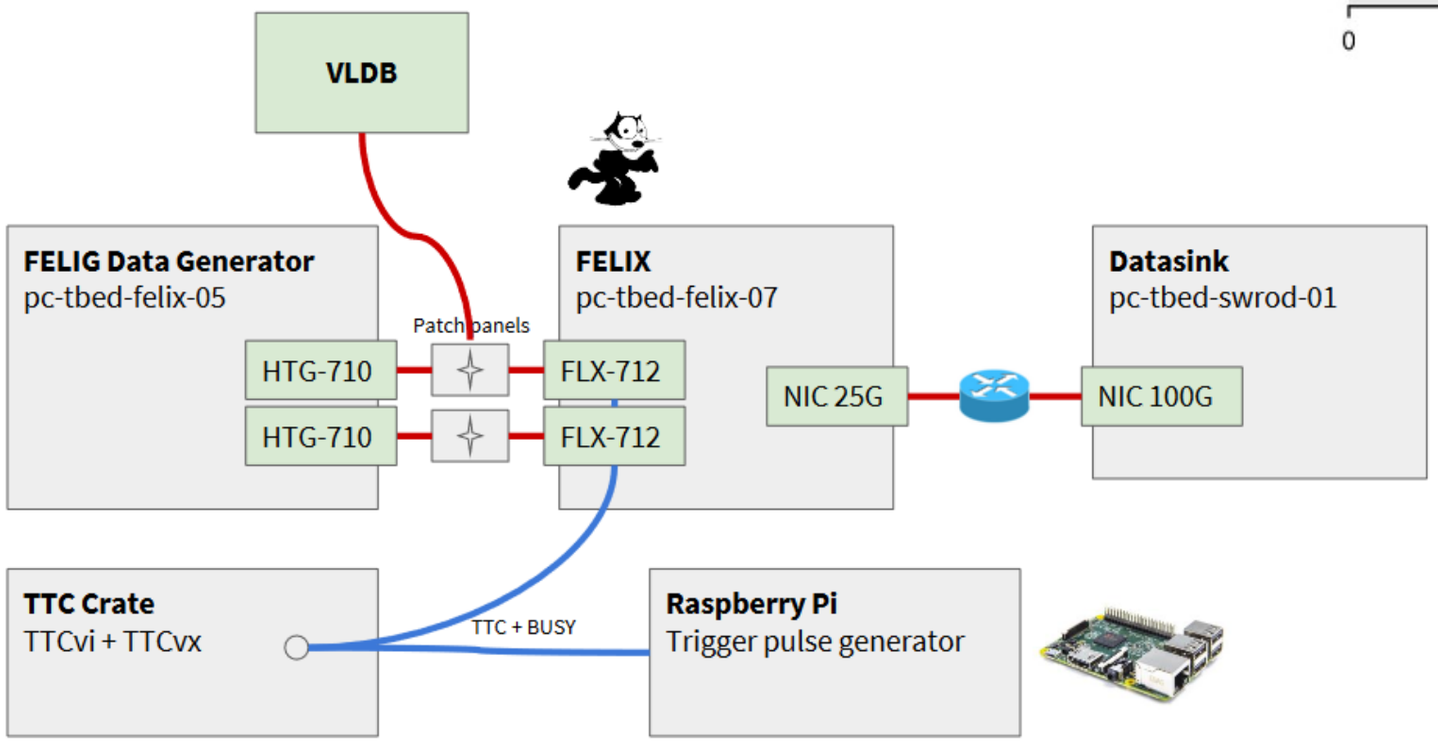
Performance, integration and commissioning

GBT mode

- BUSY signal propagation correctly handled

- Emulator rump-up demonstrated rates 50% above expectation (150 kHz)

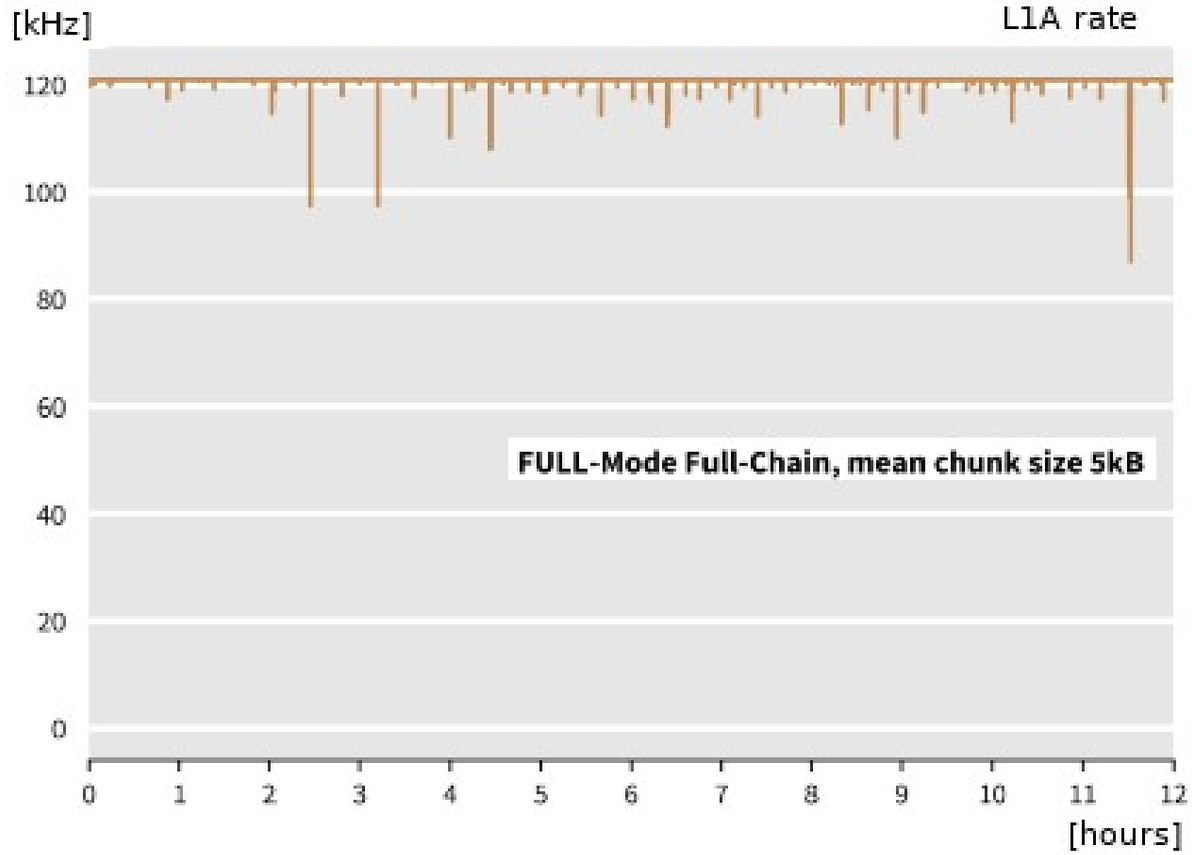


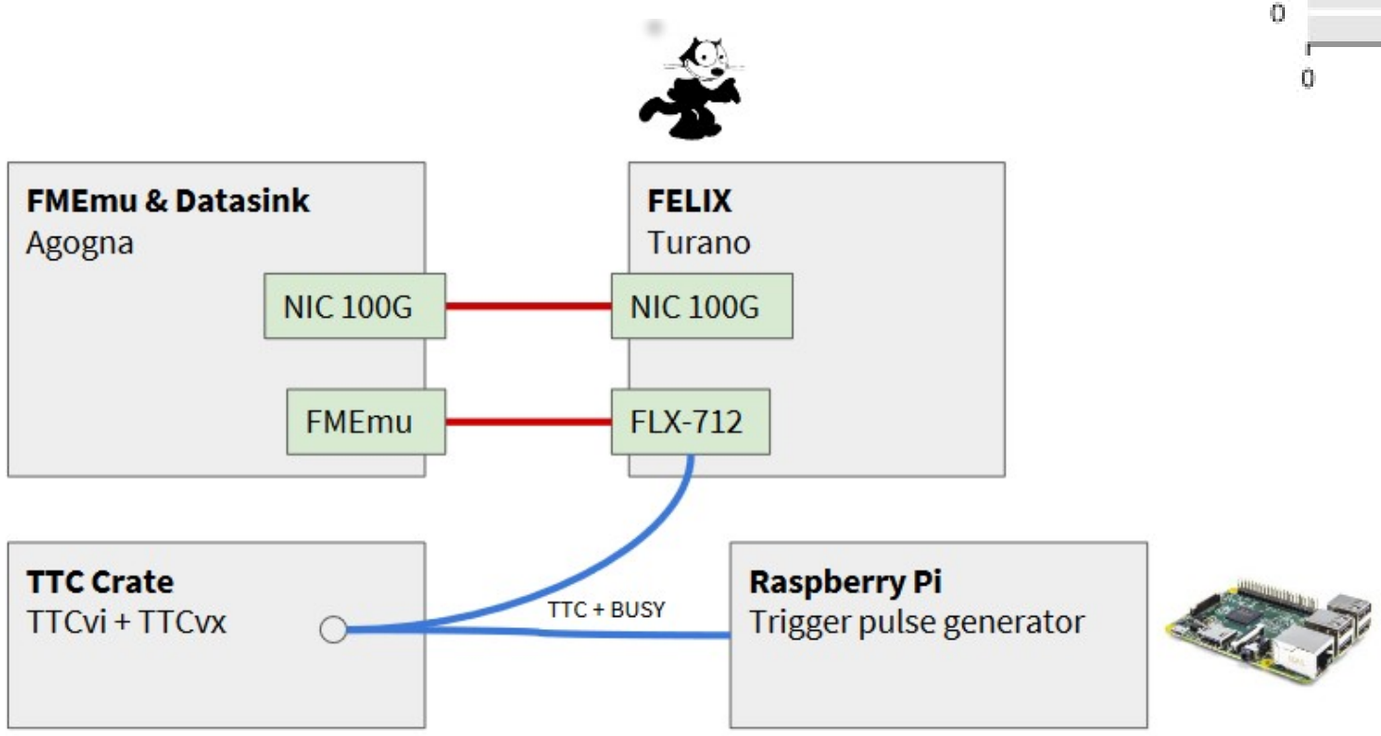


Performance, integration and commissioning

FULL mode

• stable multi-hour operation (longer than average LHC fill)

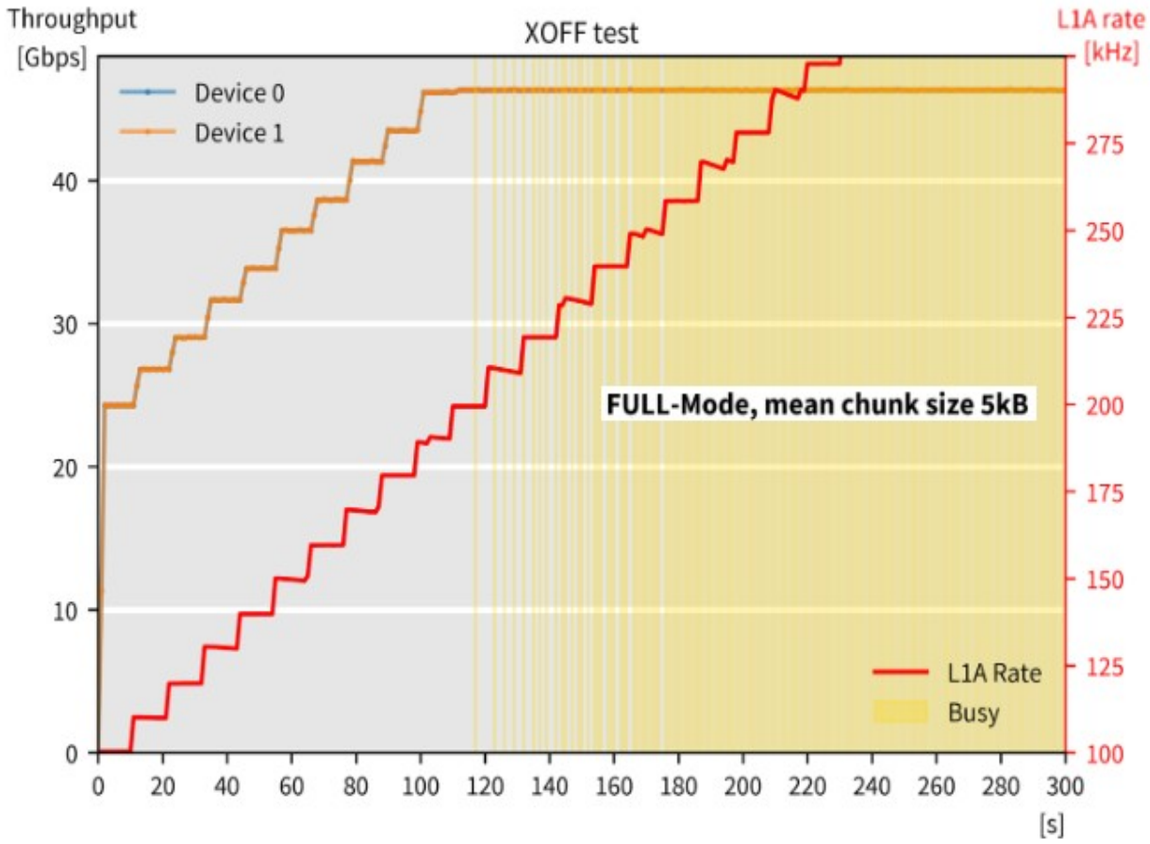


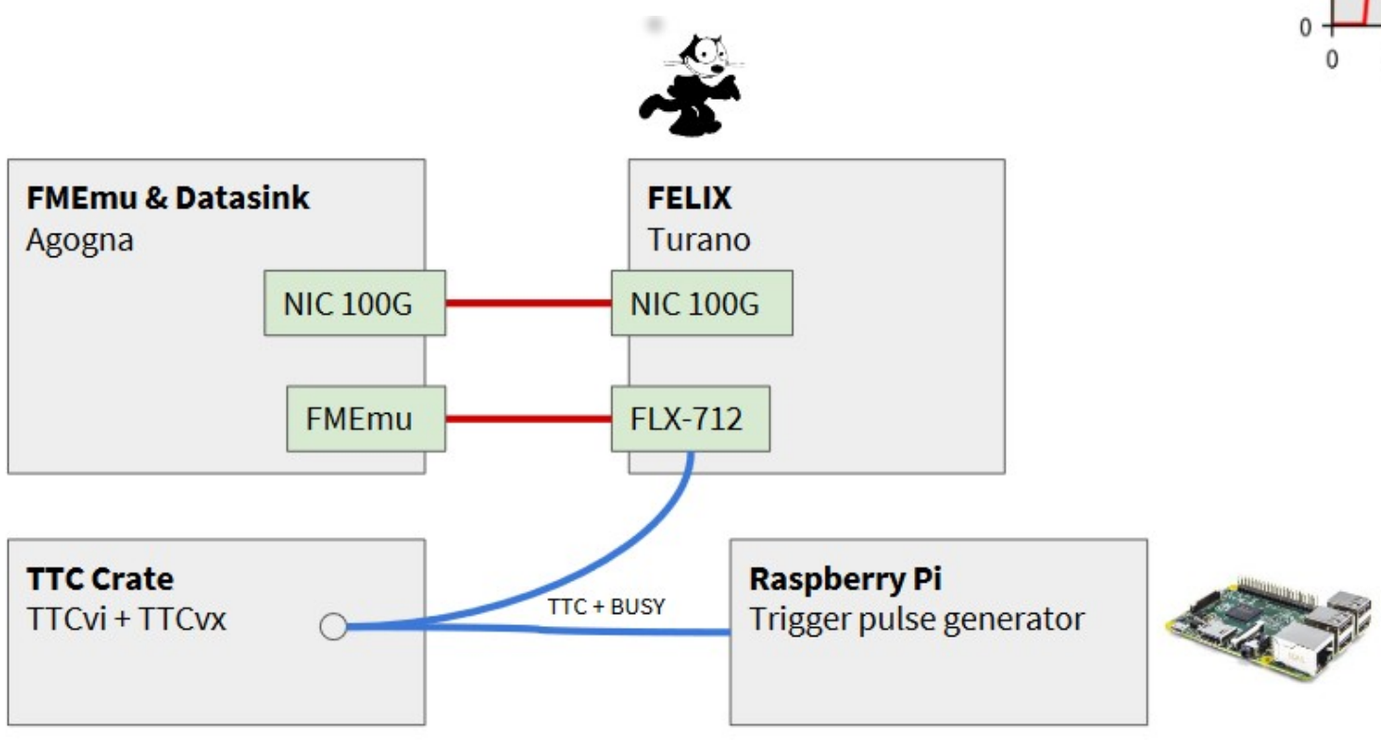


Performance, integration and commissioning

Stress test

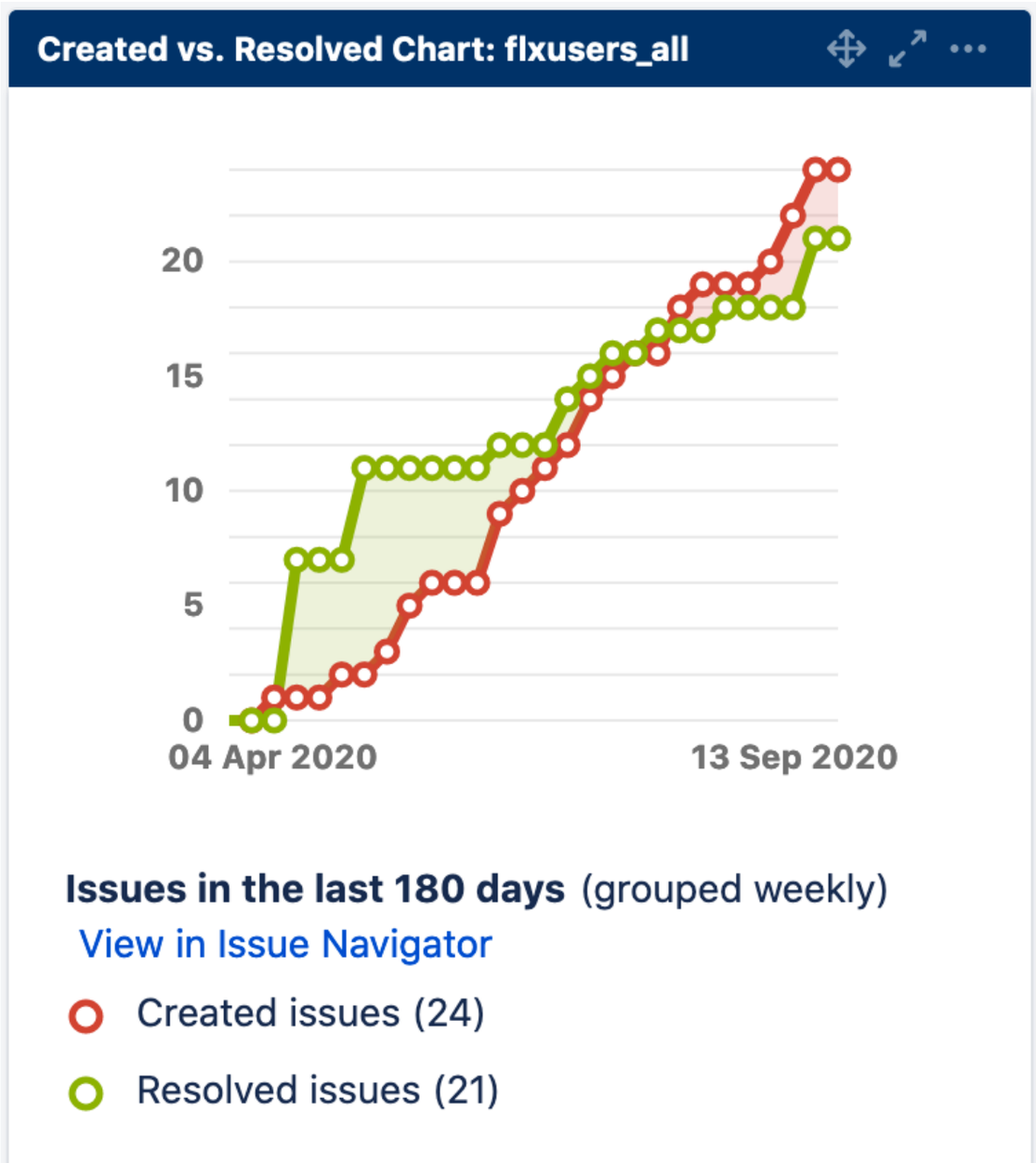
- backpressure shows up at ~ 200 kHz
- achieved ~ 300 kHz

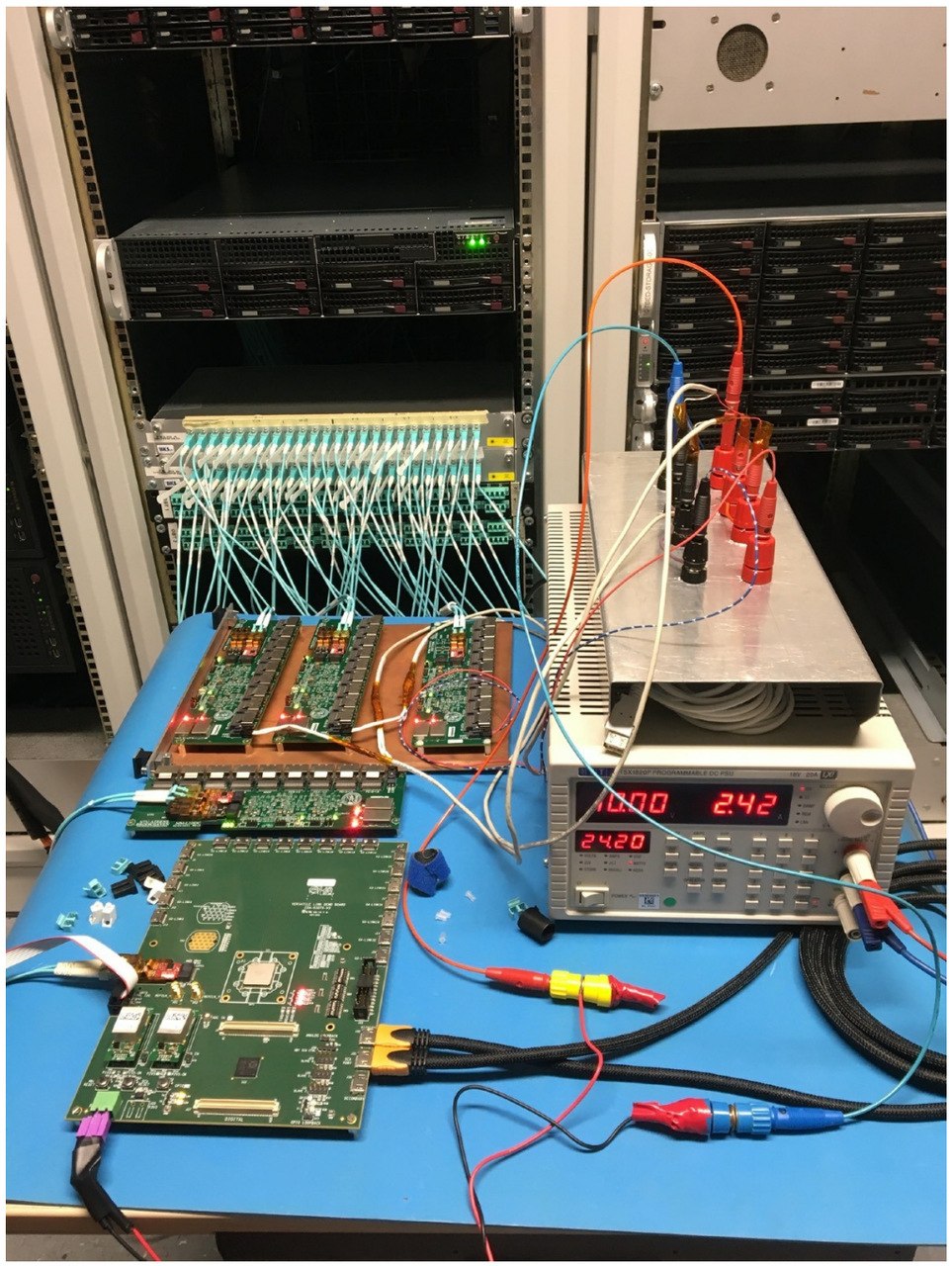




User support & integration

- Active user support → crucial:
  user issues more and more difficult to reproduce in TDAQ testbed
- Further push toward integration





Summary & Outlook

ATLAS TDAQ evolution for Run 3:

FELIX + SW ROD replace (part of) ROD + ROS system

→ more flexibility, reduced custom design

→ architecture for Run 4 and beyond

HW commissioning and deployment ongoing

(~ 200 FELIX cards already delivered)

SW development in progress

Performance tests consistently exceed Run 3 requirements

User feedback more and more relevant

Please, see also talk by Serguei Kolos on SW ROD

Thanks!

Extras

ATLAS Readout in Run 3

New Muon detectors: New Small Wheels, small “BIS7/8” RPCs
New L1 Calo trigger system
→ exploit commodity and/or common hw as soon as possible

FELIX

- Connects directly to detector FEE
- Receives and routes data from detector directly to clients
- Routes L1 trigger, clock and control signals to detector FEE
- 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 - enables detectors to insert their own processing into data path
- Previously performed in ROD hardware
- Buffers and - on request - serves data to HLT
- Interface indistinguishable from legacy readout (ROS)

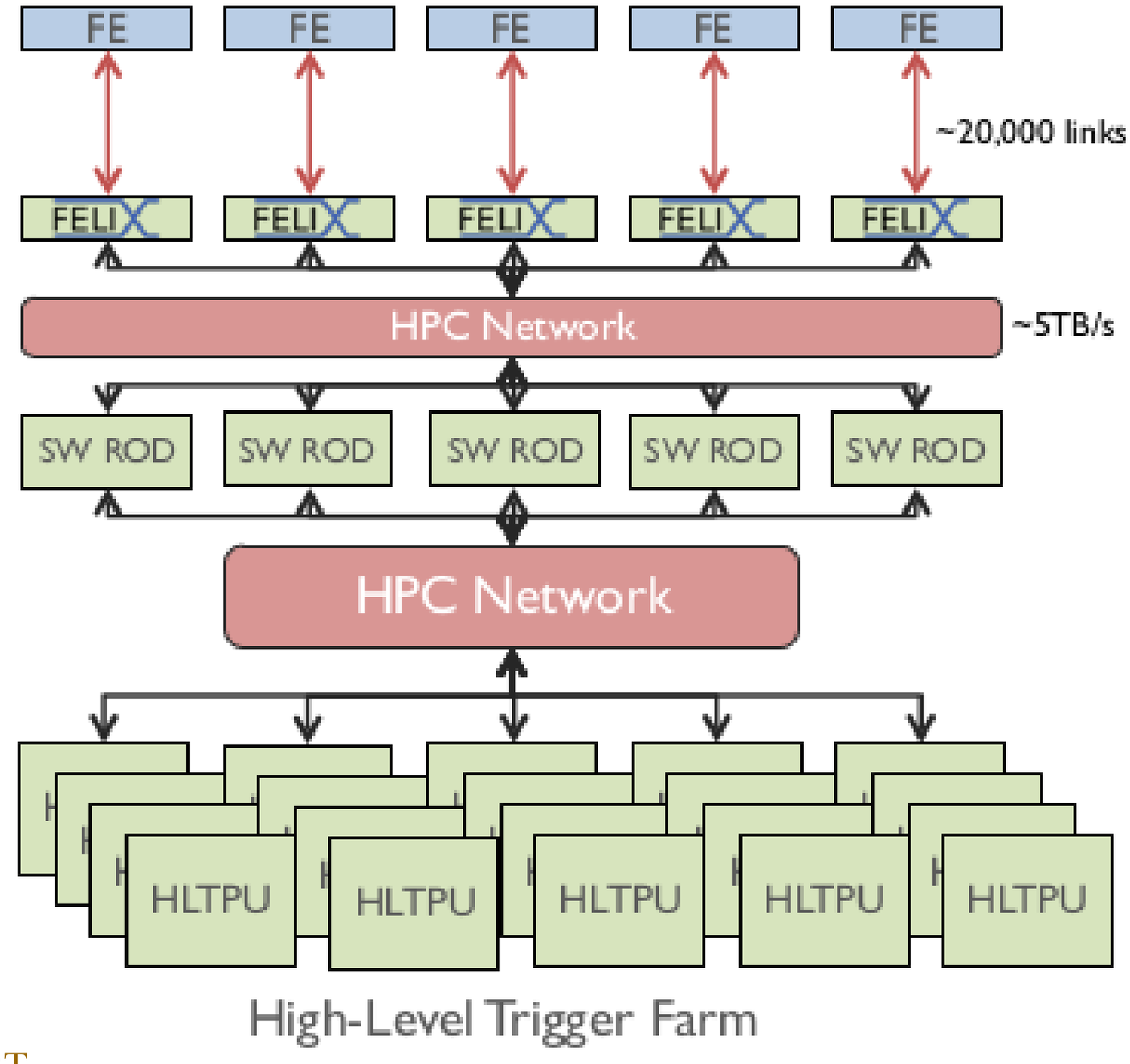
• Control and monitoring applications now distributed among servers connected to data network

Upgrade for HL-LHC (Phase-II)

~ 5 MB events, ~ 5 TB/s network bandwidth, ~ 50 GB/s recording throughput

GBT, LpGBT* or FULL mode links

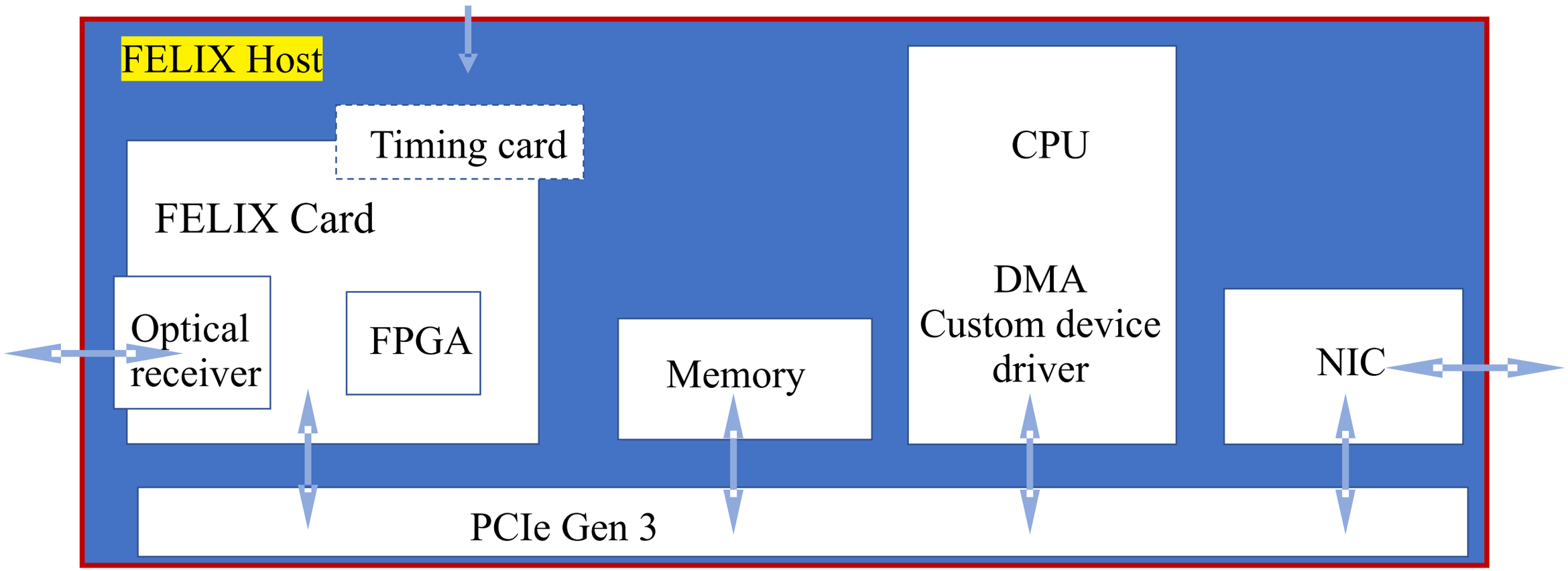
COTS network technology



Custom elx component including FELIX cards

PCs (COTS)

*LpGBT: Low-power GBT


FELIX Host
Timing card
FELIX Card
Optical receiver
FPGA
Memory
CPU
DMA
Custom device driver
NIC
PCIe Gen 3

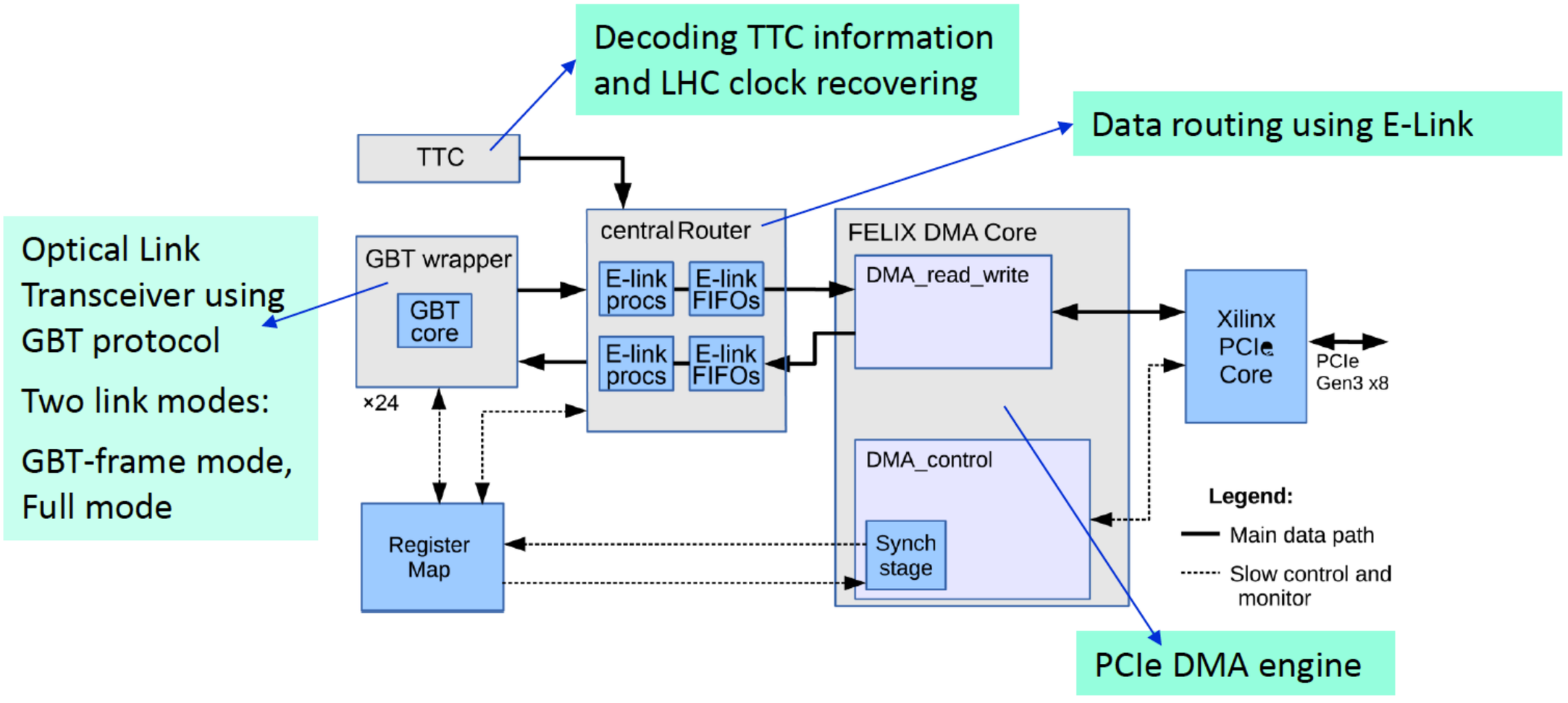
FELIX firmware design

Central router block handles e-links

e-link: data mux/demux protocol designed for ATLAS

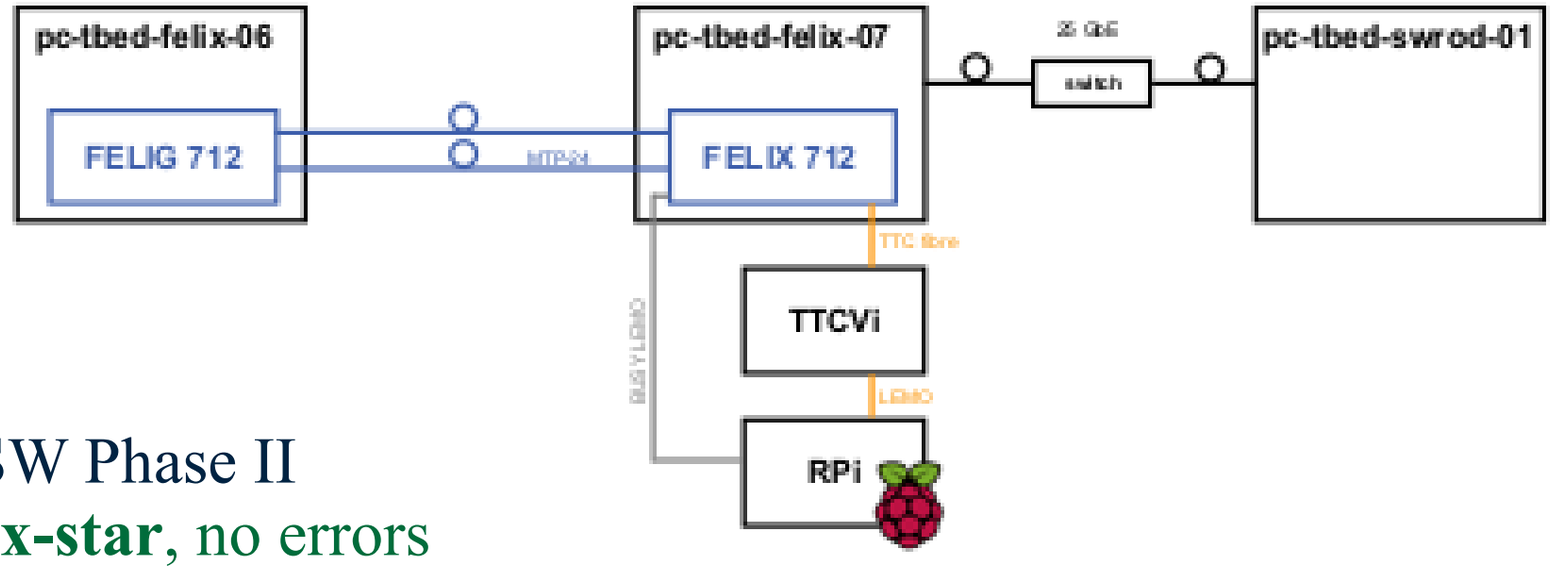
Fixed latency transmission to FE

DMA for PCIe throughput to host memory (2 × 64 Gbps)

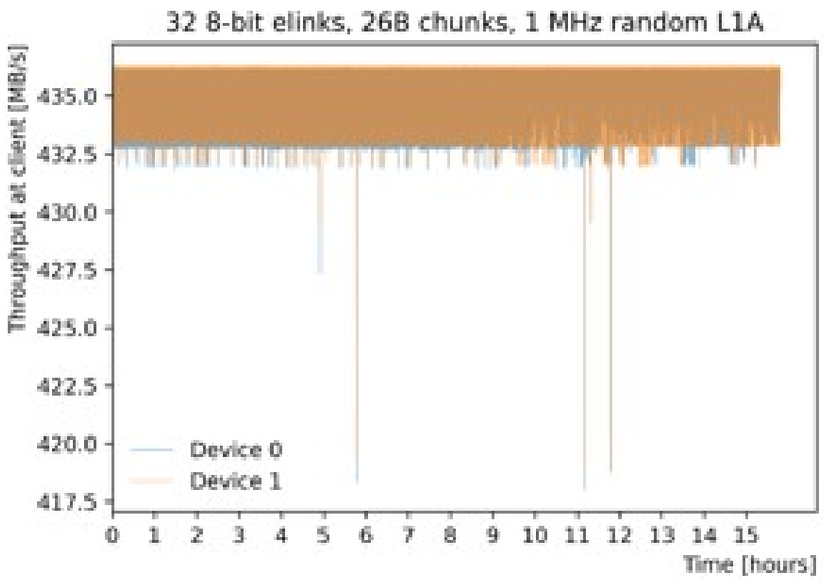


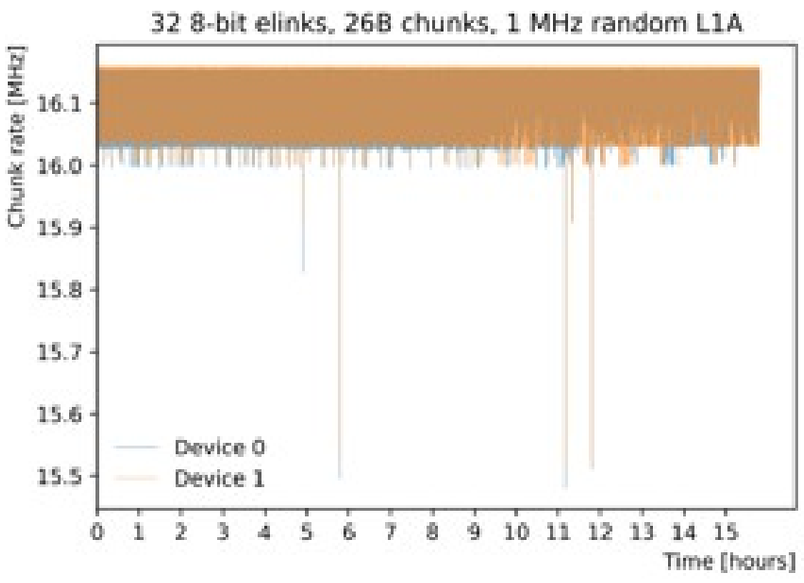
Performance & integration (Phase-II)

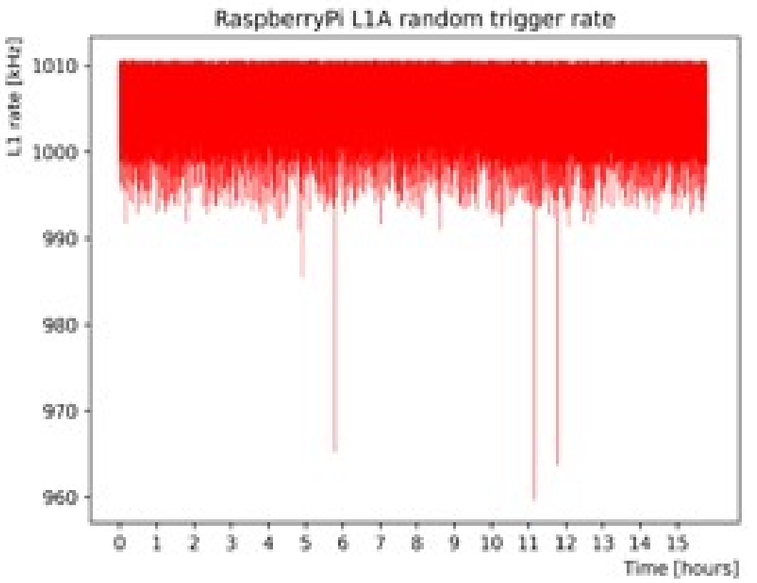
FELIX testbed:



- 32 × 260 Mb/s links
- 1 MHz random L1A for NSW Phase II
- stable transfer rate with **felix-star**, no errors
- actual test to be performed with NSW vertical slice







Integration in FELIX testbed

Integration with new trigger system (ALTI)

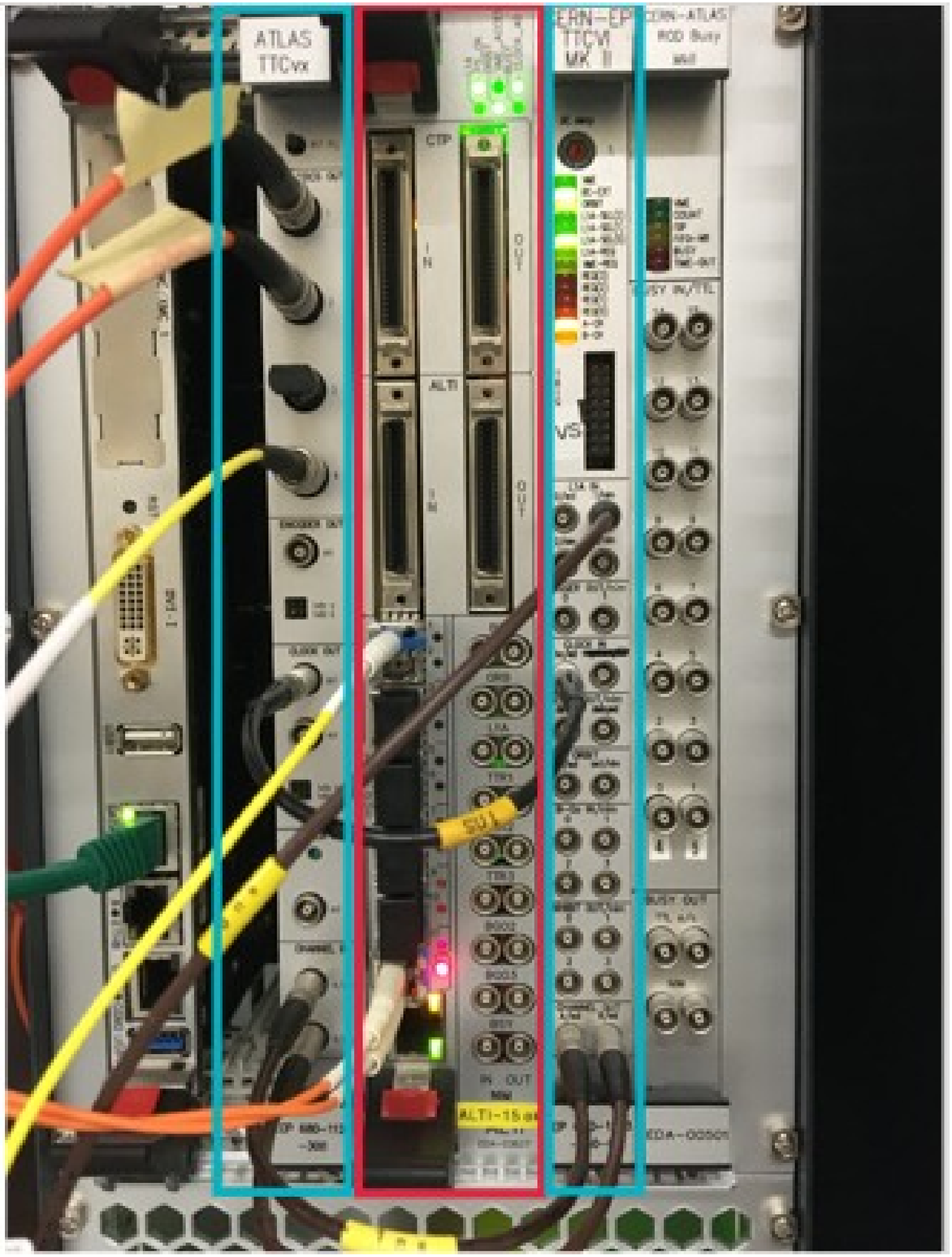
- ALTI installed and update
- learning how to operate it

Integration with swROD

- several test runs with felix sw emulator
- rerunning test with new sw

Integration with OPC-UA

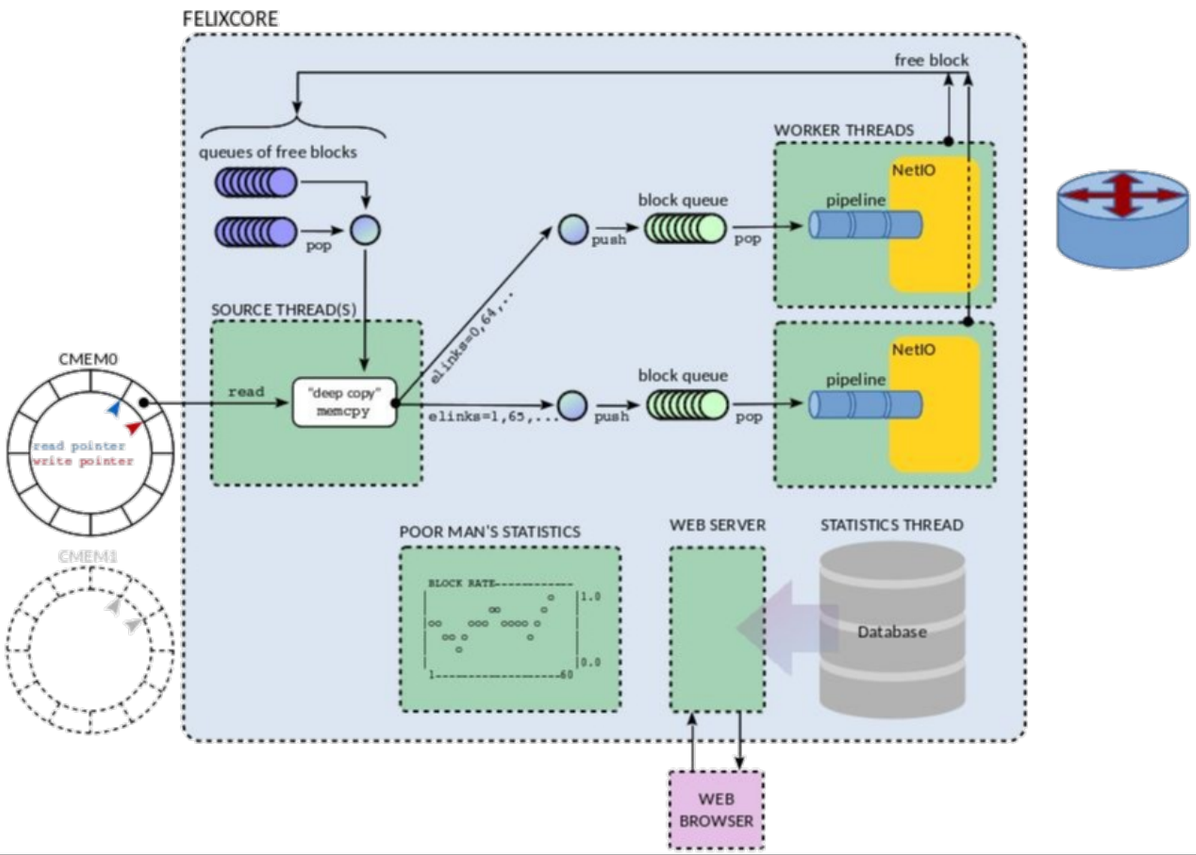
- DCS sw depends on NetIO library
- preparing joint set-up as common reference frame



Software transition

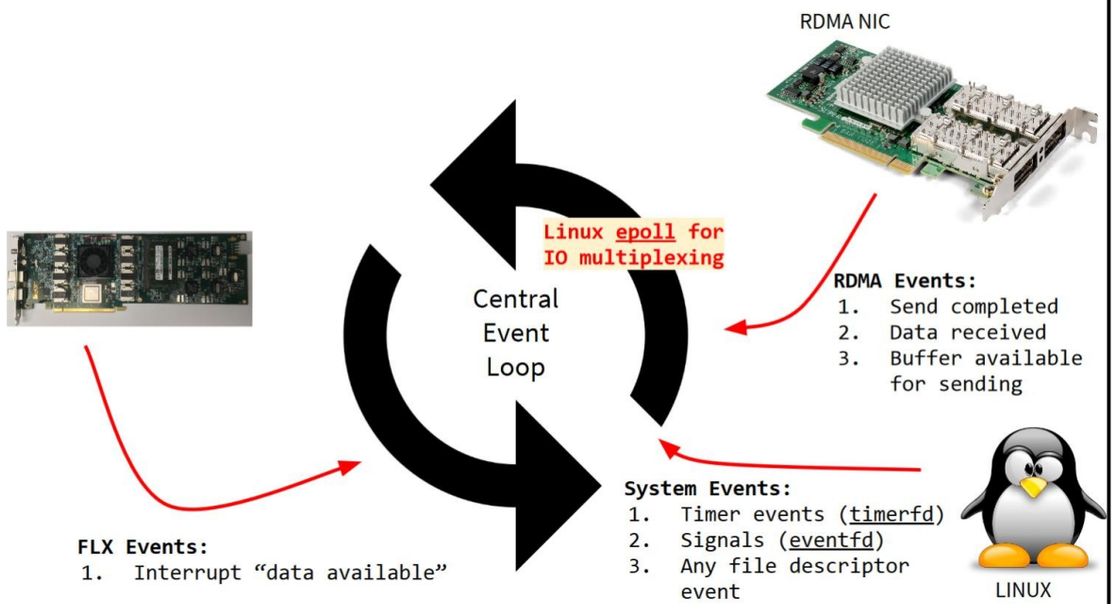
felixcore

- multiple-threaded, pipeline architecture
- networking based on “NetIO” library
- functional, minimal performance margin
- supported until all users migrated to felix-star



felix-star

- single-threaded event loop, any operation is one event
- networking based on “NetIO-next” library
- uses RDMA i.e. kernel not involved in data transfer
- higher performance
- transition in progress ...



NetIO-next performance tests

GBT test configuration:

pc-tbed-felix-06

swrod_netio_next_publisher

- 192..1152 e-links
- 1..12 publishing threads
- 40 bytes packets

netio-next
over
100 Gbps Ethernet

pc-tbed-swrod-01

swrod_test application

- 1..6 ROBs with 192 e-links per ROB
- 1..2 reading/assembling threads per ROB
- 6 HLT processing threads
- No custom processing

asyncmsg
over
40 Gbps Ethernet

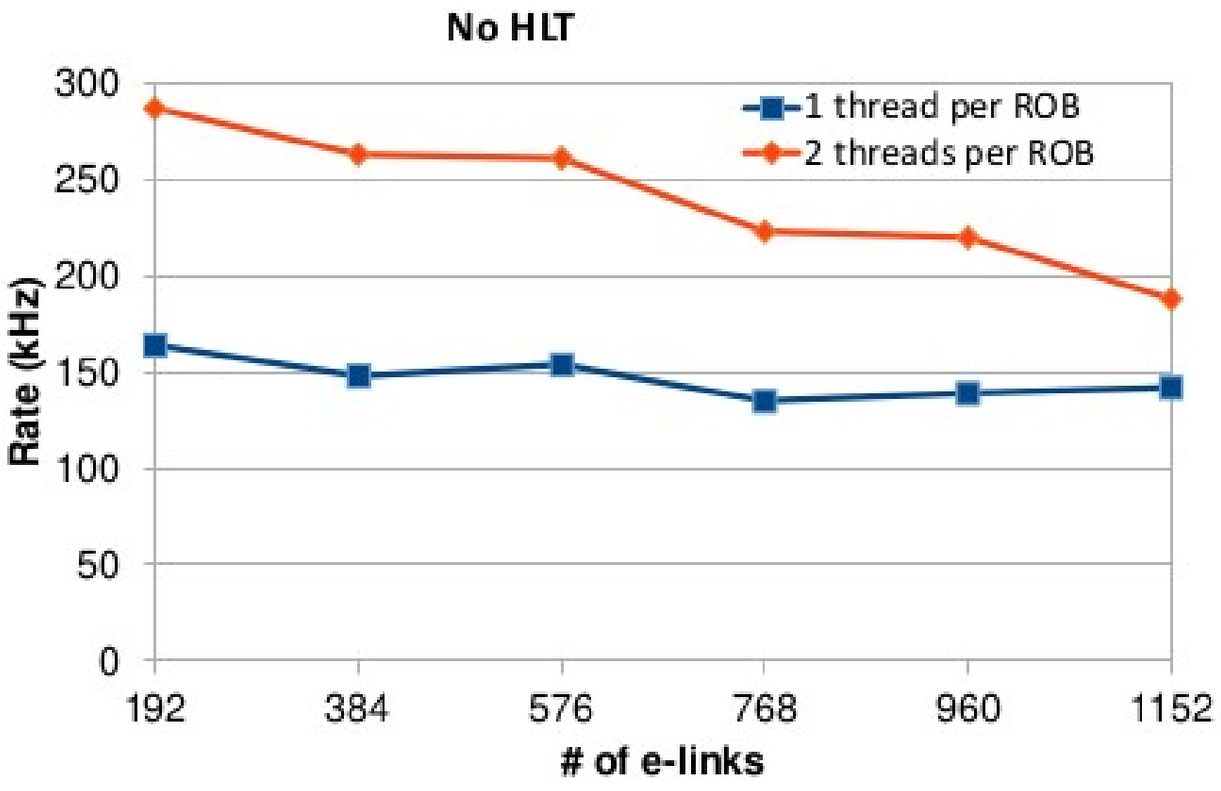
pc-tbed-net-03

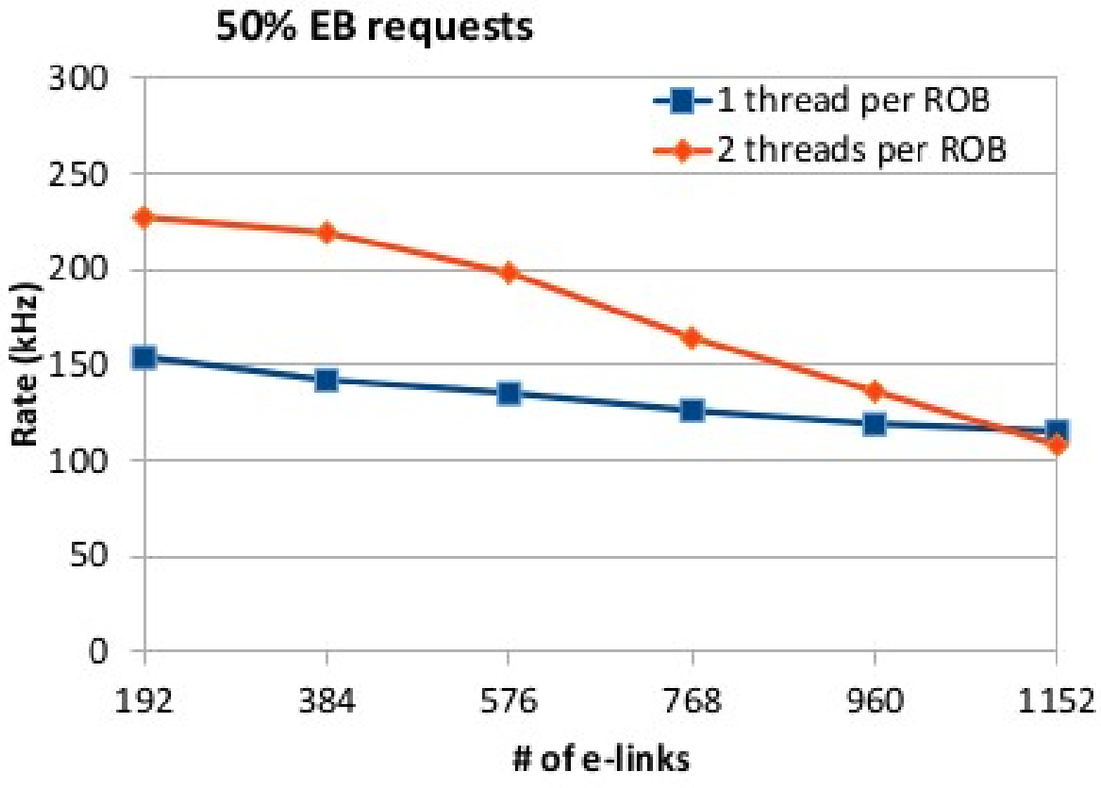
swrod_hlt_requester

- 50% EB requests
- Request rate is dynamically adjusted to the reply rate

NetIO-next performance tests

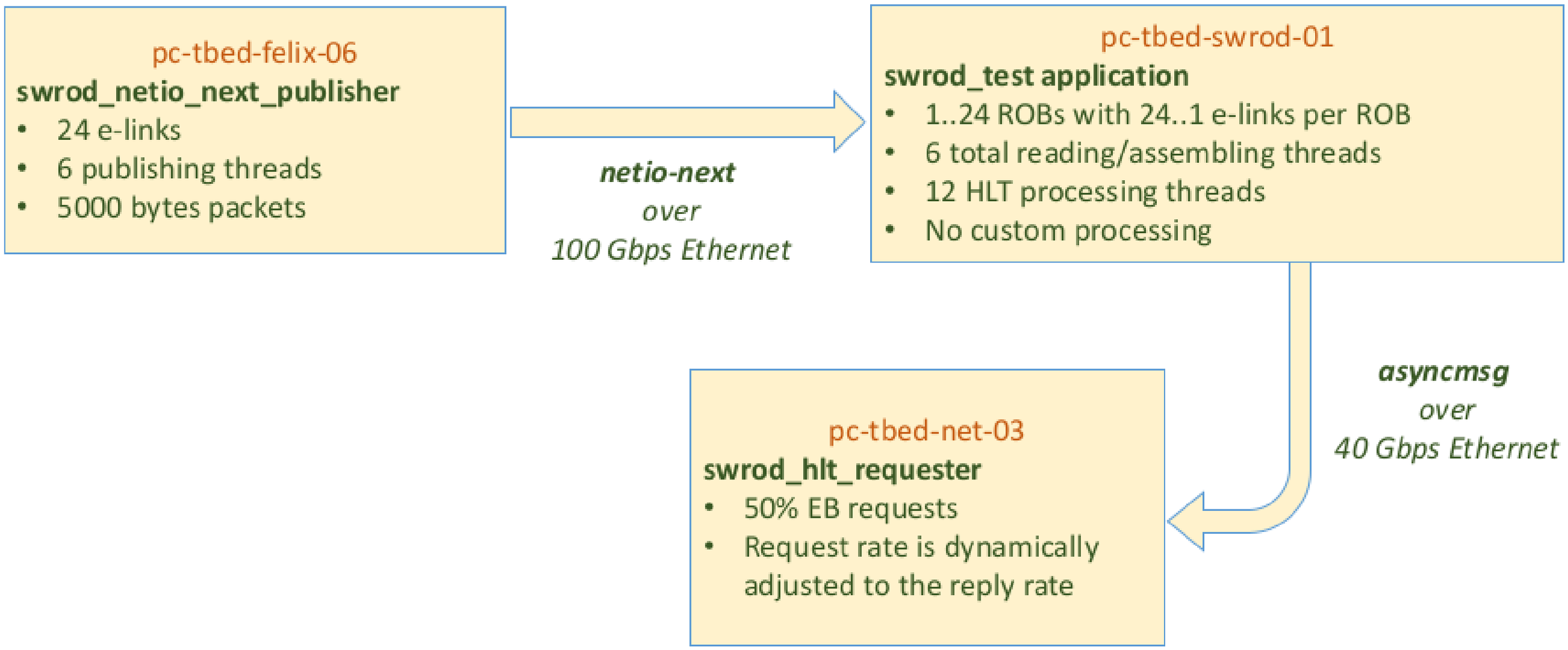
GBT test results:





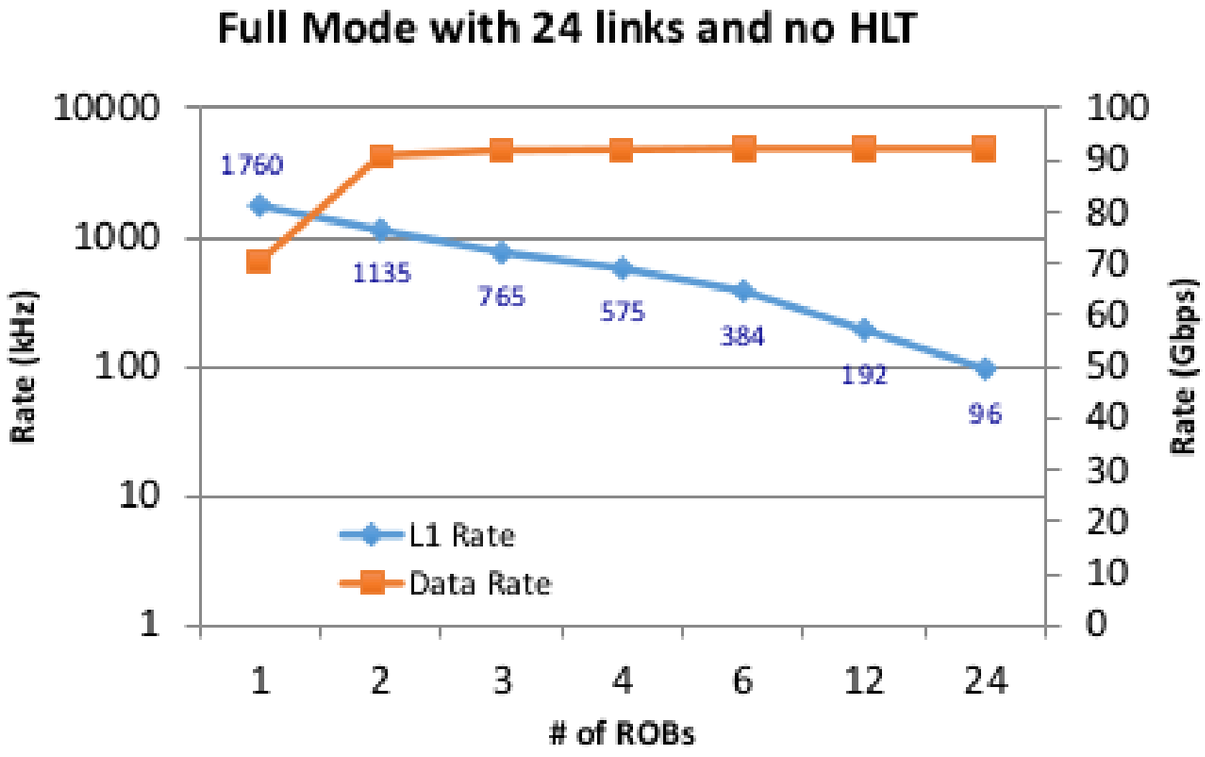
NetIO-next performance tests

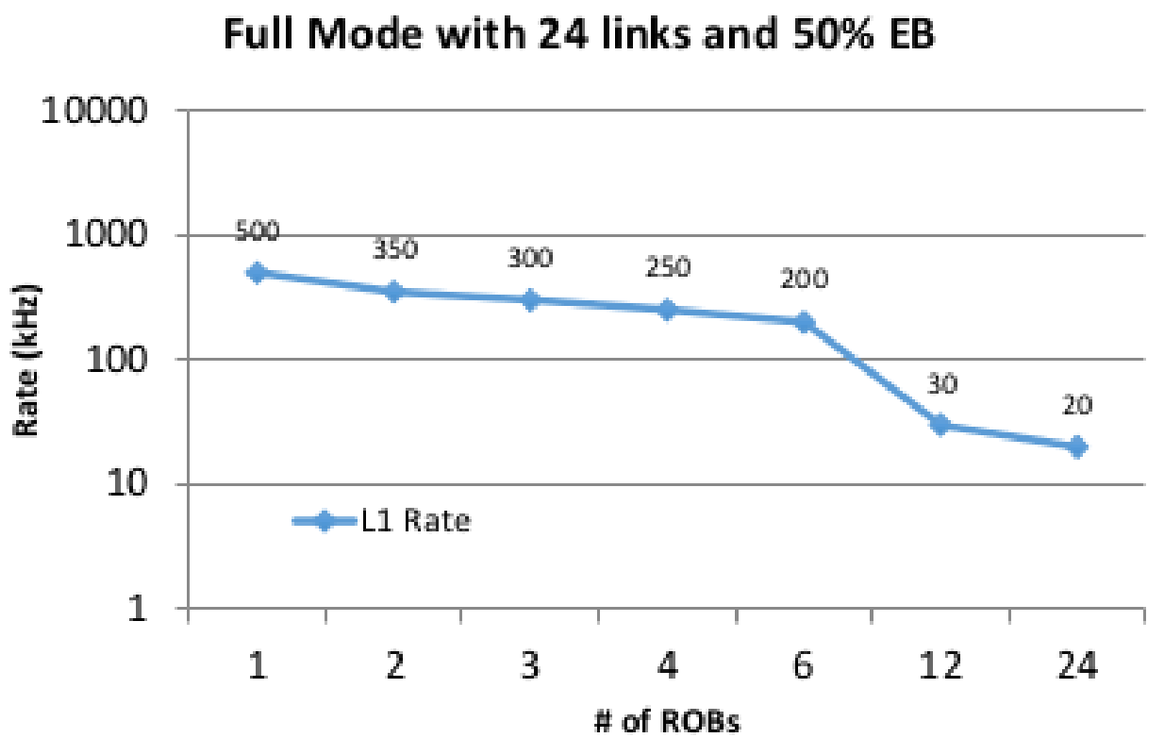
FULL-mode test configuration:



NetIO-next performance tests

FULL-mode test results:





- investigating scaling problem