FELIX
Readout System

Mengqing Wu
Radboud University Nijmegen/Nikhef
On behalf of the ATLAS TDAQ Collaboration

2024 CEPC international workshop
24 October 2024

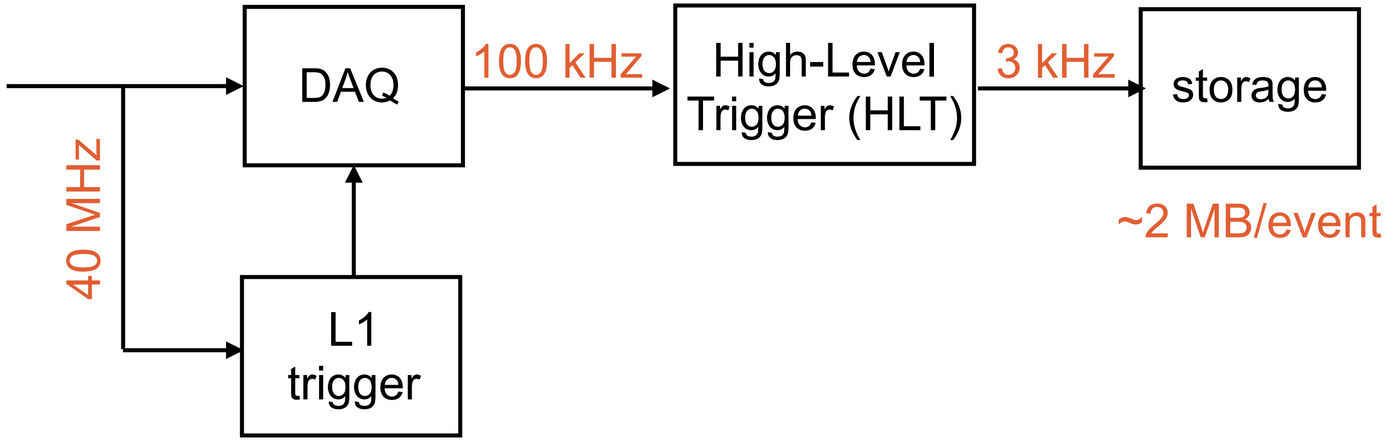
Radboud University

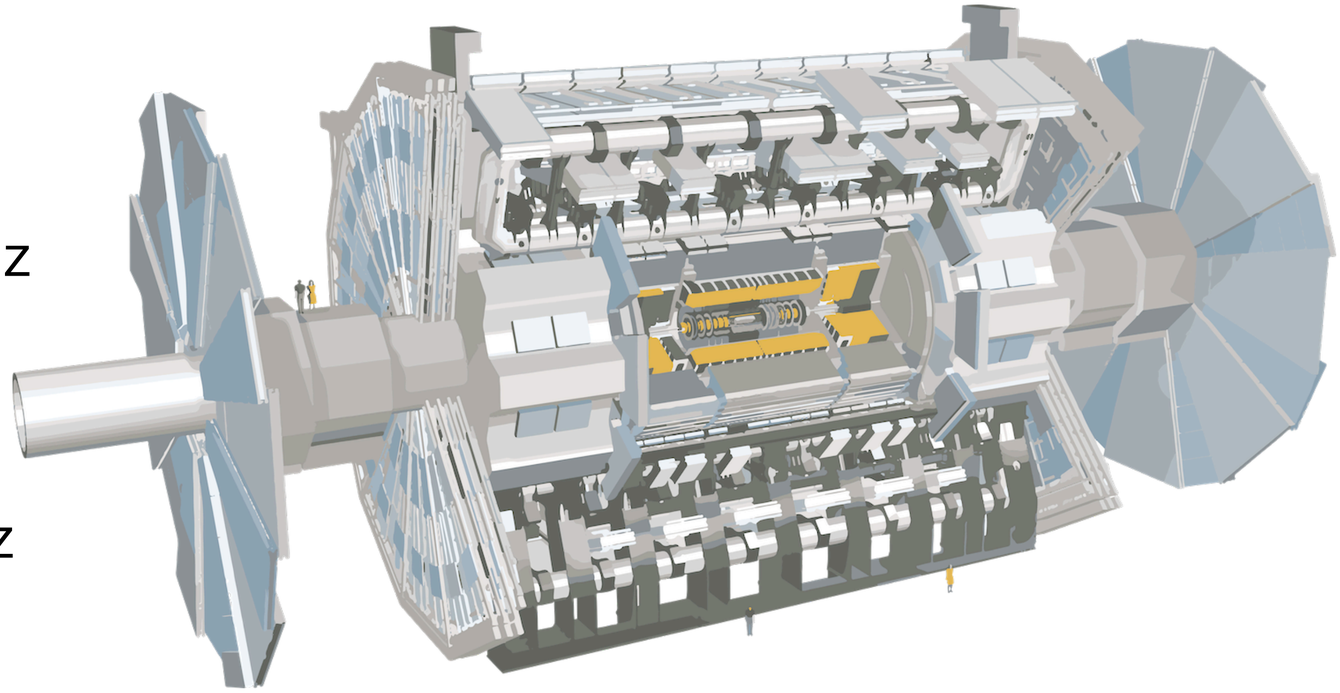
The FELIX readout system

- A generic detector readout concept, proposed by the ATLAS collaboration
  – connecting front-end serial links to commodity network
- Collaboration of different institutes
  - Open-source firmware and software
- Applications
  - At LHC: ATLAS, ATLAS Phase-II
  - Beyond LHC: NA62, protoDUNE, sPHENIX@RHIC, CBM@FAIR, LUXE

The ATLAS TDAQ architecture

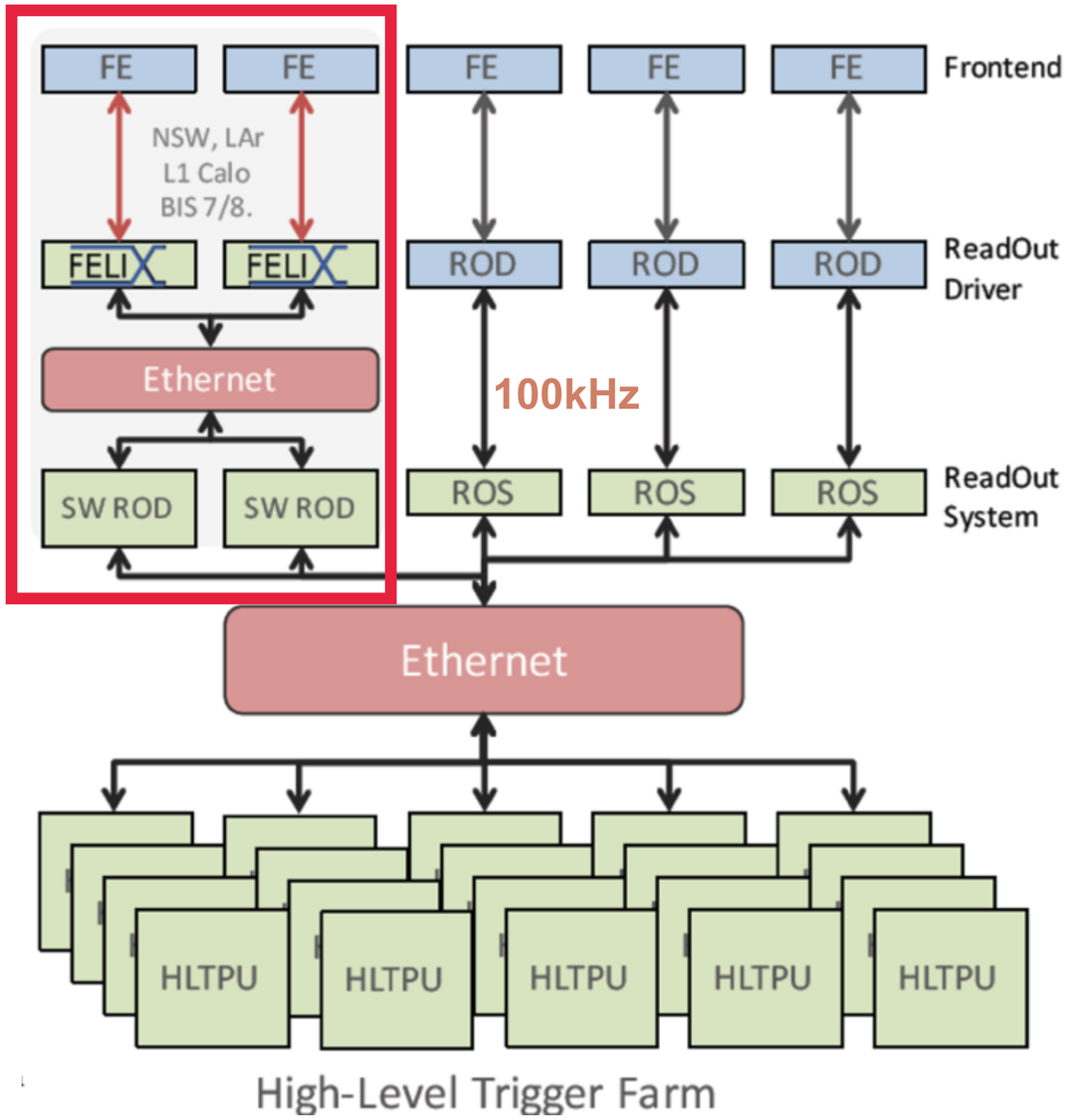
- ATLAS is a general-purpose particle detector at the Large Hadron Collider (LHC)
- LHC collides protons/ions at a 40 MHz rate
- A first level (L1) trigger, implemented in hardware, selects events at maximum rate of 100 kHz
- A “high-level” trigger, implemented in software, selects events at maximum rate of 1-3 kHz





ATLAS Readout for 2022–26

Run 3 - Now



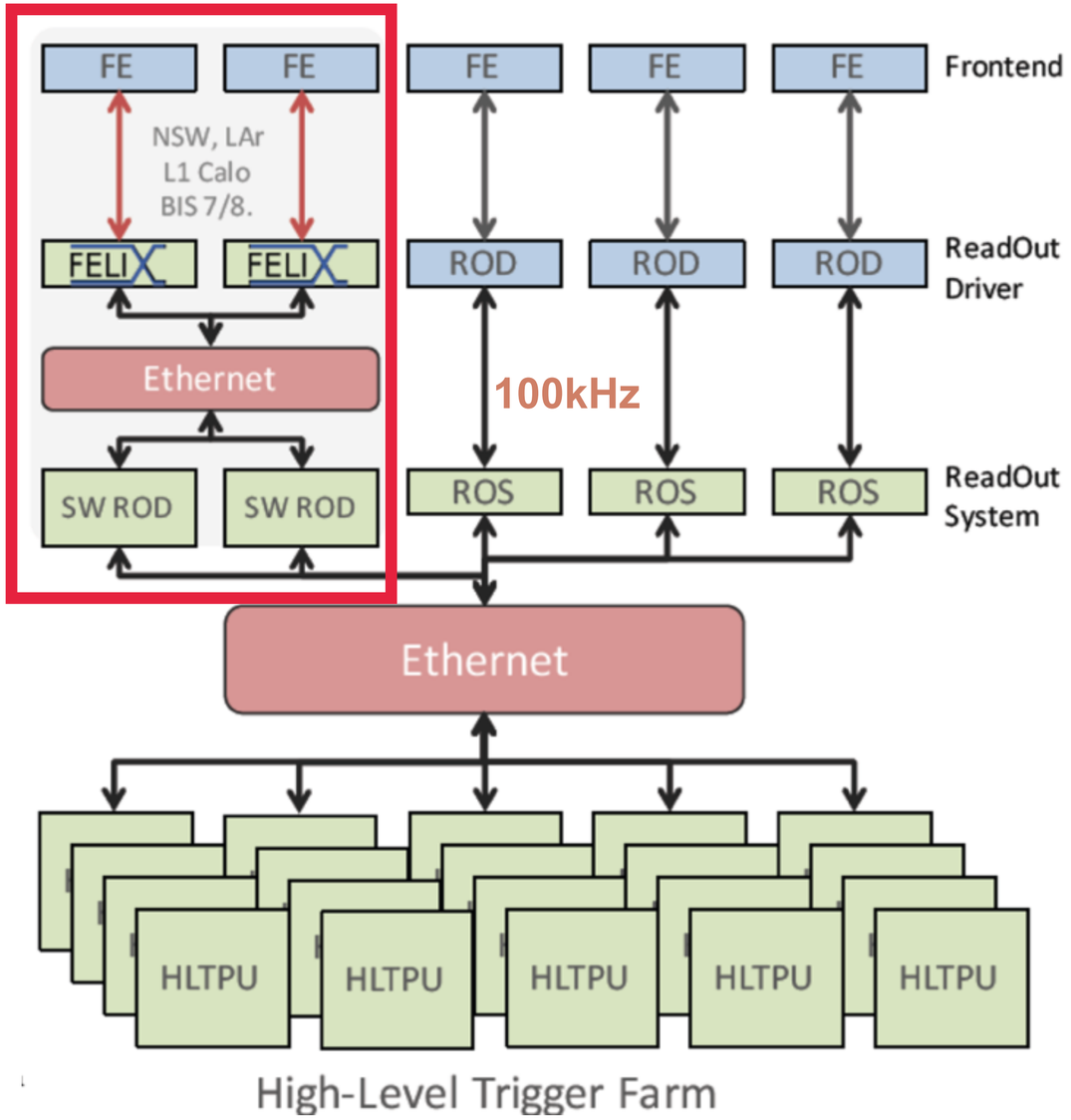
FELIX: Front-End Link EXchange

- **Custom PCIe card** hosted on commercial computer
- Interaction with FE includes readout, configuration, trigger & clock distribution, monitoring, BUSY
- Scale: ~100 cards, 60 host PCs

SW ROD: software Readout Driver

- **Software** running on commercial computer
- Builds and aggregates events, detector-specific data processing
- Scale: ~ 30 servers

Benefits of the FELIX system



- **Less** custom components
- **Less** hardware and firmware development effort
- Data transport **decoupled** from data processing
- **Industry-standard** data networks introduced earlier in the readout chain
- **Aggregation** of many links into a single high speed network link
- As a result, **less** support efforts because of common hardware platform and simplicity

The choice of a custom FPGA card

LHC Clock distribution

- During data taking the 40.079 MHZ clock signal is provided by the LHC
  - LHC clock is in sync with bunch crossing
  - All front-end and DAQ components need to be synchronised with the LHC clock
- FELIX needs an interface to the custom ATLAS TTC (Timing, Trigger and Control) system to distribute clock and L1 trigger signals to front-ends

Font-end radiation hardness

- To support data protocols used by radiation-hard front-ends
  - GBT, lpGBT ASICs and the according data protocols developed at CERN
  - TCP/IP over Ethernet not an option so far

Availability and cost of commercial solution

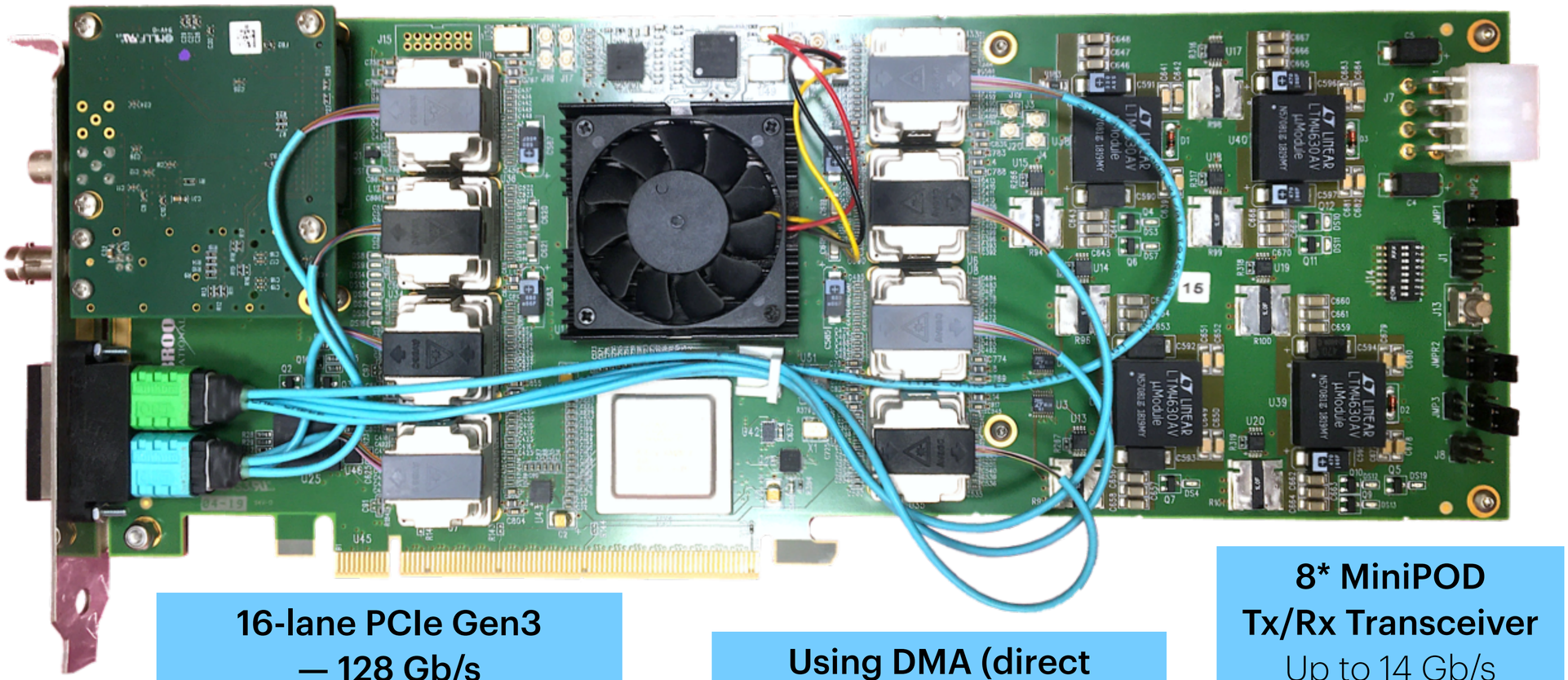
- The above constraints strongly limit the selection of commercial products

The FLX-712 card

FPGA: Xilinx Kintex UltraScale XCKU115

~300 cards produced 2020-2022

Interface to Timing Trigger and Control (TTC) systems (L1 triggers + LHC clock on a fibre) Busy output (on a LEMO)



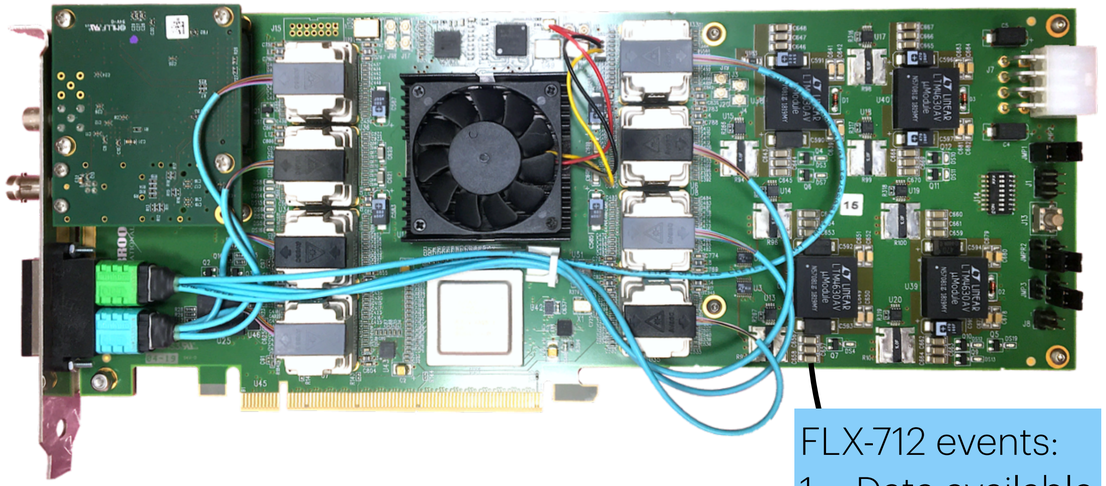
Communicate with detectors via fibres

16-lane PCIe Gen3 — 128 Gb/s

Using DMA (direct memory access) enabling data sent directly from an attached device to host server's memory

8* MiniPOD Tx/Rx Transceiver Up to 14 Gb/s

The current FELIX system



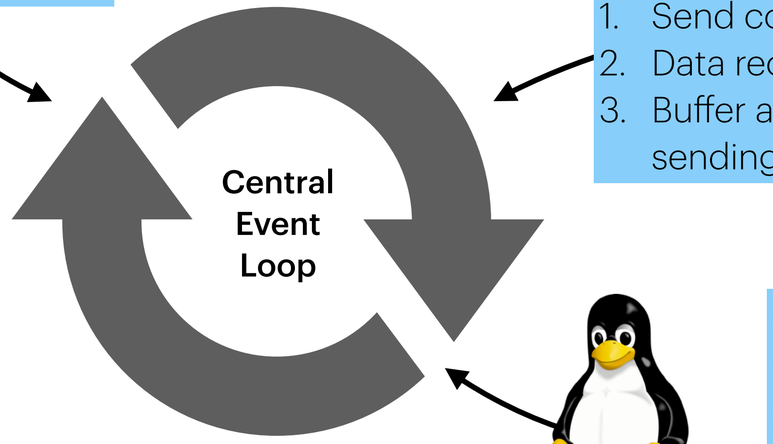
FLX-712 events:
1. Data available
2. Busy state

Firmware

- Decodes incoming data, encodes outgoing data
- Transfers data to and from a buffer in the host computer
- Come in different “flavours” corresponding to different link protocol
  - FULL: interface to other FPGA-base systems
  - GBT: interface to GBTx, a radiation hard ASIC

Software

- Transfer data over the network using RDMA technology (for low overhead transfers)
- Custom network library based on libfabric
- Runs as a daemon on FELIX servers (each hosts up to two FLX-712 cards)



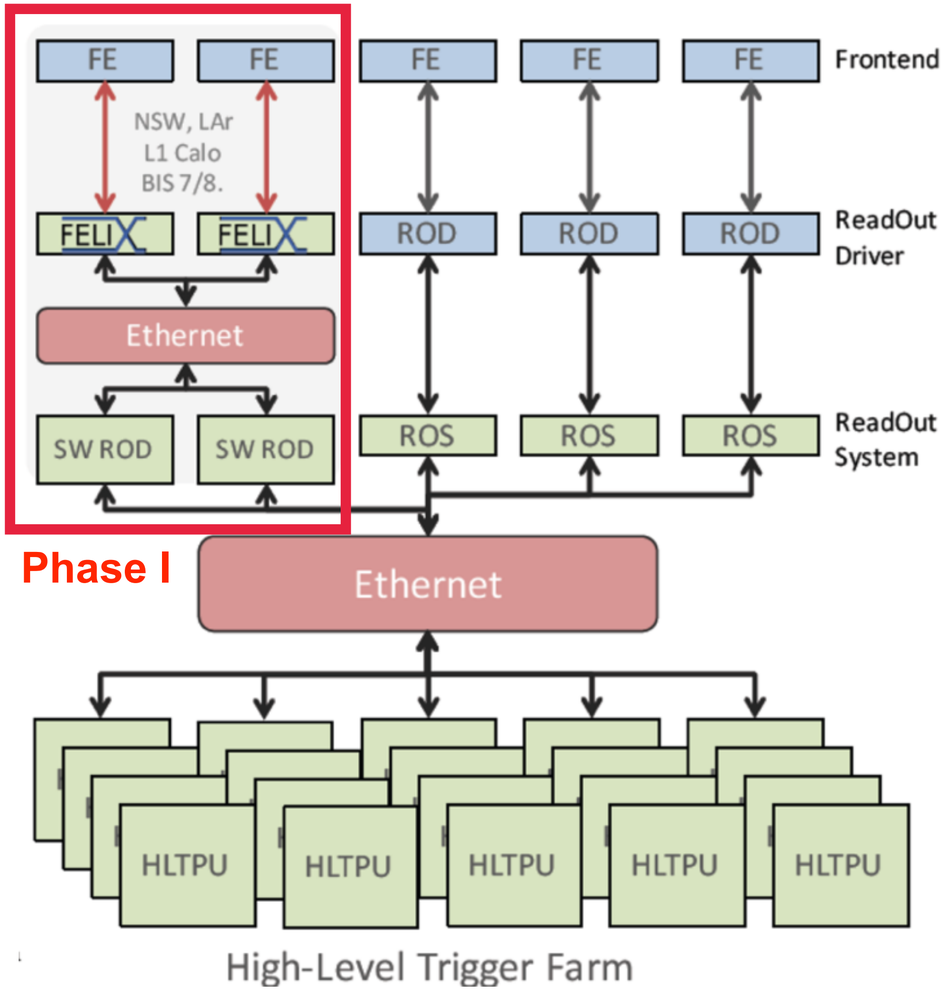
Network events:
1. Send completed
2. Data received
3. Buffer available for sending

System events:
1. Timer events
2. Signals
3. Any file descriptor event

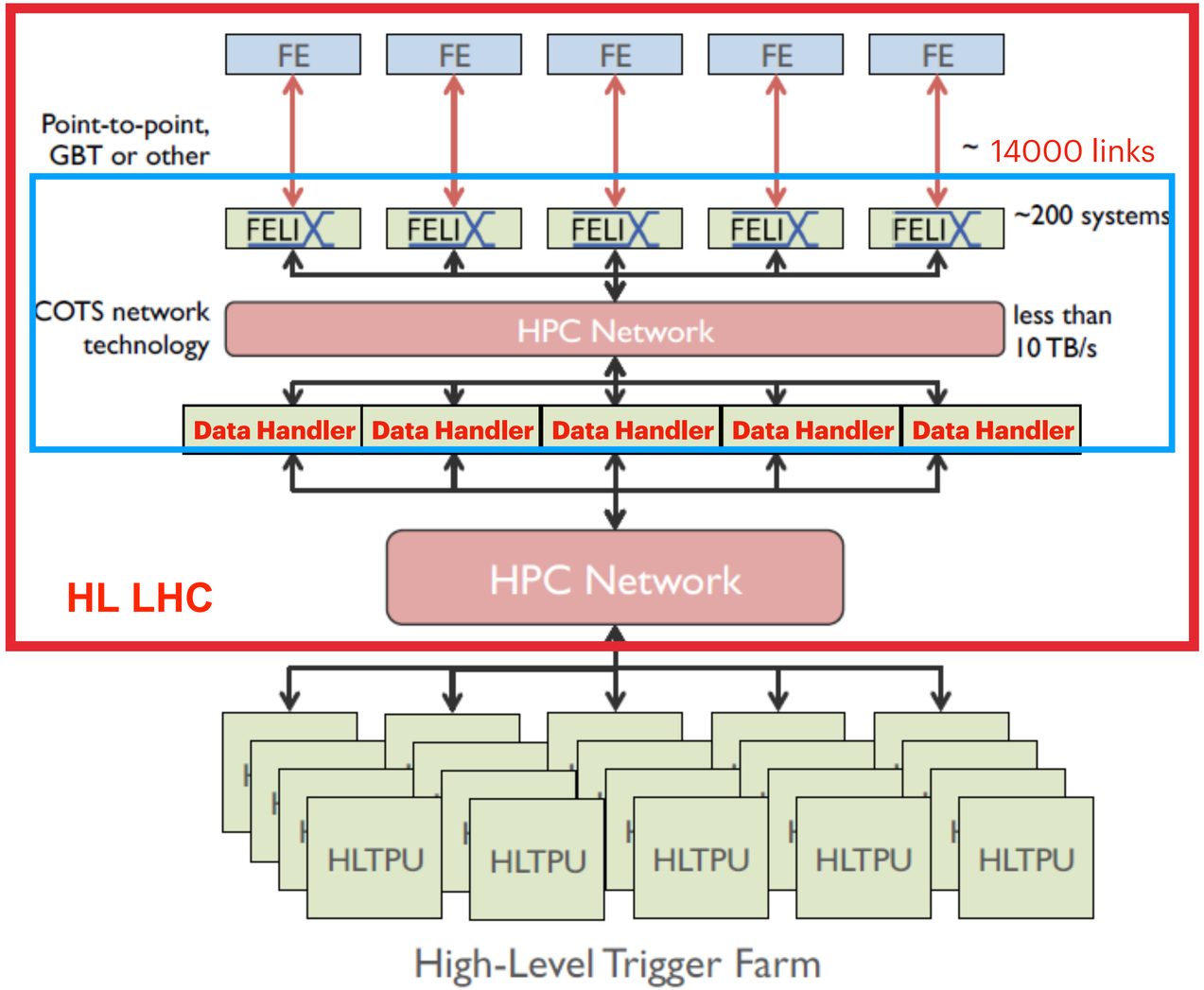
FELIX @ HL-LHC

The current FELIX development

From Phase-I to Phase-II

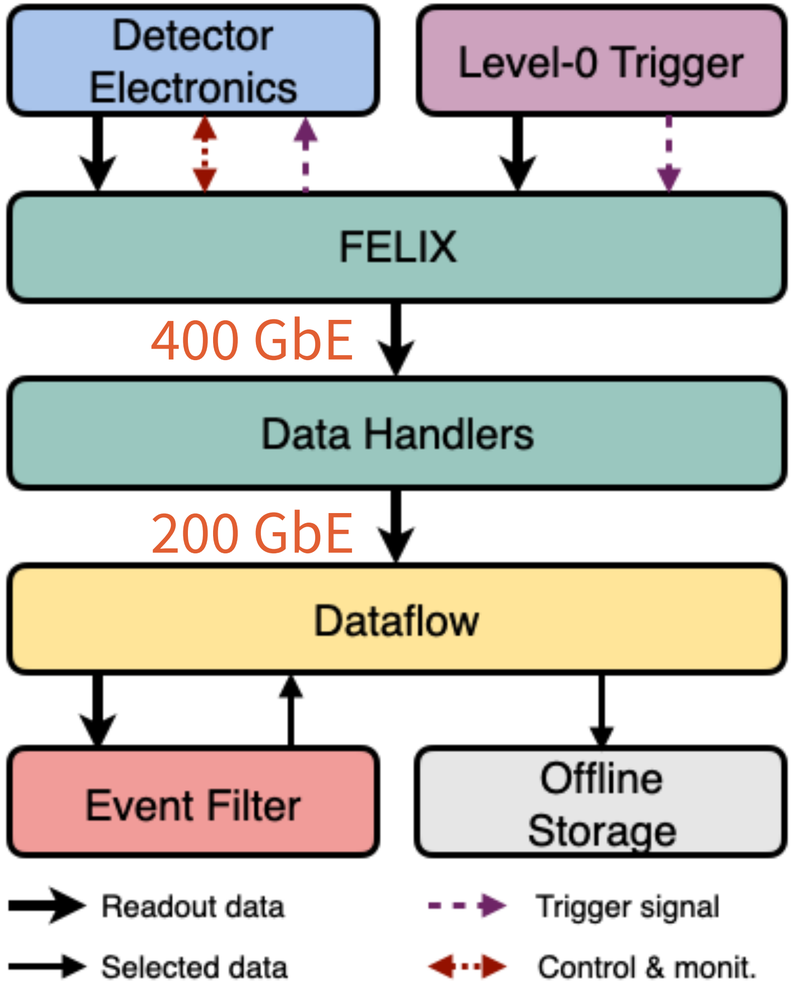


Alternative option: combine FELIX and data handler in 1 server



ATLAS Readout from 2030

Run 4 *ATLAS Phase-II Trigger see Weiming's Talk today*



Similar architecture as the current run

Run 4 conditions

- Up to **200** mean number of interactions per crossing (**~3x run3**)
- 1MHz level-0 trigger rate (**~10x run3**)
- **4.6 TB/s** total data throughput (larger event size) (**~20x run3**)

FELIX requirements

- Scale to readout of **all** sub-detectors
  - incl. all detector-specific functionalities
- Support for additional protocols
- ~14000 optical links with bandwidth up to 25 Gb/s

New FELIX hardware, firmware and software under development

The FLX-182 card – one prototype

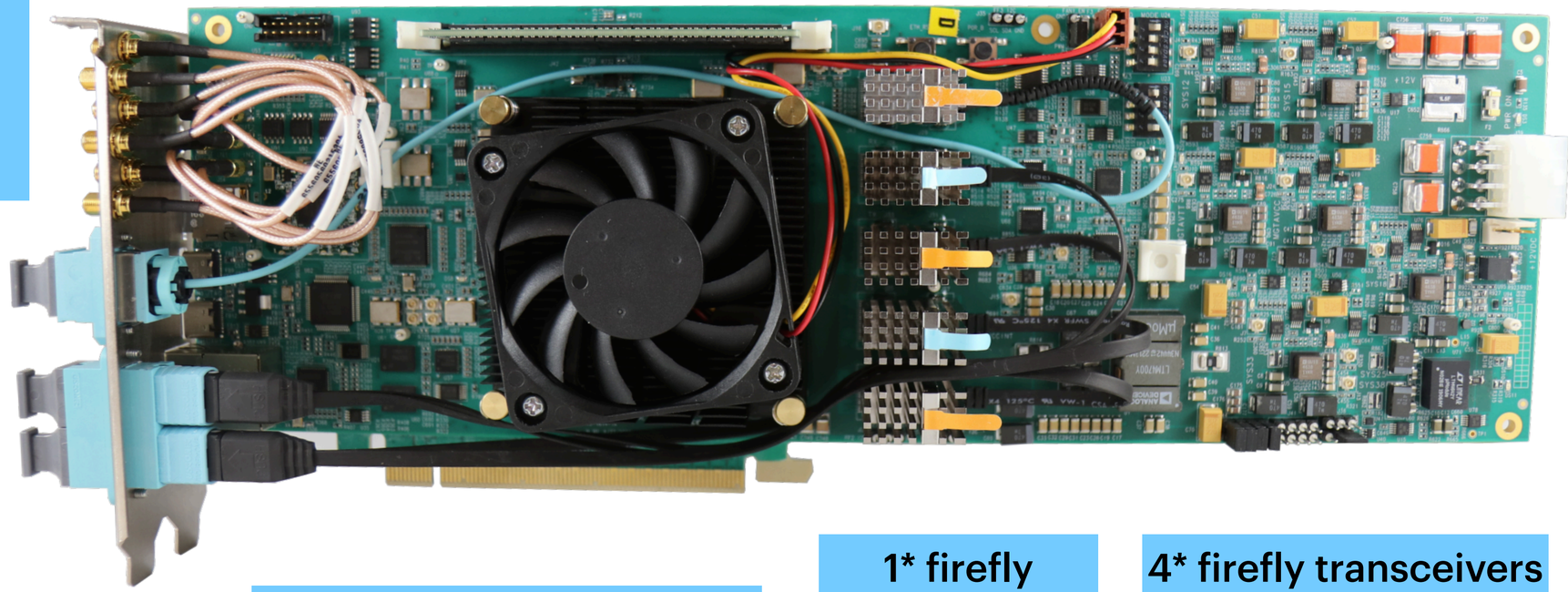
AMD Versal Prime VM1802

~46 cards produced for detector integration

Electrical interface for testing/ monitoring

TTC-LTI links

FE links



Access to the PetaLinux running on the SoC

16-lane PCIe Gen4 interface — 240 Gb/s

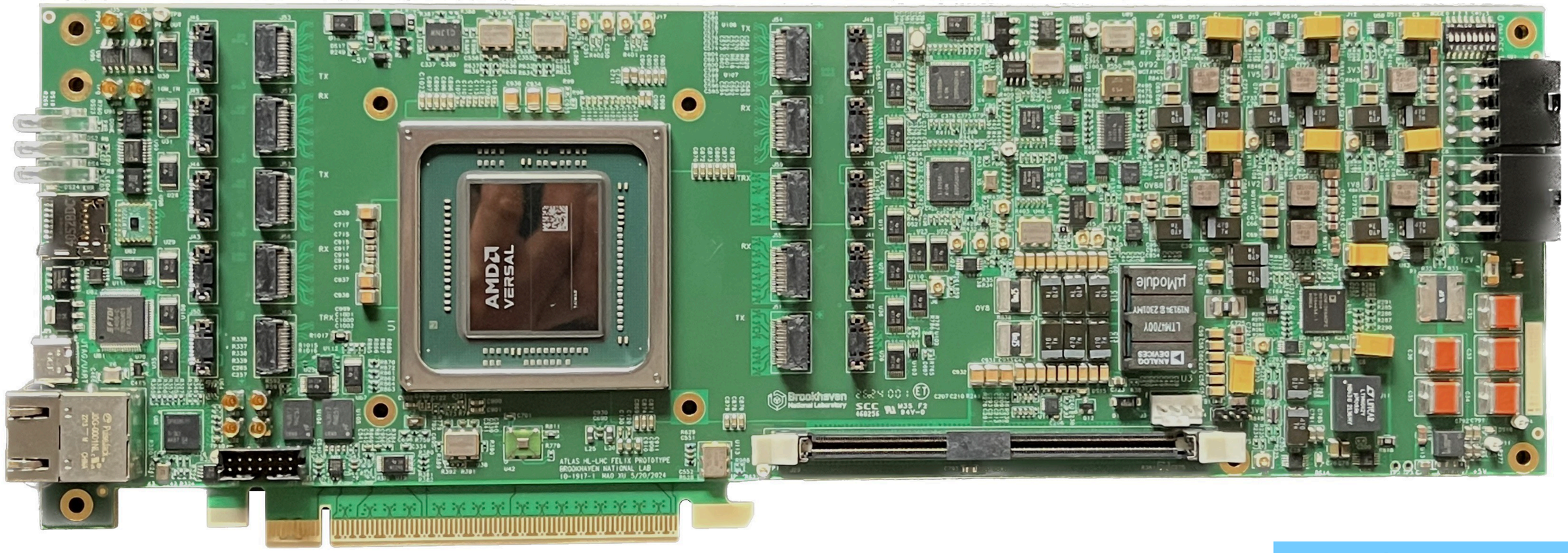
1* firefly transceiver — interface to TTC-LTI

4* firefly transceivers — support 24 bi-directional optical links (25Gb/s)

The FLX-155 card – the final design candidate

Testing stage
(1+2 cards to arrive)

AMD Versal Premium VP1552



Supports White Rabbit

16-lane PCIe Gen5 interface — 480 Gb/s

1* FireFly transceiver — for TTC-LTI

1* FireFly transceiver — for 100 GbE

8* FireFly transceivers — support 48 bi-directional optical links (25Gb/s)

Towards Phase-II

Hardware

- FLX155 is the final design candidate, FLX182 card for detector integration:
- Same interface, many common components

Firmware

- Support various data protocols incl. lpGBT and Interlaken for 25 Gbps links

Software

- Developments ongoing to scale up the current architecture to the Run-4 requirements
- Retaining RDMA technology to fully use the 400 Gbps network bandwidth

Time and clock distribution – one f/w challenge to highlight

- Detector like High Granularity Timing Detector (*HGTD*, see Mei's talk this morning) is time-sensitive, thus imposing stringent requirements on clock precision
- Clock is distributed in a long chain via several systems
- FELIX keeps the gate for clock input to the detector
- *TCLink* (by CERN High Precision Timing Distribution group) was developed for phase determinism and long-term phase stability
  - Together with *Knypaegje* (by FELIX team) FLX-182 demonstrates a spread less than 4 ps among startup phases

See M. Leguijt's poster in TWEPP2024

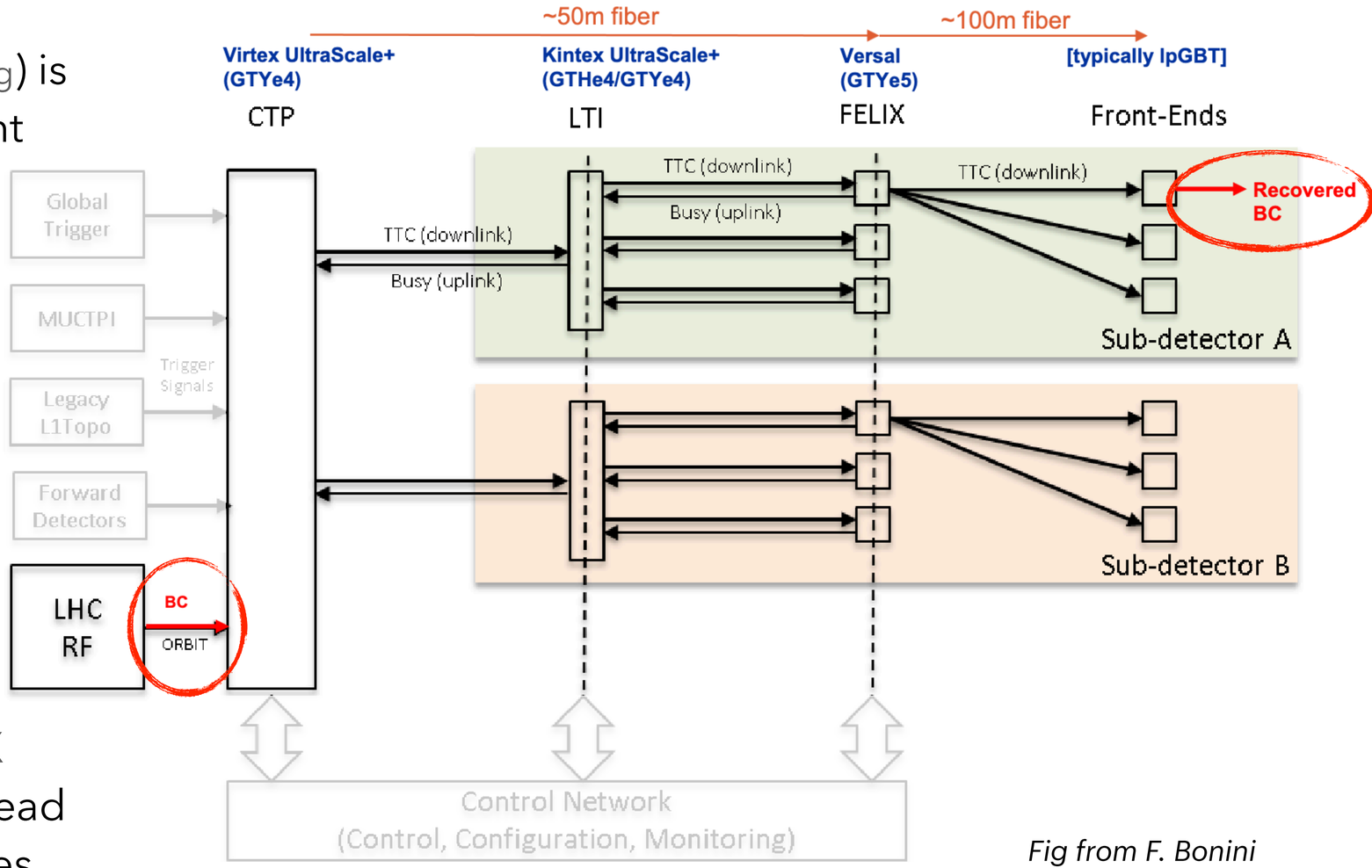


Fig from F. Bonini

Fragment Building – one s/w challenge to highlight

- Data packets from O(100) E–Links must be aggregated into one fragment
  - Total packet rate O(100)MHz
- Requires extensive and efficient use of multi-threading to utilise full power of modern CPUs

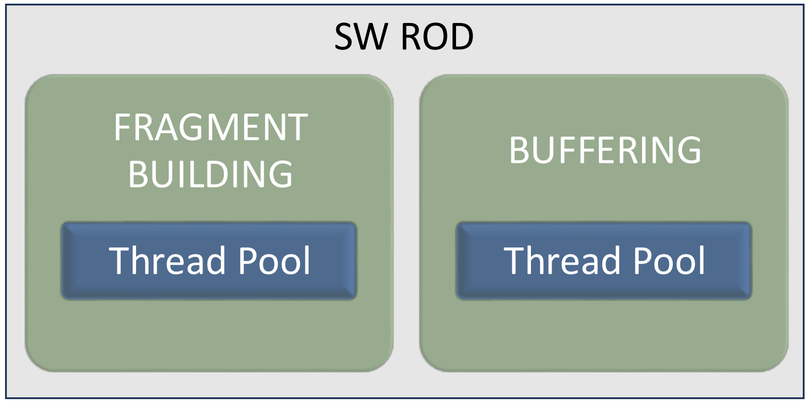


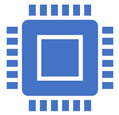
Fig from S. Kolos

SW/HW environment

Network latency is O(1) us

Worst case OS scheduler latency is O(1) ms

Can only be measured and accepted



Code quality

CPU cache-friendly

Efficient multi-threading

Scalable to many CPU cores

Under control of the SW developers

OS services

Memory management

Must be taken under control

- Memory management was the main issue that affected performance and maximum latency
- Using custom Memory Pool (single open source header) implementation, the issues have been successfully addressed

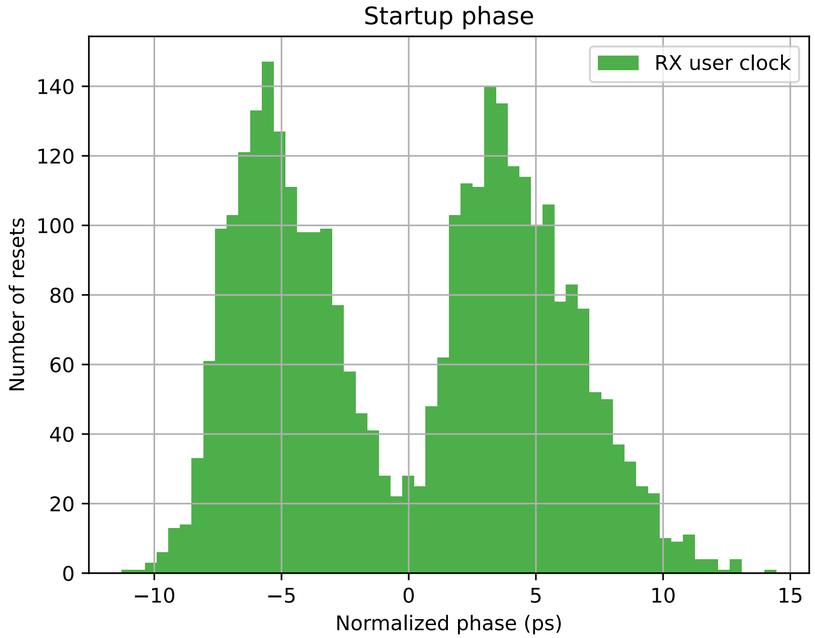
Summary

- FELIX is a versatile data acquisition platform
  - Particularly useful for experiments readout with GBT, lpGBT or similar radiation-hard technology
- The first FELIX implementation used in production
  - Successful data taking with protoDUNE-SP
  - Stable readout in ongoing ATLAS data taking
  - NA62 and sPhenix take data with FELIX
  - EIC considering FELIX for all the systems
- An evolution of FELIX for the HL-LHC phase of ATLAS is under development
  - Final design review expected in the next spring

Backup

FELIX Phase determinism

- Clock recovered by the GTYe5 transceivers on the Versal chip, shows a phase difference w.r.t. the reference clock ➔ double peak
- Knypaegje is a program that runs on the ARM processor using the correlation between eye opening and the startup phase
- Developed on FLX-182 - for proof of concept
- Final implementation is under development



knypeagje