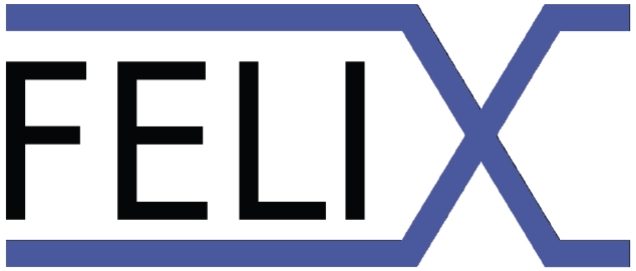


Front**E**nd **LI**nk e**X**change

FELIX
the new detector readout system for the ATLAS experiment

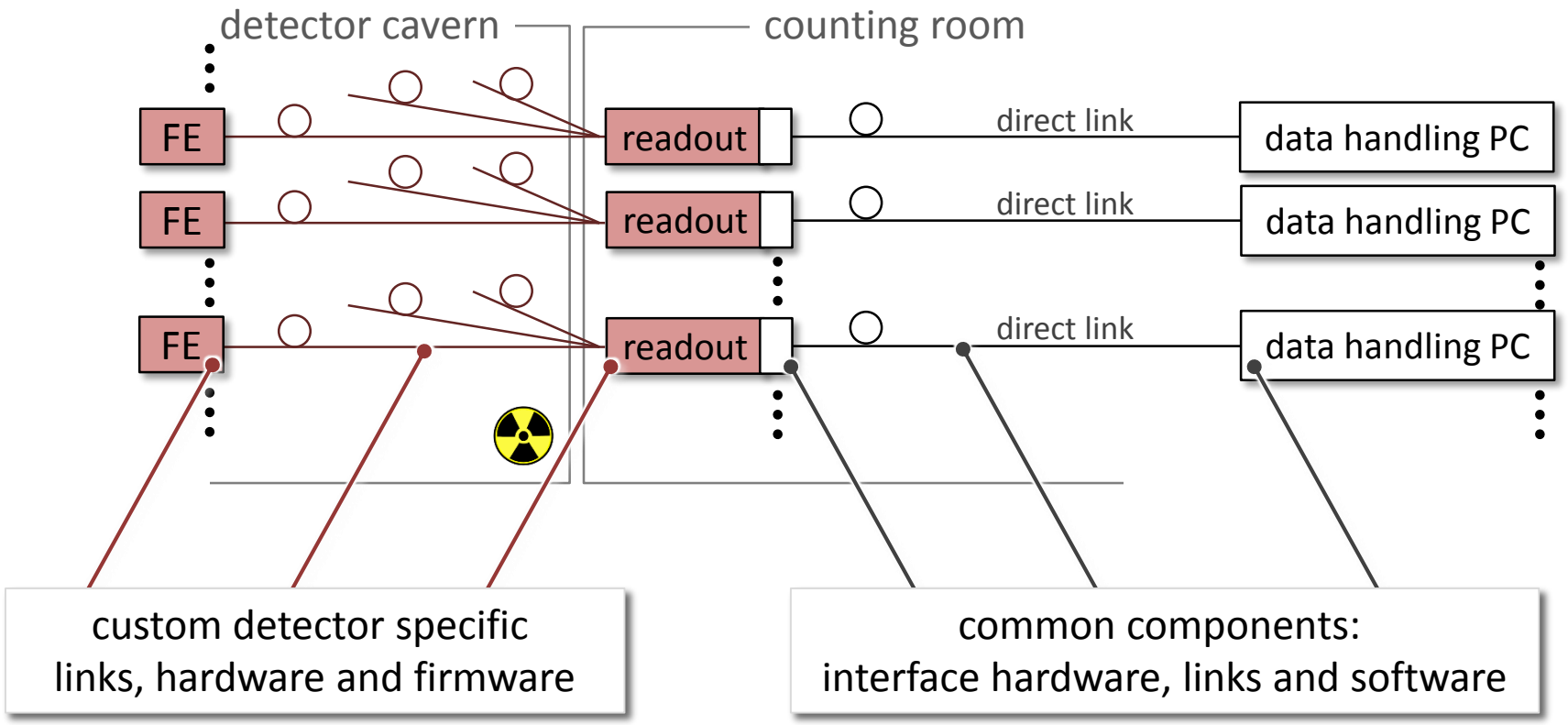
Julia Narevicius

Weizmann Institute of Science

on behalf of the ATLAS Collaboration

Introduction to ATLAS readout: today

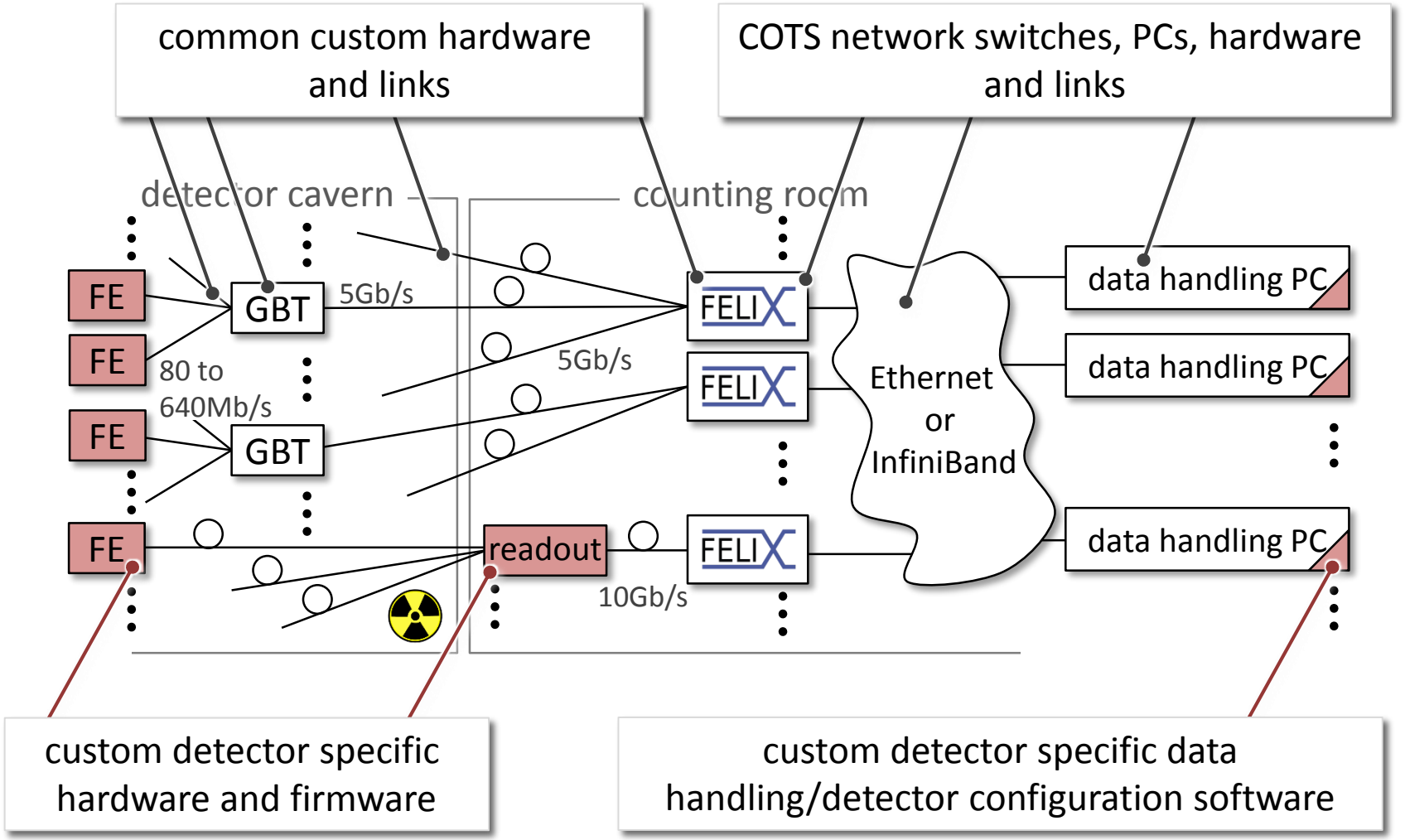
- Custom detector specific hardware and firmware solutions are used for detector readout, configuration, calibration, trigger and control
- Structure is rigid, hard to maintain and upgrade



FE = detector Front End readout electronics

Introduction to ATLAS readout: upgrade

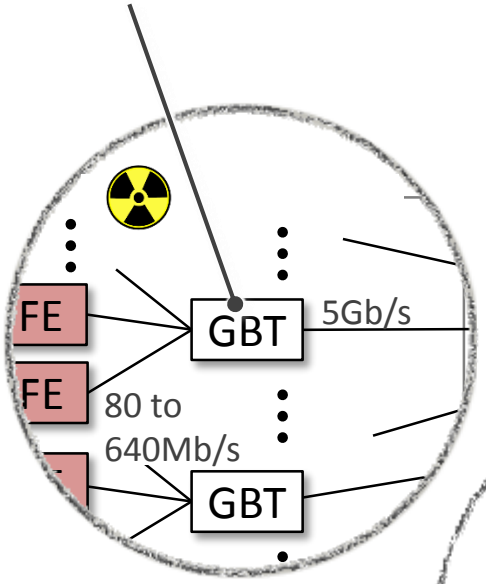
- Installation for few detectors starts in 2019
- Installation for all ATLAS detectors starts in 2024



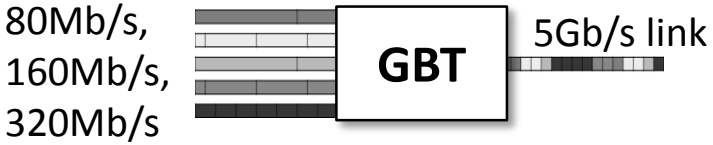
FE = detector Front End readout electronics • *GBT = Gigabit Transceiver*

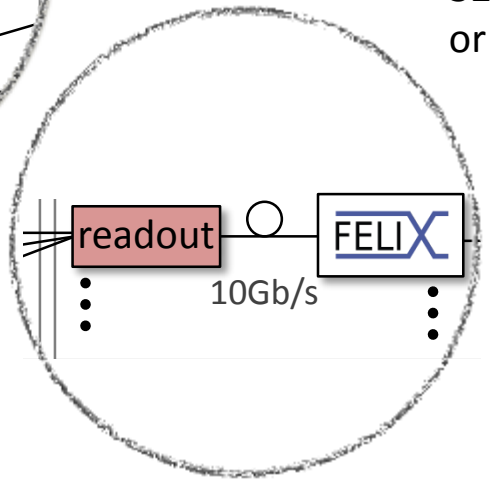
FELIX FE side interfaces

- CERN **GBT** ASIC: Gigabit Transceiver radiation hard chipset
- GBT is common custom detector independent hardware



1. A GBTx ASIC aggregates several slow "E-links", from Front-End ASICs into a 5Gb/s optical link (3.2Gb/s when using forward error correction):



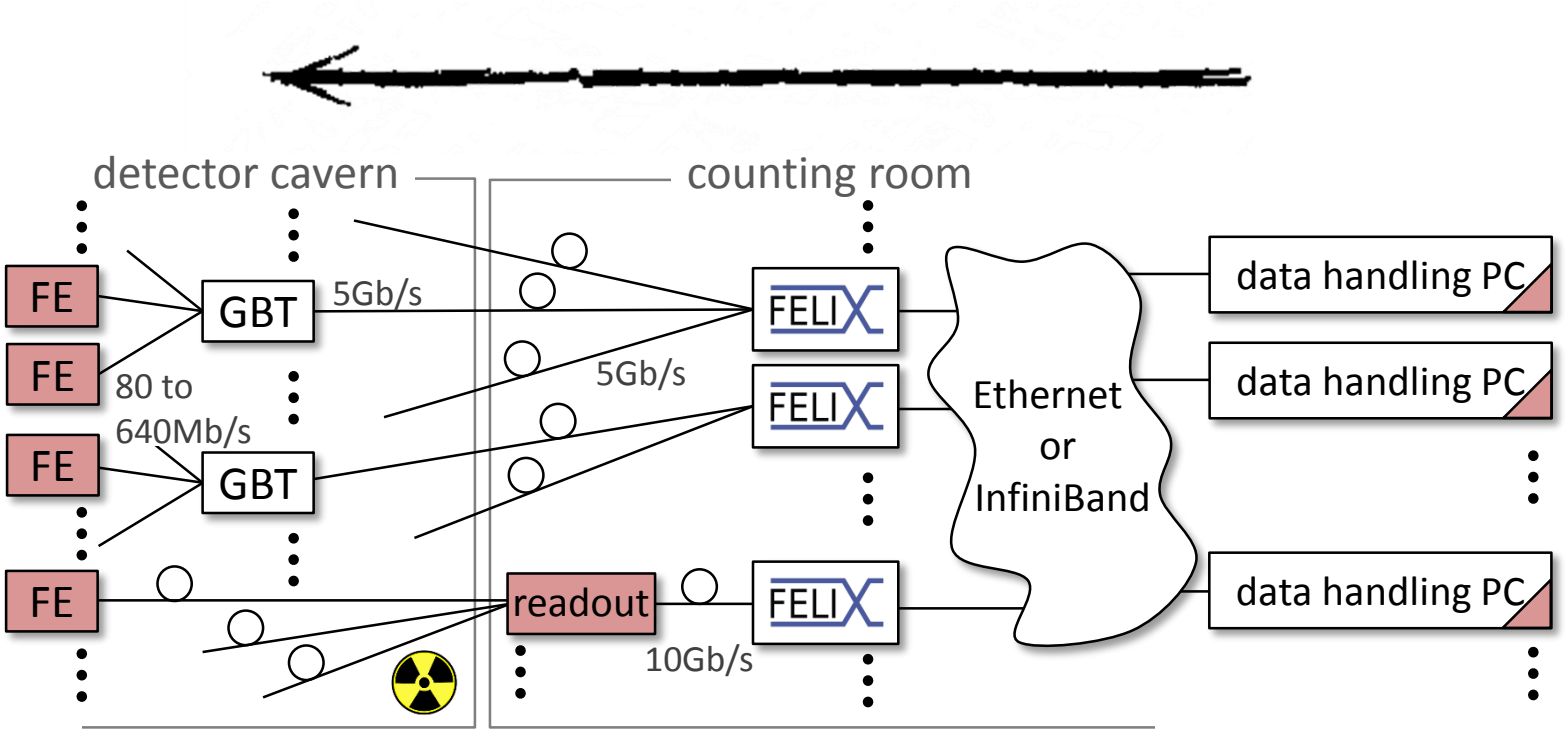


2. Direct FPGA serial transceiver to FELIX high speed links, >10 Gb/s

- *GBT = Gigabit Transceiver*
- *FE = detector Front End readout electronics*

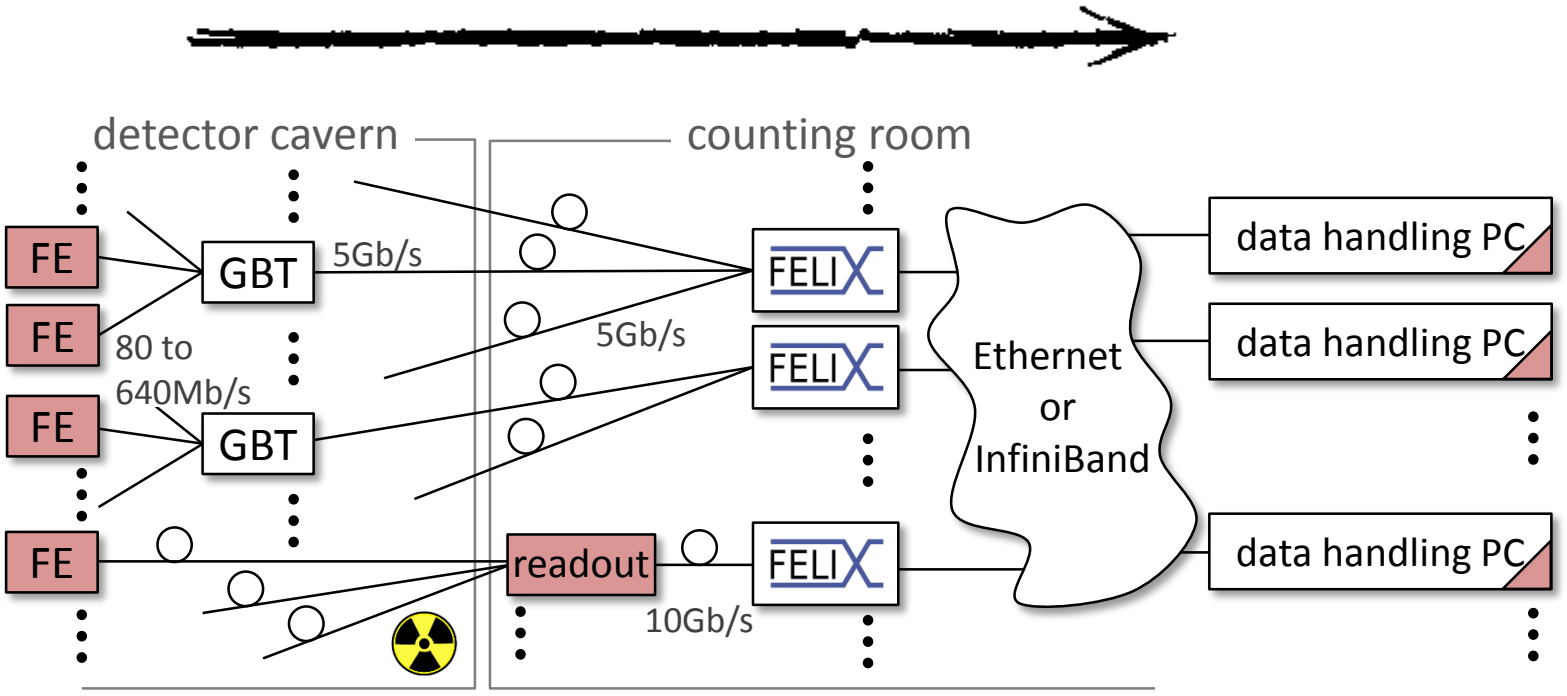
FELIX functions: to FE direction

- Routing of detector control, configuration and calibration
- Configuration of GBTs and slow control signals
- Supported data stream encoding: 8B/10B and HDLC
- Distribution of Timing Trigger and Control (TTC) information and LHC clock with low and fixed latency



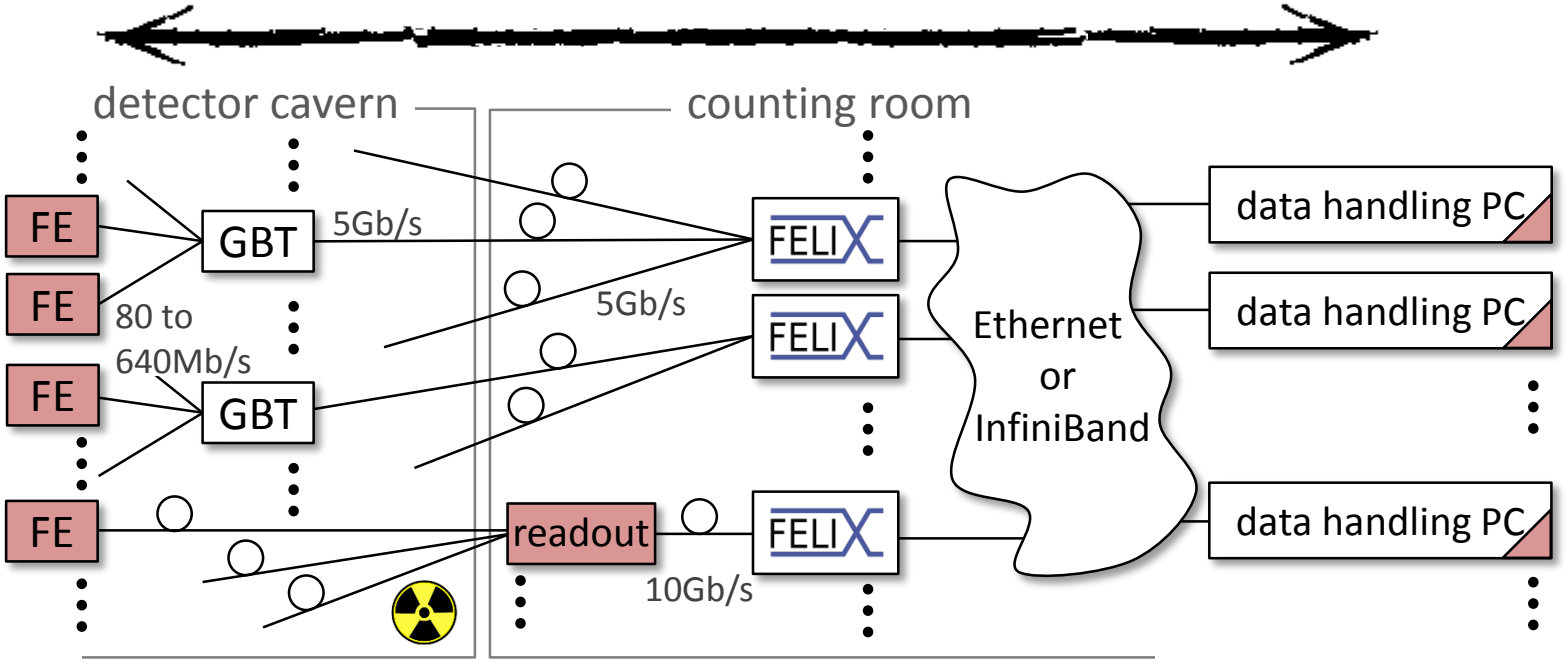
FELIX functions: to Network direction

- Routing of event data and detector monitor
- Supported data stream decoding: 8B/10B and HDLC
- Distribution of Timing Trigger and Control (TTC) information
- Unlimited number of network endpoints can subscribe to a data stream of interest
- Supported Front End 'busy' status handling



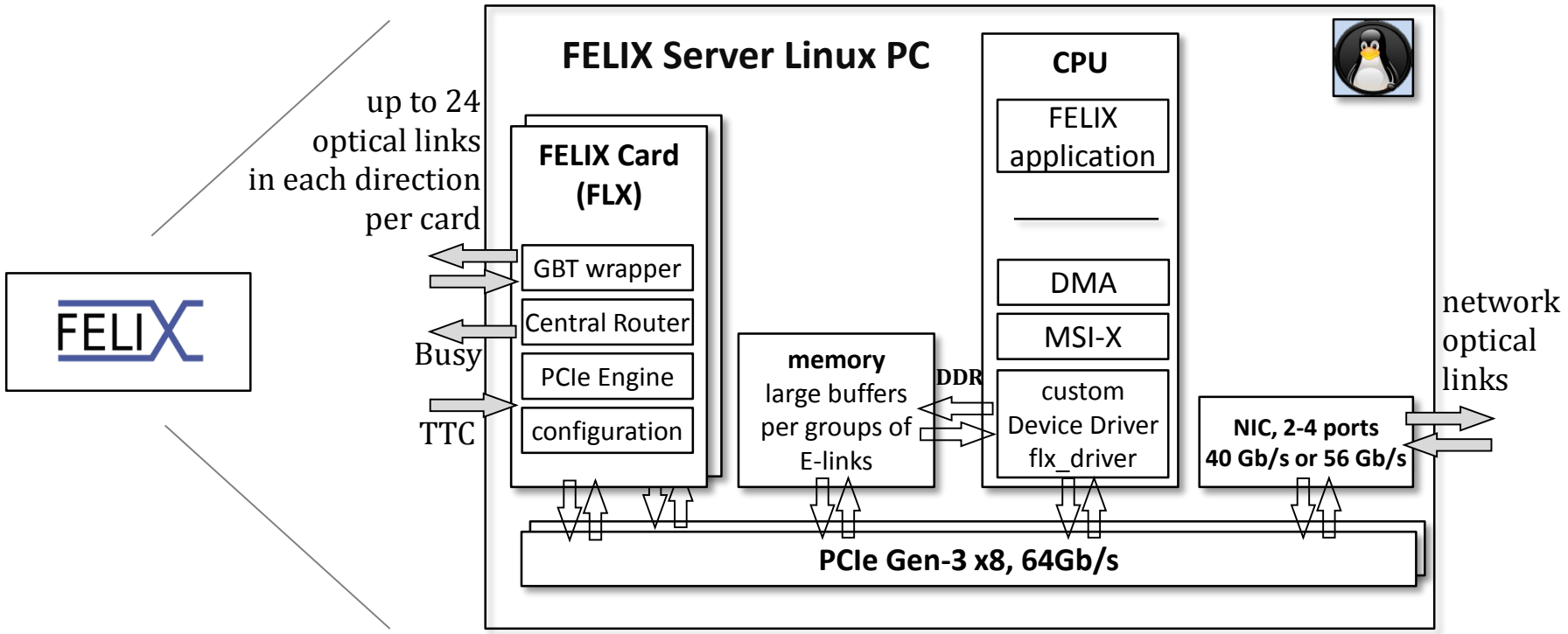
FELIX features summary

- Interfaces FE links to a high bandwidth industry standard network
- Separates GBT technology into a standard, fixed, but configurable, building block
- FELIX routes logical data flows to/from different and multiple off-detector endpoints
- Enables flexible mapping of FEs on a network
- Reduces the diversity of custom HW in ATLAS in favour of COTS HW and SW



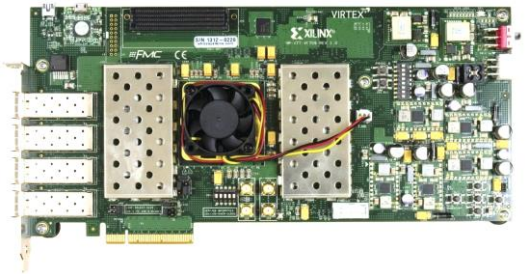
FELIX system implementation

- Server Linux PC
- Up to two PCIe interface cards with Xilinx Ultrascale FPGA, depending on bandwidth needed (2 PCIe slots Gen3 x 8 lanes)
- NIC, 40 or 100 Gb/s Ethernet interfacing or InfiniBand



FELIX system today: PCIe FPGA board HW

FLX-709 (MiniFELIX)



- Subset of the full FELIX functionality, intended for FE development support
- Xilinx VC-709: Virtex-7 X690T
- PCIe Gen 3 x 8 lanes
- 4 SFP+ connectors, card comes with optical transceivers
- FMC connector

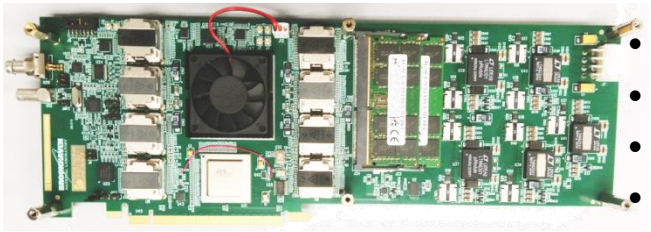
FLX-710 (FELIX)

- HiTech Global HTG-710: development platform
- Virtex-7 X690T, PCIe Gen 3 x 8 lanes
- 2x12 bidir CXP connectors
- FMC connector

TTCfx

- Custom FMC accepting Timing, Trigger & Control (TTC) optical input, used for FLX-709 and FLX-710
- Outputs: TTC clock and CH A-B info, BUSY to Lemo
- clock/data recovery: ADN2814 and jitter cleaner, TTCfxV3: Si5345

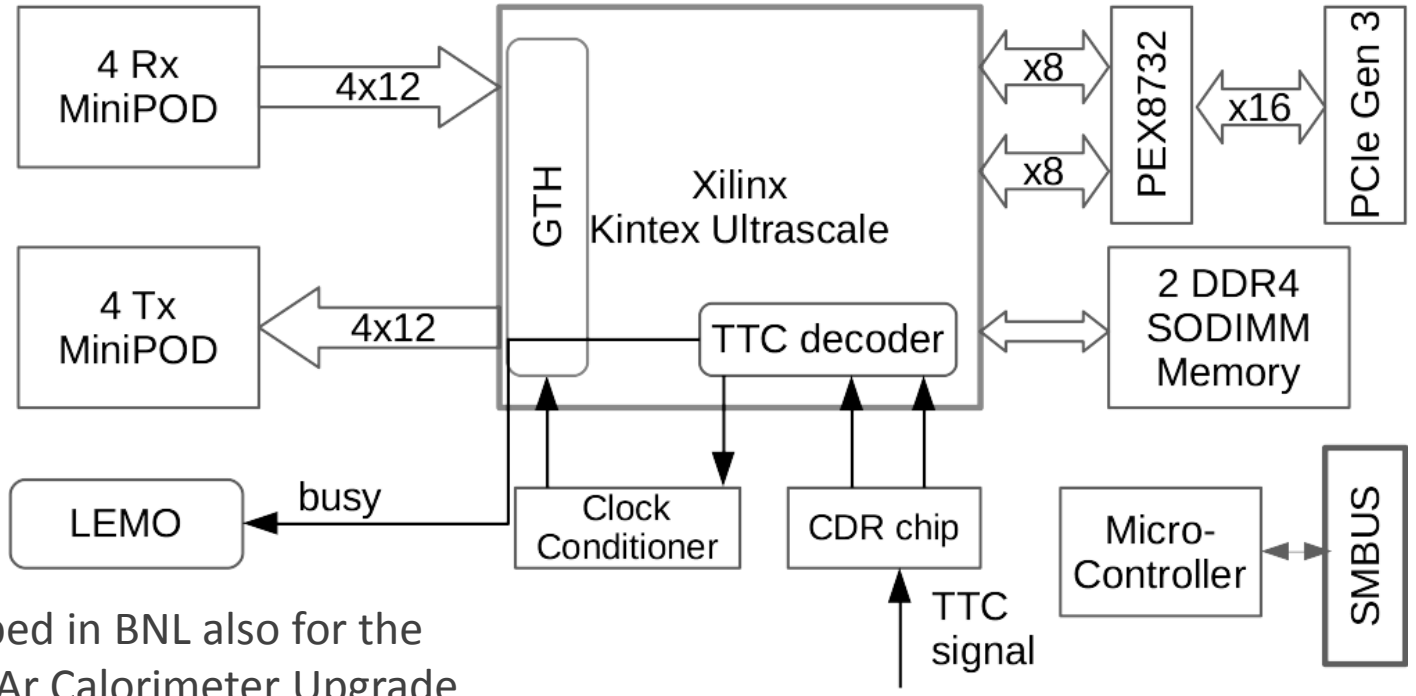
BNL-711 (FELIX candidate)



- Kintex Ultrascale XCKU115-2FLVF1924 with 4,320 18 Kb block RAM
- PLXtech PEX8732 to handle PCIe Gen3 x16 lanes interface to host
- 48-ch MiniPOD TX & RX, up to 14Gb/s per link
- Integrated Timing, Trigger & Control input and busy output
- Microcontroller for simple FPGA remote reconfiguration

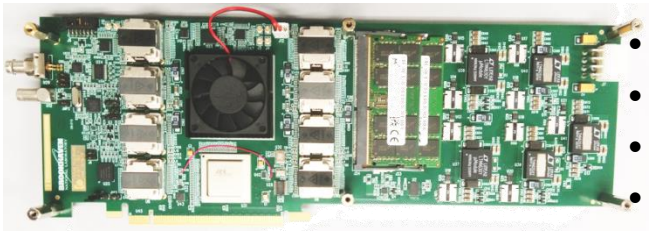
cont. FELIX system today: PCIe FPGA board HW

FELIX prototype: custom PCIe FPGA board gen3 x16, “BNL-711”



Developed in BNL also for the ATLAS LAr Calorimeter Upgrade

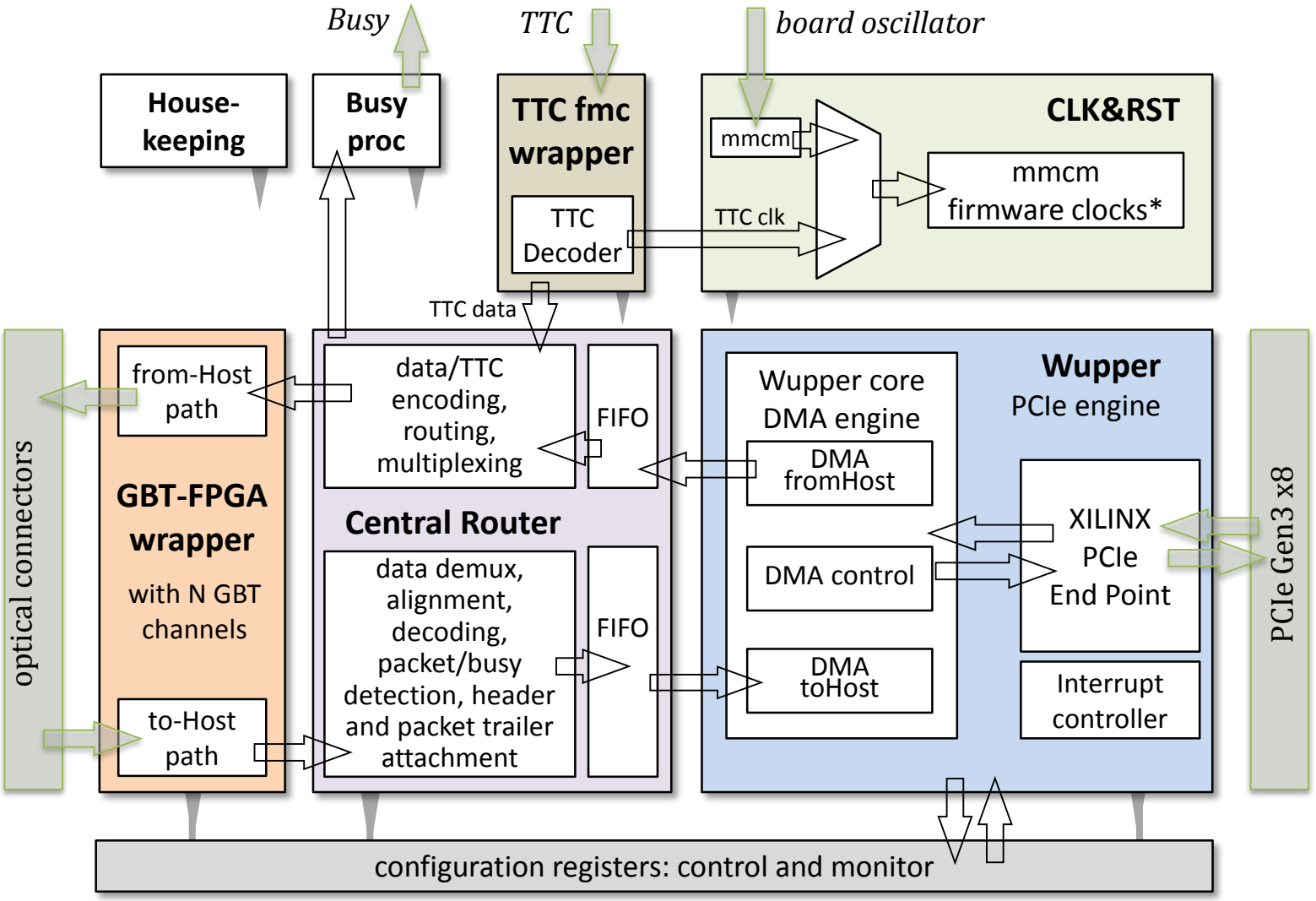
BNL-711 (FELIX candidate)



- Kintex Ultrascale XCKU115-2FLVF1924 with 4,320 18 Kb block RAM
- PLXtech PEX8732 to handle PCIe Gen3 x16 lanes interface to host
- 48-ch MiniPOD TX & RX, up to 14Gb/s per link
- Integrated TTC interface and busy output
- Microcontroller for simple FPGA remote reconfiguration

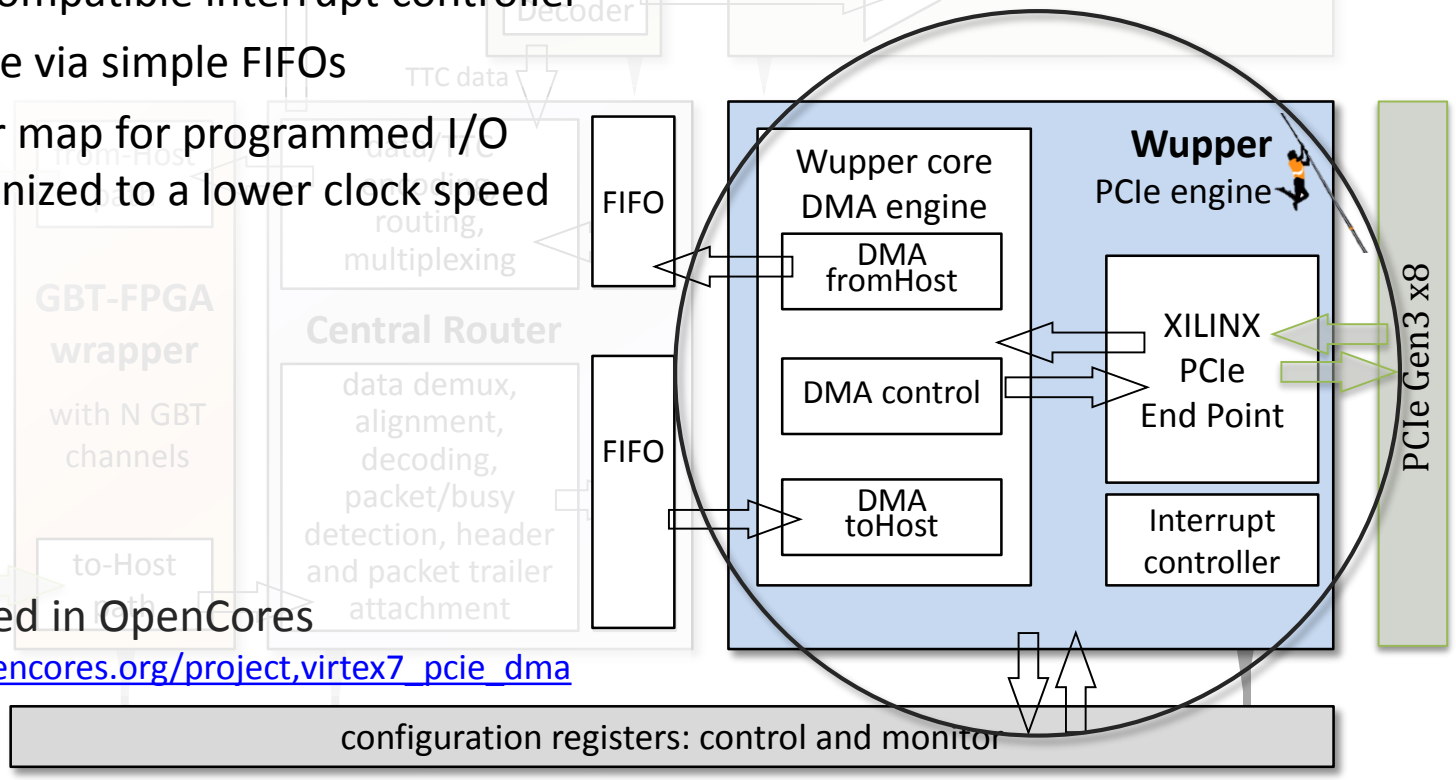
FELIX system today: PCIe FPGA board FW

Firmware Implementation diagram



cont. FELIX system today: PCIe FPGA board FW

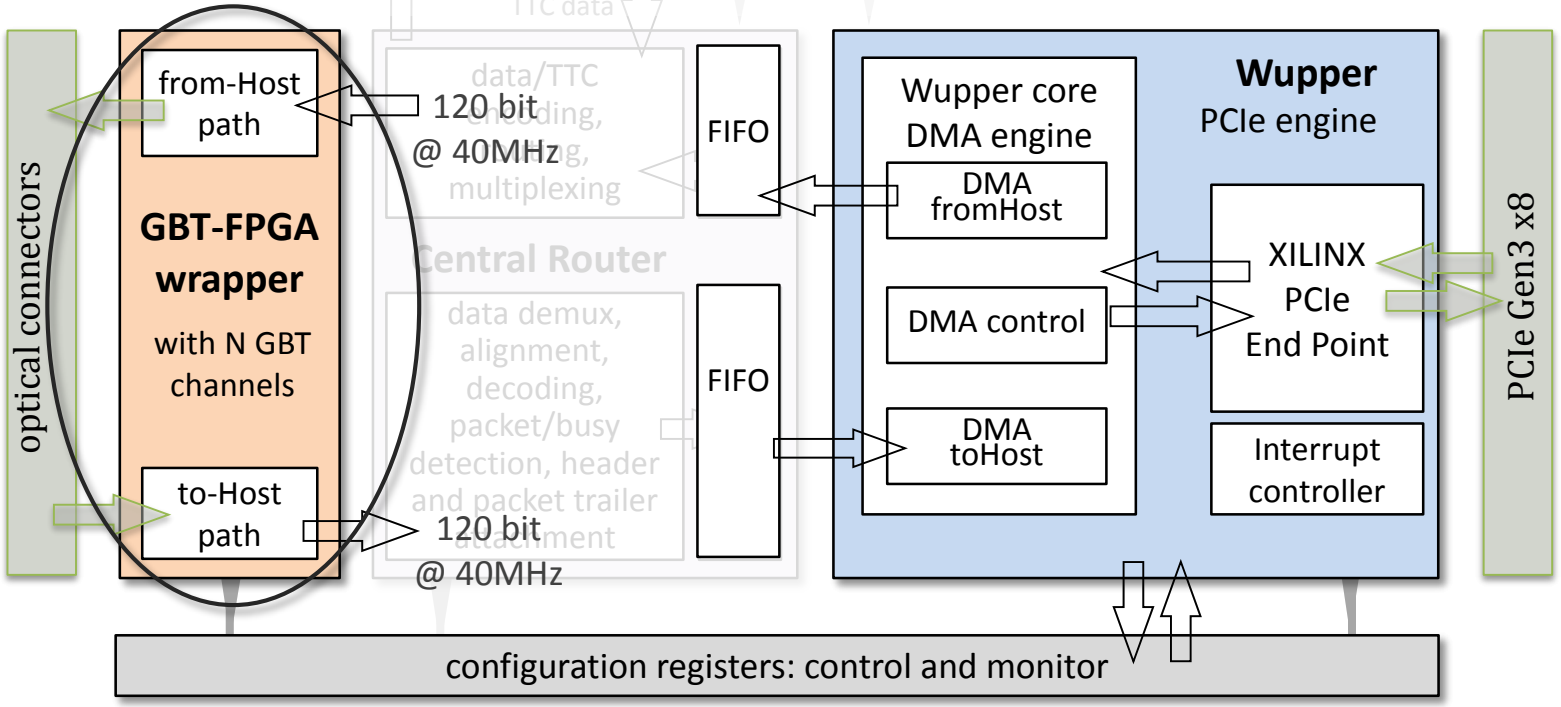
- Wupper PCIe Engine with DMA interface to the Xilinx Virtex-7 or Ultrascale PCIe Gen3 Integrated Block for PCI Express (Xilinx PG023)
- Xilinx AXI (ARM AMBA) Stream Interface (Xilinx UG761)
- Support for two core instances in the same FPGA (2 Engines Gen3 x8 lanes)
- DMA runs continuously with circular buffers
- MSI-X compatible interrupt controller
- Interface via simple FIFOs
- Register map for programmed I/O synchronized to a lower clock speed



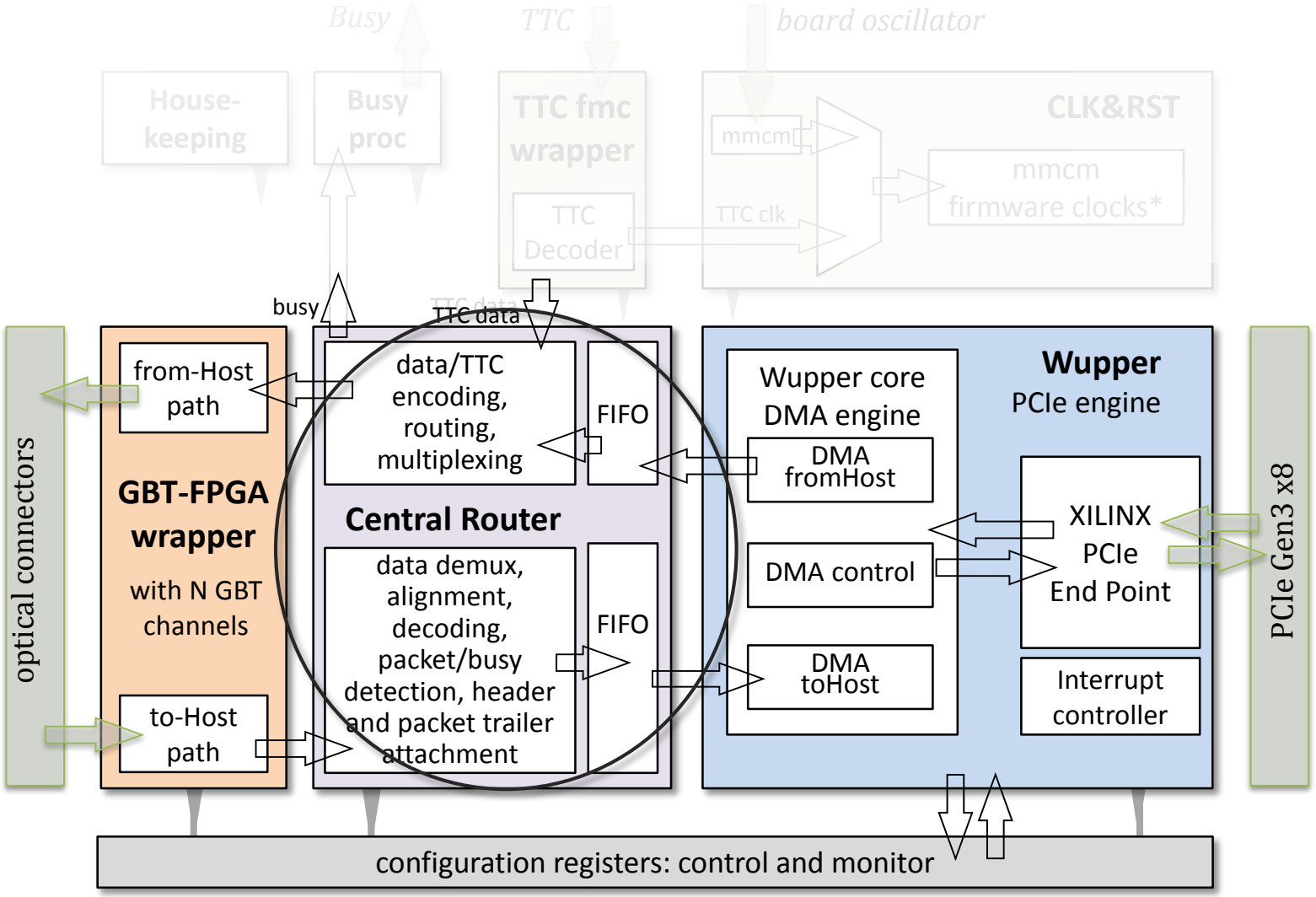
- Published in OpenCores
  http://opencores.org/project,virtex7_pcie_dma

cont. FELIX system today: PCIe FPGA board FW

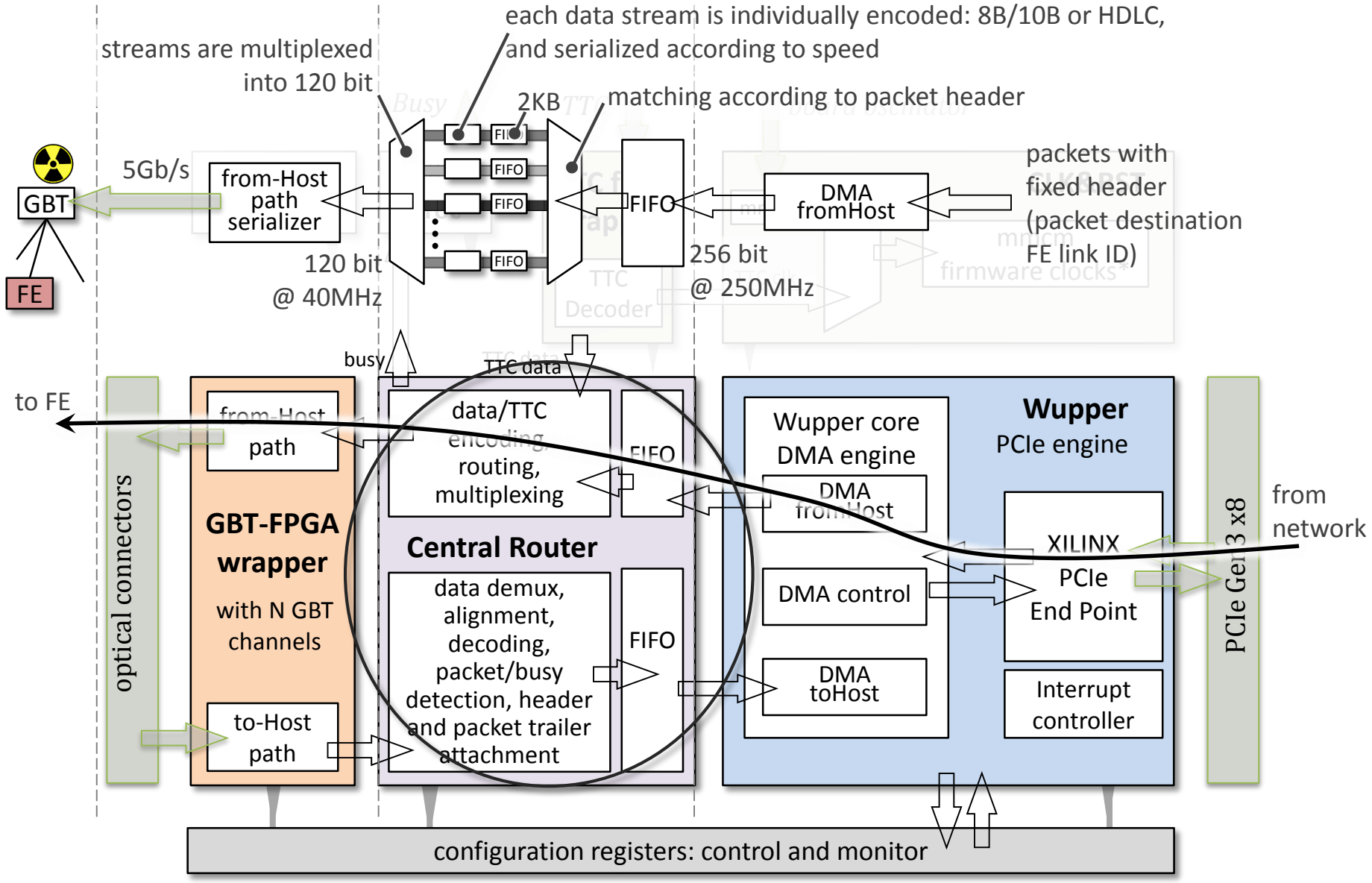
- Based on CERN GBT-FPGA implementation of the GBT protocol
- GBT-FPGA wrapper block is FPGA serial transceiver independent (Xilinx GTH, GTX, etc.)
- FPGA serial transceivers configured to have fixed and low latency
- Scalable number of channels in multiples of 4 (transceiver quads)
- Input / Output interfaces: 120-bit registers clocked at 40 MHz
- Configurable and monitored via registers



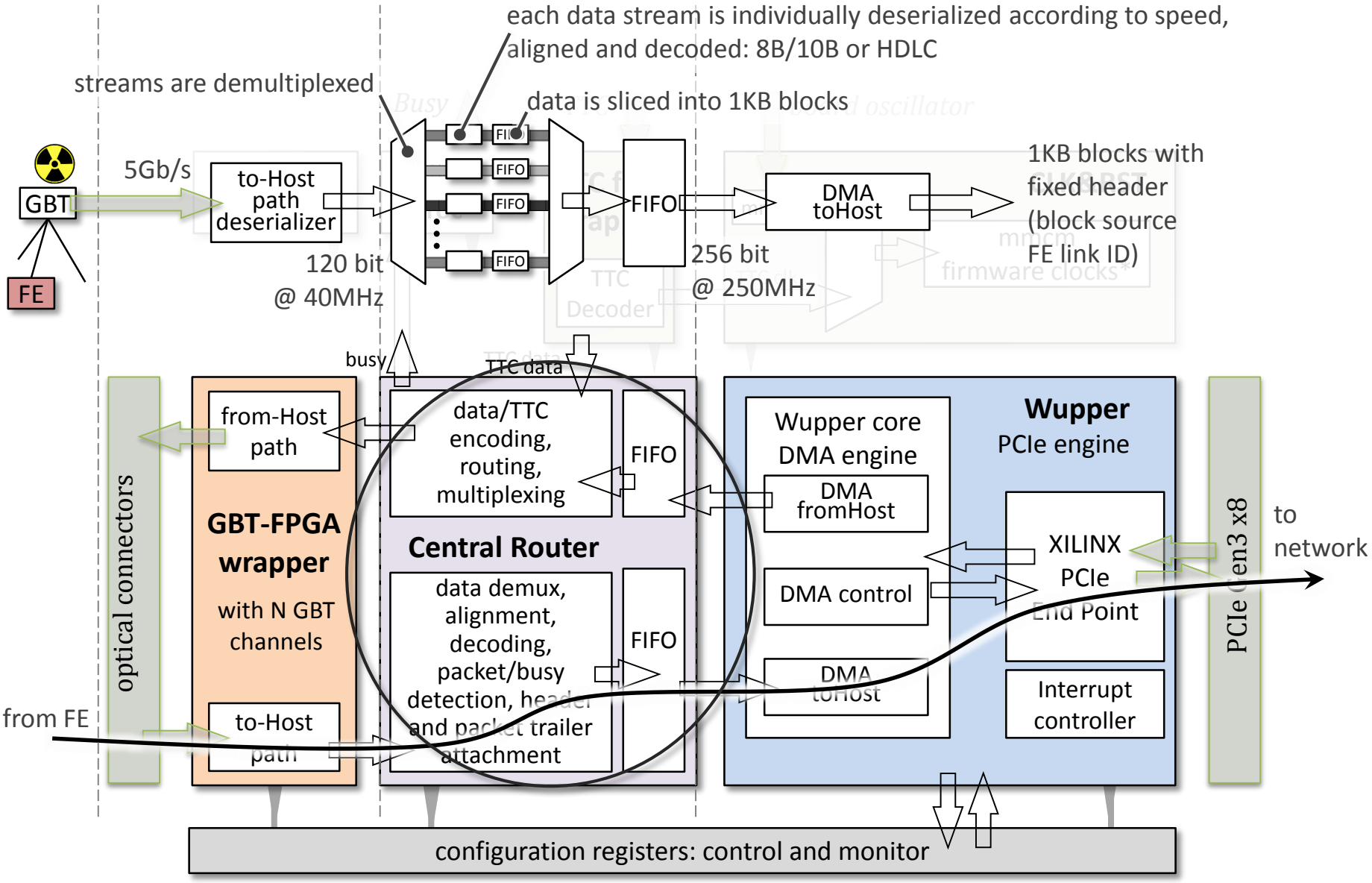
cont. FELIX system today: PCIe FPGA board FW



cont. FELIX system today: PCIe FPGA board FW

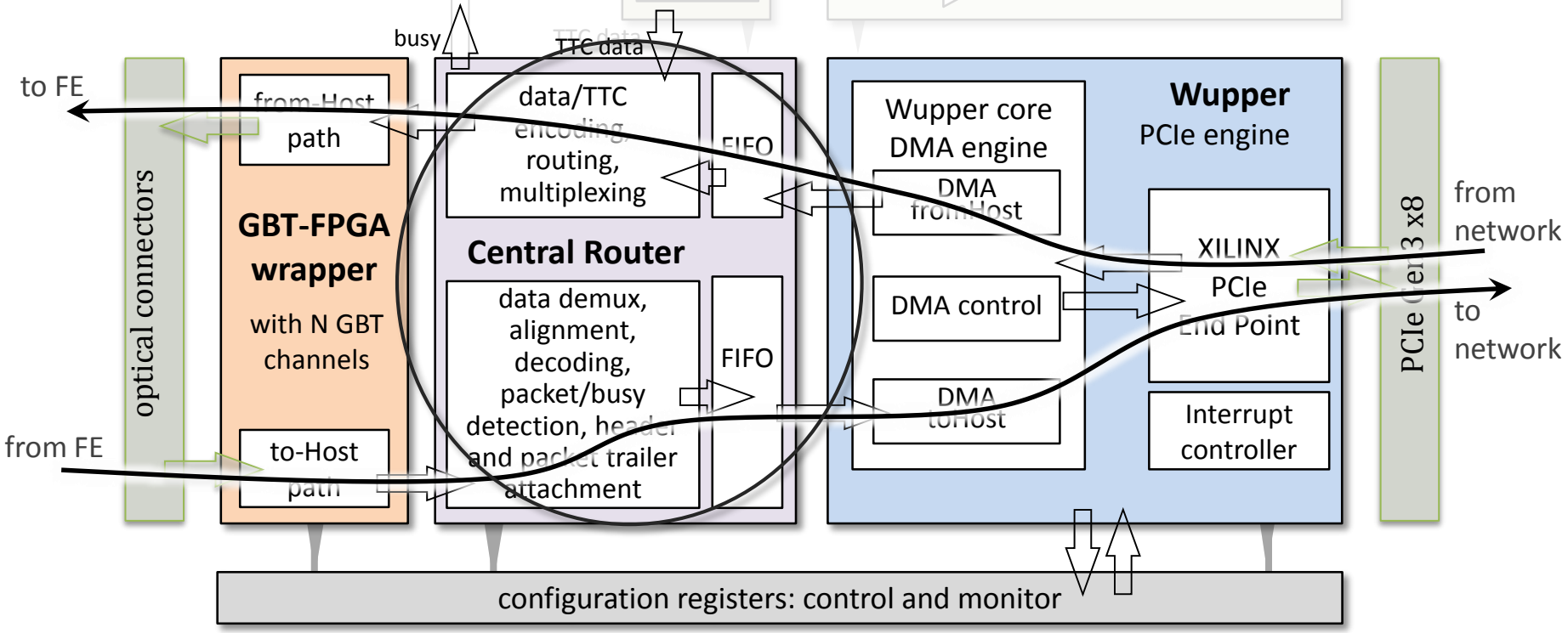


cont. FELIX system today: PCIe FPGA board FW



cont. FELIX system today: PCIe FPGA board FW

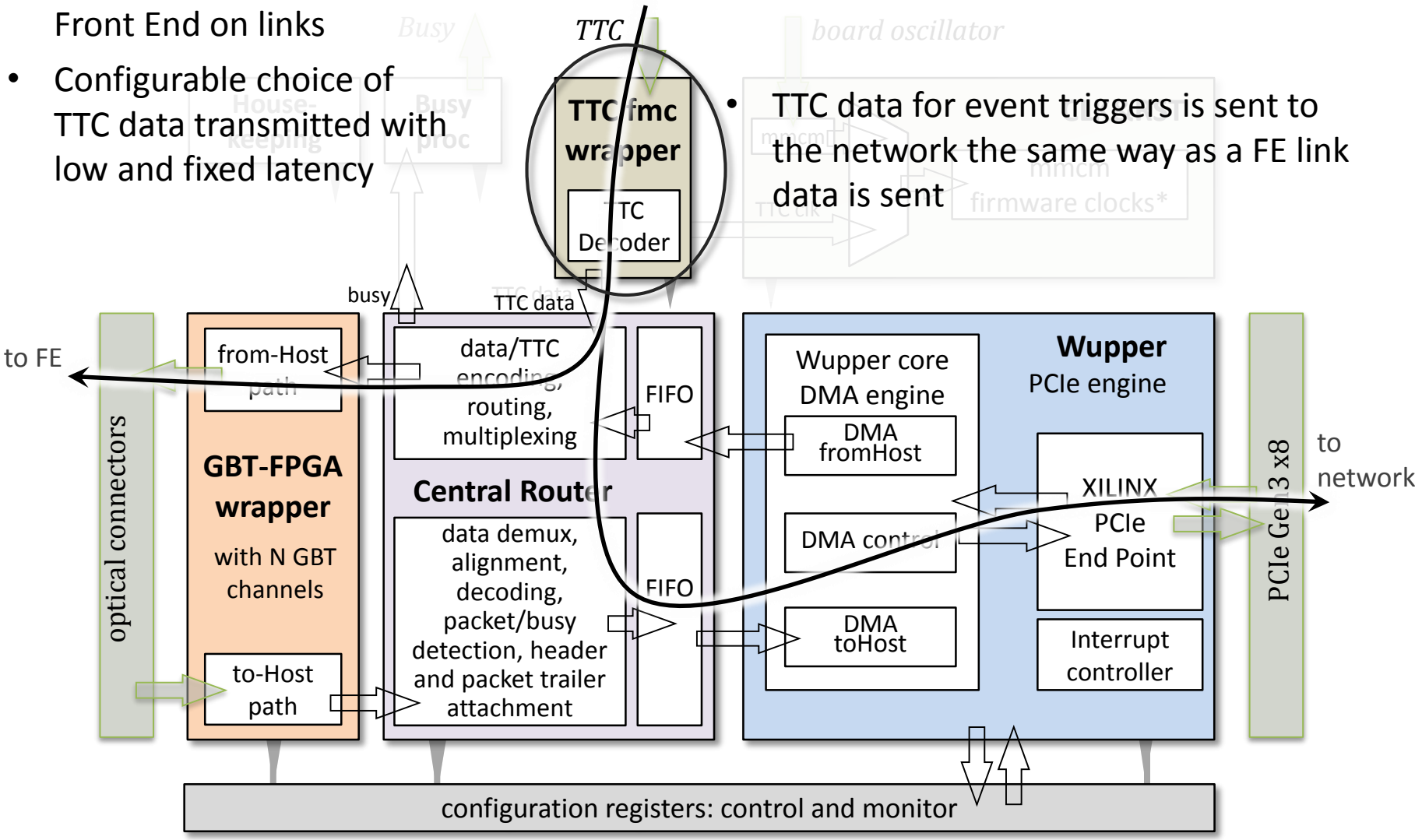
- In to-FE direction: Central Router encodes separate data streams written by DMA and multiplexes the encoded streams into corresponding link
- In from-FE direction: it demultiplexes GBT link into separate data streams, decodes each data stream according to configuration, forms 1Kbyte blocks which are read by DMA and reassembled in software according to block header and trailers
- Configurable and monitored via registers (configurable number of channels in both directions)



cont. FELIX system today: PCIe FPGA board FW

TTC = Timing Trigger and Control

- TTC data is forwarded to the Front End on links
- Configurable choice of TTC data transmitted with low and fixed latency

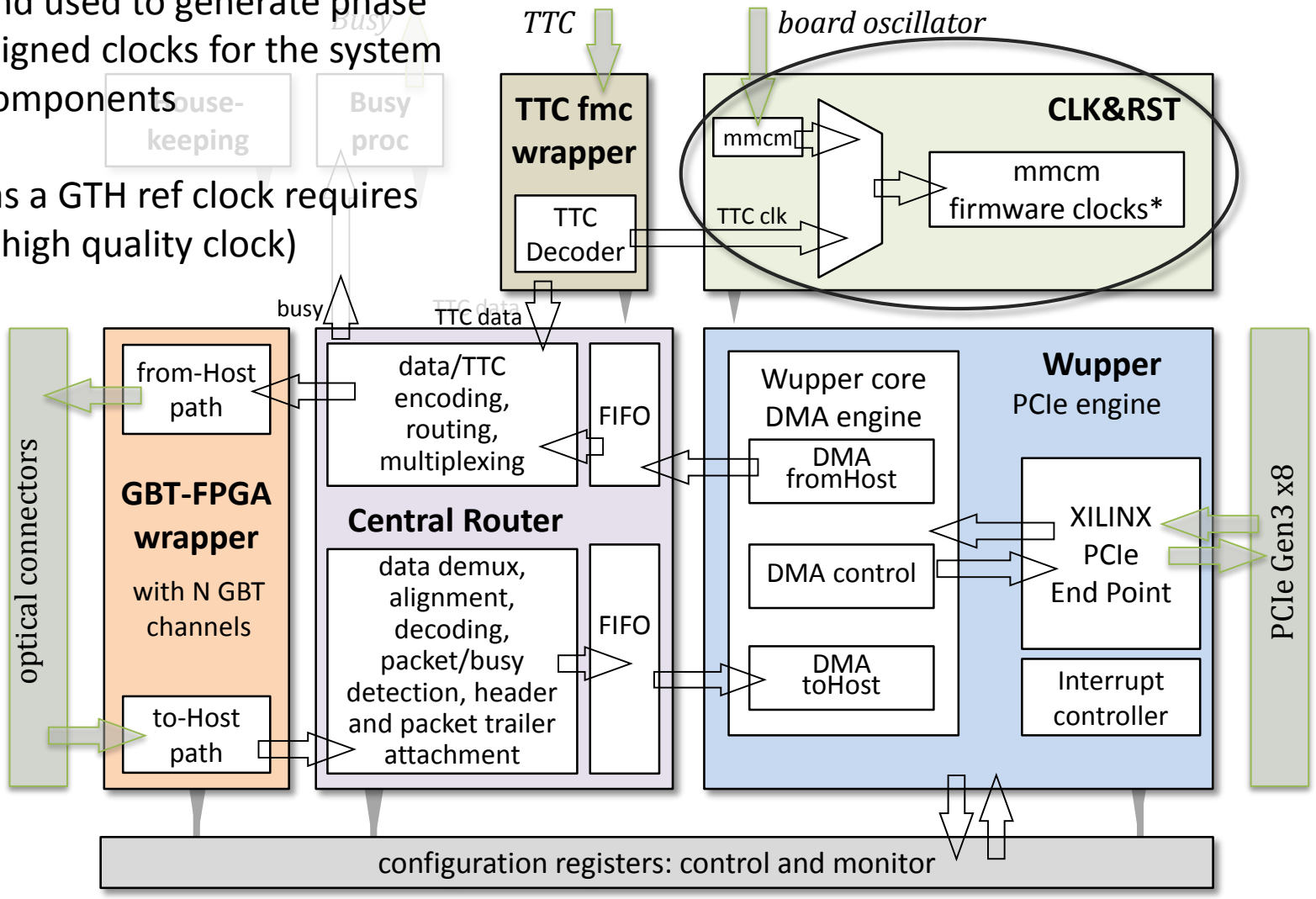


- TTC data for event triggers is sent to the network the same way as a FE link data is sent

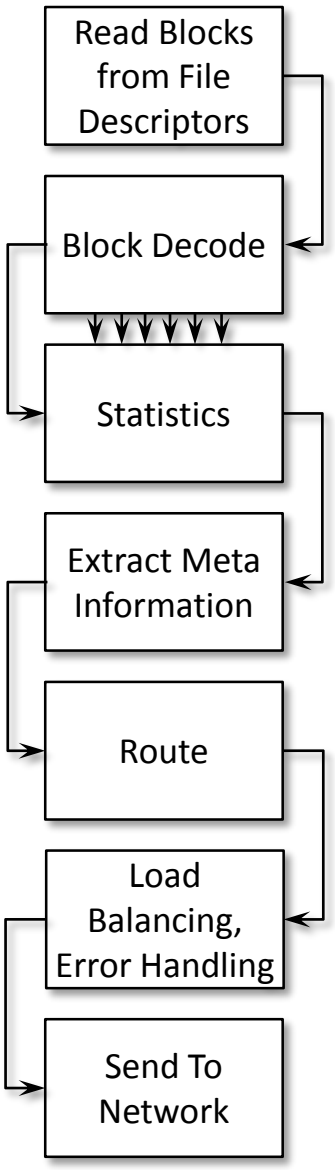
cont. FELIX system today: PCIe FPGA board FW

- Clock is recovered (x4 BC clock) and jitter cleaned (Si5345) externally and used to generate phase aligned clocks for the system components

  (as a GTH ref clock requires a high quality clock)



FELIX system today: Server SW



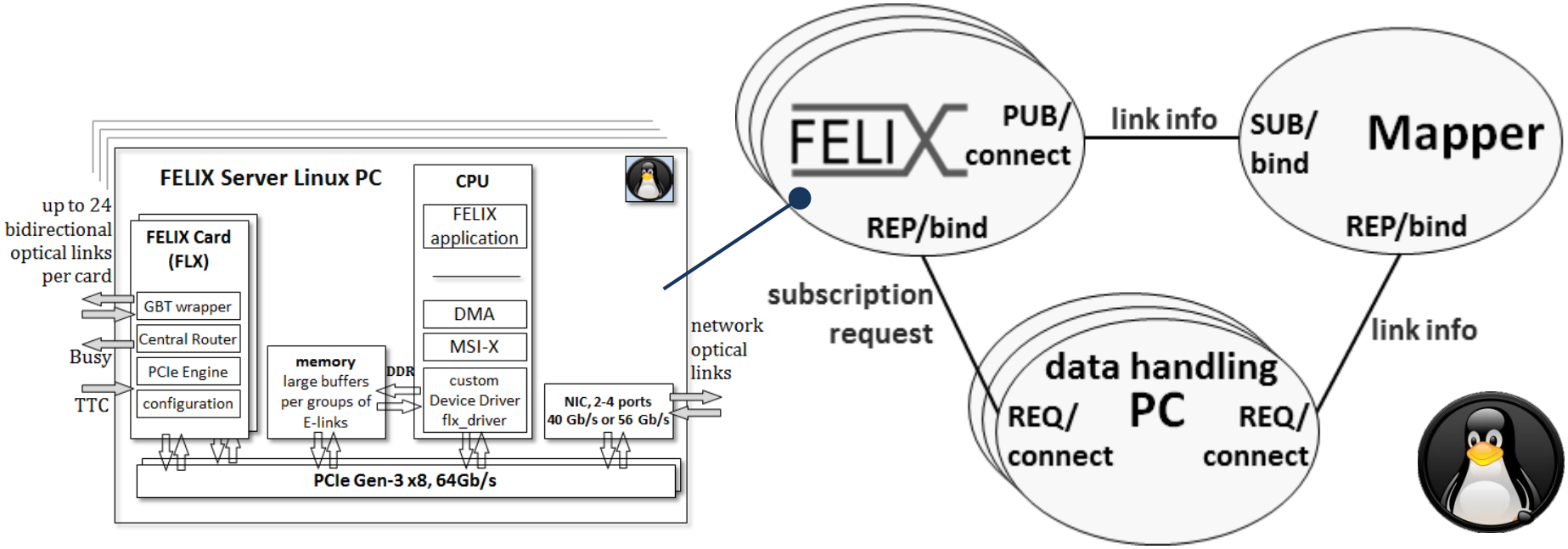
Data path FELIX application: *FELIX Core Application*

- DMA reads and transfers 1KB blocks of data over PCIe where every block contains accumulated data from one link
- Continuous DMA to circular buffer allows full use of PCIe bandwidth
- Reassemble variable sized packets from 1KB blocks
- Count processed blocks, transfer rates, etc.
- Meta-information, for example event ID, is extracted and matched against a routing table
- Route
- Distribute load among multiple systems, handle automatic failover in case of system failures
- Networking is implemented in FELIX library called "netio", which supports Ethernet and InfiniBand
- Each link in each direction is supported by its own bidirectional TCPIP socket
- Broadcasts / multicasts to links are supported
- Streams may be cloned to several network end-points

FELIX system today: Server SW

‘Mapper’ application for interface with the stateless FELIX domain:

- Startup - FELIX registers the links it handles with Mapper
- Data handling PC requests this info from the Mapper and then directly subscribes to links with FELIX
- Main application running on a FELIX Host
  - reads data from FLX card, decodes it and sends it to network subscribers
  - receives data via network, writes it to the FLX card



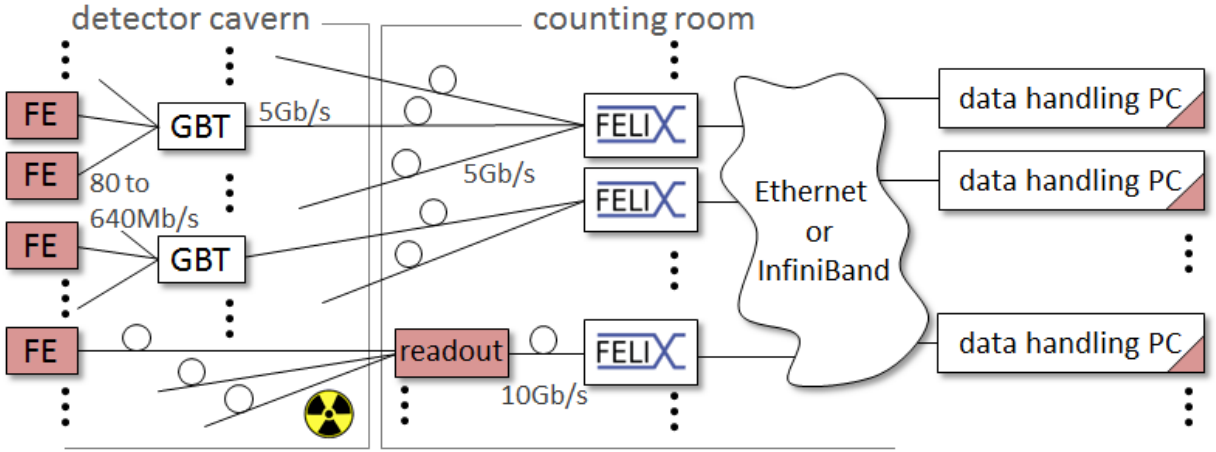
FELIX development status

- FLX-709 (Xilinx VC-709) and its software is distributed to ATLAS Sub-Detectors Front End developers
- Tests with the BNL-711 successfully accomplished, a second minor revision is planned for the coming months
- FELIX Control and monitor tools are in advanced development stage
- FELIX application is in the development stage and simple data transfers over the network are possible
- Networking layer: Initial work has been done and simple communication works. Current focus is on improving interface, performance, and looking into features like fault-tolerance
- Ongoing effort to increase overall system reliability and the number of channels

- Final Design Review is planned for October 2016.

Summary

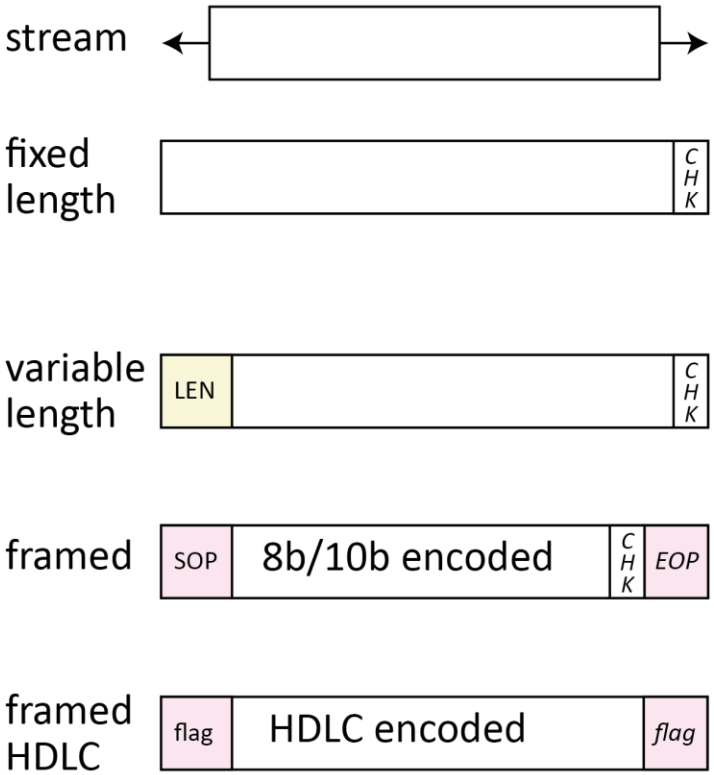
- In LHC Run-3 (installation in 2019) some detectors and trigger systems will be interfaced to the data acquisition, detector control and timing (TTC) systems by the FELIX. In LHC Run-4 (installation in 2024) this is planned for all ATLAS detectors
- FELIX is a **router** between custom serial links and a commodity network, separates data transport from data processing
- FELIX is implemented by server PCs with commodity network interfaces and PCIe cards with large FPGAs and many high speed serial fiber transceivers
- Replaces traditional point-to-point links between Front-End and the DAQ system
- FELIX provides flexibility, uniformity and upgradability and reduces the diversity of custom hardware solutions in favor of software



Back Up Slides

FELIX E-link data packet format examples

Single stream E-link:
Data can be routed to only a single end-point.



No packet boundaries, i.e. a TCP stream.
The single end-point must parse out packets.

NOT recommended.
Packet boundaries lost if bits lost on an E-link.
Fixed length packets whose framing is periodically aligned by strings of zeroes will also be implemented.

NOT recommended.
Packet boundaries lost if bits lost on an E-link.
CHK: an optional check-sum

Guarantees packet boundaries, e.g. event boundaries.
8b/10b comma, SOP and EOP symbols are not forwarded.
CHK: an optional check-sum. EOP is optional.

Guarantees packet boundaries, e.g. event boundaries.
Data outside frames are not forwarded.
Used by Slow Control Adapter ASIC.

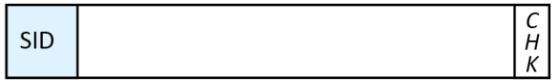
Full mode: 8b/10b, either with or without stream-id

LL_SingleStreamE-link_V01

FELIX E-link data packet format examples

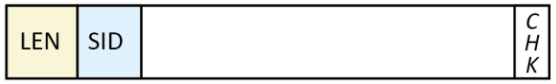
Multiple stream E-link:
Packets with different Stream IDs can be routed to different end-points.

fixed length



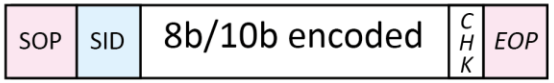
NOT recommended.
Packet boundaries lost if bits lost on an E-link.
Fixed length packets whose framing is periodically aligned by strings of zeroes will also be implemented.

variable length



NOT recommended.
Packet boundaries lost if bits lost on an E-link.
CHK: an optional check-sum

framed

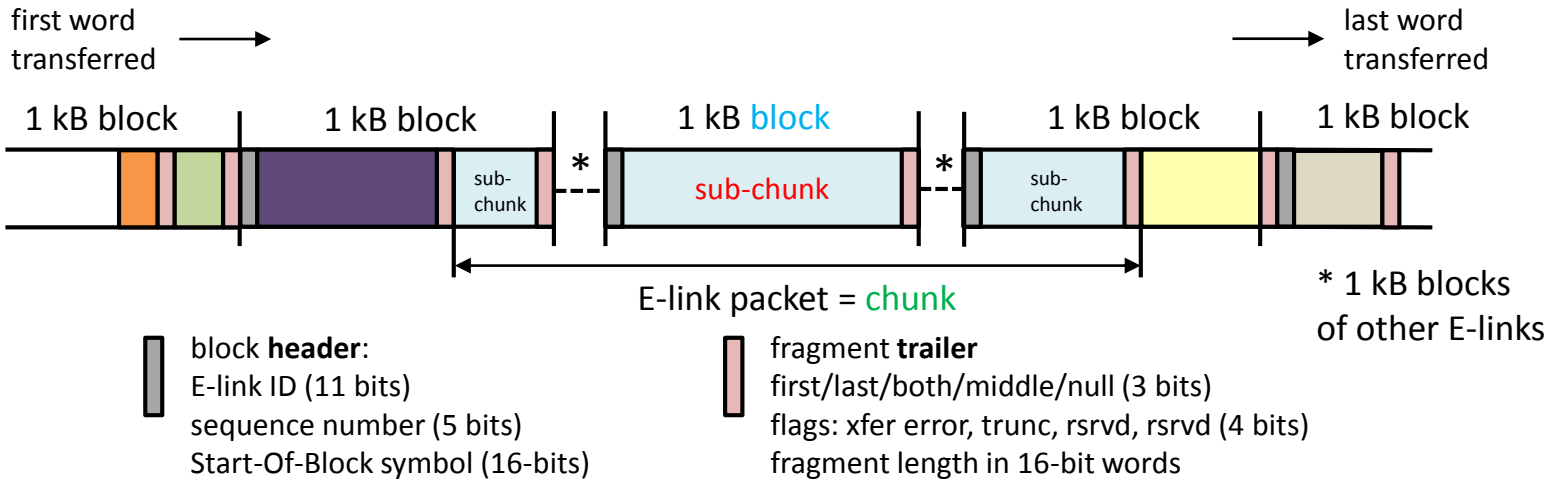


Guarantees packet boundaries, e.g. event boundaries.
8b/10b comma, SOP and EOP symbols are not forwarded.
The stream-ID is 8b/10b encoded.
CHK: an optional check-sum. EOP is optional.

Full mode: 8b/10b, either with or without stream-id

LL_MultiStreamE-link_V01

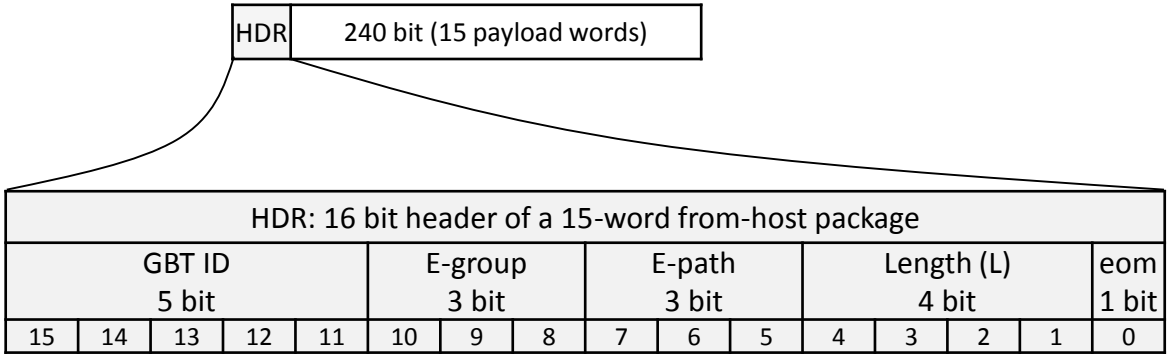
High throughput detector readout



- For each E-link fixed size blocks (1 kByte) are filled with received data
- Each block has a 4-Byte **header**: E-link ID, a sequence number and a start of block symbol
- Data packets ("chunks") received can be of arbitrary length and are subdivided (typically after e.g. 8B/10B decoding) in sub-chunks as needed to fill the blocks
- Each sub-chunk has a **trailer** with information on its length and type
- In case of low data rates time-outs will cause incompletely filled blocks to be padded and sent (the last sub-chunk in this block is then of type "null")
- Blocks are transferred using continuous DMA (Direct Memory Access) into a large (e.g. 4 GByte) circular buffer in host PC memory
- The buffer consists of contiguous memory allocated by a dedicated driver
- The DMA is controlled with two pointers, a write pointer maintained by the DMA controller in the FPGA, and a read pointer maintained by the FELIX application

FELIX to Front-End packet format

'From-Host' 32 byte (256 bit) message



The header is attached by FLX software, used by FLX card firmware

FELIX to Front-End TTC: Phase-1

FELIX decodes TTC signal, resulting in following TTC 10-bit data @ 40MHz

bit 9	bit 8	bit 7	bit 6	bit 5	bit 4	bit 3	bit 2	bit 1	bit 0
Brcst[7]	Brcst[6]	Brcst[5]	Brcst[4]	Brcst[3]	Brcst[2]	ECR	BCR	B-chan	L1A

TTC e-links towards Front-Ends, according to the corresponding detector requirements

#	speed	bit 7	bit 6	bit 5	bit 4	bit 3	bit 2	bit 1	bit 0
0	80 Mb/s							B-chan	L1A
1	160 Mb/s					B-chan	ECR	BCR	L1A
2	160 Mb/s					Brcst[2]	ECR	BCR	L1A
3	160 Mb/s					BCR	BCR	BCR	BCR
4	320 Mb/s	B-chan	Brcst[5]	Brcst[4]	Brcst[3]	Brcst[2]	ECR	BCR	L1A
5	320 Mb/s	Brcst[6]	Brcst[5]	Brcst[4]	Brcst[3]	Brcst[2]	ECR	BCR	L1A

FELIX to Network TTC: Phase-1

TTC information will be sent as a virtual e-link data to the network subscribers in predefined packets:

byte#	byte contents	
0	FMT [7:0]	
1	Len [7:0]	
2	reserved [3:0]	BCID [11:8]
3	BCID [7:0]	
4	XL1ID [7:0]	
5÷7	L1ID [23:0]	
8÷11	orbit [31:0]	
12÷13	trigger type [15:0]	
14÷15	reserved [15:0]	
16÷19	L0ID [31:0]	

FELIX system today: Server HW

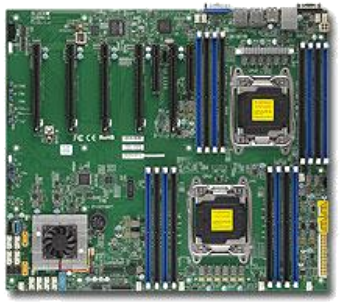
Supermicro motherboards, e.g.:

SuperMicro X9SRL-F (Nikhef)

- 1x Ivy Bridge CPU, 6 cores
- 6x PCIe Gen-3 slots
- 16 GB DDR3 Memory

http://www.supermicro.com/products/motherboard/Xeon/C600/X9SRL-F.cfm

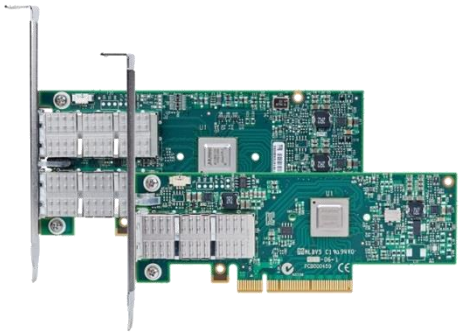
OS: SLC6



SuperMicro X10DRG-Q (CERN)

- 2x Haswell CPU, up to 10 cores
- 6x PCIe Gen-3 slots
- 64 GB DDR4 Memory

http://supermicro.com/products/motherboard/Xeon/C600/X10DRG-Q.cfm



Mellanox ConnectX-3 VPI

- FDR/QDR Infiniband
- 2x10/40 GbE

http://www.mellanox.com/page/products_dyn?product_family=119&mtag=connectx_3_vpi