TPC readout using FELIX

Frank Filthaut (for the FELIX team: Nikhef, CERN, PNNL)

ProtoDUNE DAQ review

3-11-2016

Review documentation: DUNE-doc-1846

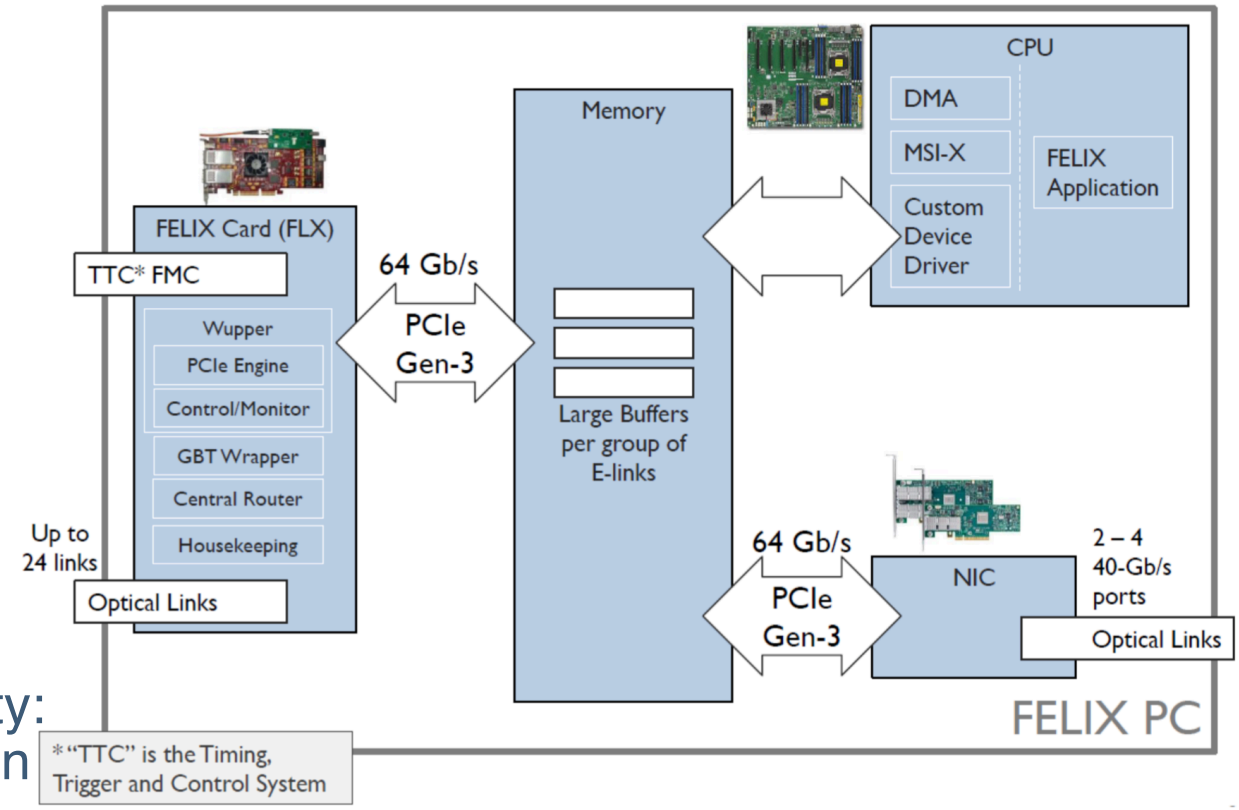
Nikhef Radboud Universiteit Nijmegen

Contents

- Design
  - Data
  - Software
- Interfaces summary
- Implementation
  - Hardware
  - Numbers
- Status
- Risks
- Plans
- Conclusion

Design

- FELIX (Front-End Link EXchange): PC-based solution using COTS hardware / networking at as early a stage as possible
- Development initially within ATLAS
  - data arriving from front-ends are de-serialised in firmware, then DMA'd to host PC memory buffer
    - e.g. 16 GB
  - software streams data to different outputs
  - add'l s/w functionality: filtering, compression



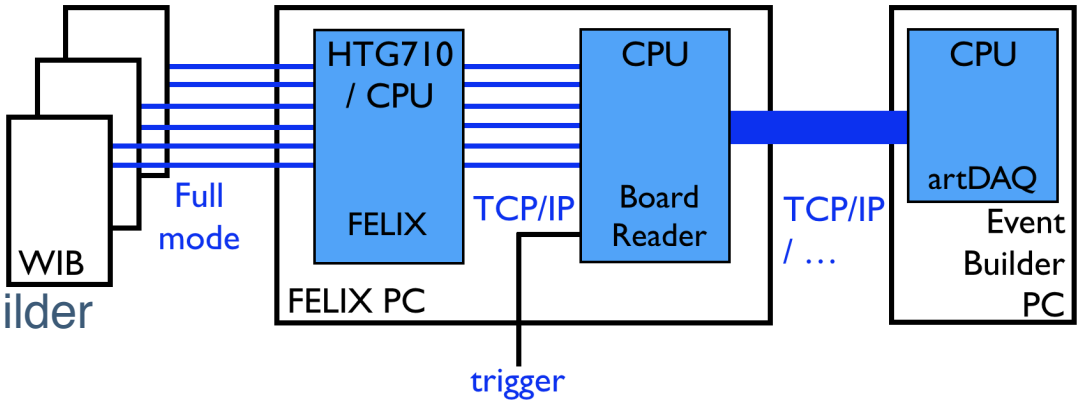
Design: data

- Transmission format: FELIX Full mode
  - 9.6 Gb/s (7.68 Gb/s payload), 32-bit words; 8b/10b encoding
- Draft data format: one 120-word frame / 500 ns (256 channels, from 4 COLDATA ASICs; 2 links / Warm Interface Board)
  - ongoing discussion: transmission of (calibrated) time-stamp information
  - (slightly) different format for FELIX than for RCE

Word	X	Y	31 30 29 28 27 26 25 24	23 22 21 20 19 18 17 16	15 14 13 12 11 10 9 8	7 6 5 4 3 2 1 0	
1	1	0	Reset_Count[23:0]			Stream_ID	WIB Header
2	1	1	WIB_Timestamp[31:0]				
3	1	1	WIB_Errors	ASIC[B]	Capture[A] Reserved FiberNo	SlotNo CrateNo	
4	1	1	Reserved	S2_Err[C]	S1_Err[C]	Checksum_B[7:0] Checksum_A[7:0]	COLDATA 1 (28 Words)
	1	1			...		
	1	1					
32	1	1	Reserved	S2_Err[C]	S1_Err[C]	Checksum_B[7:0] Checksum_A[7:0]	COLDATA 2 (28 Words)
	1	1			...		
	1	1					
60	1	1	Reserved	S2_Err[C]	S1_Err[C]	Checksum_B[7:0] Checksum_A[7:0]	COLDATA 3 (28 Words)
	1	1			...		
	1	1					
88	1	1	Reserved	S2_Err[C]	S1_Err[C]	Checksum_B[7:0] Checksum_A[7:0]	COLDATA 4 (28 Words)
	1	1			...		
	1	1					
116	0	1	CRC-32				WIB Trailer Idle
117	0	0	IDLE (filled by FELIX firmware)				
118	0	0	IDLE (filled by FELIX firmware)				
119	0	0	IDLE (filled by FELIX firmware)				
120	0	0	IDLE (filled by FELIX firmware)				
			31 30 29 28 27 26 25 24	23 22 21 20 19 18 17 16	15 14 13 12 11 10 9 8	7 6 5 4 3 2 1 0	

Design: software

- One software thread per input link, pushing the link's data onto a separate network queue (standard FELIX functionality)
- ProtoDUNE: networking to BoardReader using loopback (different process(es), possibly on different CPU)
  - BoardReader uses time-stamps (network messages from the trigger & timing system) to pass only the data in triggered time-windows (5 ms); 25 Hz ➔ 1/8 of data
  - BoardReader compresses data (further reduction by factor 4—5?)
    - lossless
    - Board Reader packages data as *artdaq* fragments and sends to event builder

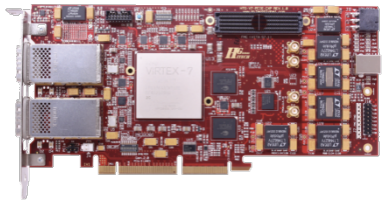


Interfaces summary

- TPC readout electronics:
  - Collect inputs from 1 APA (5 Warm Interface Boards)
  - Data format (slightly) different from that sent to RCE
- Back-end computing:
  - Send data as *artdaq* fragments to EventBuilder: 10 Gb/s ethernet
  - *artdaq* BoardReader integrated in FELIX system (even if not part of FELIX itself)
  - Also a natural place for online monitoring, to be investigated
- Timing/trigger:
  - Network messages from the trigger & timing system
- Constraints:
  - Data throughput per CPU limited to ~ 50 Gb/s by Gen-3 8-lane PCIe ➔ need 2 systems
    - 3 when using one VC709 card / PC, possibly 1 when using BNL-711 card

Implementation: hardware

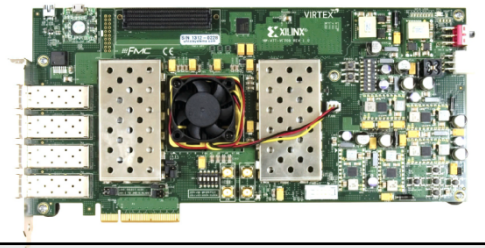
FLX-710: HiTech Global HTG-710 w/TTCfx



- → FELIX development
- → Virtex-7 X690T
- → 2x12 bidirectional CXP connectors
- → PCIe Gen3 x8

or

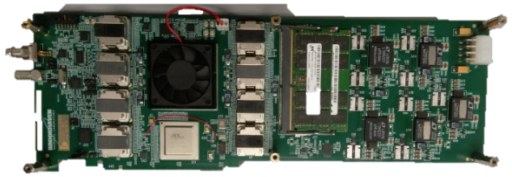
FLX-709: Xilinx VC-709 w/TTCfx



- → Subset of full FELIX, intended for FE development support
- → Virtex-7 X690T
- → 4 SFP+ connectors
- → PCIe Gen3 x8

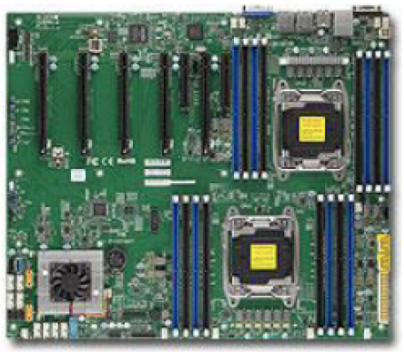
or

FLX-711 from BNL



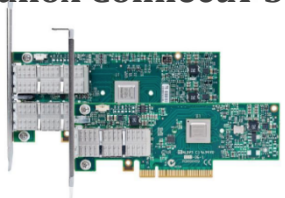
- → FELIX phase 1 prototype
- → TTC input ADN2814 + Si5338
- → Xilinx Kintex **Ultrascale** XCKU115
- → 48 duplex optical links (based on MiniPODs)
- → PCIe Gen3 x16

SuperMicro X10DRG-Q



- → 2x Haswell CPU, up to 10 cores
- → 6x PCIe Gen3 slots
- → 64 GB DDR Memory

Mellanox ConnectX-3 VPI



- → 2x 10/40 GbE

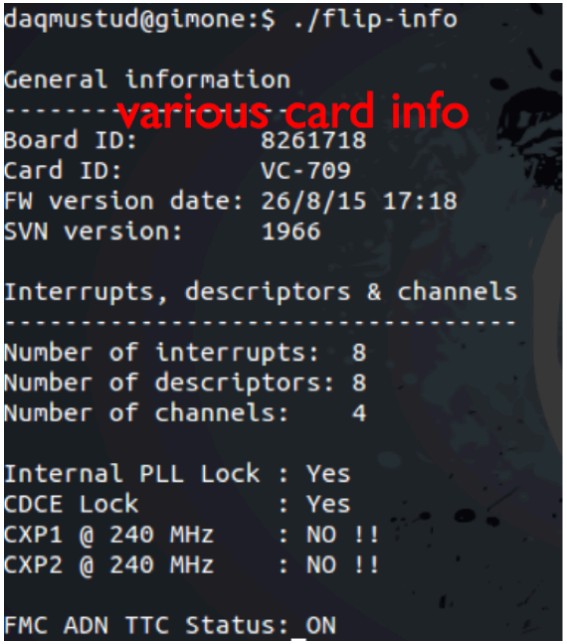
Implementation: numbers

Quantity	Number
Channels	2560 (1 APA)
Input links	10 (= 5 WIBs × 2 links/WIB)
FELIX systems	2 (1 VC-709 + 1 HTG-710)
Input rate (payload) [Gb/s]	76.8 (= 30.7 + 46.1)
Input rate (data) [Gb/s] (*)	75.5 (= 30.2 + 45.3)
Board Reader output data rate [Gb/s] (**)	2.4 (= 0.96 + 1.44)

- Nominal design: two systems using commercial FELIX cards
- (*) ignore 2 out of 120 data words (2 filled by FELIX firmware)
- (**) assume that BoardReader can handle both trigger filtering and data compression
  - readout fraction: 25 Hz × 5 ms time window = 0.125
  - compression: factor 4

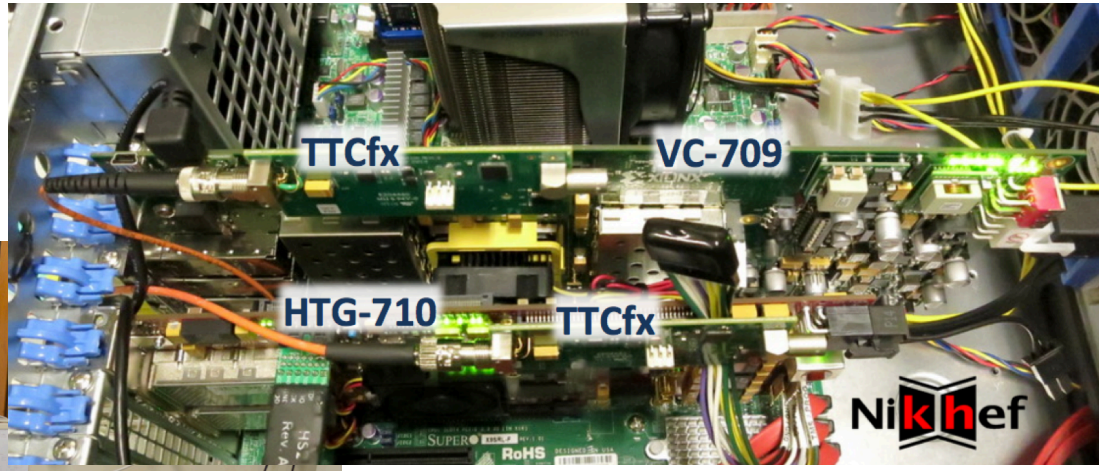
Status

- Hardware (except BNL-711 card) available
  - in the Nikhef case, shared with core FELIX setup
  - 2nd BNL-711 version (minor improvements & bug fixes) under test, 3rd version to be produced soon
- Optical links (on the BNL-711 card) tested to run up to 12.8 Gb/s
- PCIe Gen-3 16-lane throughput of 101.7 Gb/s measured (from the BNL-711 card)
  - close to earlier measurements using 1 or 2 PCIe Gen-3 8-lane connections: ~ 50 Gb/s per PCIe interface, single shot DMA transfers
- First tests of FELIX “Full mode” done
- Extensive suite of software tools



Status

- Nikhef FELIX “playground” and detail
  - TTC not relevant to ProtoDUNE



Risks

- Trigger filtering too time-consuming for BoardReader running on the same CPU (unlikely)
  - run on a separate CPU (dual CP host) or in a separate PC (requires high-bandwidth NIC, which may otherwise not be needed)
- (Efficient) data compression too time-consuming for BoardReader running on the same CPU (compression algorithms being studied)
  - run on a separate CPU or PC (or offload BoardReader)
    - or a mixture of BoardReaders running on the same and different PCs
  - run in firmware (wire by wire, as for RCE; affects firmware development)

Plans

- Emulation of data sent by WIB in FELIX “Full mode” to be done in integrated environment (Nikhef, CERN; February 2017)
- Addition of BoardReader application (Nikhef, CERN; July 2017)
  - buffering, reception of trigger time-stamp, trigger filtering
  - data compression
    - starting to evaluate (offline) different libraries (e.g. Zstd) for software compression
  - packaging as *artdaq* fragments
  - testing of throughput and memory usage; choice of configuration
- Pursue alternative of compression in firmware (PNNL; end 2017)
- Note: (ATLAS) FELIX Final Design Review on November 11

Conclusion

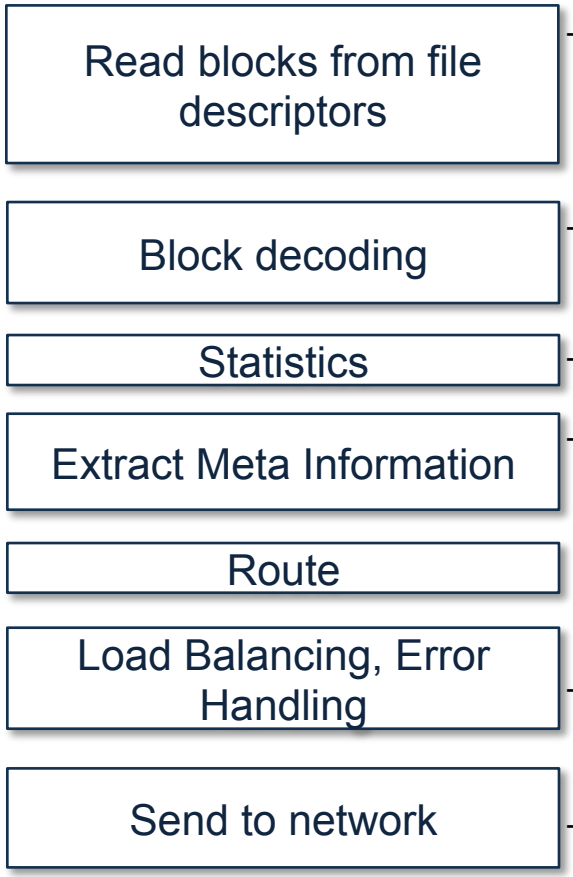
- Very versatile design based on (solely or mostly) COTS hardware & networking
  - Nominal design (2 systems) should allow for a configuration that comfortably fits bandwidth limitations. In case of unforeseen limitations (biggest unknown: compression), hardware can be easily added
    - at most a few PCs
  - Relies heavily on core FELIX development taking place within ATLAS
    - ongoing (but but very advanced) project
  - On the ProtoDUNE specific side, the work is mostly on software (BoardReader)
    - apart from the option of compressing data in firmware

Finally…

- This presentation (and actual work) targets operation at 25 Hz trigger rate
- However, the system will be versatile enough to adapt to other configurations (e.g., trigger-less readout as desired for DUNE, or higher trigger rates)
  - of course these other configurations may require additional hardware

(Generic) FELIX Software data path

- (from Andrea Borga)



→ DMA transfers and reads 1KByte blocks of data, encoded for the transfer over PCIe. Every block contains accumulated data from one E-link.

→ Decode into variable sized packets for transmission over network.

→ Count processed blocks, transfer rates, etc.

→ Meta-information, for example event ID, is extracted and matched against a routing table.

→ Distribute load among multiple systems, handle automatic failover in case of system failures.

→ Asynchronous message service 'netio' for data exchange with network hosts.

Interfaces summary

- FELIX based readout:
  - Collect inputs from 1 APA (5 Warm Interface Boards)
  - Apply trigger filtering (based on network messages from the trigger & timing system) and data compression
  - Package as *artdaq* fragments and send to event builder
- Constraints:
  - Input link speed (in *Full mode*): 9.6 Gb/s ➔ 2 links / WIB
  - Data throughput per CPU limited to ~ 50 Gb/s by Gen-3 8-lane PCIe ➔ need 2 systems
    - 3 when using one VC709 card / PC, possibly 1 when using BNL-711 card