

### FrontEnd LInk eXchange

# **Detailed Description**

### Andrea Borga Nikhef on behalf of the FELIX group



Neutrino DAQ workshop CERN 17-05-2016

# FELIX development team



- John Anderson
  - Soo Ryu
- Jinlong Zhang



- Hucheng Chen
- Kai Chen
- Weihao Wu
- Francesco Lanni
- Frans Schreuder
- Andrea Borga
- Henk Boterenbrood
- Jos Vermeulen





- Giovanna Lehmann
- Wainer Vandelli
- Benedetto Gorini



- UNIVERSITÄT PADERBORN Die Universität der Informationsgesellschaft
  - Joern Schumacher



Mark Donszelmann



ROS/ROD effort on software framework:

- Daniel Guest
- Daniel Whiteson

- WEIZMANN INSTITUTE OF SCIENCE
- Julia Narevicius
- Alex Roich
- Lorne Levinson

- Will Panduro (RHUL)
- Joern Schumacher (CERN / Paderborn)
- Wainer Vandelli (CERN)

Gordon Crone (UCL)

# FELIX at a glance

#### Interfaces multiple GBT (GigaBit Transceiver) serial links to a high bandwidth standard network (40 Gbps Ethernet or 56 Gbps InfiniBand) then routes logical data flows to/from different off-detector endpoints.



- Separates GBT technology into a standard, fixed, but configurable, building block for several detectors.
- FELIX does not need to know the running state of the experiment or detailed data formats high-reliability: always running.
  - Enables different logical data flows to be handled by different off-detector end points.
    - In the past done by using separate physical interconnects.
- Reduces the amount of custom HW in ATLAS in favour of COTS HW and SW.

# FELIX: Front-End Link eXchange



- $\rightarrow$  Routing of event data, detector control, configuration, calibration and monitor
- $\rightarrow$  Connects ATLAS detector frontends to the ATLAS DAQ system, both up-links and down-links
- $\rightarrow$  SW based data processors and handlers
- $\rightarrow$  All FELIX components are independently upgradable (PCs, FPGAs, NICs)
- $\rightarrow$  Integrated Timing Trigger and Control (TTC) and LHC clock distribution
- $\rightarrow$  Scalable architecture
- $\rightarrow$  Flexible mapping of FE GBTs to data handlers
- → Detector independent
- $\rightarrow$  GBT Normal mode (with FEC, 3.2 Gb/s)
- $\rightarrow$  GBT wide mode (no error correction, 4.48 Gb/s)
- $\rightarrow$  "Full mode" (~10 Gb/s, protocol to be defined)

### FELIX: Front-End Link eXchange



ANY

# FELIX server PC components



# FELIX demo and development platform



### FELIX today



### **FELIX Flow overview**



### FELIX Flow: From-FrontEnd



# The GBT protocol (simplified)

- The *GBT protocol* is a data serializier (with line encoding)
- The *GBT frame* is 120 bits @ 40 MHz
  - 4 bits are reserved for a header (data / no data)
  - 4 bits are always available for the To FrontEnd direction
- GBT wide mode, 4.48 Gb/s (112 bit @ 40MHz) payload
- 8B/10B symbols (<u>not</u> for line encoding) used to delimit stream boundaries, Start Of Frame (SOF), End Of Frame (EOF), comma



#### 2KByte FIFO.

 In the To-Host direction accumulates 1KByte blocks of corresponding E-link data. Transfers complete 1KByte blocks.

**E-link:** variable-width logical link (80, 160 or 320 Mb/s) on top of the GBT protocol. Can be used to logically separate different streams on a single physical link.

**E-group:** a group of e-links (any combination of 16 bits)

**E-proc:** in a e-group handles every e-link toward its FIFO buffer

Neutrino DAQ workshop CERN 17-05-2016

### **FELIX Flow: To-Host**



# FELIX Flow: To-Host

Attached by FLX card firmware, used by FLX software



Neutrino DAQ workshop CERN 17-05-2016

### **FELIX Flow: To-Host**



# 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, sequence number, 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

A. Borga

#### Neutrino DAQ workshop CERN 17-05-2016

### FELIX Flow: From-Host



### FELIX Data Flow: From-Host



Neutrino DAQ workshop CERN 17-05-2016

### FELIX Flow: To-FrontEnd



# The GBT protocol (simplified)

- The *GBT protocol* is a data serializier (with line encoding)
- The GBT frame is 120 bits @ 40 MHz
  - 4 bits are reserved for a header (data / no data)
  - 4 bits are always available for the To FrontEnd direction



- *IC (Internal Control)* bits are used to configure the GBTx chip itself
- *EC (External Control)* bits are used to configure devices external to the GBTx

### FELIX... halftime







# FLX PCIe card firmware: zoom-in



# FLX PCIe card firmware: configuration



# Wupper\*: PCIe Engine for FELIX



• Developed for use in FELIX

A. Borga

 Published as OpenSource (LGPL) on OpenCores <u>http://opencores.org/project,virtex7\_pcie\_dma</u>

- PCIe Engine with DMA interface to the Xilinx Virtex-7 PCIe Gen3 Integrated Block for PCI Express (PG023)
- Xilinx AXI (ARM AMBA) Stream Interface (UG761)
- MSI-X compatible interrupt controller
- Applications access the engine via simple FIFOs
- Register map for programmed I/O synchronized to a lower clock speed

\* The person performing the act of bongelwuppen, the version from the Dutch province of Groningen of the "famous" Frisian sport Fierljeppen (canal pole vaulting) <u>https://www.youtube.com/watch?v=YP32iWoqjnQ</u>

Neutrino DAQ workshop CERN 17-05-2016

# FLX PCIe card firmware: data path



### FELIX Flow: Software



# FELIX software: control, monitor

#### FELIX low-level tools: *FLXtools*

FELIX firmware configuration through register map	→ configure the FLX card via the internal registers mapped into the host	various card infoGeneral informationGeneral informationGeneral informationBoard ID:8261718Card ID:VC-709
FELIX 'housekeeping' functions	memory.	FW version date: 26/8/15 17:18 SVN version: 1966
	$\rightarrow$ Control FLX external devices (e.g.	Interrupts, descriptors & channels
	clock chips) via I2C and SPI.	Number of interrupts: 8 Number of descriptors: 8
System test and debugging	$\rightarrow$ Performance testing, data flow control	Number of channels: 4
	and verification.	Internal PLL Lock : Yes CDCE Lock : Yes
Complete user toolset	$\rightarrow$ collection of tools to control, monitor,	CXP1 @ 240 MHZ : NO !! CXP2 @ 240 MHZ : NO !!
	configure different FLX cards.	FMC ADN TTC Status:_ON

- \$ flx-dump-blocks -n 100 -f output.blocks
- \$ flx-config \$ flx-dma-test
- **FLXtools set**
- \$ flx-eeprom
- \$ flx-i2c
- \$ flx-info
- \$ flx-init
- \$ flx-irq-test
- \$ flx-spi
- \$ flx-reset
- \$ flx-throughput

oqmustud@gimone:\$ ord model HTG 710 vitch I2C address: .st of available d evice	./flip-i2c list -d 1 <sup>0x70</sup> HW periphe levices: monitor (FL) Model	ral Ktools)	Addres	s				
.ock_RAM .ock_SYS .ock_CXP1 .ock_CXP2 IC_ADN IC_TEMP_SENSOR	ICSBN4Q001L IDT ICSBN4Q001L IDT IDT BN3Q001 IDT BN3Q001 ADN2814 (on TTCfx FMC) TC74 (on CRORC TEST FMC)	0:0 1:0 2:0 3:0 4:0 4:0	0x6e 0x6e 0x6e 0x6e 0x40 0x40 0x4a		CDT d		8L (5	
(P1_1X (P1_RX (P2_TX (P2_RX )R3-1	AFBR-83PDZ AFBR-83PDZ AFBR-83PDZ AFBR-83PDZ SRAM-MT16JTF25664HZ	GBT CHANNEL 0	ALIGNMEN	Va T STATUS	rious 3 4	card s (F	status LXtool 6	es Is) 7
		Aligned   Ye	s Yes	Yes Y	'es Yes	Yes	Yes Y	/es

dagmustud@gimone:\$ ./flip-info

# FELIX software: support

#### FELIX support software (WupperCodeGen, fel, FELIX E-link configurator)

Automatic generation of a register map

Firmware development tools

Firmware test and debugging

 → Automation software (Jinja based) to handle register map consistently between firmware (HDL), software (C headers), and documentation (Latex) and beyond (OKS, ...).
 → Configuration of the internal data emulators to verify the functionality 'off-line'.

• Communication with FLX card, data read and analysis.





# FELIX software: data path

#### FELIX application: FELIX Core Application



### **FELIX Flow: Networking**



Network-side packet handling



- Networking is implemented in a FELIX library called "netio", which has different backends to support, Ethernet, InfiniBand and potentially more in the future.
- For Ethernet, FELIX uses TCP for communication with network endpoints.
- Using TCP has the advantage of guaranteeing no packets loss. InfiniBand RDMA connections guarantee reliable delivery.
- Note that although TCP sockets are bi-directional, GBT streams are not; a TCP socket is associated with only one direction of a GBT transfer.
- Streams may be cloned to several network end-points.
  For clones marked with QoS as less-than-best effort, some packets may not be sent, i.e. "sampling".
- Broadcasts / multicasts to E-links are supported.

# FELIX Flow: Timing



White Rabbit provides **sub-nanosecond accuracy** and picoseconds precision of **synchronization** for large distributed systems. It also allows for deterministic and reliable data delivery.

White Rabbit allows you to **precision time-tag measured data** and lets you trigger data taking in large installations while at the same time using the same **network to transmit data**.

http://www.ohwr.org/projects/white-rabbit/wiki

http://www.ohwr.org/projects/white-rabbit/wiki/WRUsers

# FELIX hardware: White Rabbit integration

#### RabbitFX: a white Rabbit Mezzanine for *FELIX*



- → For the FELIX development in ATLAS a TTC (Timing, Trigger and Control) FMC mezzanine was developed by the collaboration in order to interface commercial (development) cards to the TTC system
- → Nikhef has also expertise in the area of the White Rabbit timing protocol
- → Already developed a simple FMC for another (Neutrino) experiment demonstration
- → Develop a new small FMC equipped with the hardware necessary to make a platform "white rabbit ready"
- $\rightarrow$  SFP input (optional)
- → Basically a DAC (Digital To Analog Converter) and a VCXO (Voltage Controlled Xtal Oscillator

Neutrino DAQ workshop CERN 17-05-2016

# Other White Rabbit nodes examples



Central Logic Board (KM3NeT)



KC705+SoftPLL (KM3NeT test)



CUTE-WR (LHAASO)







Neutrino DAQ workshop CERN 17-05-2016

# FELIX firmware: White Rabbit core



http://www.ohwr.org/projects/white-rabbit/wiki/WRReferenceDesign

# the White Rabbit collaboration



A. Borga

Neutrino DAQ workshop CERN 17-05-2016

# FELIX development status

- FLX-709 (Xilinx VC-709) and its software reached a development status sufficient to be distributed to ATLAS Sub-Detectors Front End developers.
- Tests with the BNL FLX-711 successfully accomplished, a second revision with minor bug fixes is planned for the coming months.
- FLXtools for control and monitor of FELIX are in the advanced development stage.
- FELIX application is the development stage. 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.
- Ongoing effort to increase the number of input channels supported.
- Ongoing effort to implement more functionalities: control of GBTx ASIC via its IC port, HDLC encoding of E-link data to/from the SCA ASIC (for DCS), routing of data From-Host.
- Final Design Review planned for October 2016.
- FELIX is not (yet) a "product"!

# FELIX reference documents

- FELIX Preliminary Design Review: <u>https://indico.cern.ch/event/395378/</u>
- DEBS2015 paper: http://cds.cern.ch/record/2014753
- TWEPP2015 paper [JINST 11 C01055]: http://iopscience.iop.org/article/10.1088/1748-0221/11/01/C01055/pdf
- <u>http://www.ohwr.org/projects/white-rabbit/wiki</u>
- <u>http://www.ohwr.org/projects/white-rabbit/wiki/WRUsers</u>



# backup slides

# FELIX development cards: FLX







#### FLX-710 (FELIX)

- HiTech Global HTG-710
- Virtex-7 X690T
- PCle Gen 3 x 8 lanes
- 2x12 bidir CXP connectors
- FMC connector

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

#### TTCfx

- Custom FMC accepting TTC input
- Outputs TTC clock and CH A-B info
- V1: ADN2814 + CDCE62005

Neutrino DAQ workshow2: ADN2814 + Si5338 JV, Münich Muon Week, 15-Oct-2015 CERN 17-05-2016

A. Borga

ļ

39

# FELIX PC: motherboard and NIC



OS: SLC6



Supermicro motherboards, e.g.:

#### SuperMicro X9SRL-F

(Nikfef)

- 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

#### 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

Neutrino DAQ workshop CERN 17-05-2016

### FELIX To-Host data boundary format

FELIX Chunk trailer format			04-Jul-1	.4												
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Туре		Туре		mod1	mod0											
first	1	1	0	rsrvd	rsrvd		length: 2^11 16-bit words $\rightarrow$ 2^(11+1) bytes = 4KB									
middle	1	0	0	rsrvd	rsrvd											
last, even	1	0	1			lengt										
both, even	1	1	1		Trunc-1	validity refers to this chunk only, i.e. CRC error, or bad 8b/10b encoding										
last, odd	0	0	1	valiu=1	Trunc=1											
both, odd	0	1	1													
		0	0	0	null=0	lengt	h: 2^11	16-bit w	vords, e	xcl this <b>v</b>	I this word; length=0 means only this word					
control	0			0	err=1	error flags, e.g. GBT errors, loss of link sync, etc.										
CONTION				1	rsrvd=0											
				1	rsrvd=1											
reserved	0	1	0													

### FELIX GBT-FPGA



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



**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	SID C H K	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	LEN SID	NOT recommended. Packet boundaries lost if bits lost on an E-link. CHK: an optional check-sum
framed	SOP SID 8b/10b encoded $\begin{bmatrix} C \\ H \\ K \end{bmatrix}$ EOP	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

# FELIX flx-tools

#### Collection of tools to control, monitor, configure different FLX cards.

flx-config	<b>flx-config</b> : access to configuration registers of the card:
flx-dma-test	\$ flx-config list
flx-dump-blocks	\$ flx-config set GBT_EMU_ENABLE=1
flx-i2c	\$ flx-config store my_config.dat
flx-info flx-init	\$ flx-config load my_config.dat
flx-irq-test flx-spi	<b>flx-init:</b> Initializes an FLX device. No output = everything is fine.
flx-reset flx-throughput	<b>flx-info:</b> Display various status information of an FLX device \$ flx-info GBT
	\$ flx-info CXP
	<b>flx-dump-blocks:</b> Dump blocks from FLX into a file for analysis. Example: Dump 100 1kB-blocks
	\$ flx-dump-blocks –n 100 –f output.blocks

See '-h' for full list of options. \$ flx-reset -h

# FELIX FPGA: internal data multiplexing to PCL



#### **FELIX** Operation modes implementation TTC network 40Gb/s

- GBT Normal mode (with Forward Error Correction, 3.2Gb/s)
- GBT wide mode (112 bit, 4.48Gb/s)
- "Full mode" (9.6GB/s)
- TTC distribution

fiber

Busy

FELIX

(N GBT channels)

FPG/

other' link 🗸

... A ...

GBTx

#### GBT mode, 3.2 Gb/s (80 bit @ 40MHz), 5 bidirectional E-groups

- each E-group has up to 8 E-links
- each E-group rate is up to 320 Mb/s (8bit @ 40MHz)
- available E-link rates: 80, 160, 320 Mb/s (640 Mb/s)





#### GBT wide mode, 4.48 Gb/s (112 bit @ 40MHz), 7 to-Host, 3 From-Host E-groups

- each E-group has up to 8 E-links
- each E-group rate is up to 320 Mb/s (8bit @ 40MHz)
- available E-link rates: 80, 160, 320 Mb/s (640 Mb/s)



![](_page_48_Figure_0.jpeg)

#### To-Host Full mode, ~9.6 Gb/s (under discussion), From-Host: standard GBT links

- single E-link at full link bandwidth, can use 8b/10b encoding
- supports BUSY-ON and BUSY-OFF symbols (K-characters)
- From-Host "IC" link is used for flow control, XON and XOFF

![](_page_48_Figure_5.jpeg)

![](_page_49_Figure_0.jpeg)

#### TTC mode (FEC), 3.2 Gb/s (80 bit @ 40MHz), 5 From-Host E-groups

- each E-group has up to 8 E-links
- each E-group rate is up to 320 Mb/s (8bit @ 40MHz)
- available E-link rates: 80, 160, 320 Mb/s

![](_page_49_Figure_5.jpeg)

### Digital Dual Mixer Time Difference (DDMTD)

![](_page_50_Figure_1.jpeg)