A new approach to interfacing on-detector electronics

B. Gorini[1], D. Francis[1], F. Lanni[2], L. Levinson[3], N. Neufeld[1], J. Vermeulen[4]

[1] CERN, [2] BNL, [3] Weizmann Institute, [4] Nikhef

✱ **Start reading here.**
Arrows point to extra info.

Goals

- Define a readout architecture with the best opportunities to take advantage of future advances in electronics and information technology:
  - Many, many core CPUs, GPUs, Intel PHI, FPGA co-procs
  - N x 100Gb/s networking
  - SW tools for parallelization
- Reduce dependence on custom electronics and its inherent problems of long term support by experts.
- Encourage migration from custom HW → COTS HW → SW.
- Define an architecture that is scalable and flexible, since not known exactly what future requirements will be.
- Provide a communication path for TTC, detector control, configuration and monitoring

Advantages

- Read out topology can be configured to match the bandwidth and computing power required by the actual detector occupancy, and the performance demanded to handle the event request rate by the High Level Trigger.
- Allows the endpoint builder to concentrate on the end-point functionality
- The off-detector endpoints need not implement GBT hardware interfaces.
  This means that they may not need to include FPGAs.
- A significant part of the read out logic would be shared and maintained centrally.
- Enables an easily scalable read out system
- Off-detector endpoints do not need to be mapped one-to-one to the geographical areas serviced by a GBT.
- Any GBT can be connected to any FELIX, just reconfigure
- Can upgrade to faster GBT, faster COTS network independent of data processor upgrades

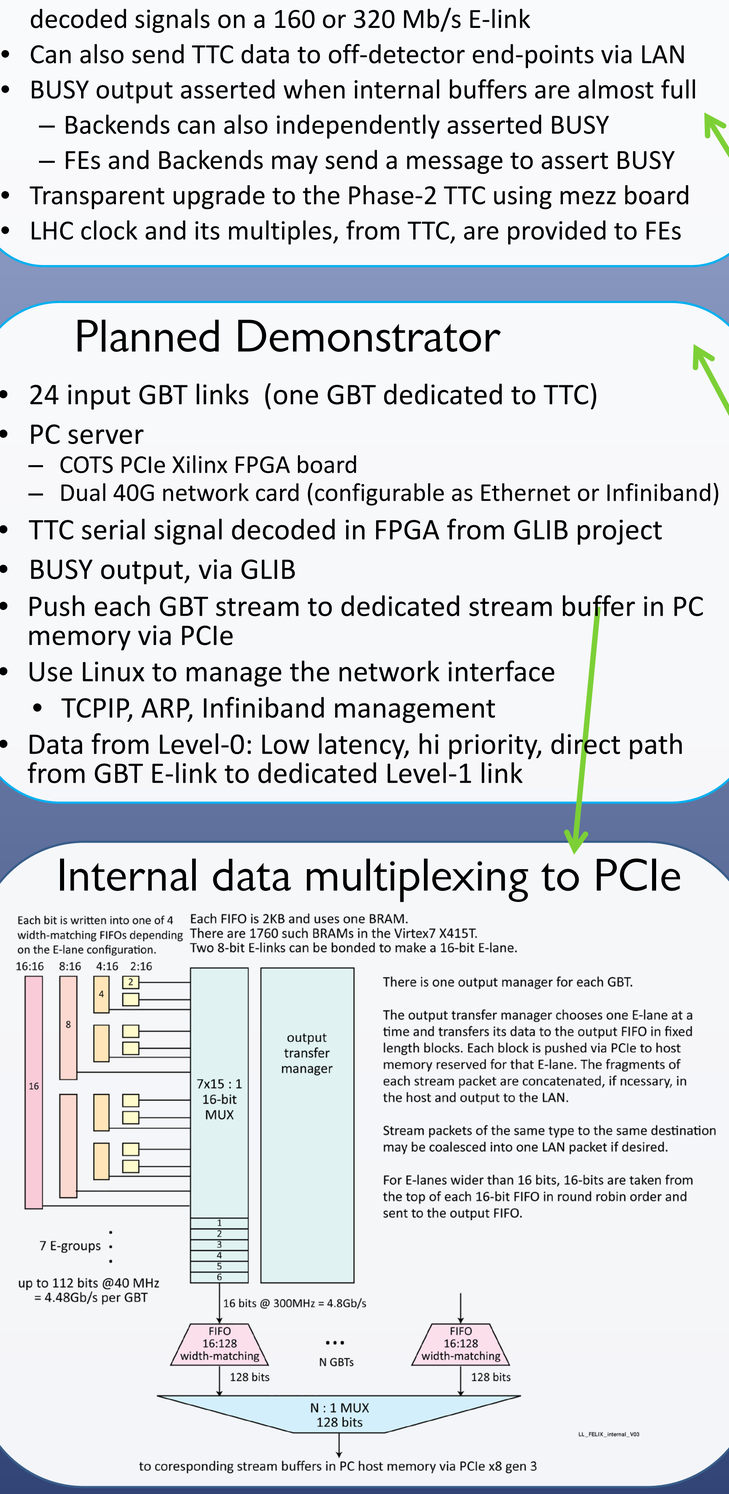
TTC and BUSY

- Dedicated connection from TTC to send TTC data to GBTs
- Chan A & B on an 80Mb/s E-link, or
  decoded signals on a 160 or 320 Mb/s E-link
- Can also send TTC data to off-detector end-points via LAN
- BUSY output asserted when internal buffers are almost full
  - Backends can also independently asserted BUSY
  - FEs and Backends may send a message to assert BUSY
- Transparent upgrade to the Phase-2 TTC using mezz board
- LHC clock and its multiples, from TTC, are provided to FEs

Planned Demonstrator

- 24 input GBT links (one GBT dedicated to TTC)
- PC server
  - COTS PCIe Xilinx FPGA board
  - Dual 40G network card (configurable as Ethernet or Infiniband)
- TTC serial signal decoded in FPGA from GLIB project
- BUSY output, via GLIB
- Push each GBT stream to dedicated stream buffer in PC memory via PCIe
- Use Linux to manage the network interface
  - TCPIP, ARP, Infiniband management
- Data from Level-0: Low latency, hi priority, direct path from GBT E-link to dedicated Level-1 link

Internal data multiplexing to PCIe

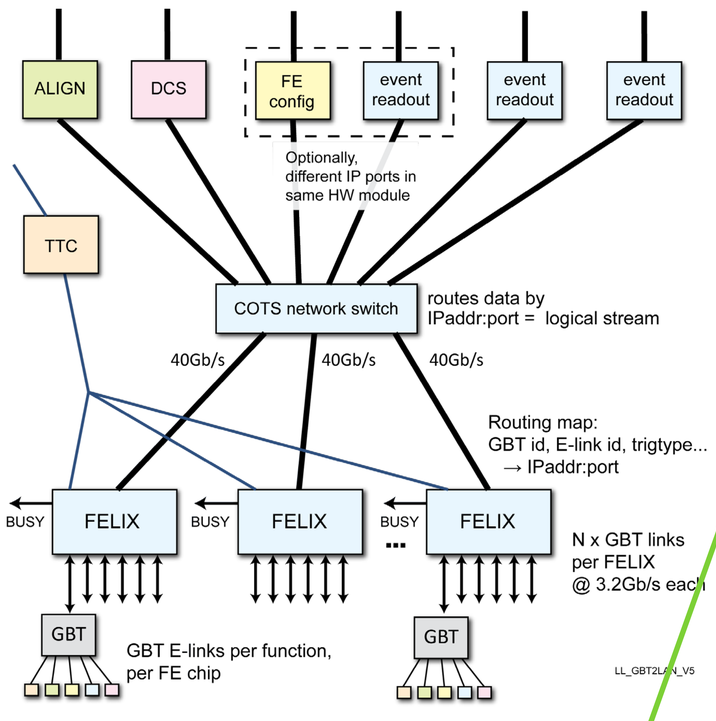


FELIX – Front End Link eXchange

FELIX, a new approach to interfacing on-detector electronics separates the Front End link interfacing from the Front End data processing. It eliminates the static point-to-point connections between the Front Ends and the Read Out system. **Goals**

This allows use of high bandwidth commercial network technology for data aggregation, transport and distribution, while preserving the ability of each detector to perform its specific (and evolving) data processing. COTS PCs may often be used for data processing.

Several custom, radiation-hard links based on the CERN **GBT** ASIC are aggregated to a hi-speed (>40G) Ethernet or Infiniband switched network which provides flexibility, adaptability and upgrade-ability.



GBT links aggregate many slower serial links, called E-links, onto one fast serial link. Different E-links may carry different kinds of logical data, "streams": events, TTC, configuration, calibration, slow control and monitoring, etc. Packets belonging to different **streams** of logical data may flow in the same E-link. A single fibre pair at the Front end can carry all the needed services.

FELIX maps streams to network endpoints; endpoints can then be dedicated to specific functions without regard to the physical link topology. The advantages of interfacing data streams from front end links to a switched network include flexibility in routing and aggregating data, easier load balancing and easy scalability. FELIX is not detector specific, but streams have many configurable options.

Splitting off Detector Control System, DCS, data streams preserves the ability of the DCS to operate independently of the DAQ system.

TTC and **Busy** are special endpoints.

Status:

- Expected to be recommended by ATLAS Readout Upgrade Working Group for deployment after LHC Long Shutdown 3
- Deployment after LHC Long Shutdown 2 for ATLAS Muon NSW readout and LAr upgrade trigger readout
- LHCb is implementing a functionally compatible module which would allow sharing of firmware HDL and higher level software modules.
- A **demonstrator** is being built from a PC server and commercial FPGA and network PCIe cards.

To handle the need for transmitting Regions-of-Interest subsets of the Front End data determined by the **Level-0 trigger** to a Level-1 processor, direct low latency GBT-to-GBT links are provided.

Some **calibration** procedures require commands to be sent to a Front end at a fixed time interval before a calibration trigger. This is provided by transfer lists with guaranteed relative BC clock timing.

Deterministic, but not dynamic, **event routing** is also provided.

Deterministic routing of events

In addition to routing by the source E-link, data can be routed by:

- The low bits of the event-id, for load balancing
  - Being deterministic, the high level trigger knows where to find the event fragments for a given event.
- Trigger type, in order to route events for specific processing or monitoring.

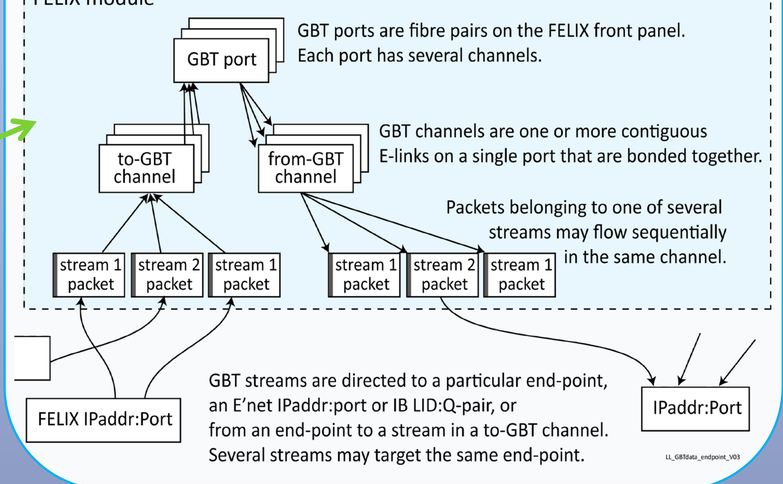
GBT: Gigabit Transfer ASIC

- Developed by CERN
- Radiation hard
- 3.2 Gb/s throughput with error correction, 4.48 without
- Aggregates many slow (80, 160 or 320Mb/s) serial data streams (E-links) onto a single link
- 120-bit frames synchronous with LHC clock, fixed latency
- GBT ASIC provides links in both directions
  ⇒ one GBT can transport event, configuration, control and monitoring streams
- Provision for LHC clock and transfer of timed signals with fixed, adjustable, latency
- Future GBT ASIC planned with double the bandwidth

Stream attributes

- Decode, or only check, input optional 8b/10b encoding
- Variable length or fixed length packets
  - var ⇒ length field or use of 8b/10b start & end symbols
- Opportunistic coalescing of packets with a given stream type destined to the same end-point, or not
- Clone stream to multiple end-points
- TCP stream or UDP datagram
- Multicast or point-to-point
- QoS level
- Checksum, or not, at end of packet

Mapping data between GBT streams & network end-points



Direct, low latency paths for trigger

- For transferring Lvl-0 data to Lvl-1 processors, a Direct Output Low Latency stream will be routed to an output GBT link rather than a network endpoint.
  - Several input streams may be combined, in parallel, onto one GBT output link dedicated to Level-1 input.
- Similarly a Direct Input Low Latency stream is used to inform FEs which data is requested for Lvl-1 processing.
- The transfer is GBT-to-GBT, entirely within the FELIX FPGA board, using "cut-through" routing to minimize latency.

Data transfer lists with guaranteed relative BC clock timing

- Primarily for calibration sequences
- A list of data items with relative delays (in BCs) can be sent to one or more output E-links.
  - There is one list per GBT, stored in FELIX.
- The absolute time when the transmission begins is not defined by the list.
- An activation command can be sent via the network, or via a bit in the 8-bit TTC broadcast packet.
  - With TTC activation, all enabled lists in all FELIX's are activated on the same BC.

Acknowledgement: We would like to acknowledge important requirements and criticisms provided by members of the ATLAS Read Out Upgrade Working Group.

TWEPP 2013, Perugia, Italy