ATLAS DAQ

Jinlong Zhang

Argonne National Laboratory

CPAD Instrumentation Frontier Meeting, October 5-7, 2015, University of Texas at Arlington

Introduction

- ATLAS DAQ system evolution
- New readout architecture
- Front-End LInk eXchange (FELIX)
- Future system capacity
- Summary

ATLAS TDAQ Operating Parameters

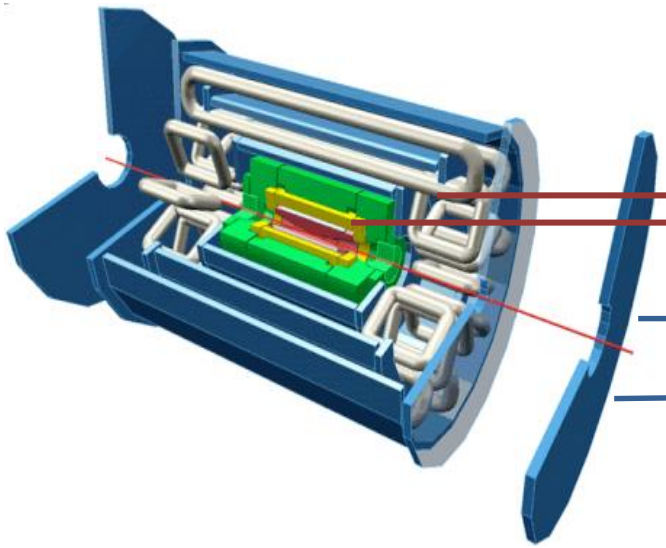
	# Trigger levels	Rates (kHz)		Event Size (MB)	Network Bandwidth (GB/s)	Storage	
						GB/s*	kHz
Run 1	3	L1	75	~1	10	0.5	~0.4
		(L2+EF)	~0.4				
Run 2	2	L1	100	~2	50	1	1
		HLT	1				
Run 3	2	L1	100	~2	50	1	1
		HLT	1				
Run 4	3	L0	1000	~5	2000	~30	10
		L1	400				
		HLT	10				

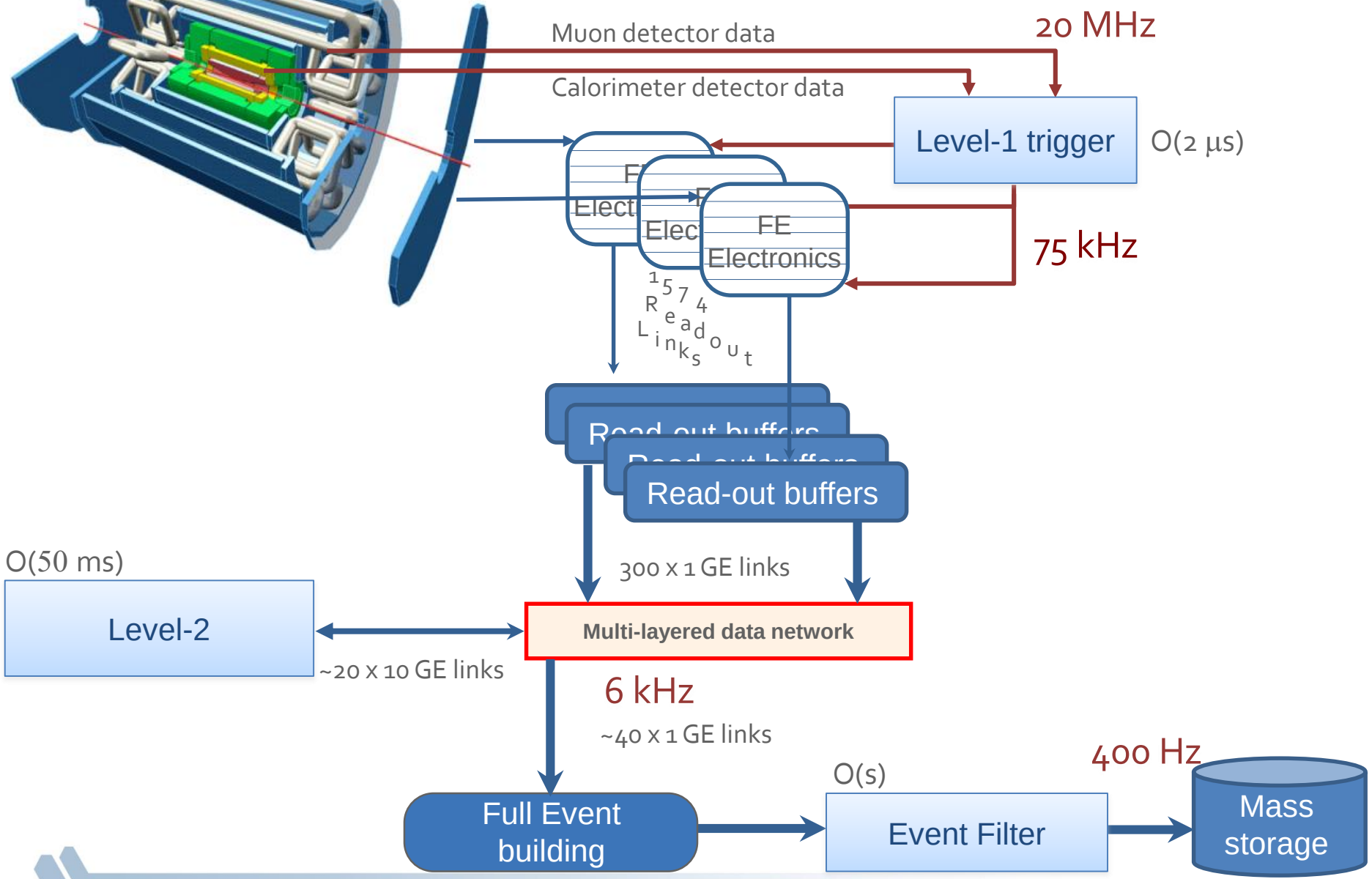
- **Major architecture overhaul for Run 4**

DAQ Terminology in General

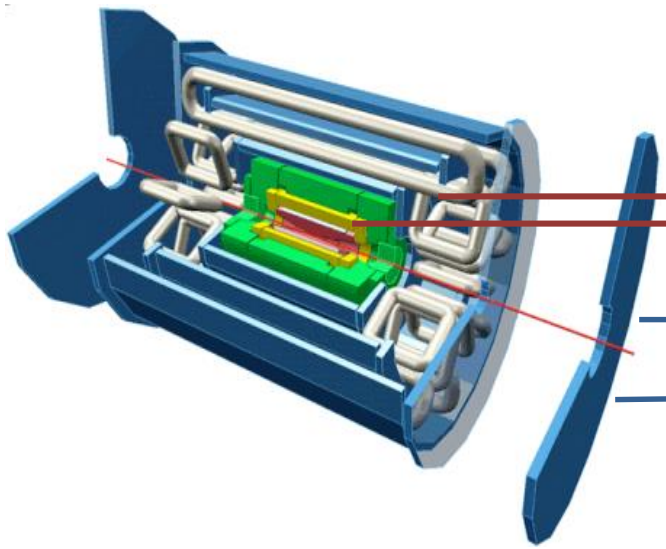
- Data transmission from the detector (Readout System)
- Data aggregation across multiple sources (Event Building)
- Data processing (Event Filter/HLT)
- Data logging (Mass Storage)
- Control (to steer the acquisition) and Configuration (to setup the hardware and software)
- Monitoring for system health and data integrity

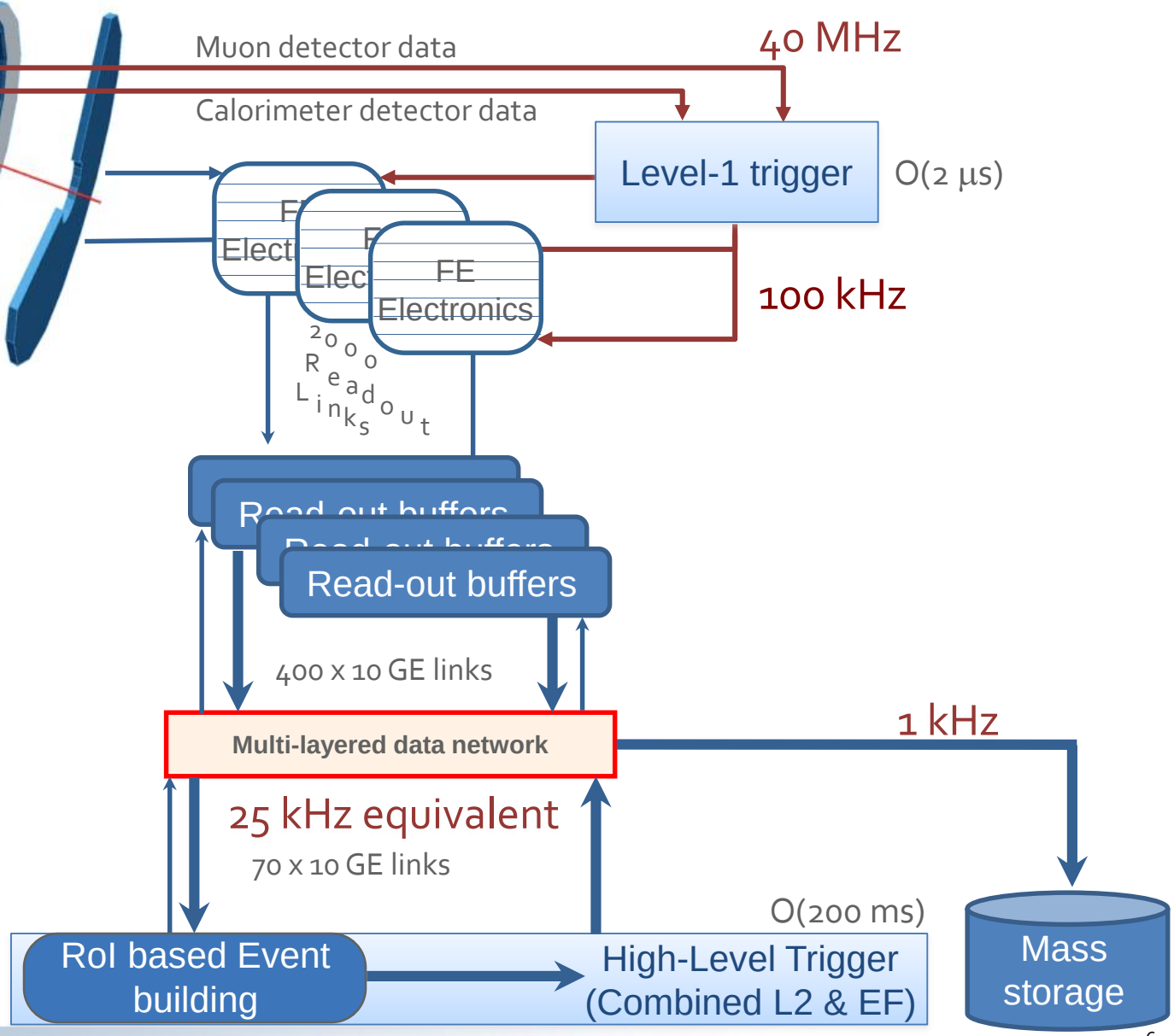
Run 1




Muon detector data
20 MHz
Calorimeter detector data
Level-1 trigger
O(2 μs)
FE Electronics
FE Electronics
FE Electronics
75 kHz
1574 Readout Links
Read-out buffers
Read-out buffers
Read-out buffers
300 x 1 GE links
O(50 ms)
Level-2
~20 x 10 GE links
Multi-layered data network
6 kHz
~40 x 1 GE links
Full Event building
O(s)
Event Filter
400 Hz
Mass storage

Run 2




Muon detector data
40 MHz
Calorimeter detector data
Level-1 trigger
O(2 μs)
FE Electronics
100 kHz
2000 Readout Links
Read-out buffers
400 x 10 GE links
Multi-layered data network
1 kHz
25 kHz equivalent
70 x 10 GE links
O(200 ms)
RoI based Event building
High-Level Trigger (Combined L2 & EF)
Mass storage

Hardware

ROBINNP (ALICE C-RORC)
Xilinx V6 LX240T FPGA
3 QSFP+ transceivers
2 x 4 GB DDR3 1066 MHz RAM
PCIe 8 lane Gen2 interface

ReadOut System (ROS)

~100 PCs, each with
2 ROBINNP
4 x 10GBE

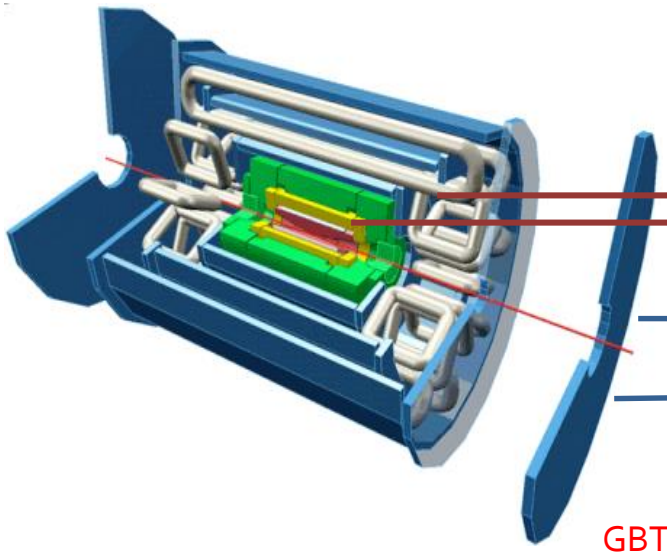
MCT

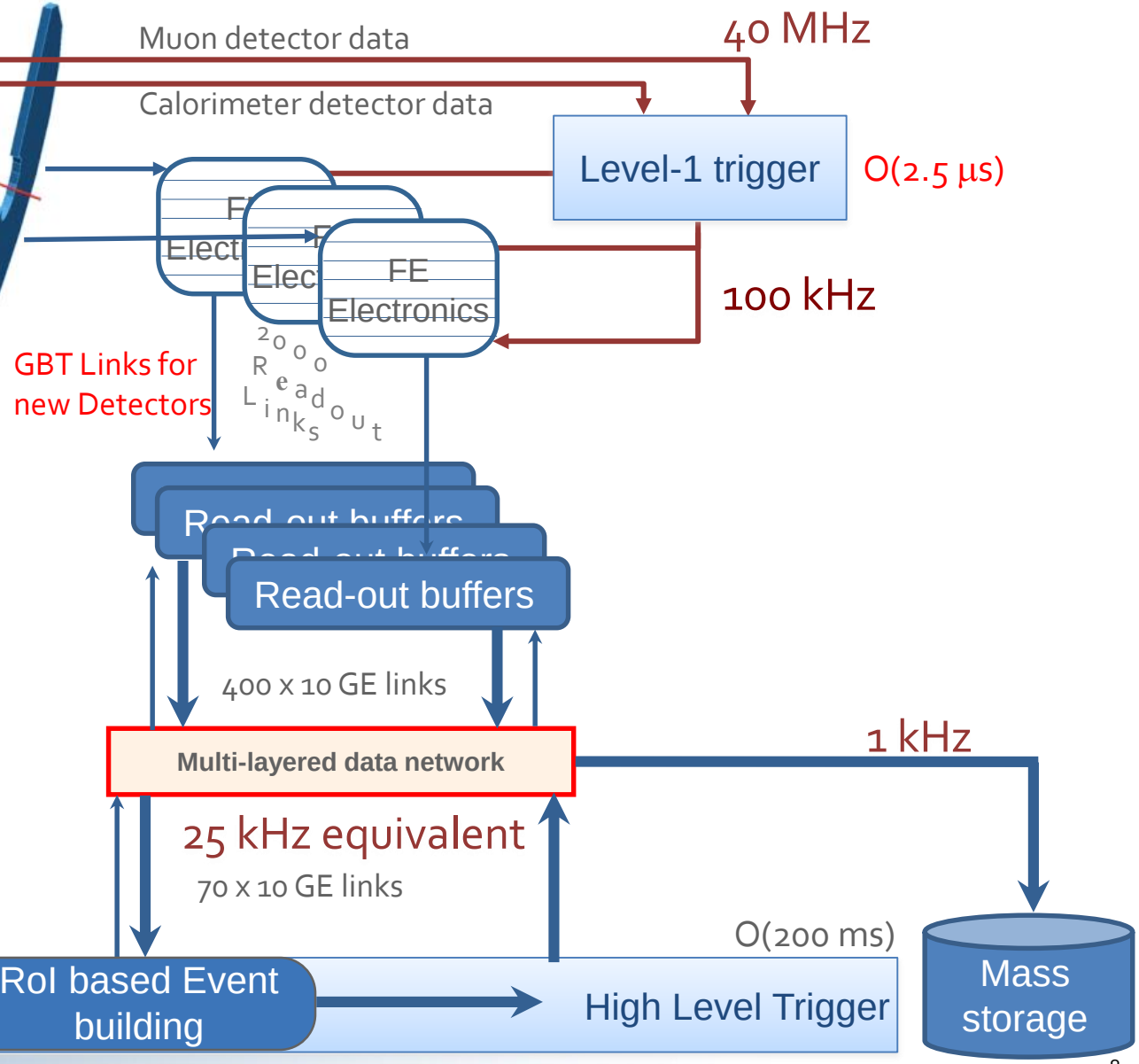
Network (1 GBE, 10GBE)
With
Redundancy &
Multi Chassis Trunking
(Brocade)

HLT Farm

~2000 Multi-core PCs
with rolling replacement

Run 3

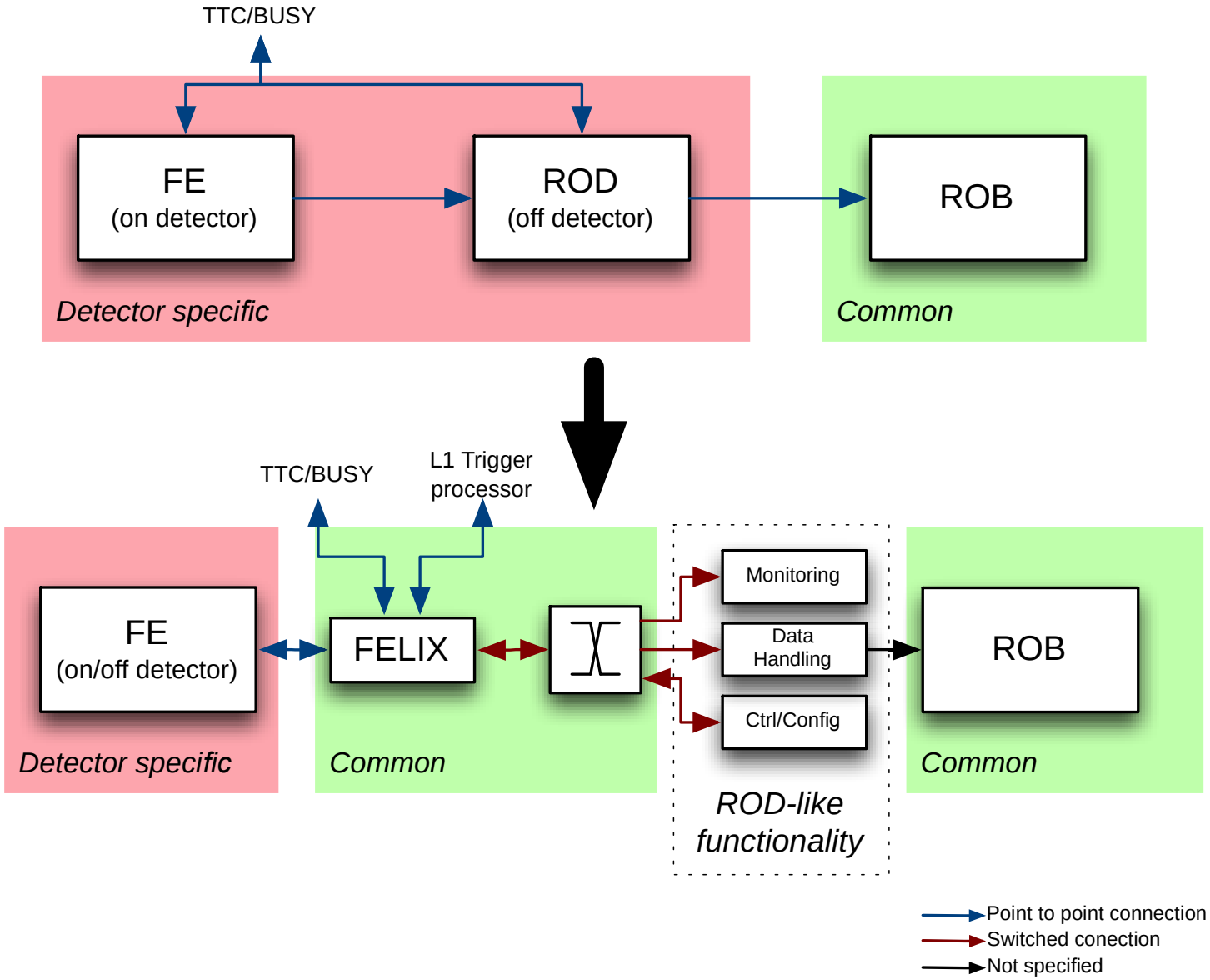



Muon detector data
40 MHz
Calorimeter detector data
Level-1 trigger
O(2.5 μs)
FE Electronics
100 kHz
GBT Links for new Detectors
200 Read-out Links
Read-out buffers
400 x 10 GE links
Multi-layered data network
1 kHz
25 kHz equivalent
70 x 10 GE links
O(200 ms)
RoI based Event building
High Level Trigger
Mass storage

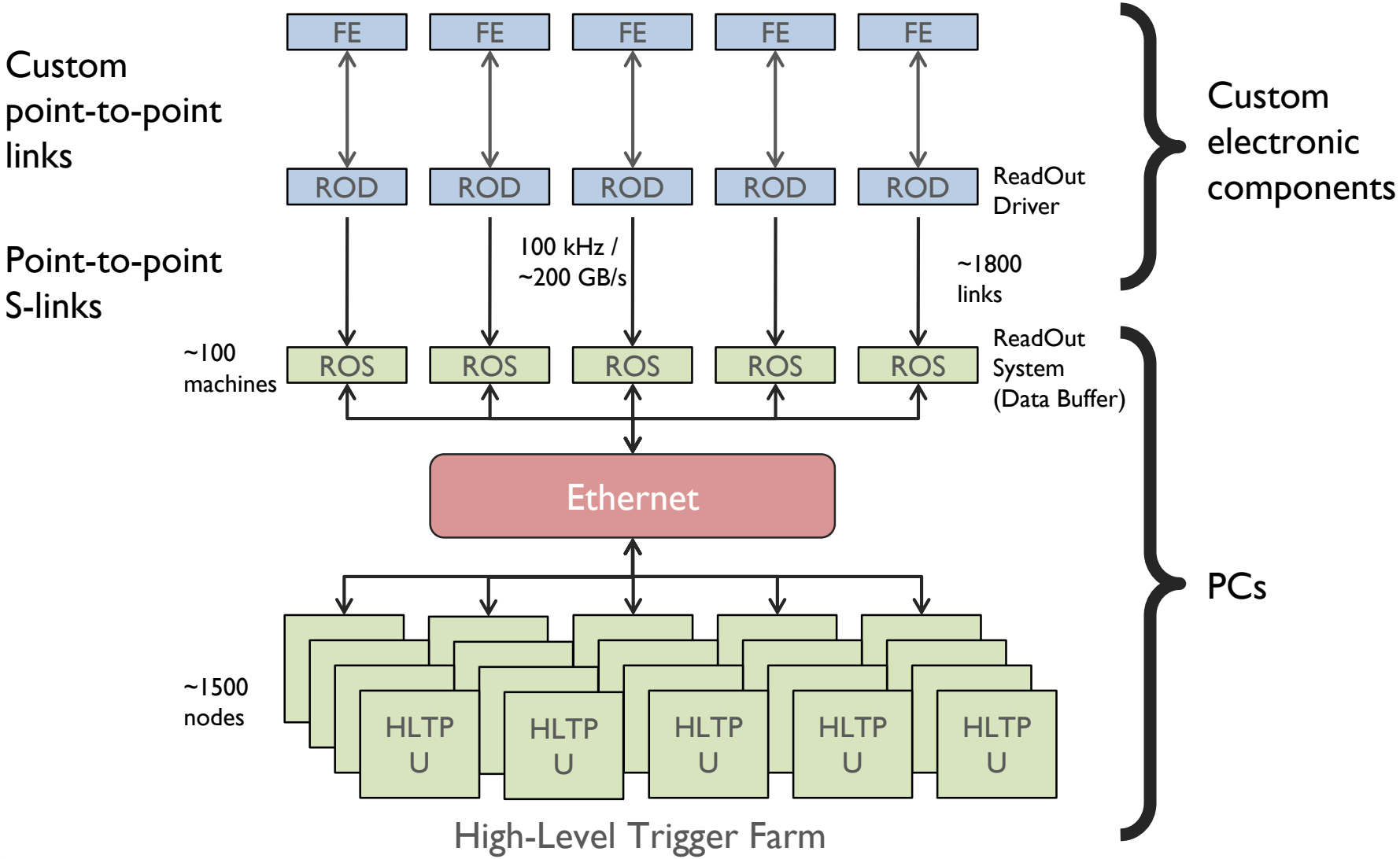
Readout Evolution Motivation

- **Higher level of commonality between detectors**
  - **A common object providing functionalities today implemented in detector-specific back-end custom electronics (ROD)**
- **Increased use of COTS components**
  - **all ROD-like functionality (including data processing) could most likely be implemented in standard computers by Phase-II**
- **Performance scalability built-in**
  - **Programmable connectivity between detector FE and DAQ**
- **Capability to disentangle ROD-like functions from hardware implementation**
  - **Different granularity for monitoring, control, data handling …**
  - **DCS and DAQ traffic separation**

Readout Evolution Schematic



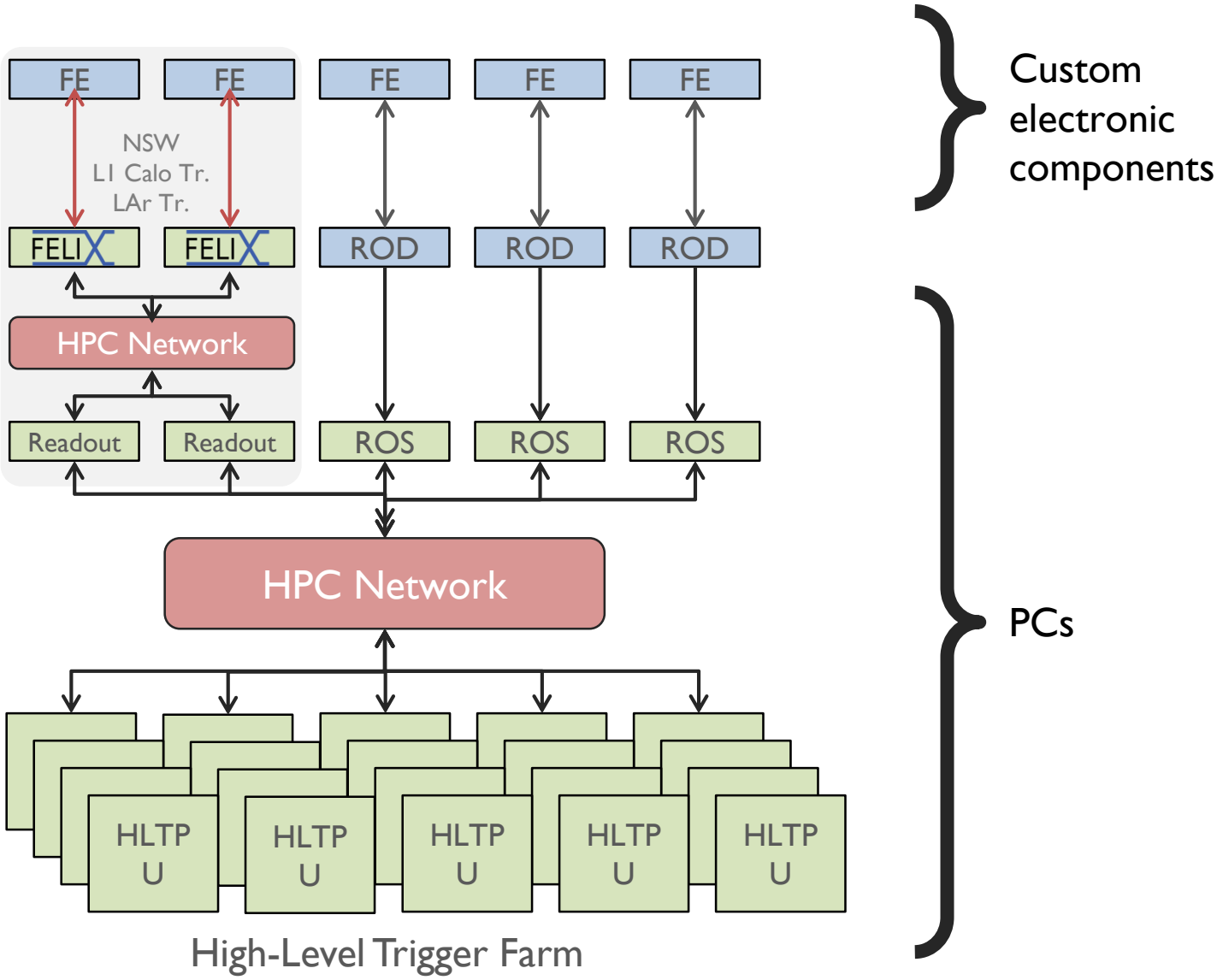
Readout (Present)



Readout (Run 3)

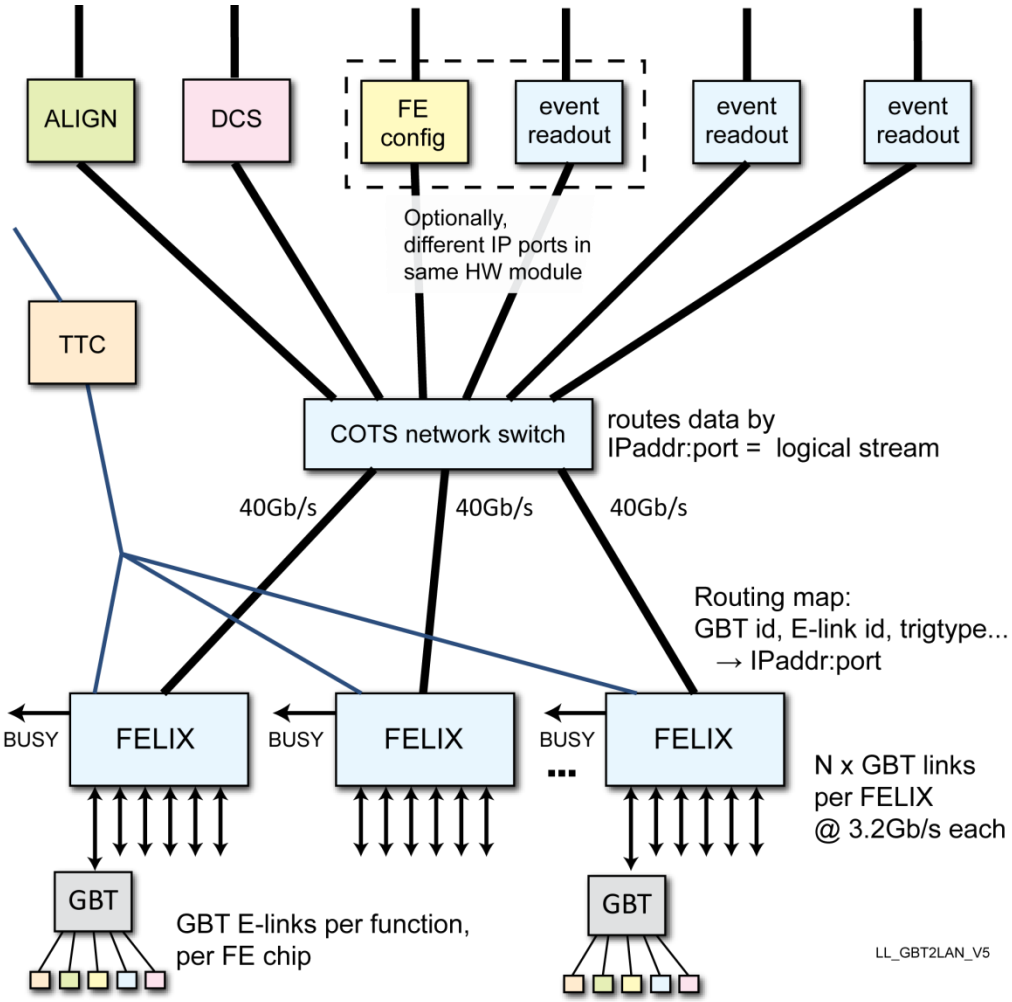
Point-to-point
(GBT, ad-hoc)

40 Gb Ethernet,
Infiniband

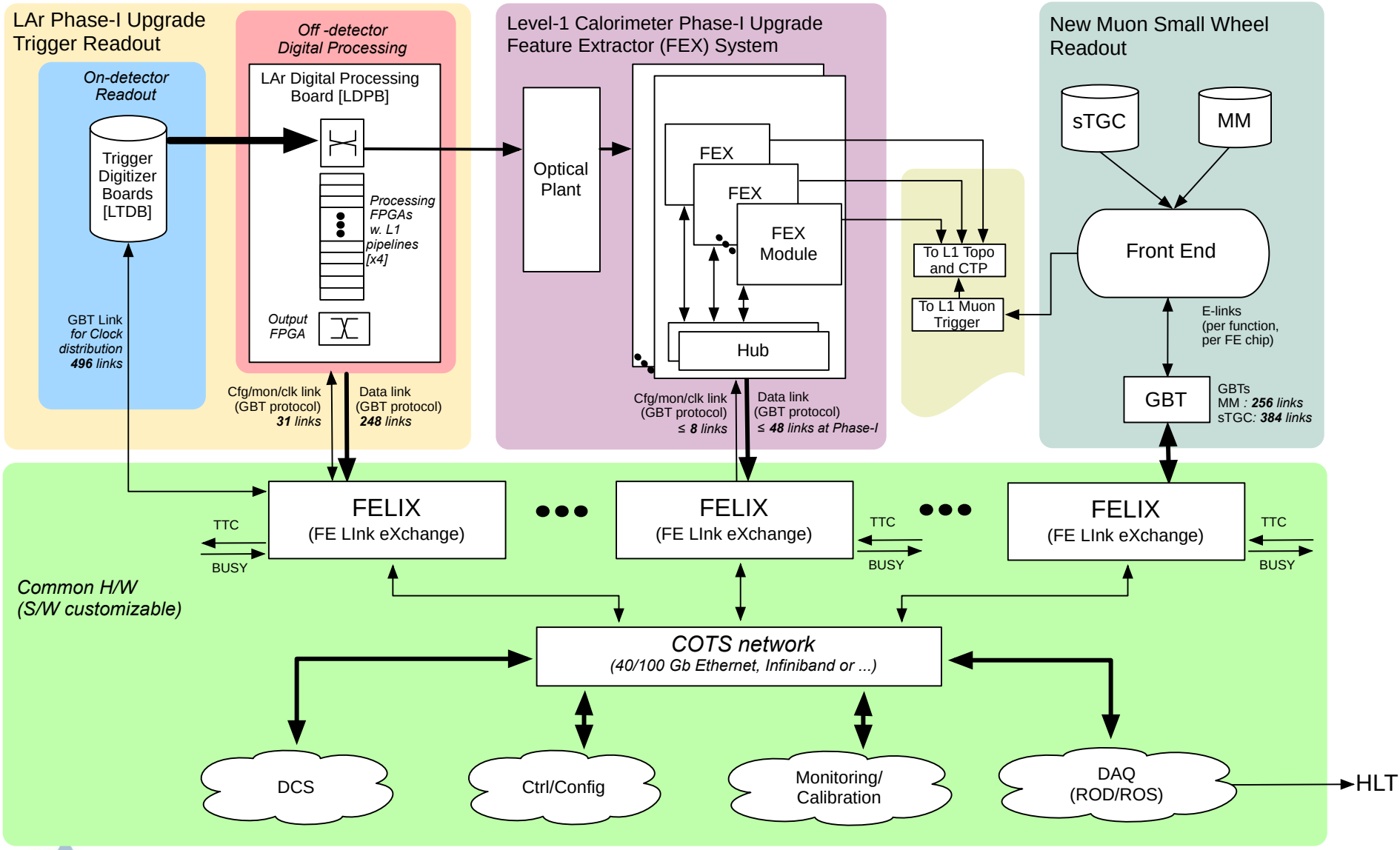


FELIX

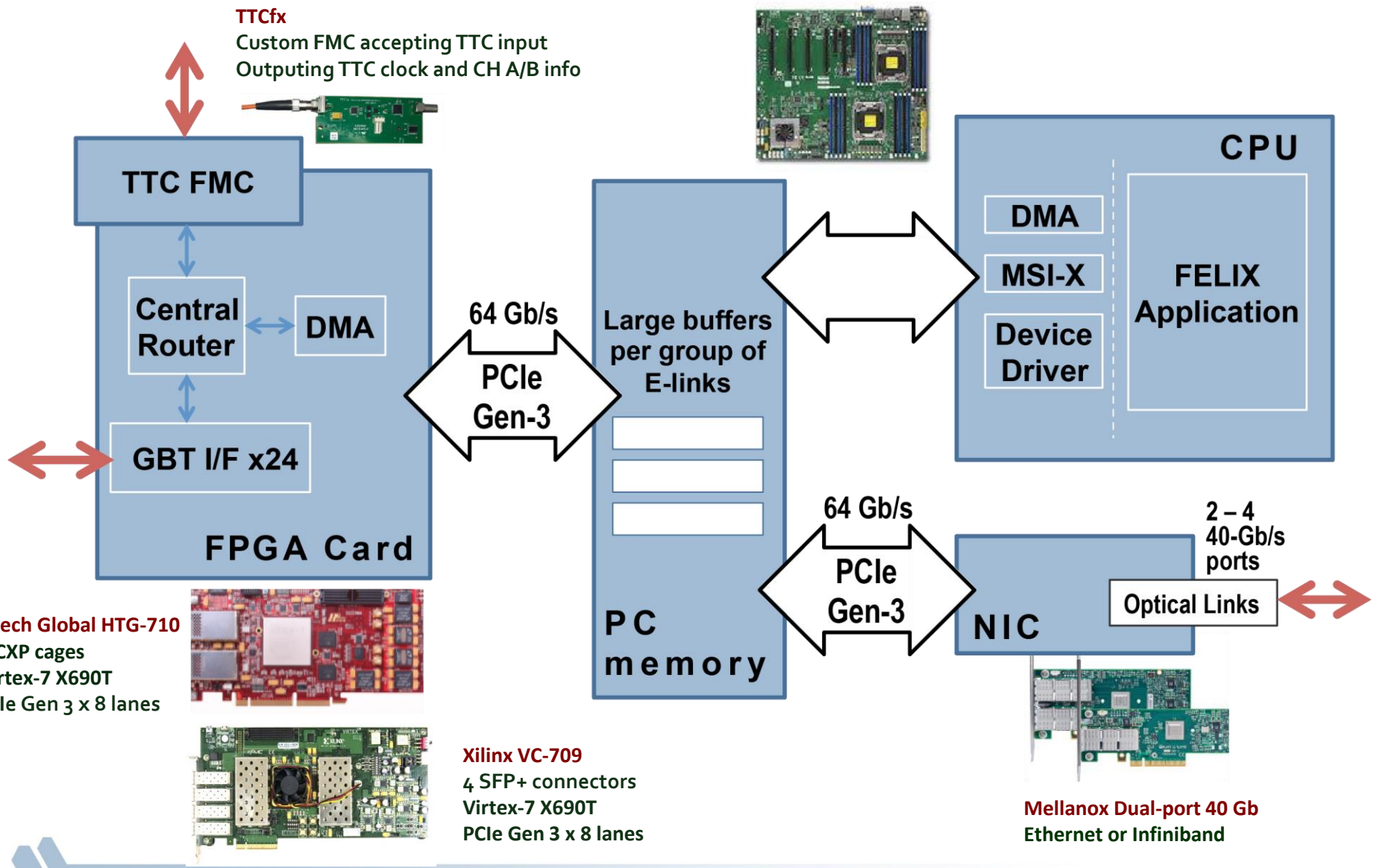
- Enabling transition from custom hardware to COTS as early as possible
- Using high level switch protocols of high speed and large bandwidth
- Configurable and flexible data routing and error handling, without relying on detector specific hardware
- Direct low latency paths between links
- Universal ATLAS-wide TTC/BUSY handling as for now
- Command scheduling with guaranteed timing for calibration



FELIX Deployment in Phase-I

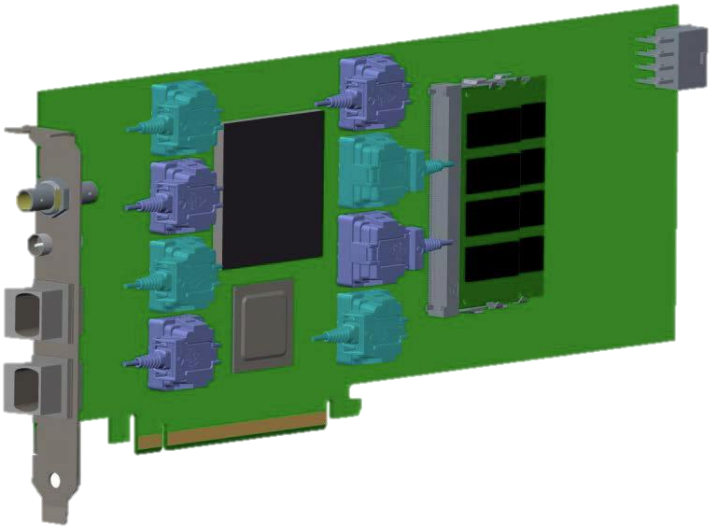


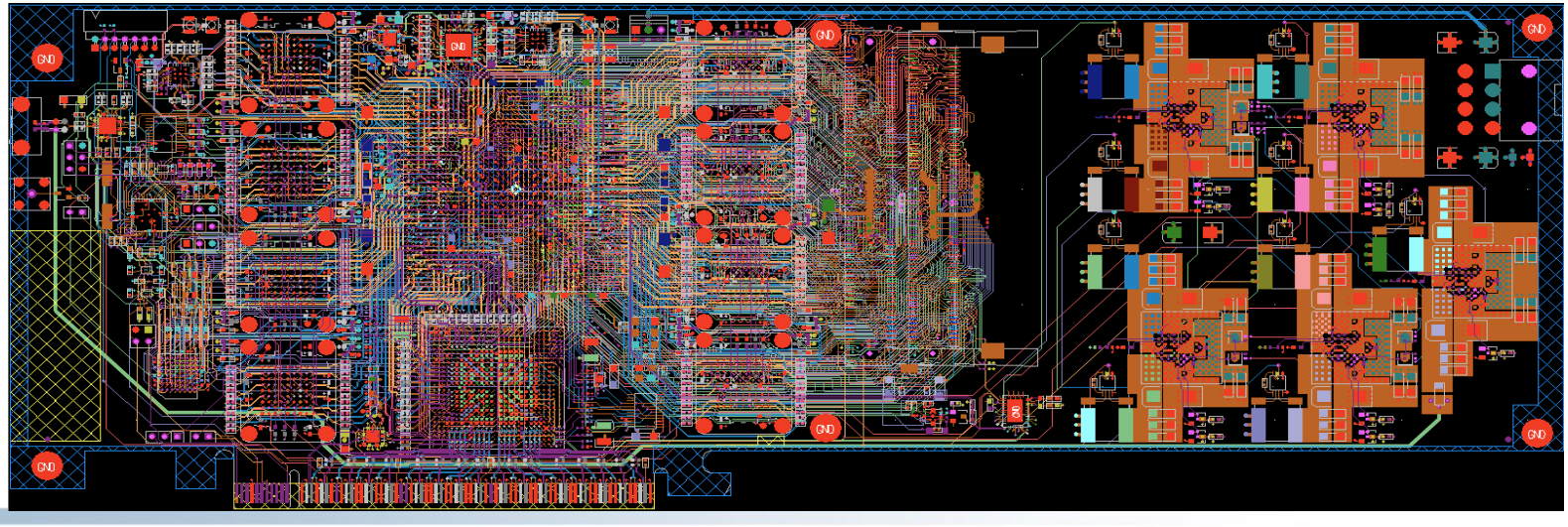
FELIX Development



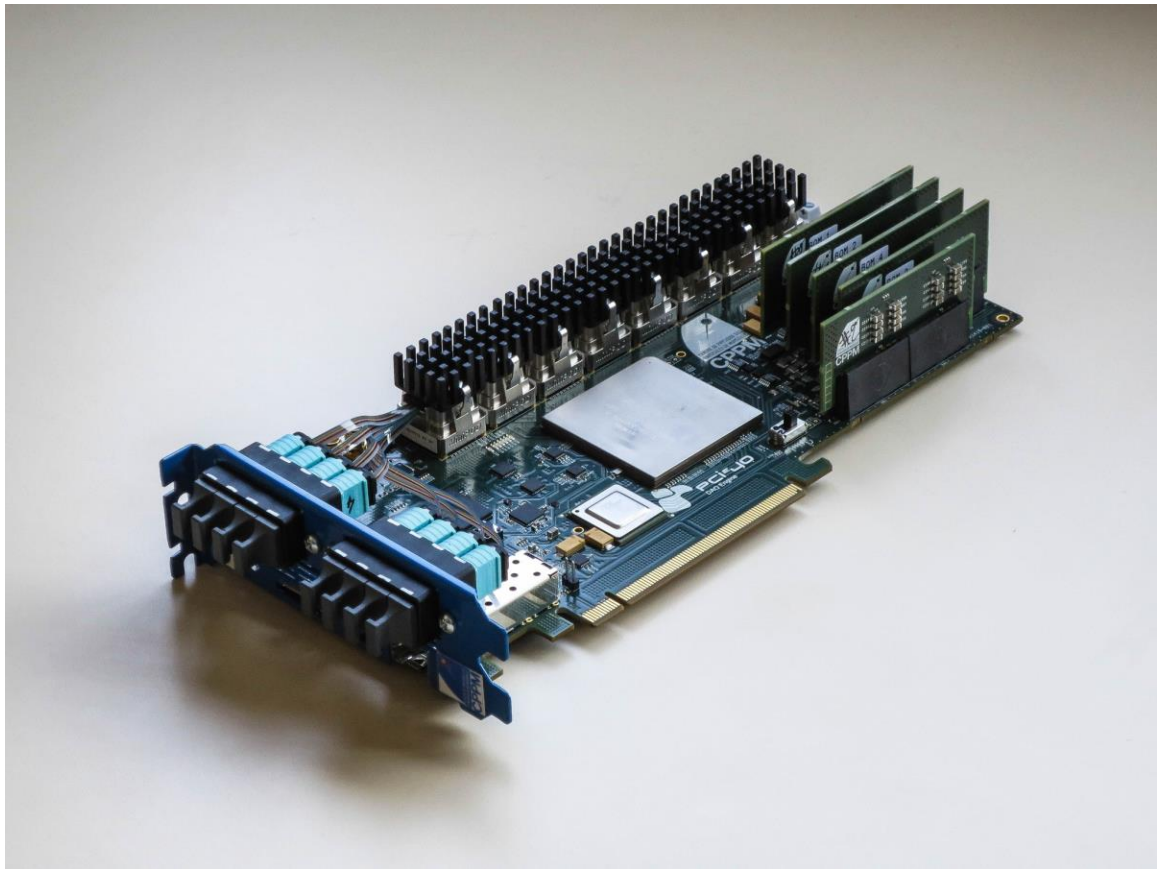
Hardware Candidate

- Xilinx Kintex Ultrascale FPGA
  - 64 GTH transceivers (16.375 Gb/s)
- 4 MiniPOD TX and 4 MiniPOD RX
  - 48 bidirectional optical links (14 Gb/s)
- PCIe Gen3 x16 lanes
- Two DDR4 SODIMM up to 16GB
- Onboard: Clock conditioner, TTC receiver, BUSY lemo output





It Looks Like



This is actually LHCb PCIE40

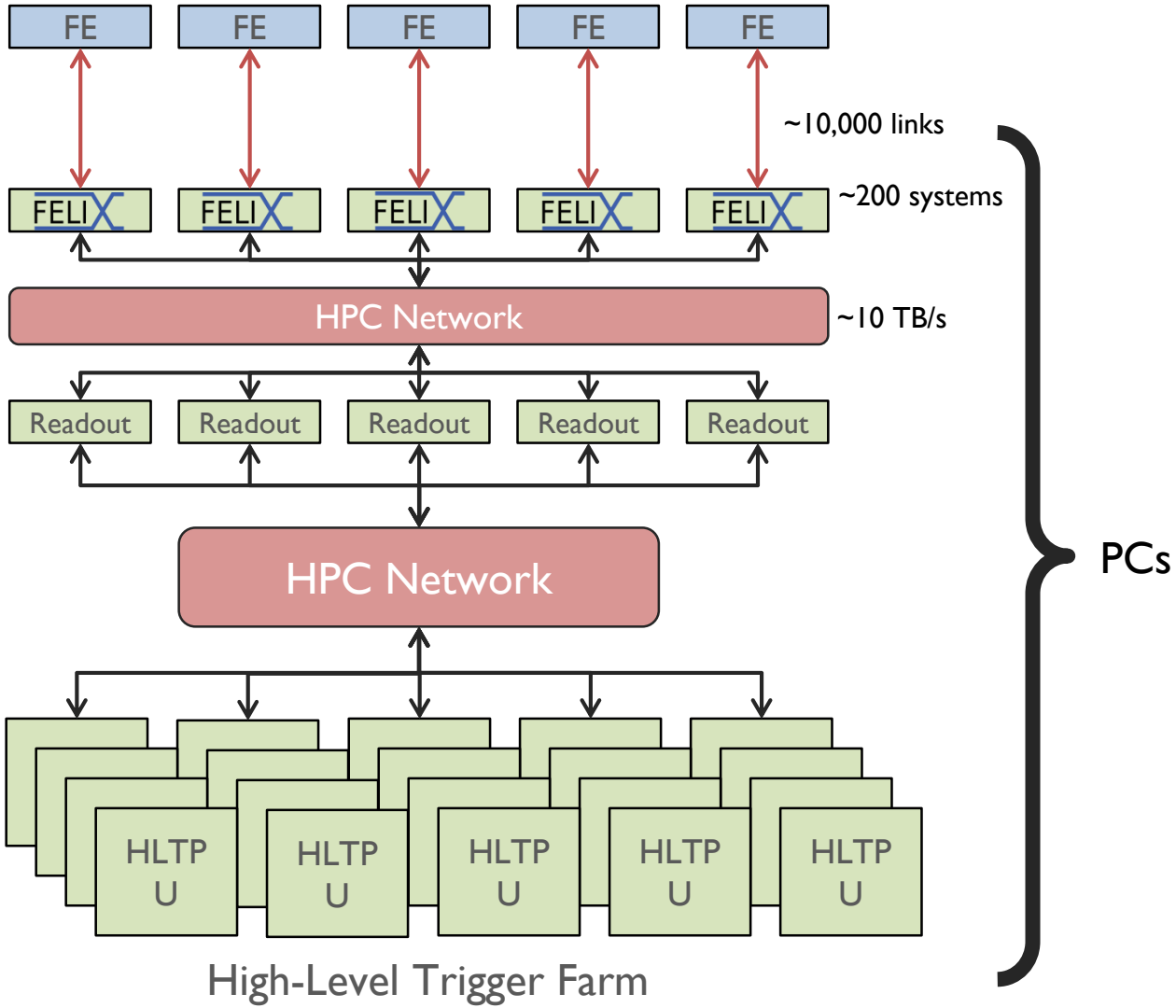
Trends to Future

- Higher and higher trigger rates
  - Triggerless not yet possible (W. Smith's challenge at ECFA workshop 2014)
- PC-based single-stage data aggregation
  - Ethernet or InfiniBand
  - PCIe 4
- Network bandwidth becoming very affordable
  - Changing from the philosophy of "move minimal amount of data"
  - Capability for full event building @ L1A rate (even decouple from HLT)
- Heterogeneous HLT computing (ASIC/FPGAs, GPGPUs, …)
- Tight integration with offline
  - From the blur boundary to the full fusion?
  - Utilization of online resources during non-beam time (e.g., Sim@P1)

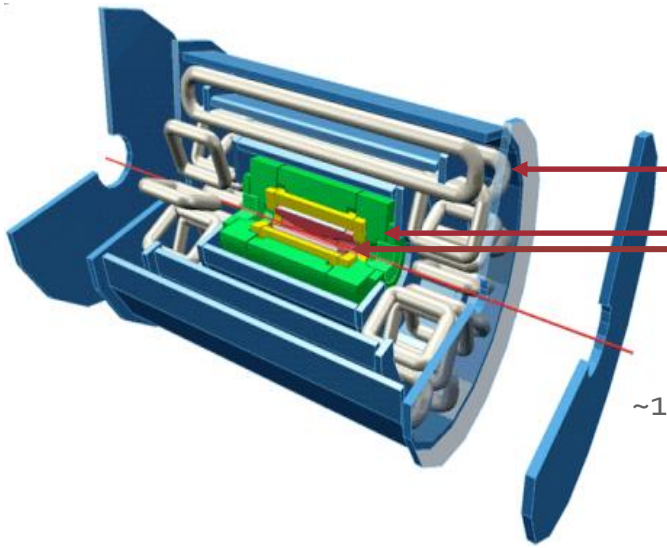
Readout (Run 4)

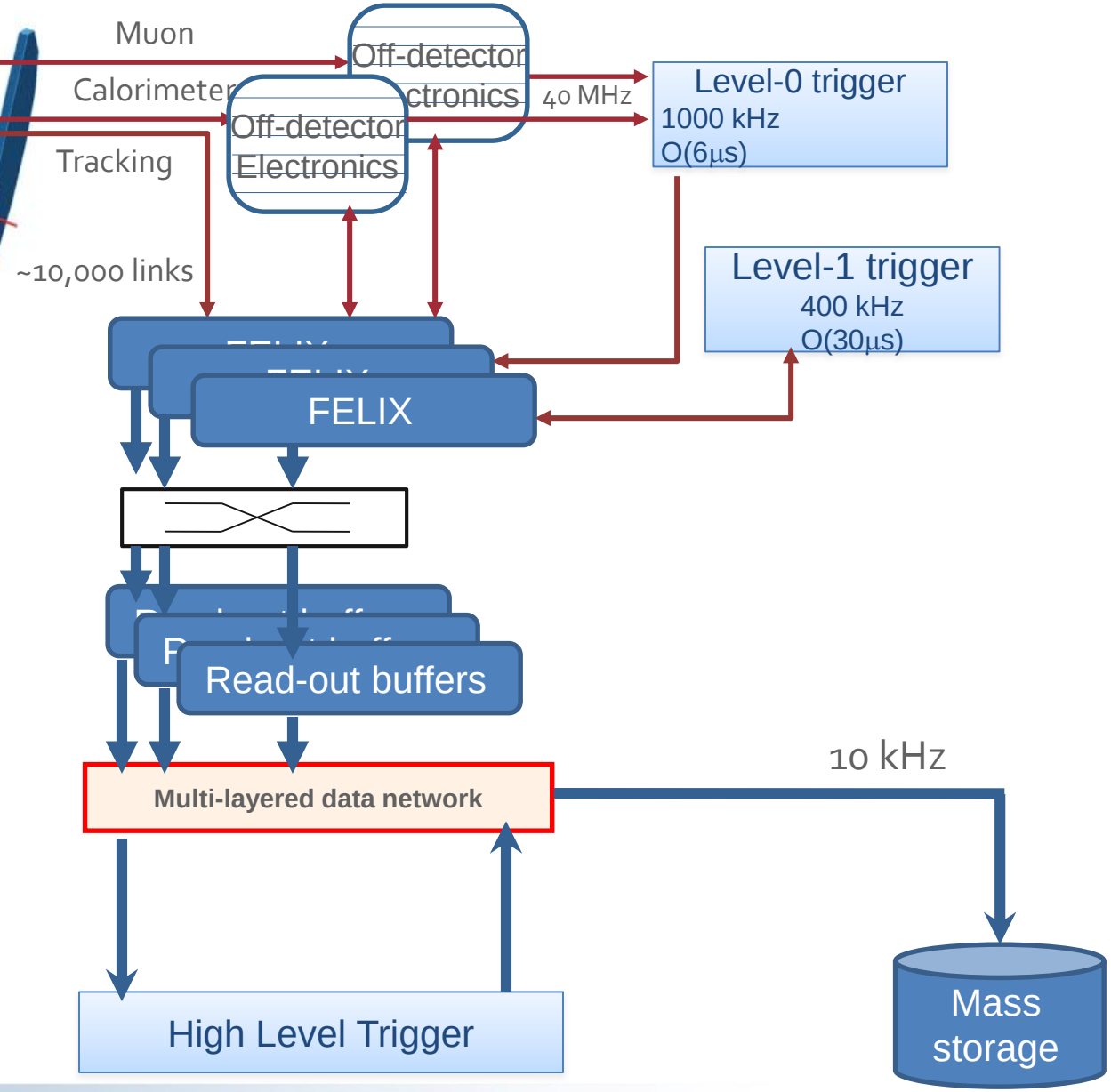
Point-to-point
(GBT, LpGBT,ad-hoc)

COTS network technology



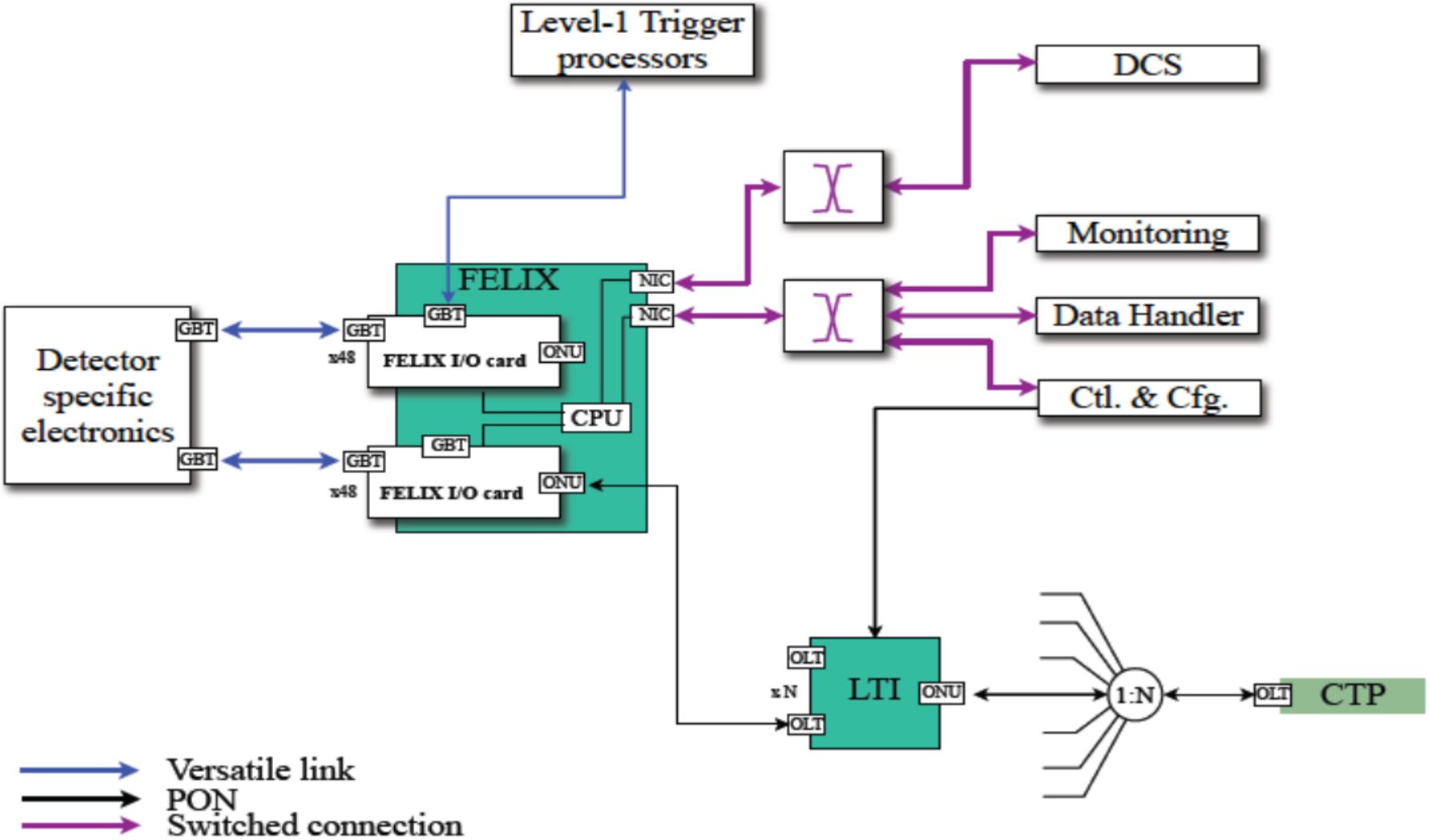
Run 4





Readout Parameters	
Links from detectors	11000
FELIX I/O card	~450
FEXLIX servers	~230
Low latency links	~900
FELIX NICs (100 Gb/s)	~250
"Readout"	~400
"Readout" NICs (100 Gb/s)	~500

Readout in L0/L1 Scheme



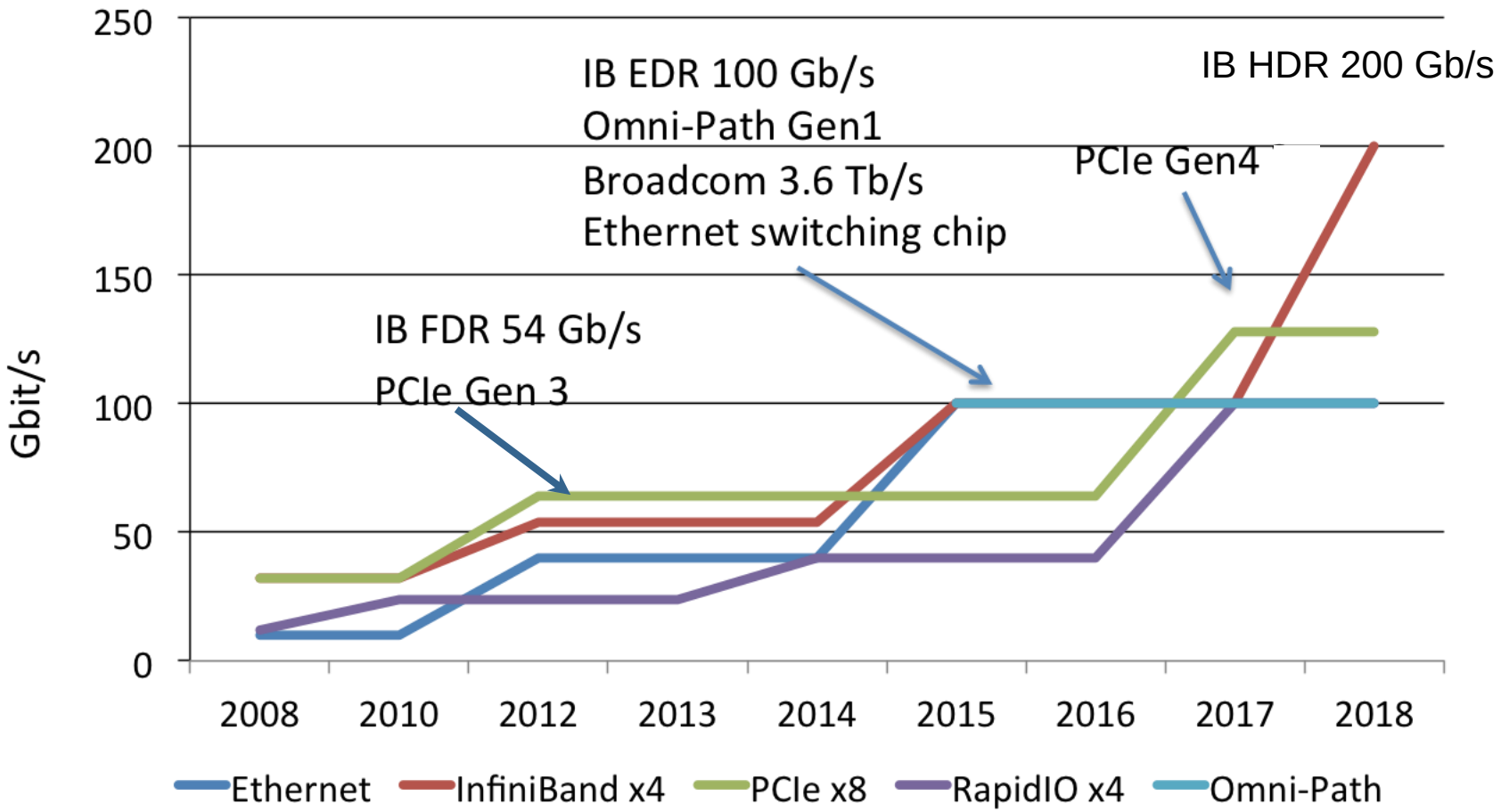
Link to Upstream

http://www.xilinx.com

	Type	Max Performance[1]	Max Transceivers	Peak Bandwidth[2]
Virtex UltraScale+	GTY	32.75	128	8,384 Gb/s
Kintex UltraScale+	GTH/GTY	16.3/32.75	44/32	3,268 Gb/s
Virtex UltraScale	GTH/GTY	16.3/30.5	60/60	5,616 Gb/s
Kintex UltraScale	GTH/GTY	16.3/16.3	64	2,086 Gb/s
Virtex-7	GTX/GTH/GTZ	12.5/13.1/28.05	56/96/16[3]	2,784 Gb/s
Kintex-7	GTX	12.5	32	800 Gb/s
Artix-7	GTP	6.6	16	211 Gb/s
Zynq UltraScale+	GTR/GTH/GTY	6.0/16.3/32.75	4/44/28	3,268 Gb/s
Zynq-7000	GTX	12.5	16	400 Gb/s

- **Readout system will utilize these serDes speeds or faster, so**
- **GBT, even lpGBT (to be used for Phase-II) be modest**
  - **Lightweight protocol being considered in some cases**

Link in Downstream



- **Network for ~500 of 100 GBE links not a problem in 2024 (Phase-II)**
- **PCIe Gen4 expected in later 2017**

Event Filter Computing

- A factor of O(50) in HLT computing power needed wrt to Run 1
  - Processing time extrapolations to pileup of 200
  - Run 4 rejection requirements similar/better than in Run 1-2 (10k/400k vs. 1k/100k)
- Heterogeneous Event Filter farm
  - Full event tracking @ 100 KHz with special hardware (FTK++, GPU, etc)
  - General purpose PCs (performance increase of a factor 10 in ~10 years)
  - Integrated solutions (e.g., Open Compute Project)?

Event Filter	
Processing power	6 MHS06
Racks	80
Mother boards/rack	40

Summary

- Apology for not covering Control & Configuration and Monitoring
- Current ATLAS DAQ system performing well while Phase-II upgrade progressing as planned
  - Initial Design Review in Q1 2016
  - Technical Design Report planned in Q4 2017
- Increased use of Commodity hardware
  - Transit as early as possible from custom rad-hard links to commodity network