ATLAS DAQ Upgrade

Jinlong Zhang

Argonne National Laboratory

CPAD Instrumentation Frontier Meeting, October 8-10, 2016, California Institute of Technology

Introduction

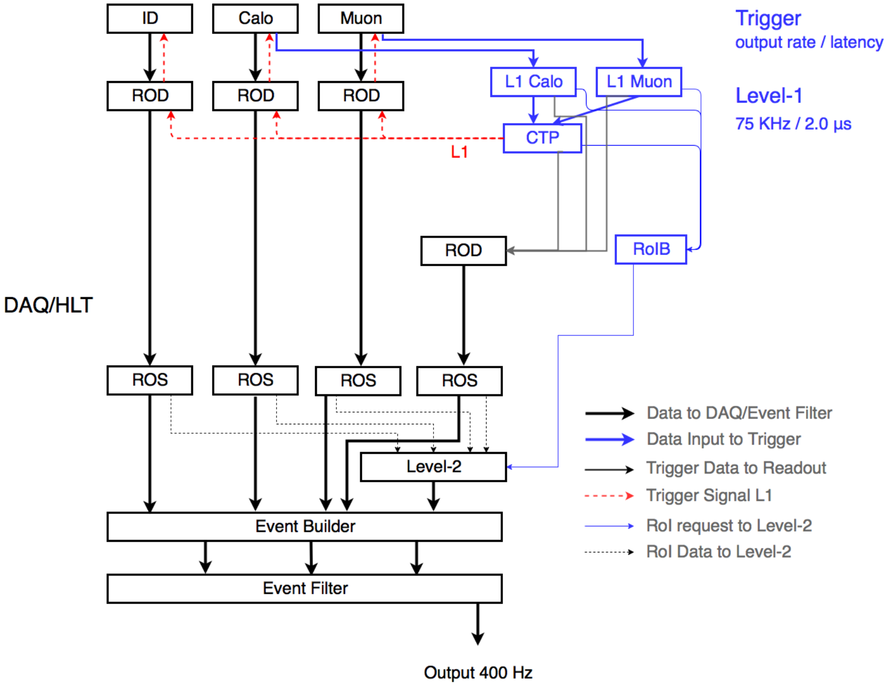
- ATLAS DAQ system
- Evolution and Phase-0 upgrades (Run 1&2)
- Phase-I upgrades (Run 3)
  - —Front-End LInk eXchange (FELIX)
- Phase-II upgrade (Run 4)
  - —Core functionalities
- Summary

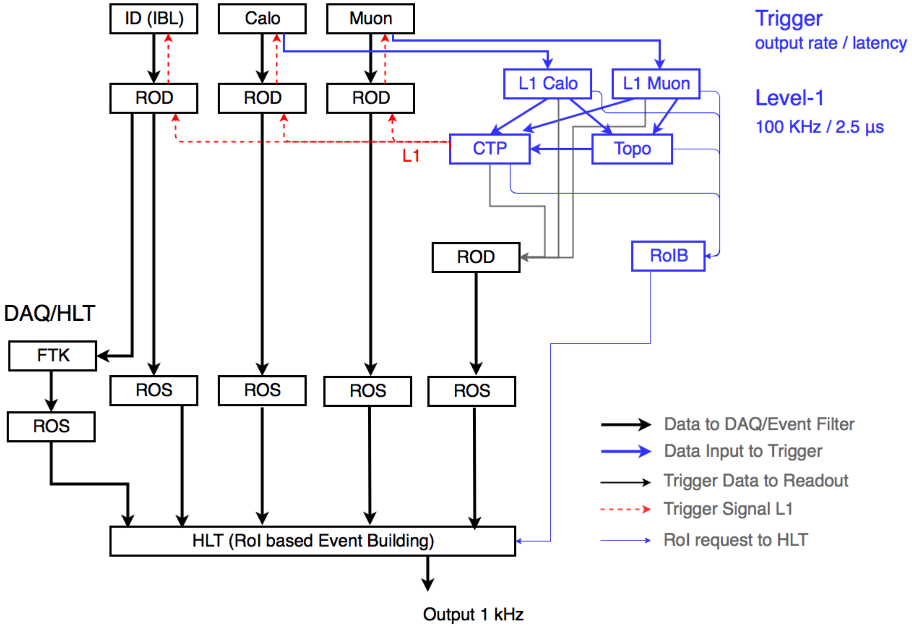
ATLAS TDAQ Operating Parameters

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

- **Major architecture overhaul for Run 4**
  - **Different options being studied**

Run 1 & 2 TDAQ Architecture

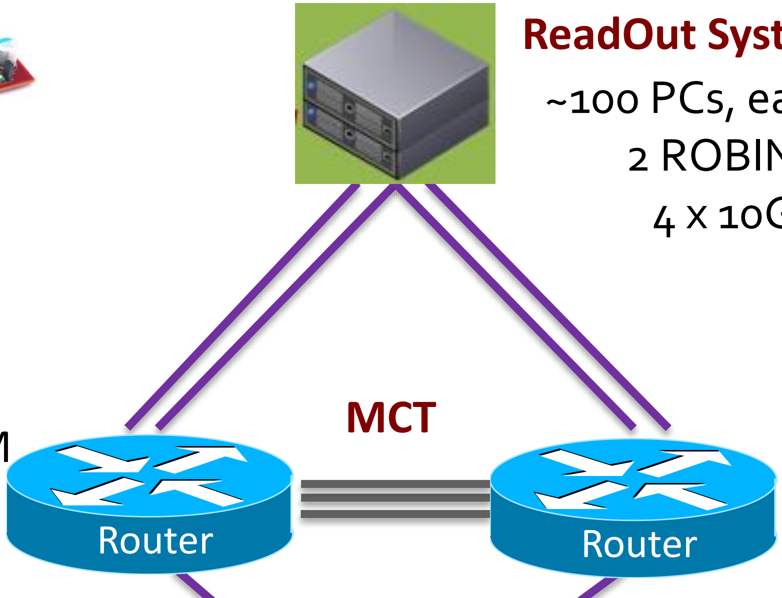




- **Level 2 and Event Filter merged as High Level Trigger (HLT)**
- **Topo trigger, new Central Trigger Processor (CTP), Fast TracKer (FTK)**
- **Readout System (R0S), Region of Interest Builder (RoIB) evolution**

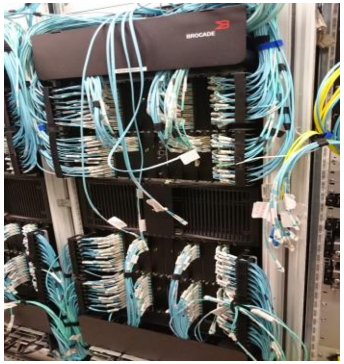
Hardware

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

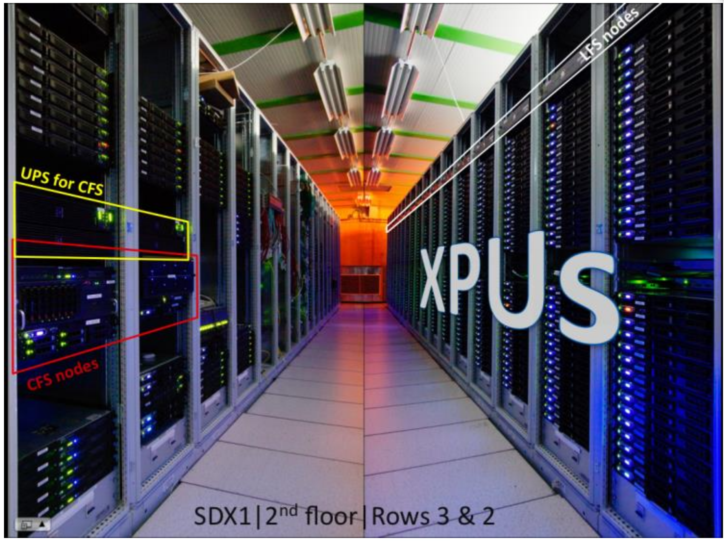


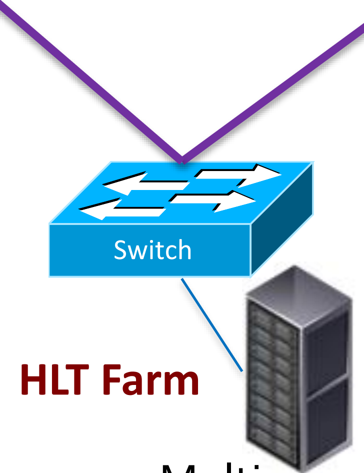
ReadOut System (ROS)

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



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

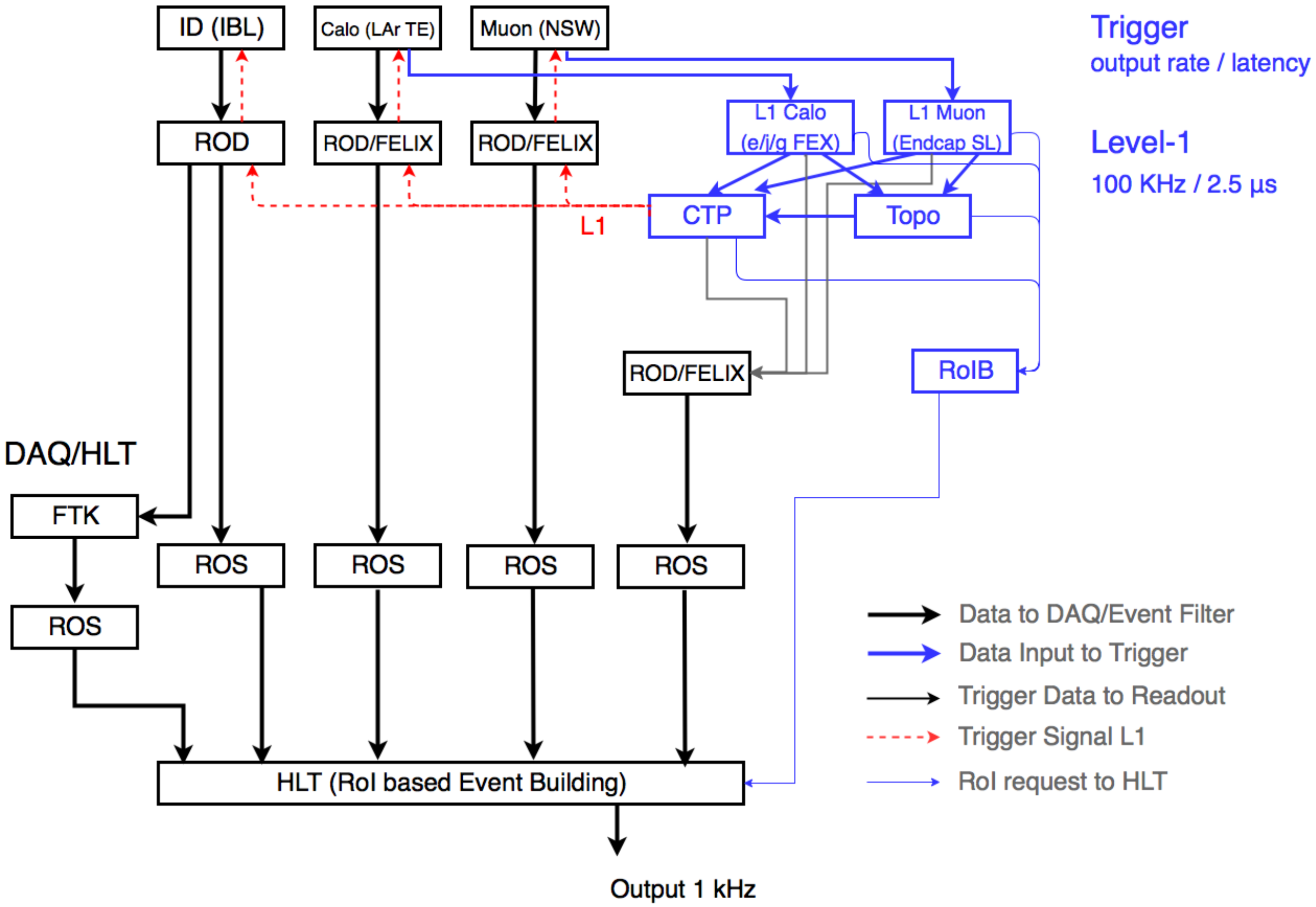




HLT Farm

~2000 Multi-core PCs
with rolling replacement

Run 3 TDAQ Architecture

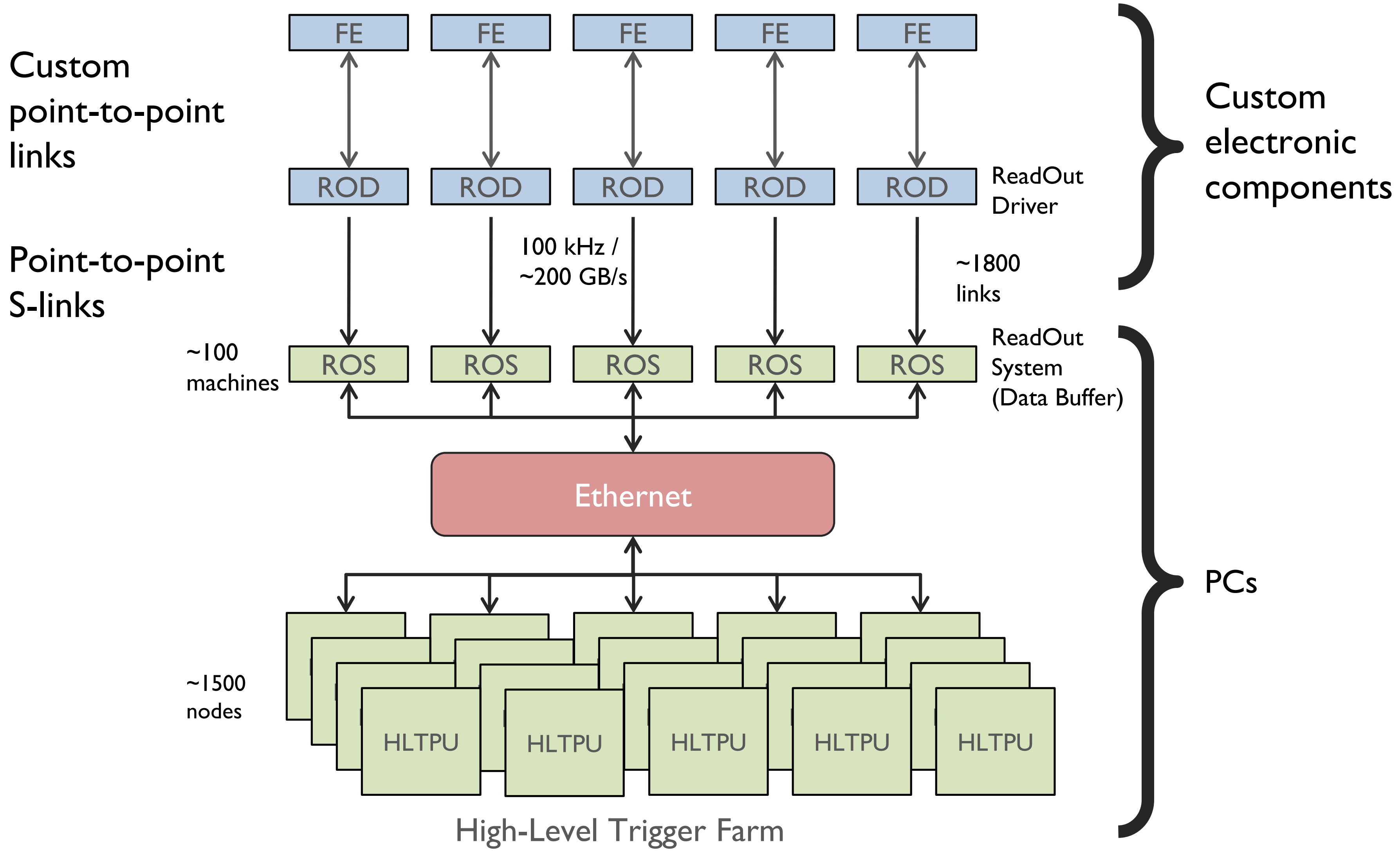


- **Level-1 Calorimeter trigger (L1CALO) with fine granularity LAr data**
- **Level-1 Muon trigger (L1Muon) with New Small Wheel data**
- **New readout with Front-End Link eXchange (FELIX)**

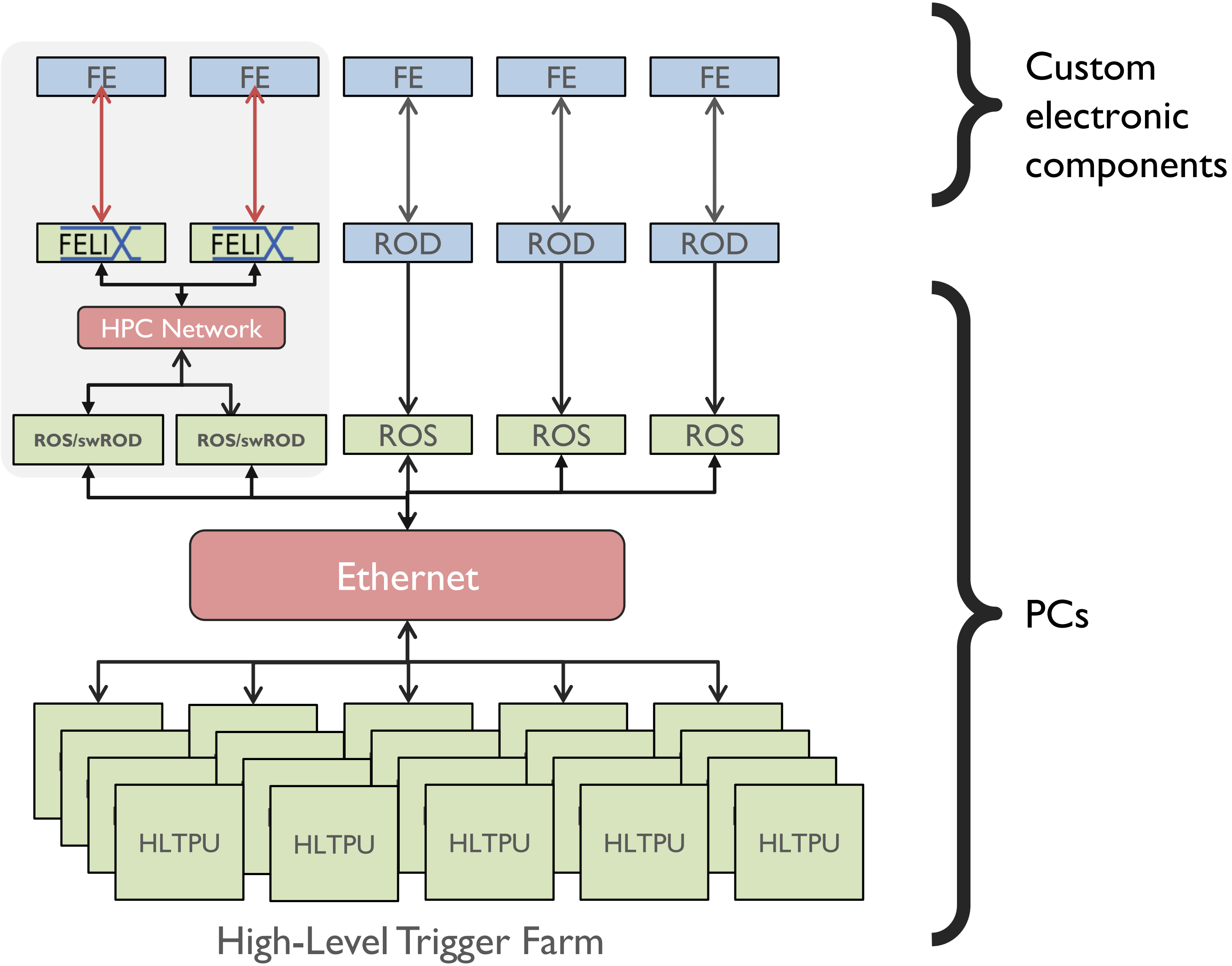
Readout Evolution Motivation

- 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 (Run 1&2)

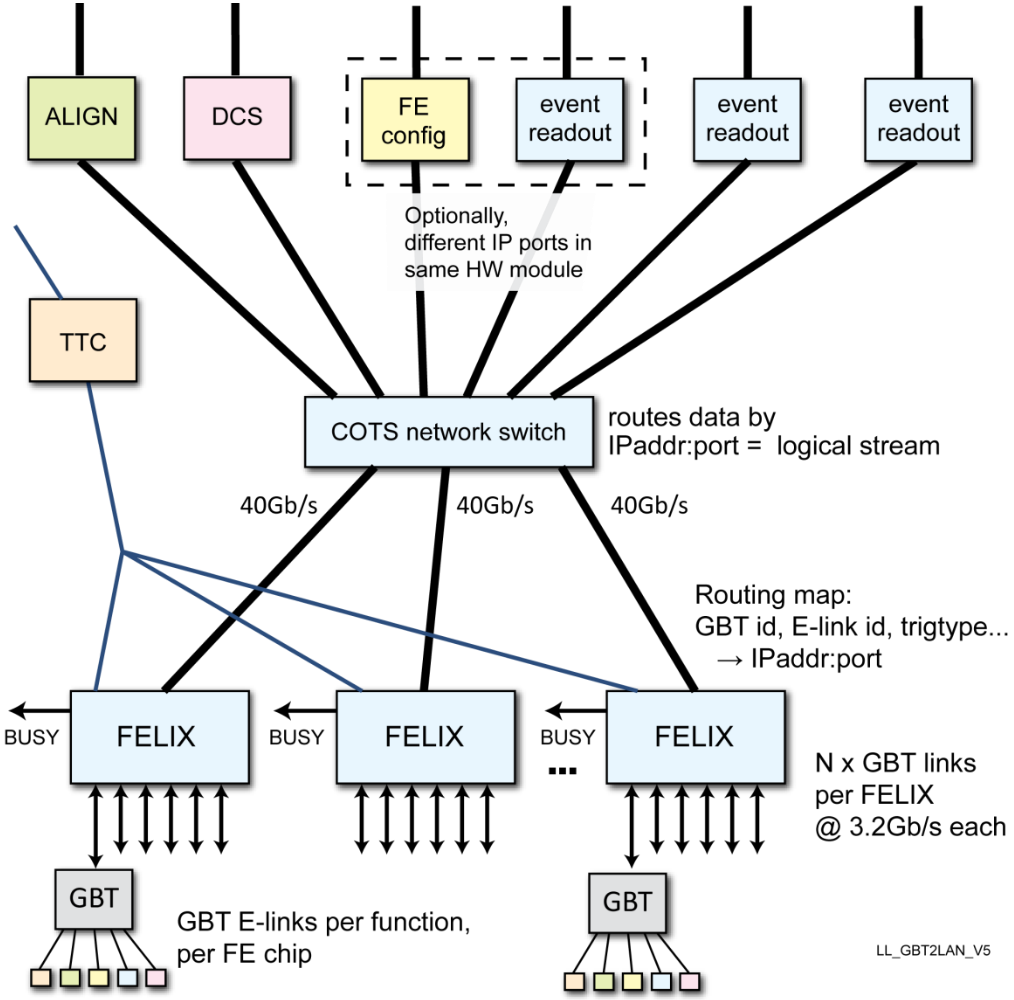


Readout (Run 3)

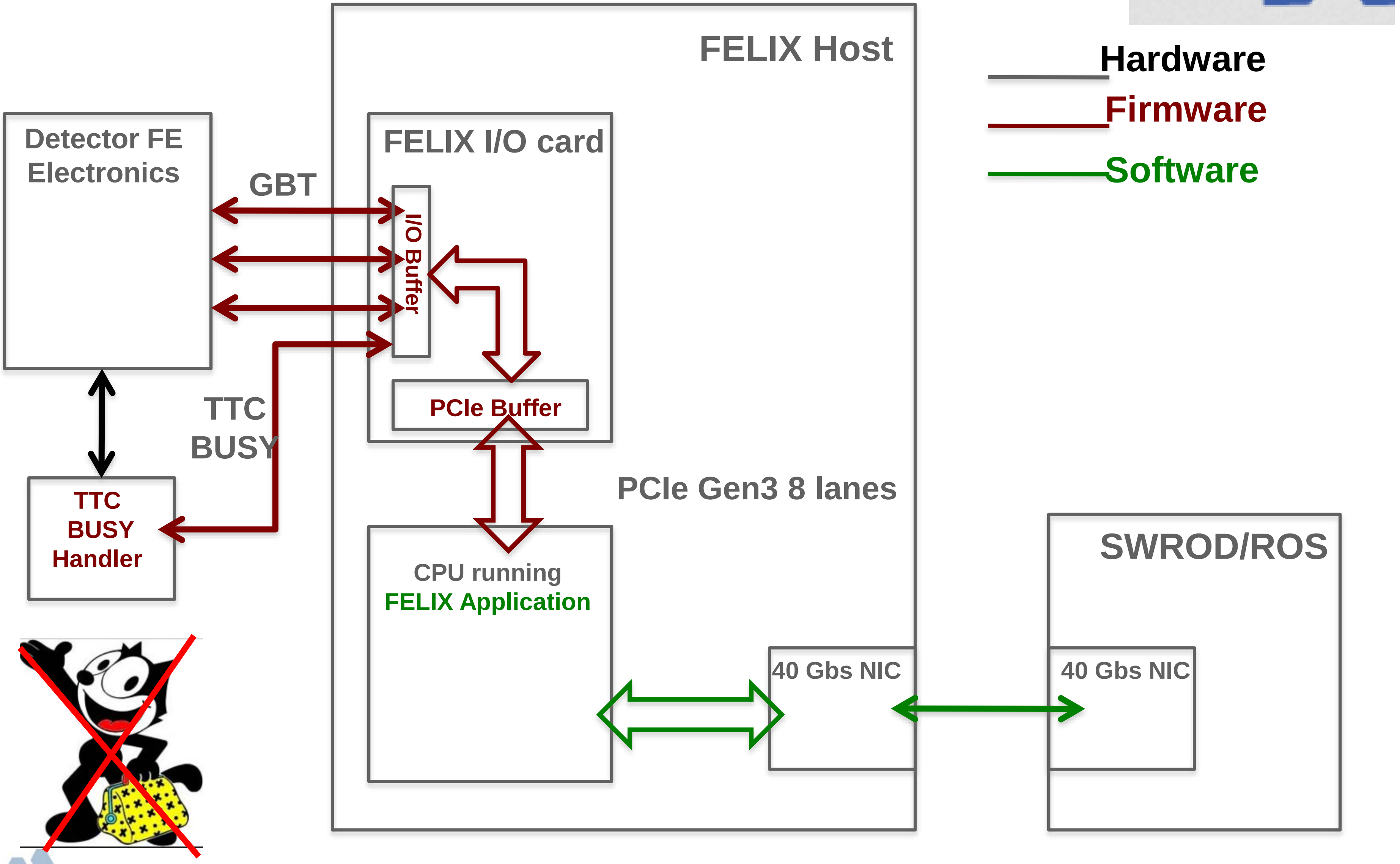


FELIX

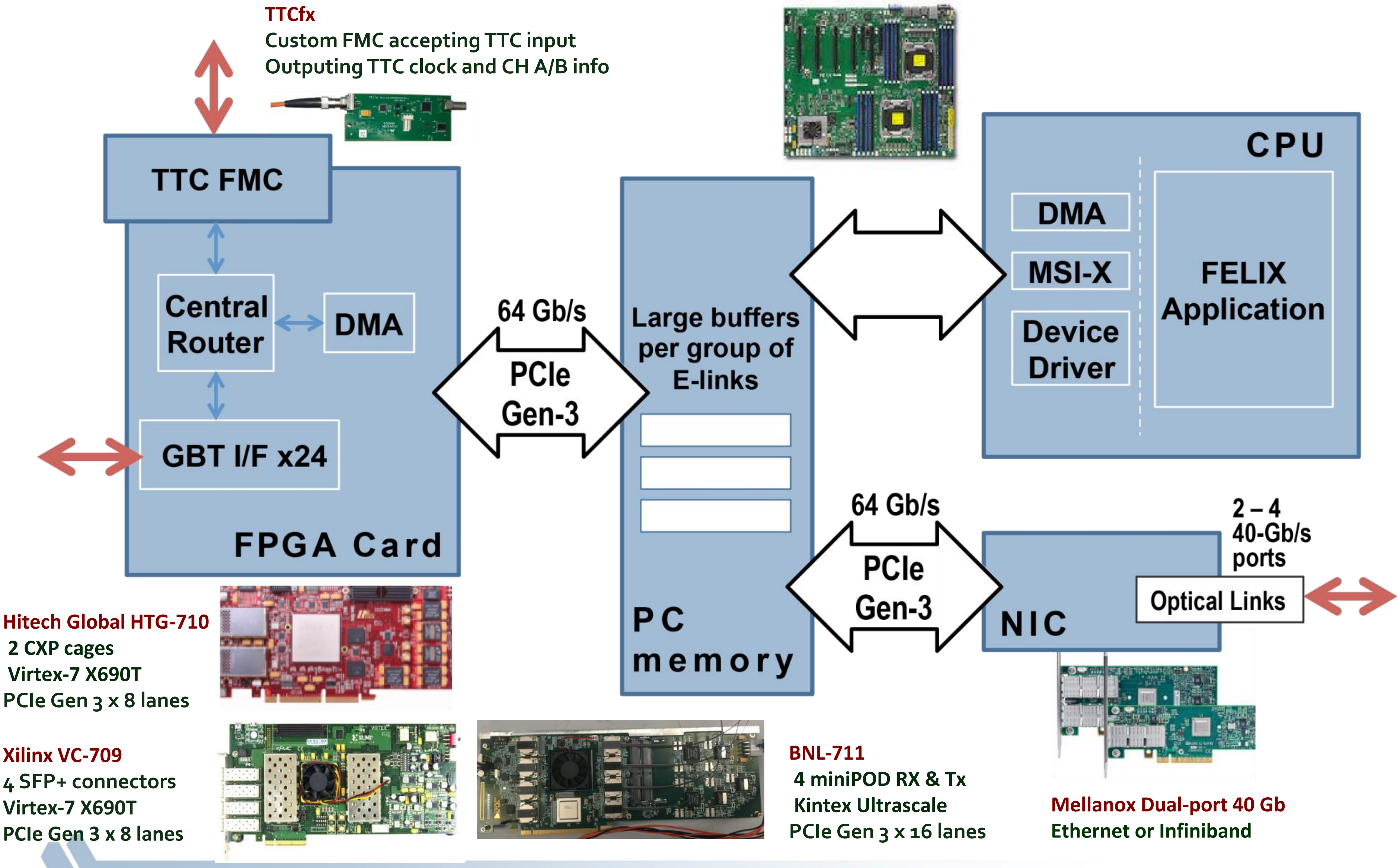
- Enabling transition from custom hardware to COTS as early as possible
- Using high level switch protocols of high speed and large bandwidth
- Configurable and flexible data routing and error handling, without relying on detector specific hardware
- Direct low latency paths between links
- Universal ATLAS-wide TTC/BUSY handling as for Run 1&2
- Command scheduling with guaranteed timing for calibration


ALIGN
DCS
FE config
event readout
event readout
event readout
Optionally, different IP ports in same HW module
TTC
COTS network switch
routes data by IPaddr:port = logical stream
40Gb/s
40Gb/s
40Gb/s
Routing map: GBT id, E-link id, trigtype... → IPaddr:port
BUSY
FELIX
BUSY
FELIX
BUSY
...
FELIX
N x GBT links per FELIX @ 3.2Gb/s each
GBT
GBT E-links per function, per FE chip
GBT
LL_GBT2LAN_V5

FELIX as a System



FELIX Development



TTCfx
Custom FMC accepting TTC input
Outputing TTC clock and CH A/B info

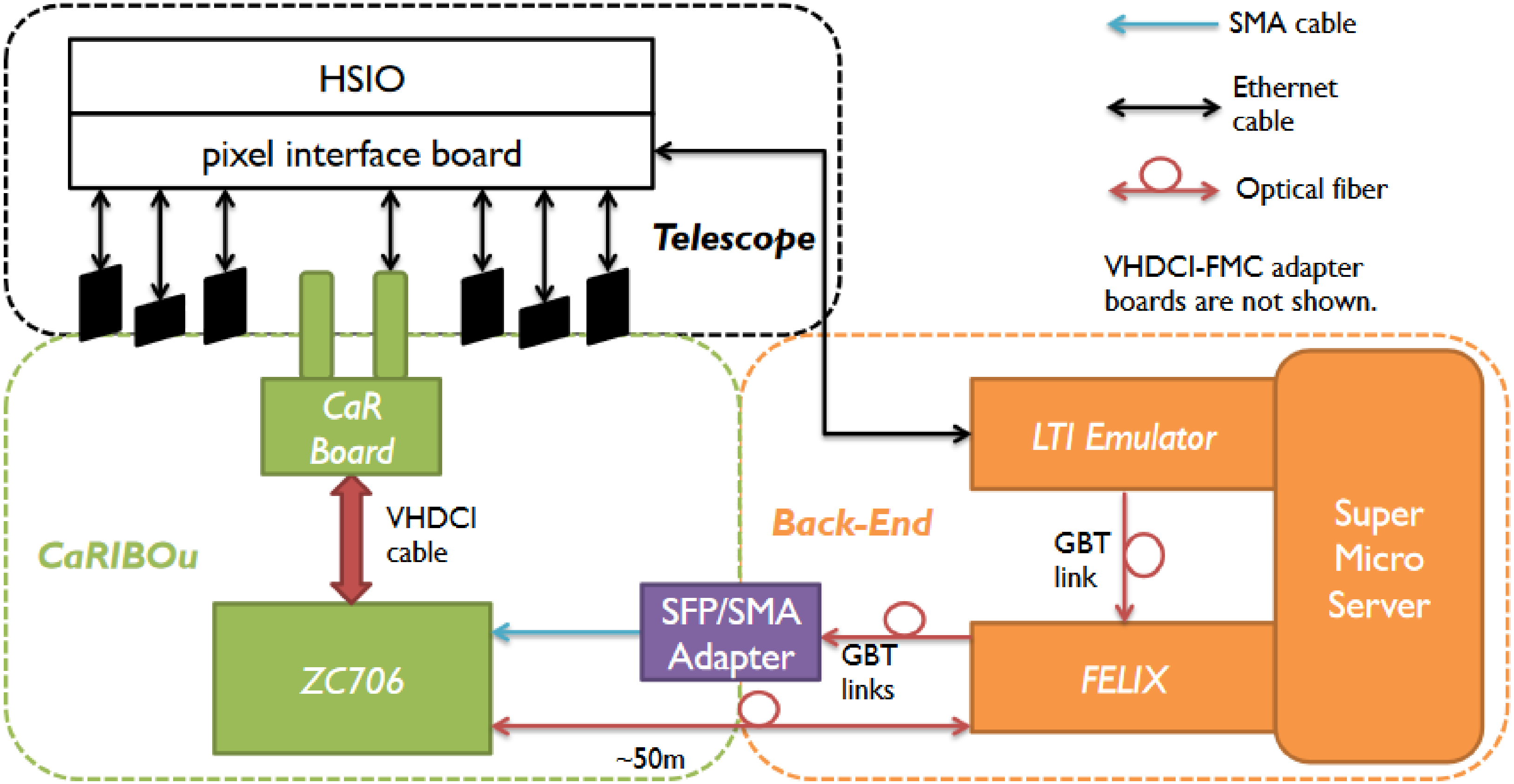
Hitech Global HTG-710
2 CXP cages
Virtex-7 X690T
PCIe Gen 3 x 8 lanes

Xilinx VC-709
4 SFP+ connectors
Virtex-7 X690T
PCIe Gen 3 x 8 lanes

BNL-711
4 miniPOD RX & Tx
Kintex Ultrascale
PCIe Gen 3 x 16 lanes

Mellanox Dual-port 40 Gb
Ethernet or Infiniband

FELIX in Action



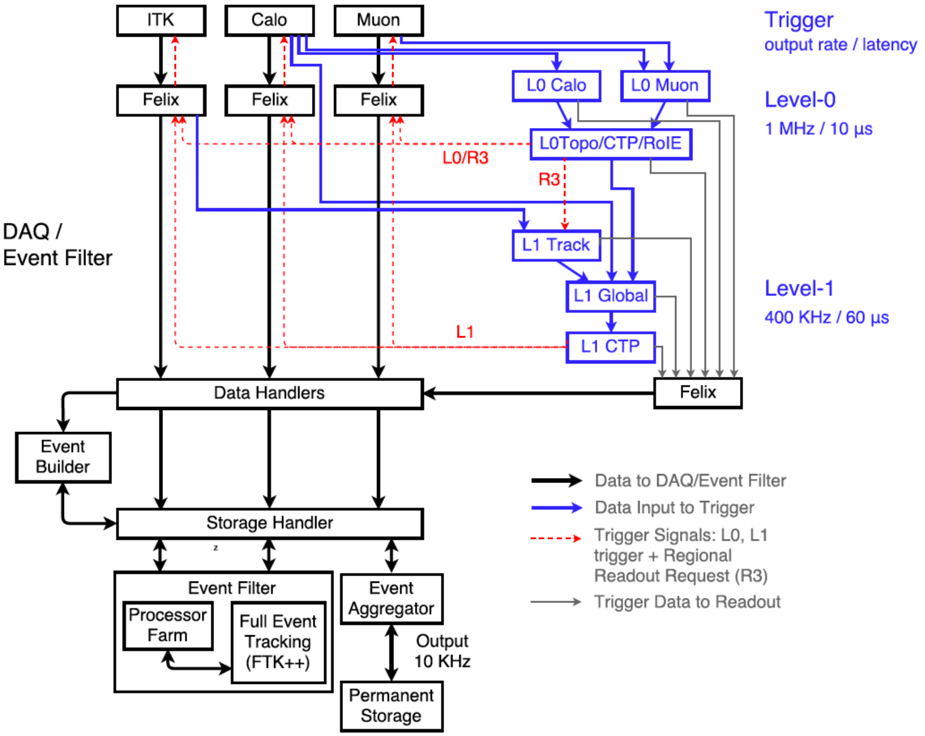
- Current ZC706 firmware supports to interface one DUT.
- System clock & TTC commands are from LTI emulator.
  - For this test platform, an Ethernet cable connects one RJ45 from HSIO to LTI emulator. The emulator extracts clock and commands from it.
  - This makes FELIX & CaRIBOu system to be synchronized with the telescope readout.

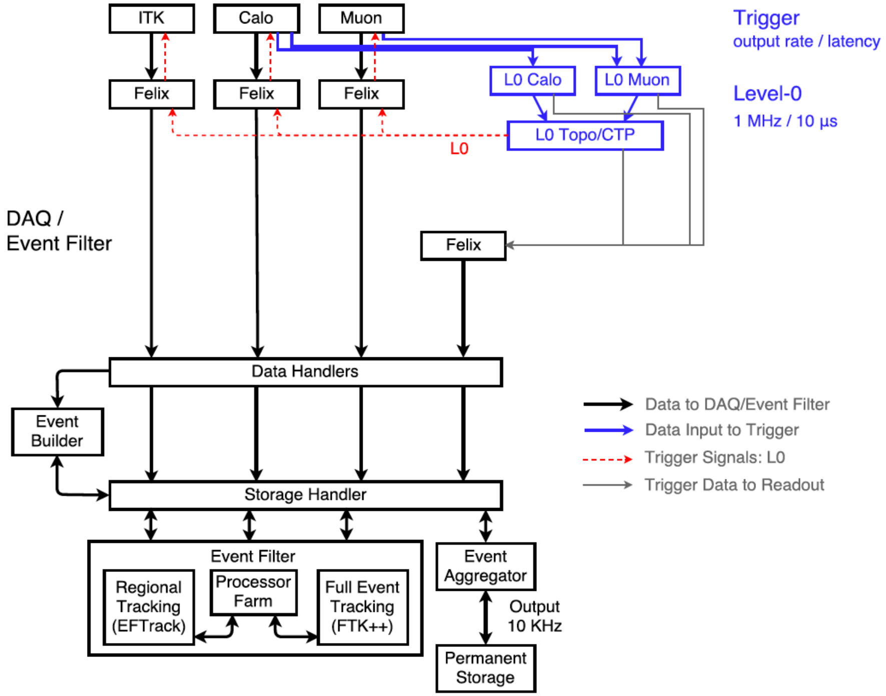
3

Trends to Future

- Higher and higher trigger rates
  - Triggerless not yet possible
- 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

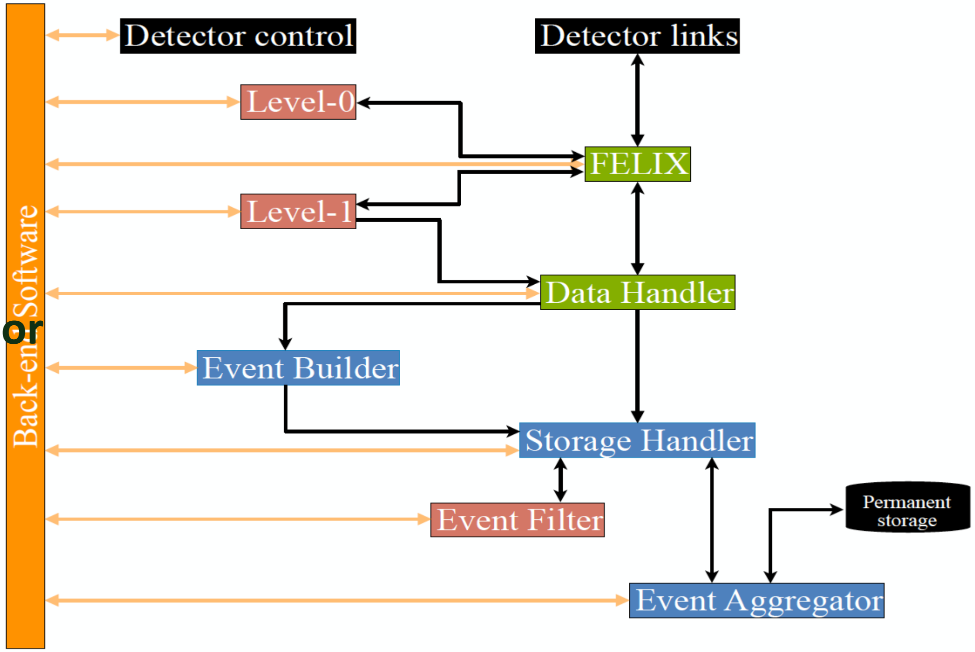
Run 4 TDAQ Architecture





- **Two major TDAQ architecture options being studied (L0/L1 with different operating parameters, L0-only)**
- **Though no big difference in the DAQ architecture except the rate/throughput**

Architecture in View of DAQ

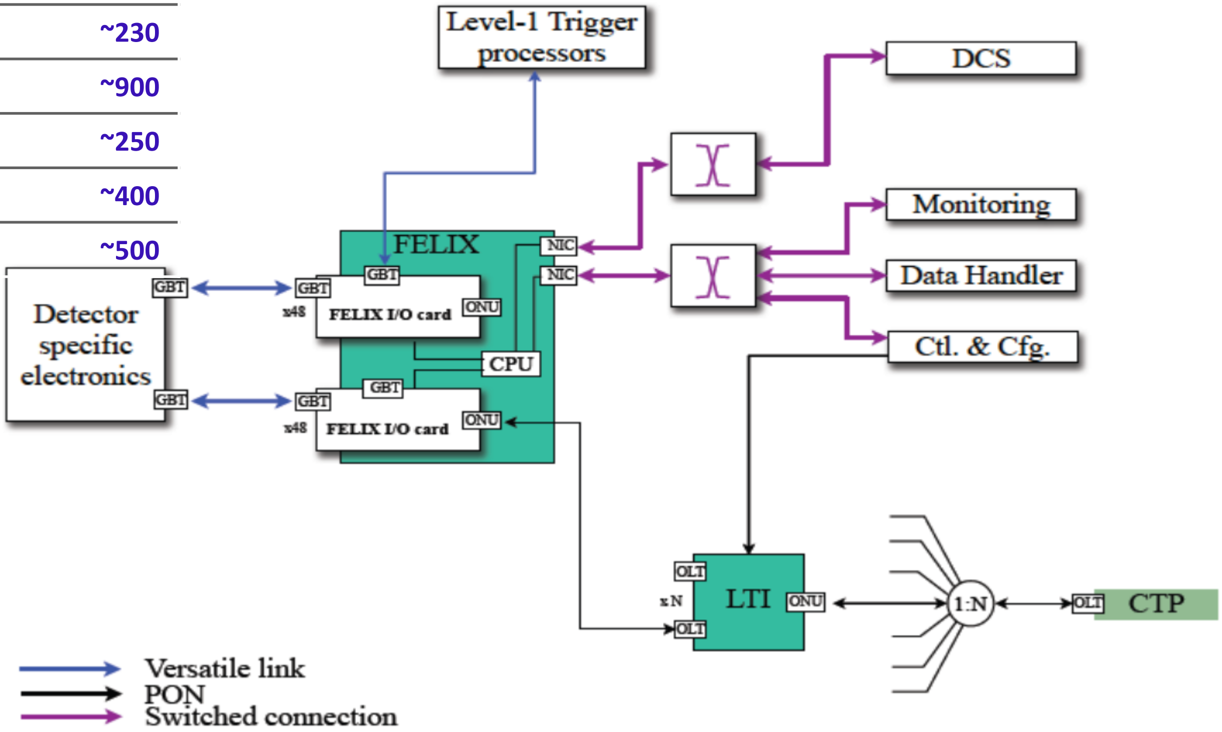


- **Standard architecture**
  - **Readout infrastructure to transport data out of the detector**
  - **Dataflow infrastructure to build events and buffer during event filtering**
- **Introduce a large storage area before filtering**
  - **High-level interface between dataflow and event filtering**
    - **To allow for a heterogeneous farm (accelerator , tracking devices, …)**
  - **Decoupling event filtering operation from LHC cycle**
    - **To take advantage of inter-fill periods for best use of compute resources**

Readout (L0/L1 Scheme)

Readout Parameters in L0/L1

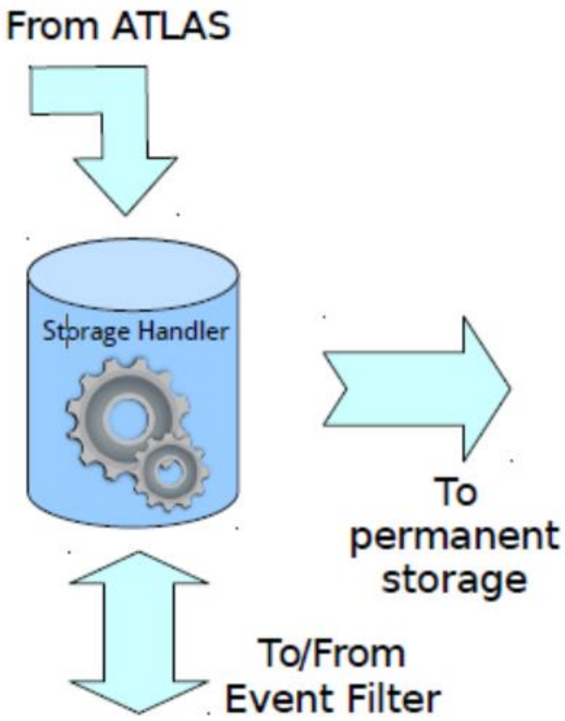
Link from detectors	11000
FELIX I/O card	~450
FELIX PC servers	~230
Low Latency links	~900
FELIX NICs (100 Gbps)	~250
Data Handler servers	~400
Data Handler NICs (100 Gpbs)	~500



- FELIX extended to all ATLAS detector subsystems, possibly with new hardware/firmware/software implementation, and with low latency link to trigger processors and detector specific firmware/software if needed
- Data Handler implementing detector specific data processing, with software on commodity PCs

Storage Handler

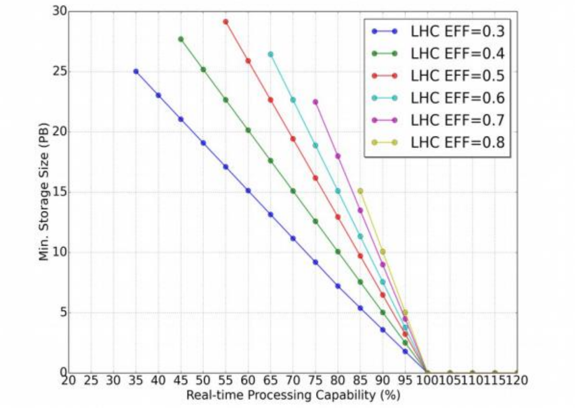
- Core dataflow infrastructure
- To decouple DAQ and Event filtering operation with large buffer area
- To offload data movements to distributed file system infrastructure
- Still to provide
  - Data bookkeeping
  - Event assignment
  - Load balancing
- Not need to have dedicated storage for accepted events
  - Event Aggregator to fetch and aggregate events on their way to permanent storage



Storage Requirement

- Capacity is a trade-off
  - Volume vs asynchronous processing
- Depends on
  - LHC duty cycle and efficiency
  - Considered timescale
- Several tens of PB for a single cycle
  - 20 – 60 PB

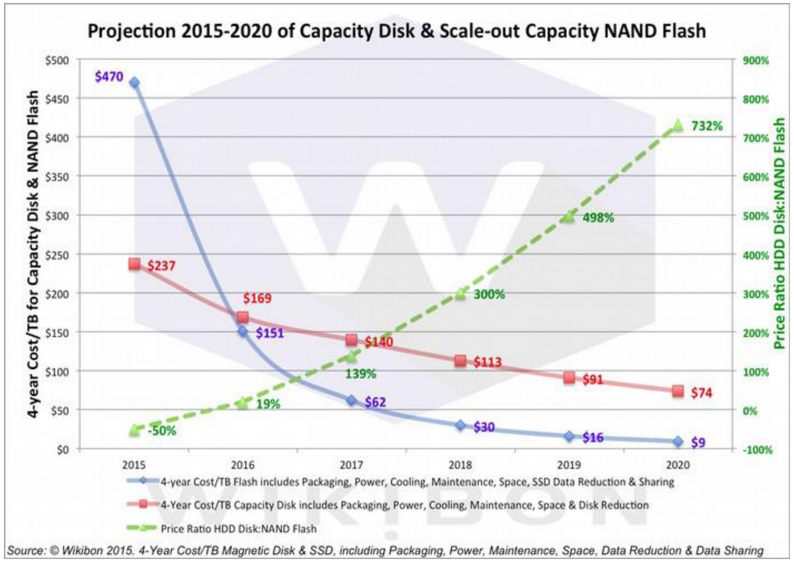
Parameter	L0/L1	L0
Input from detector	2 TB/s	5 TB/s
Output to tracking processors	<0.5 TB/s	< 1 TB/s
Output to Farm	< 2 TB/s	< 2 TB/s
Output to offline	50 GB/s	50 GB/s



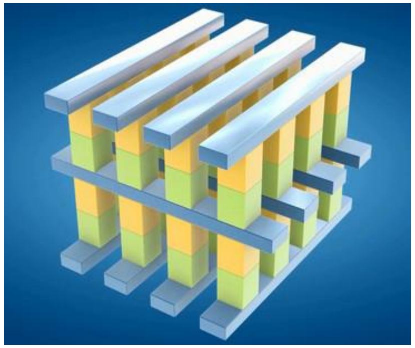
- Throughput is real challenge
  - Especially considering spinning hard drives
- Exacerbated by evolution of drive characteristics
  - Capacity growing much faster than I/O capability
- Assuming 10 TB/drive
  - 50 PB → 5000 drives
- Assuming (optimistic) 100 MB/s/drive
  - 5 TB/s → 50000 drives

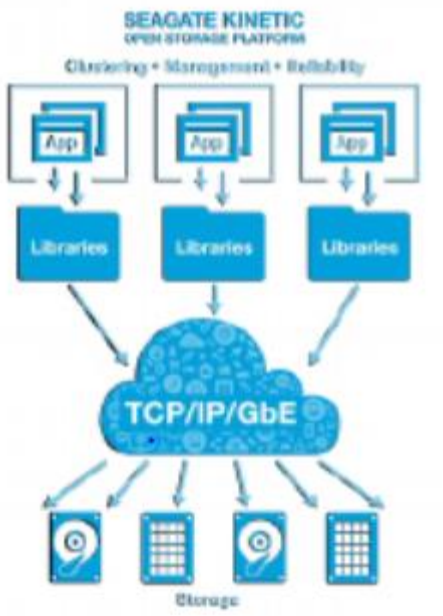
Storage Evolution

- Real world example exists with current Technology (Backblaze)
- We should look at storage technologies 10 years from now
- Evolution of existing technologies
  - Consumer NAND drive getting cheaper than spinning drive
  - Lustre and GPFS
- New technologies
  - 3D XPoint
- Innovations in the storage stack
  - Seagate Kinetic, …



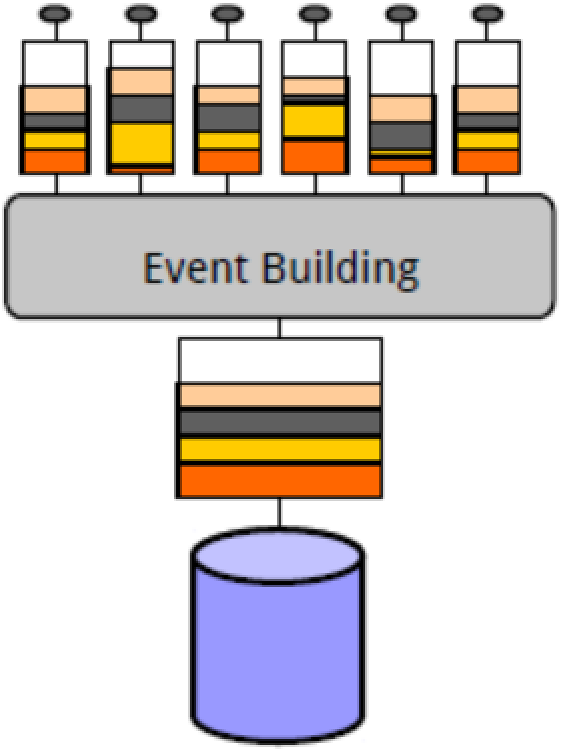
Source: © Wikibon 2015. 4-Year Cost/TB Magnetic Disk & SSD, including Packaging, Power, Maintenance, Space, Data Reduction & Data Sharing



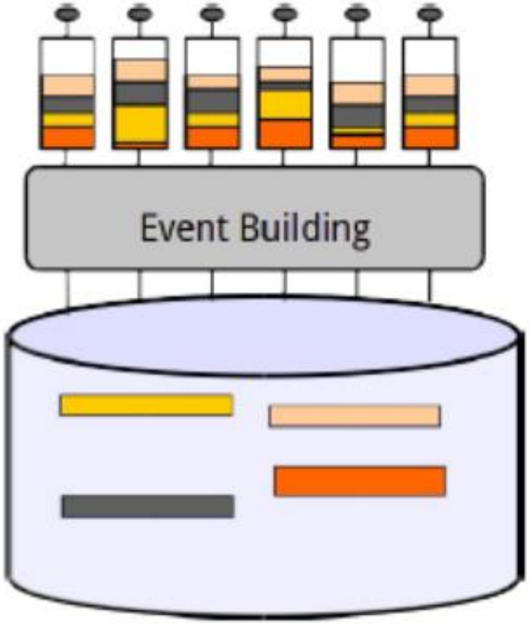


Event Building

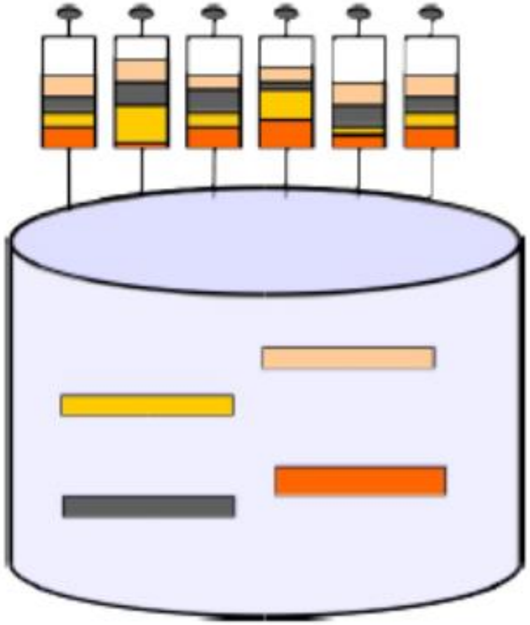
- Aggregating partial data fragments into a coherent unit
  - Convenient format for event filtering and necessary for offline transmission
- Methodology concerning
  - Necessity to gather all pieces together ?
    - Run2 event building taking place only for accepted events
  - Machinery to access or discard any piece
  - Physical vs Logical event building



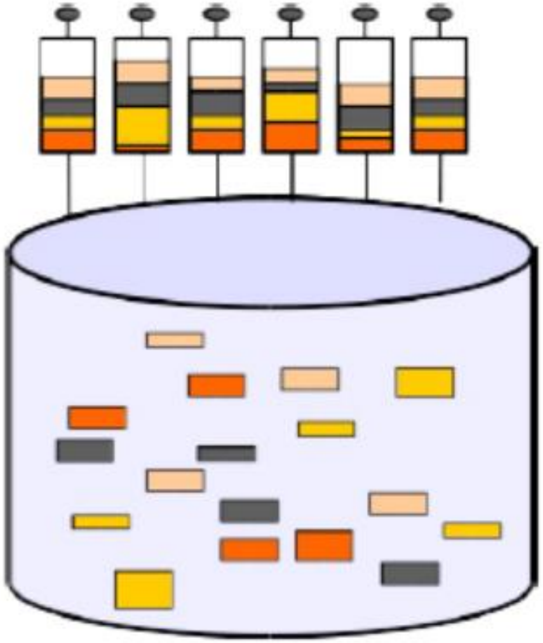
Event Building



- Physical EB with dedicated resources
  - Possible isolation of EB specific network challenges
  - Event level data compression



- Physical EB offloaded to storage
  - Possible optimization
  - Depending on storage performance and implementation



- Logical EB
  - Aggregation of information on fragment location
  - Physical data still fragmented
  - Key-value database

Event Filter Implementation

- Expecting Event Filter to include different technologies
  - Run1/2/3 with homogeneous processor farm (well, FTK)
  - Run4 processor farm aided by accelerators
    - Full event tracking @ 100 kHz with special hardware (FTK++, GPGPU, etc)
    - Possibility to utilize arising technologies
- Clear interface allowing various processing implementations
  - Files vs events vs object storage
  - HLT in Run 1/2
    - Requesting data fragments from Readout system with detector knowledge (cabling, partitioning, etc)
    - Using offline software with a software layer coupled it to the DAQ environment
- Expecting event processing to be RoI-based

Event Filter Computing

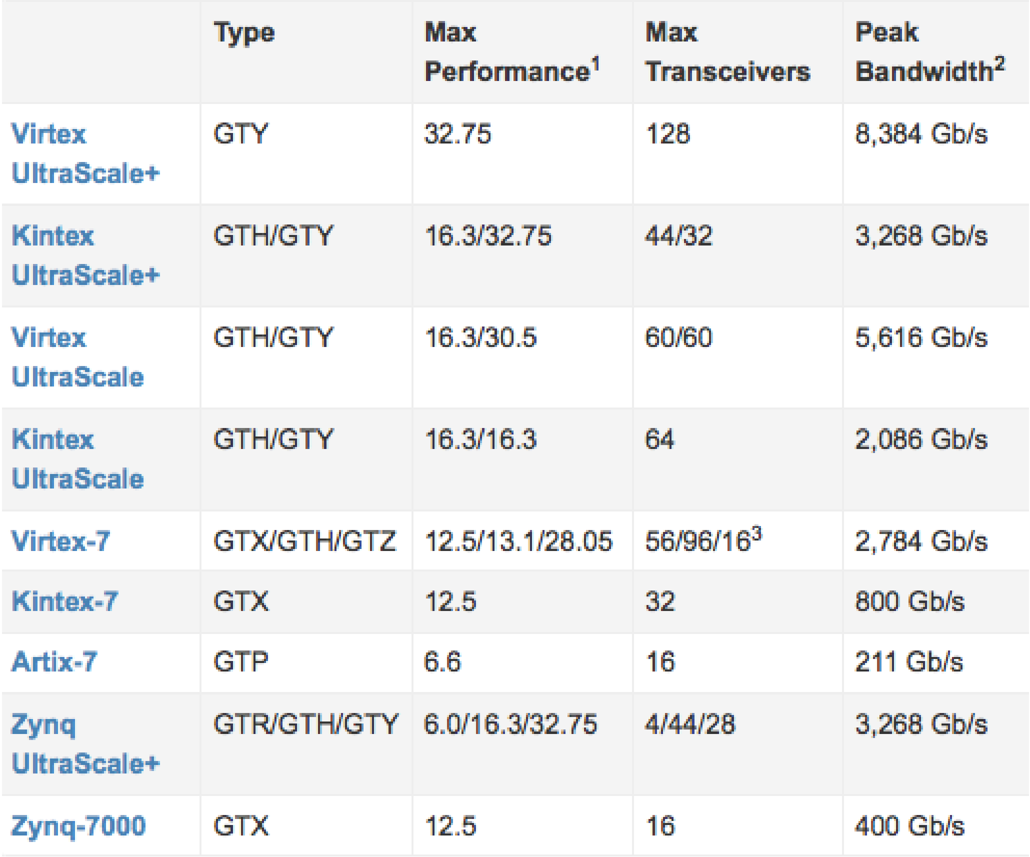
Parameter	L0/L1	L0	Run 2
Input Rate	400 kHz	1 MHz	100 kHz
Computing Power	11 MHS06	> 11 MHS06	0.8 MHS06
Computing power for tracking	5 MHS06	5 MHS06	

Summary

- Not covering software which is a key component of DAQ
- Current ATLAS DAQ system performing well while upgrades progressing as planned
  - Phase-I projects on schedule
  - Phase-II upgrade Technical Design Report in Q4 2017
- Increased use of Commodity hardware
  - Transit as early as possible from custom rad-hard links to commodity network (FELIX)
  - Take advantage of arising technologies

Link to Upstream

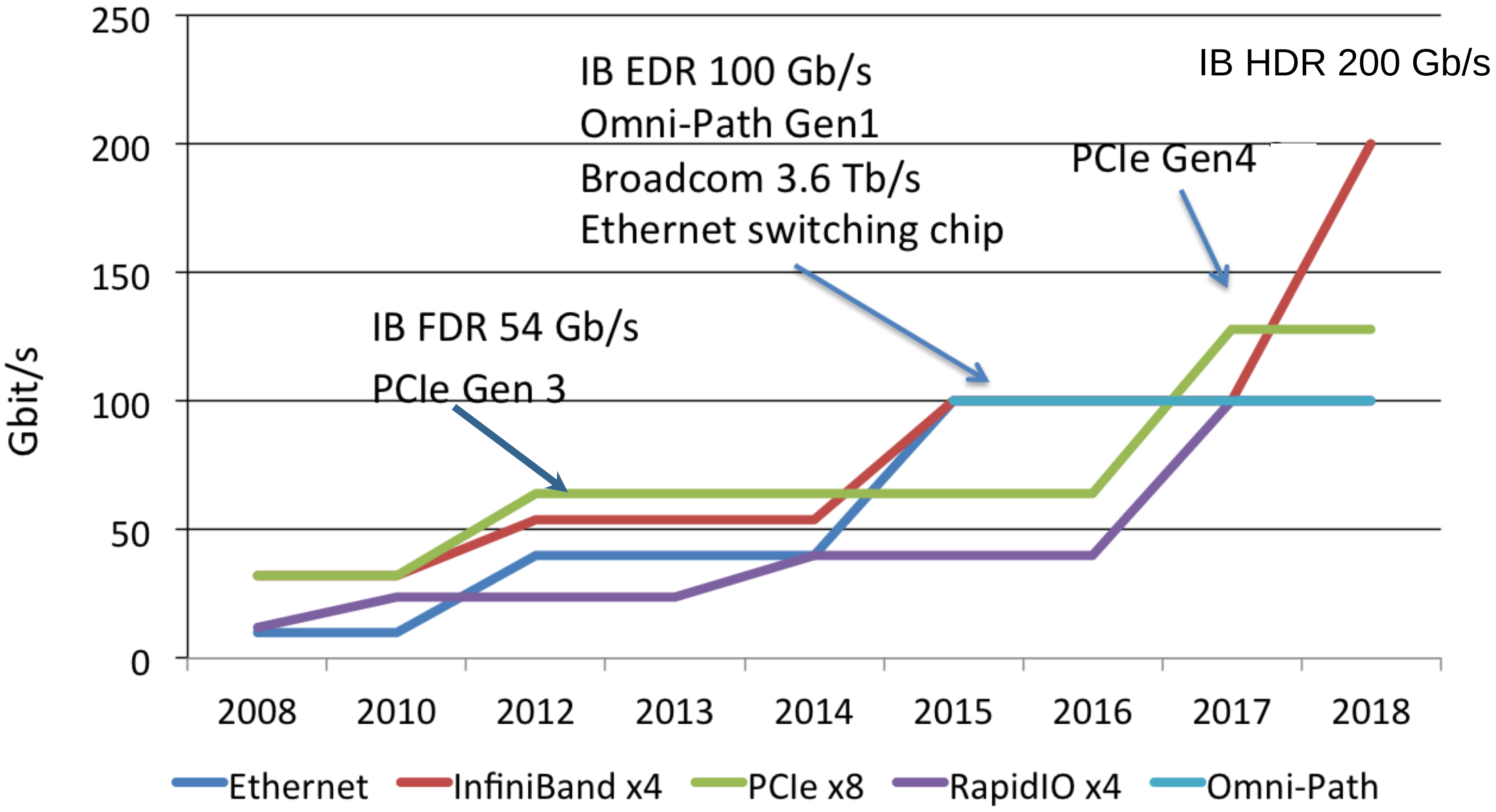
http://www.xilinx.com



	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**