Overview of future DAQ for Alice, LHCb and **Atlas**, **CMS**

ACES 2016, CERN

Frans Meijers – CERN EP-CMD

- DAQ parameters
- Read -Out
- Event Building
- HLT facility
- Storage
- Concluding Remarks

Acknowledgements

- Alice:
  - P.VanDeVyvre
- Atlas:
  - D.Francis, W.Vandelli
- CMS:
  - E.Meschi, A.Bocci, S.Cittolin, A.Racz, S.Erhan, V.Innocenti, M.Hanssen, J.Hegeman, R.Mommsen, P.Zejdl, A.Holzner
- LHCb
  - N.Neufeld

- IT department
  - O.Barring, M.Girone

Schedule TDR

- Alice
  - 2015-Q2 TDR online-offline computing system O2
- LHCb
  - 2014-Q2 TDR trigger and online upgrade
- Atlas
  - 2013 LOI
  - 2015-Q3 Phase-II Upgrade scope
  - 2016-Q2 ***IDR (Initial Design Report)***
  - **2017 TDR for TDAQ**
- CMS
  - 2014 Phase-II TP
  - 2015-Q3 Phase-II Upgrade scope
  - **2019 TDR for Trigger**
  - **2020 TDR for DAQ**

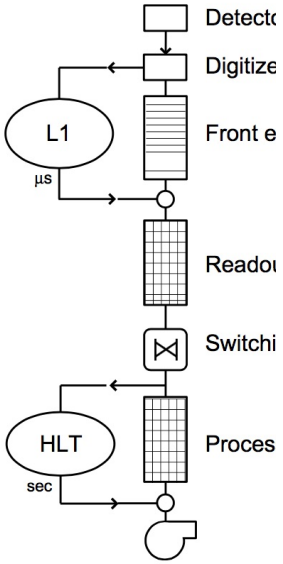
DAQ PARAMETERS

DAQ Parameters

	Run	# HW/SW trigger levels	Level-x accept rate	Event size	EVB Effective thru	storage
Alice (Pb-Pb)	3	1 / 0	50 kHz	60 MB	0.5 TB/s	80 GB/s
LHCb	3	0 / 1	40 (30) MHz HLT 20 kHz	0.1 MB	4 TB/s	2 GB/s
Atlas	4	1 or 2 / 1	L0/L1 400 kHz or L0 1 MHz HLT 10 kHz	~ 5 MB	5 TB/s	50 GB/s
CMS	4	1 / 1	L1 750 kHz HLT 7.5 kHz	~ 5 MB	4 TB/s	40 GB/s

- Note:
  - Alice: HLT does extensive data reduction (factor 6) before EVB
  - Atlas: HW trigger L0/L1 or L0 under discussion

CMS L1 / DAQ / HLT



- Same two-level architecture as current system
  - L1 hardware trigger: 40 MHz clock driven, custom electronics
  - High Level Trigger (HLT): event driven, COTS computing nodes

Table 7.1: DAQ/HLT system parameters.

	LHC Run-I	LHC Phase-I upgr.	HL-LHC Phase-II upgr.	
Energy	7-8 TeV	13 TeV	13 TeV	
Peak Pile Up (Av./crossing)	35	50	140	200
Level-1 accept rate (maximum)	100 kHz	100 kHz	500 kHz	750 kHz
Event size (design value)	1 MB	1.5 MB	4.5 MB	5.0 MB
HLT accept rate	1 kHz	1 kHz	5 kHz	7.5 kHz
HLT computing power	0.21 MHS06	0.42 MHS06	5.0 MHS06	11 MHS06
Storage throughput (design value)	2 GB/s	3 GB/s	27 GB/s	42 GB/s

CMS Phase-II detector R/O parameters

Sub-det	# links on- 2 off-detector	Type (Gbps)	use	Data reduction	Event size (Mbyte)	#DAQ links (100 GBps)
TK-outer	13 k 2 k	GBT (4 G) GBT (9 G)	DAQ + Trig 20% + 80%	On-det	05. – 0.6	100
TK-pixel	1 k	lpGBT (9 G)	DAQ	On-det	0.7 – 1.0	200
ECAL-barrel	12 k	GBT (3 G)	streaming	Off-det	1.2	200
HCAL	2 k	GBT (3 G)	streaming	Off-det	0.2	40
HGCAL	9 k	lpGBT(9 G)	Streaming?	On-det?	1.2	200
Muons DT	6 k	GBT (3 G)	streaming	Off-det	0.1	20
Muons CSC	1 k	GBT (3 G)	DAQ+Trig 50%+50%	Off-det	0.1	20
Trigger						20
EVB					4.2-4.6	800

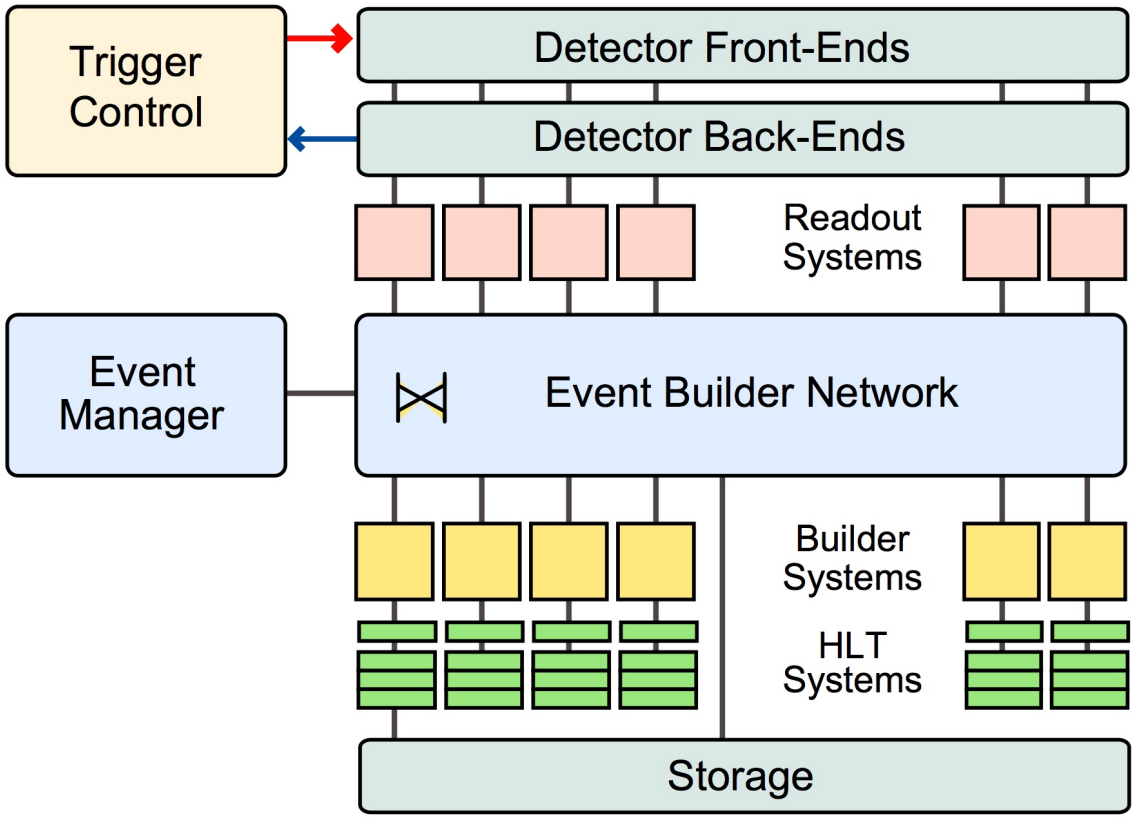
Atlas Phase-II

System	Buffering	L0 Readout	L1 Readout
ITk	On-detector	RoI/Full	Full
LAr	Off-detector	RoI	Full
Tile	Off-detector	RoI	Full
MDT	Off-detector	None	Full
Legacy MDT	On-detector	None	Full
NSW	On-detector	None	Full
RPC	Off-detector	None	Full
Thin Gap Chamber (TGC)	Off-detector	None	Full

- Under discussion: L0/L1 1 MHz / 400 kHz or L0 1 MHz
- Event size ~5 MB

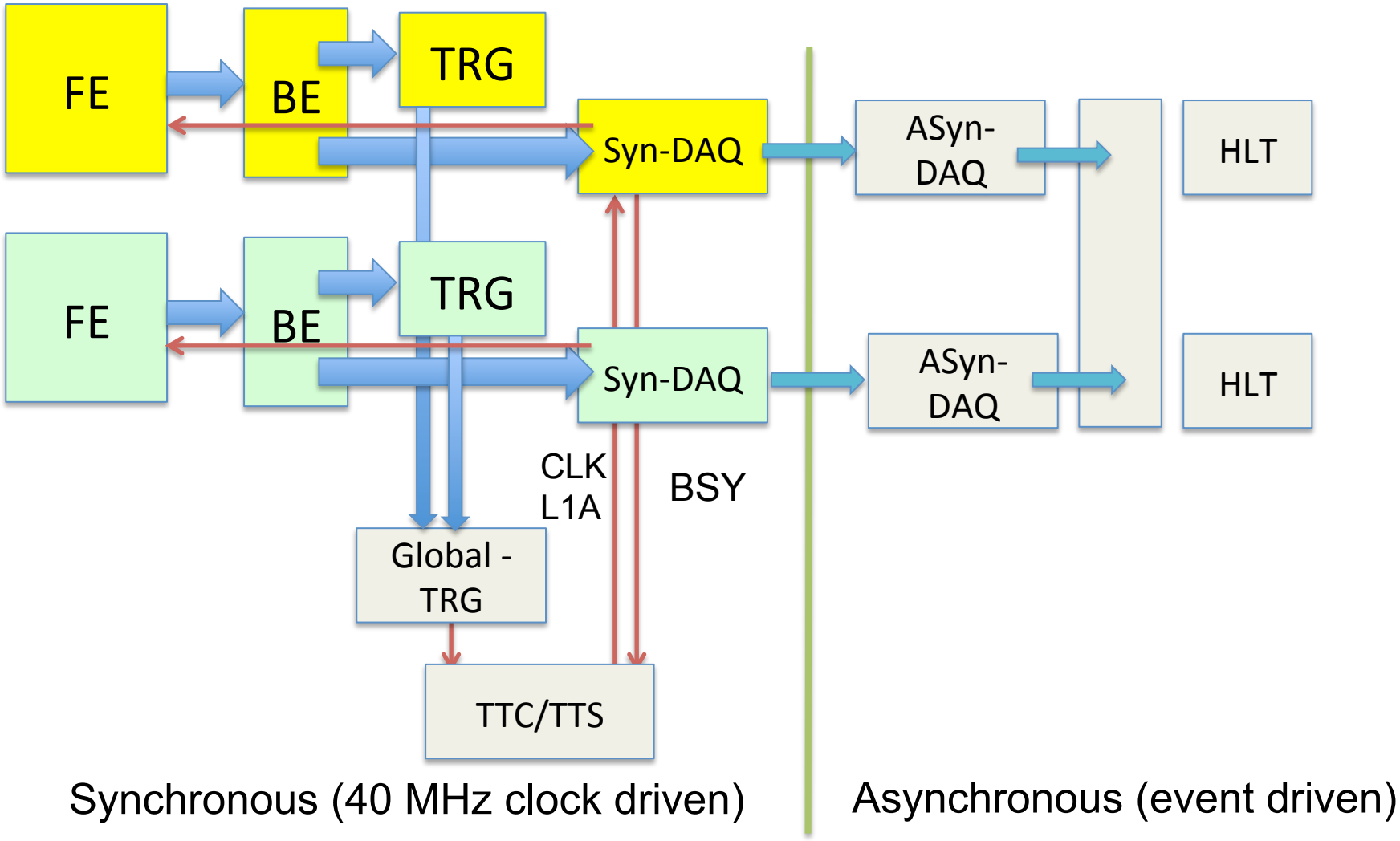
DAQ to 1st order

- All similar



READ-OUT

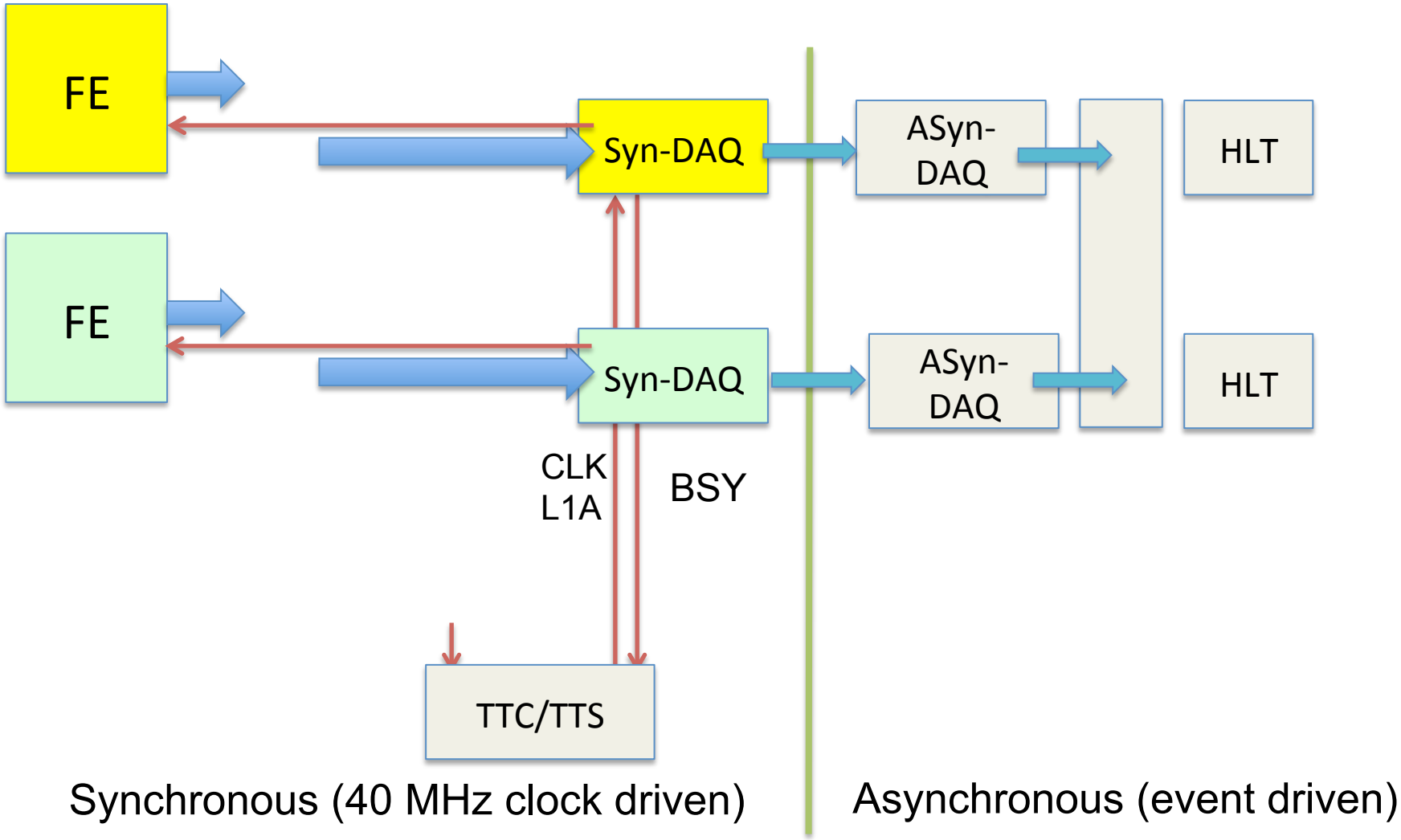
Trigger DAQ


FE
BE
TRG
Syn-DAQ
ASyn-DAQ
HLT
FE
BE
TRG
Syn-DAQ
ASyn-DAQ
HLT
CLK
L1A
BSY
Global - TRG
TTC/TTS
Synchronous (40 MHz clock driven)
Asynchronous (event driven)

High level design – readout unit

- Synchronous domain (40 MHz clock driven)
  - Towards front-end electronics via serial point-to-point links 5-10 Gbps (mostly GBT)
  - Receive DAQ data
  - TTC and throttle
  - Optional Send / receive “DCS”
  - Optional Transmit to Trigger electronics (same or separate links as DAQ)
  - forward trigger selected data in case of “streaming” DAQ
- Asynchronous domain (event driven)
  - Data aggregation (concentrator)
  - Buffer
  - Sub-detector specific processing (data reduction, processing, especially suited for channel-by-channel processing)
  - Protocol conversion, transmit / receive standard commercial network

LHCb Trigger DAQ



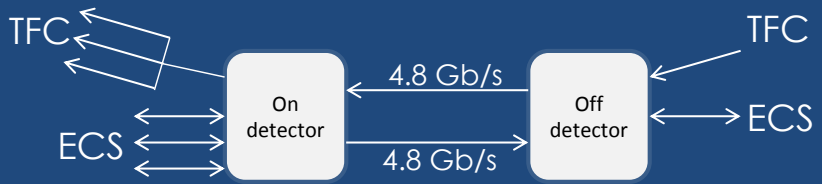
LHCb

Architectural choices

Data compression on front-end driven by link cost:
We will need ~ 15,000 links (4.8 Gbit/s)

Try to be flexible & scalable:
4 of 6 sub-detectors will use FPGAs in front-ends

Combine slow & fast controls (ECS & TFC) in duplex links

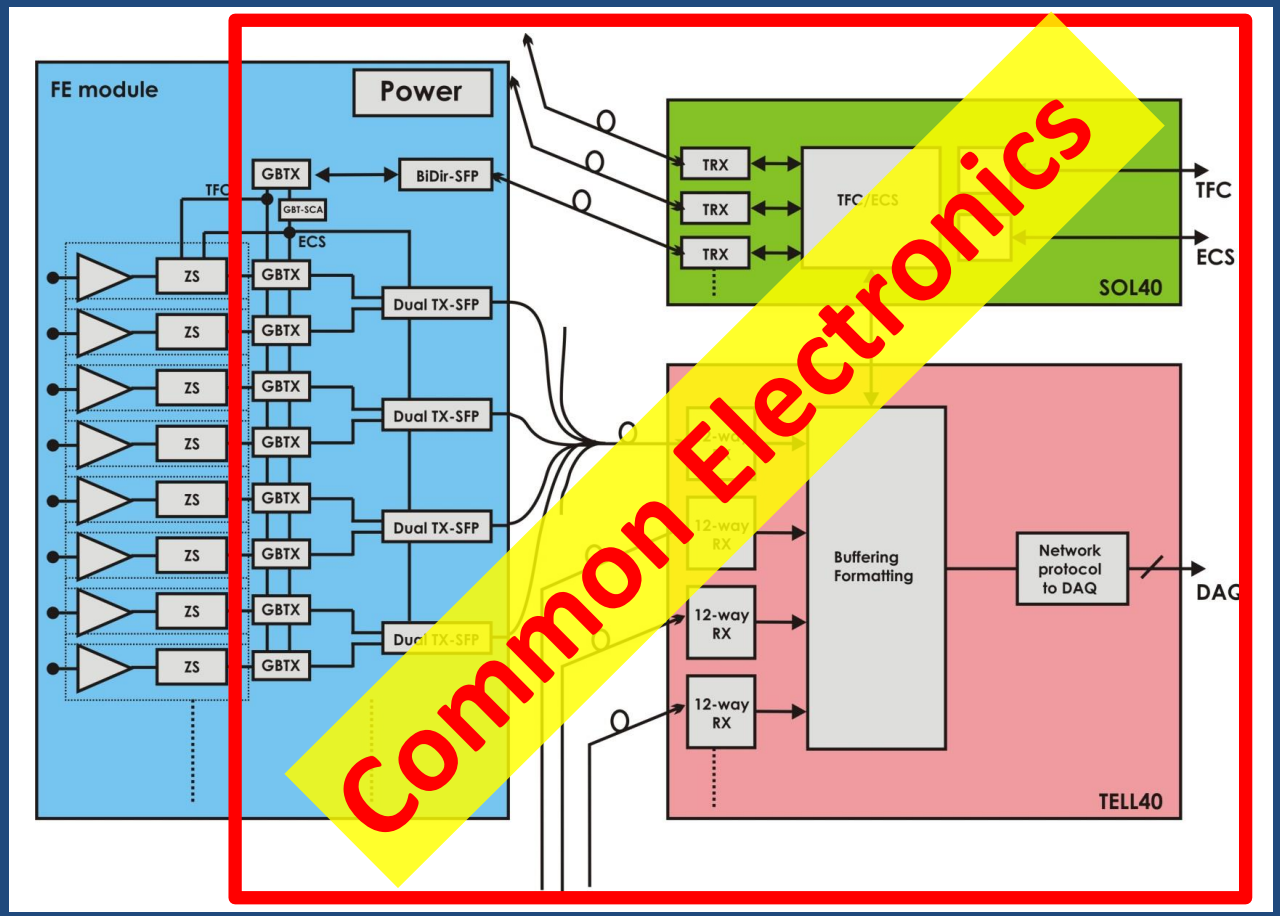


Simplex data links

Baseline choice for backend electronics is PCIe format

LHCb

Generic Implementation



LHCb

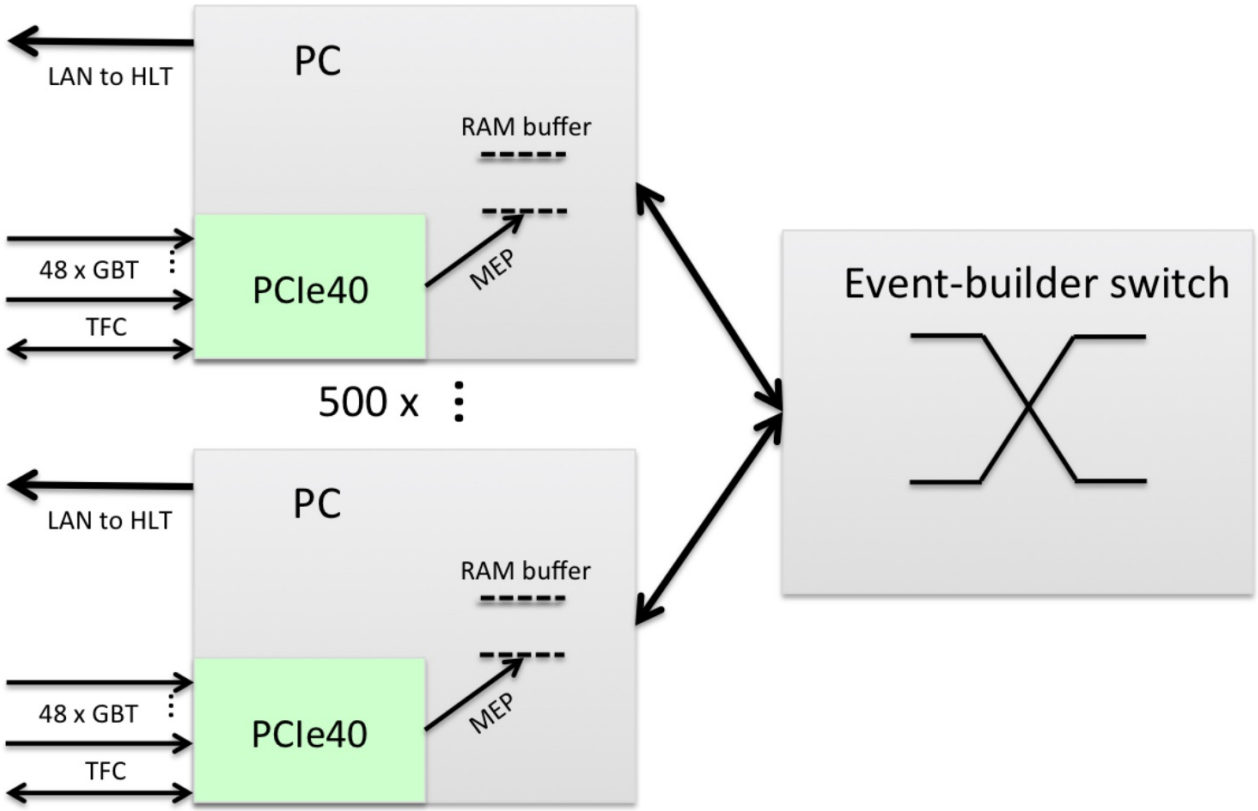
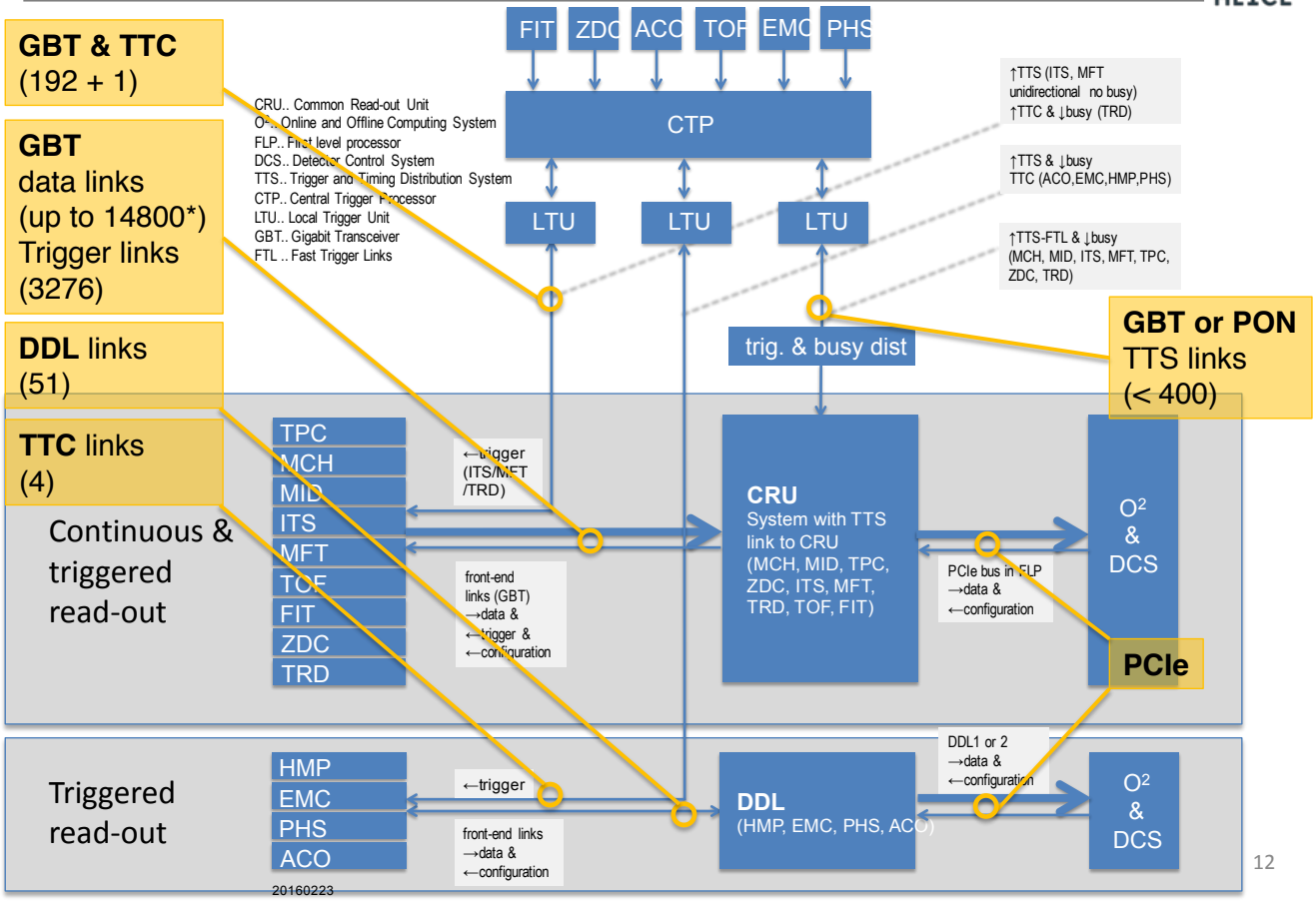


Figure 3.7: The PCIe based readout system. The PCIe40 readout boards are directly connected to the event-builder PCs through 16-lane PCIe edge-connector.

- Read-out Unit = PC + PCI40 card
- PCIe-g3 100 GbE, so ~25 GBT links providing 4 Gbps

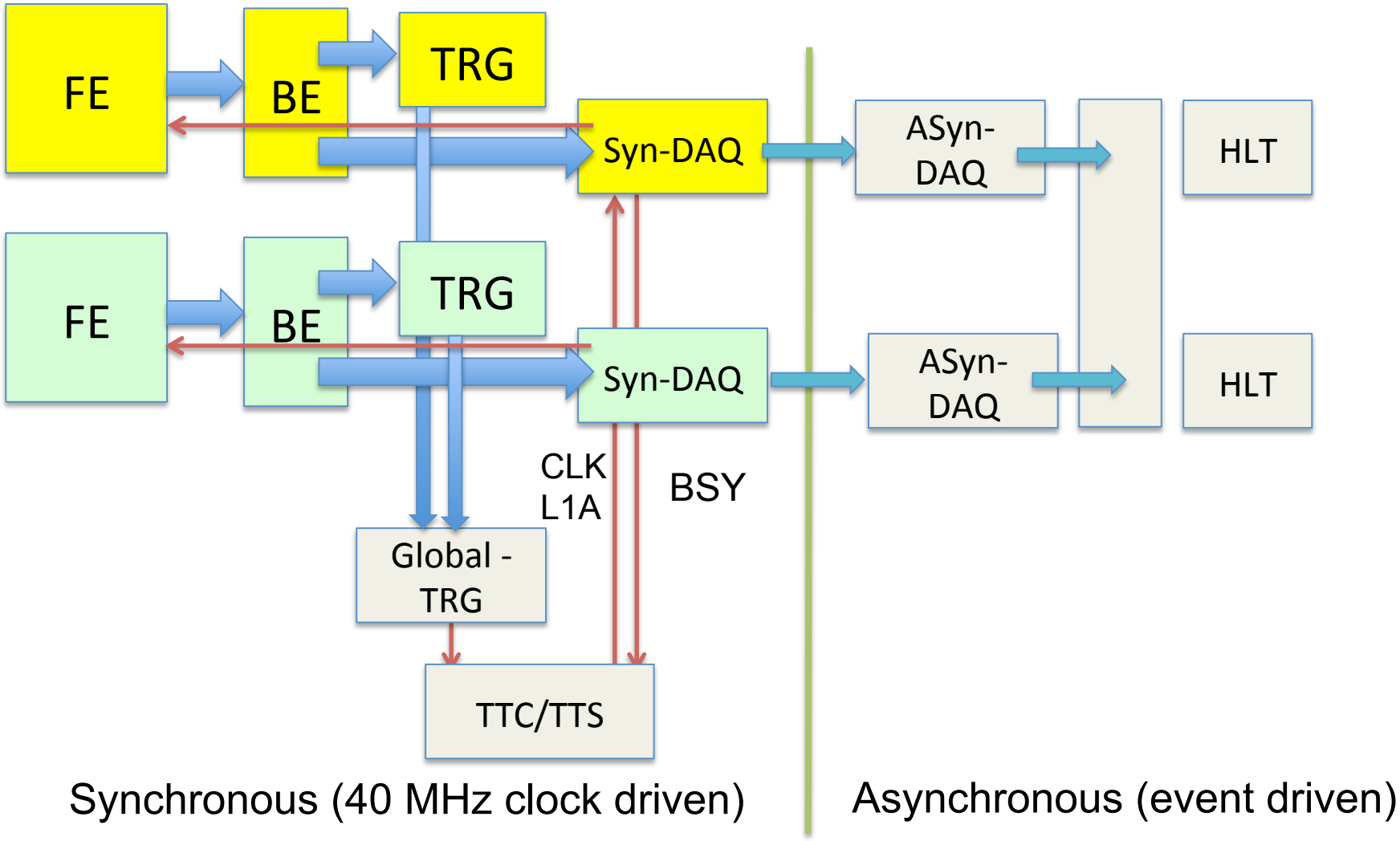
Alice Common Readout Unit

Upgrade architecture overview

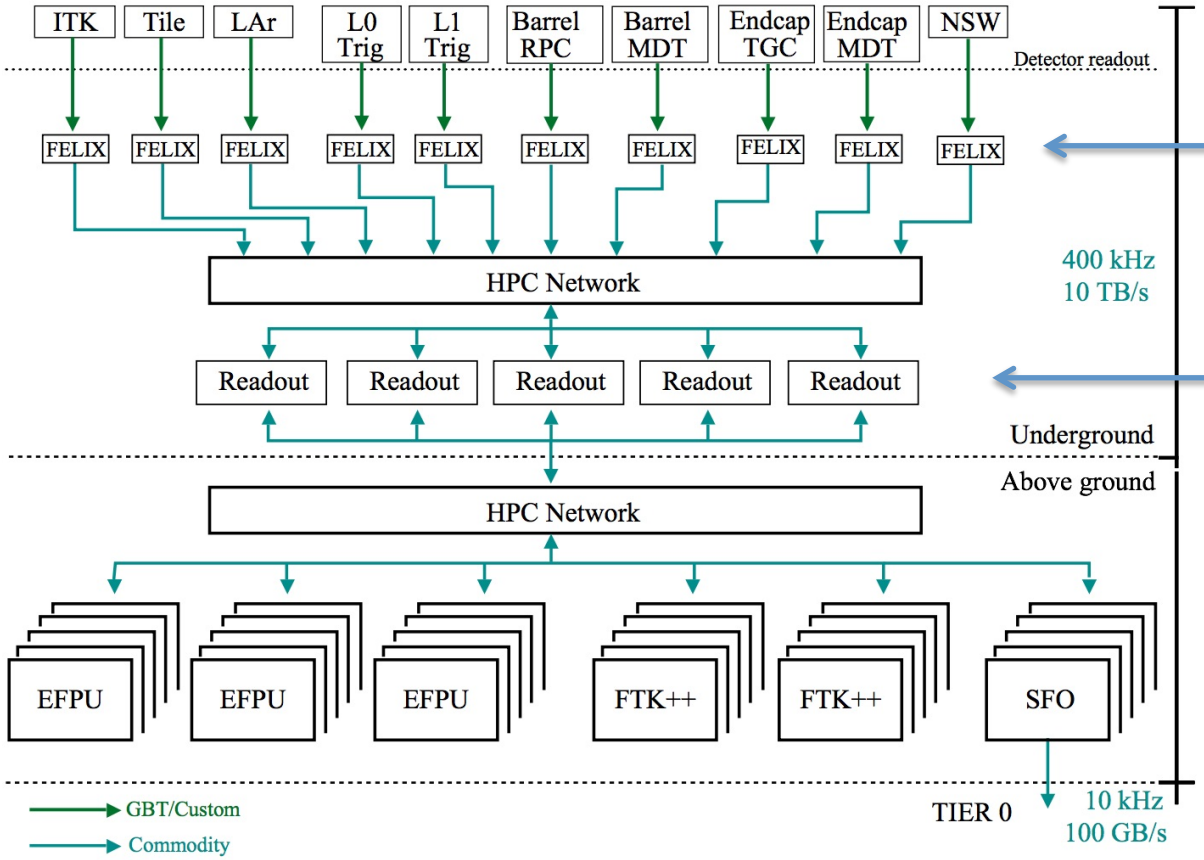


- Use PCI40 from LHCb

Atlas Trigger DAQ



Atlas DAQ "Readout"



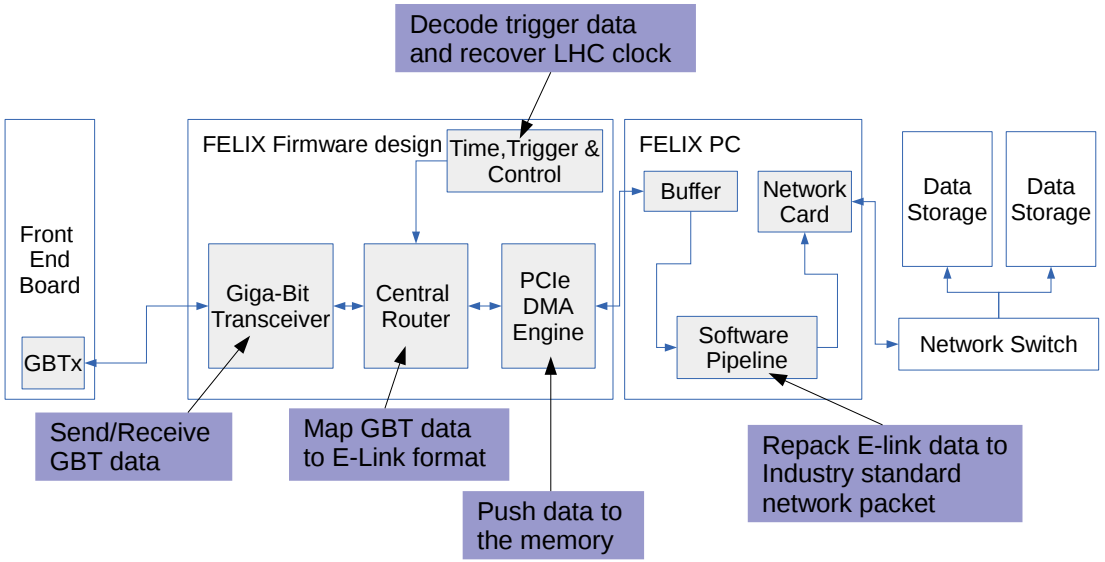
FELIX: common HW platform FPGA+PC

Additional Layer Common PC platform "ROS"

- Subdet specific processing
- Monitoring, calibration

Atlas FELIX implementation for Phase-I

FELIX Functional View



Aug 6, 2015 DPF meeting 2015, Ann Arbor 9

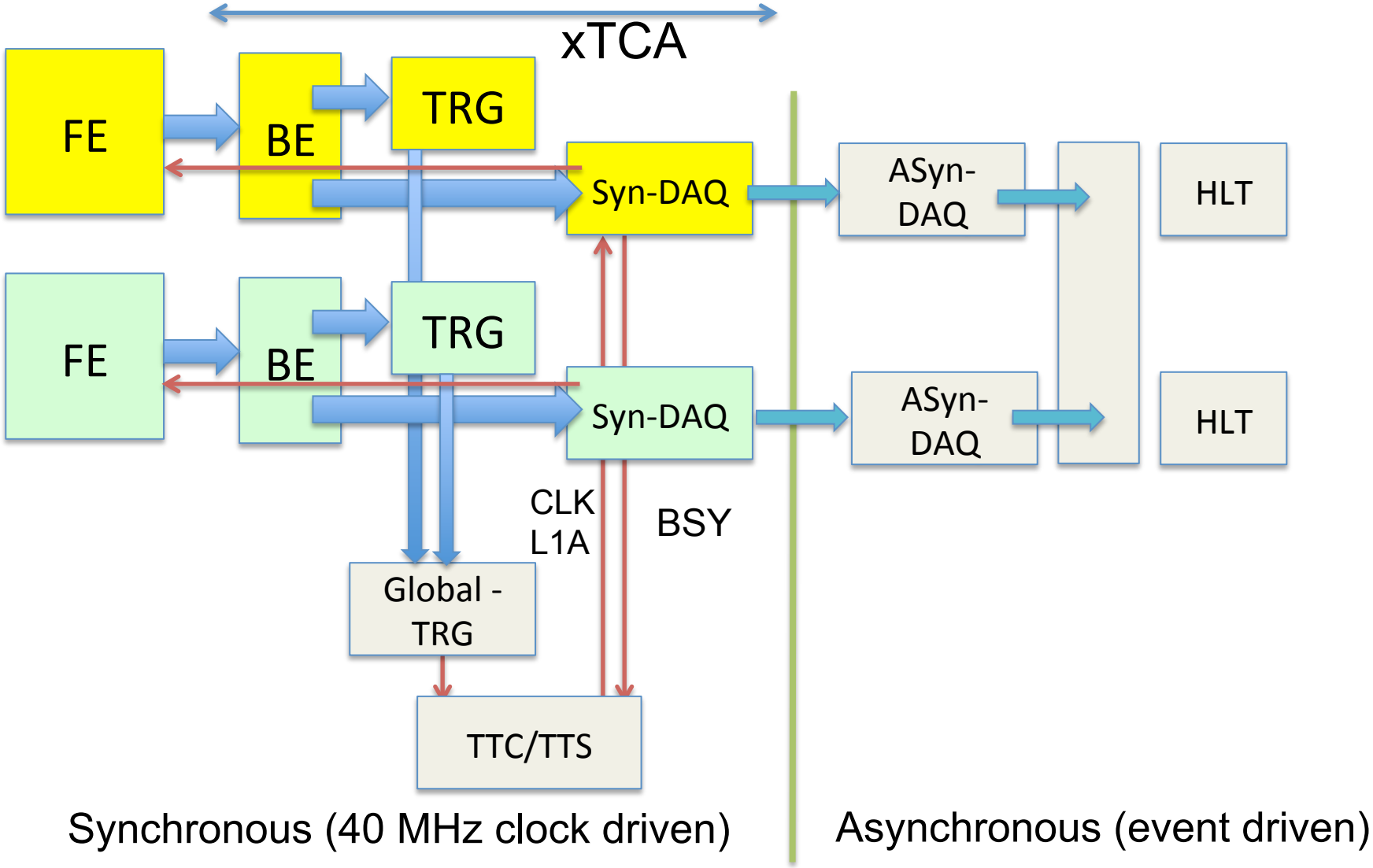
- Atlas FELIX looks similar to PCI40 architecture of Alice / LHCb

Atlas FELIX for phase-II

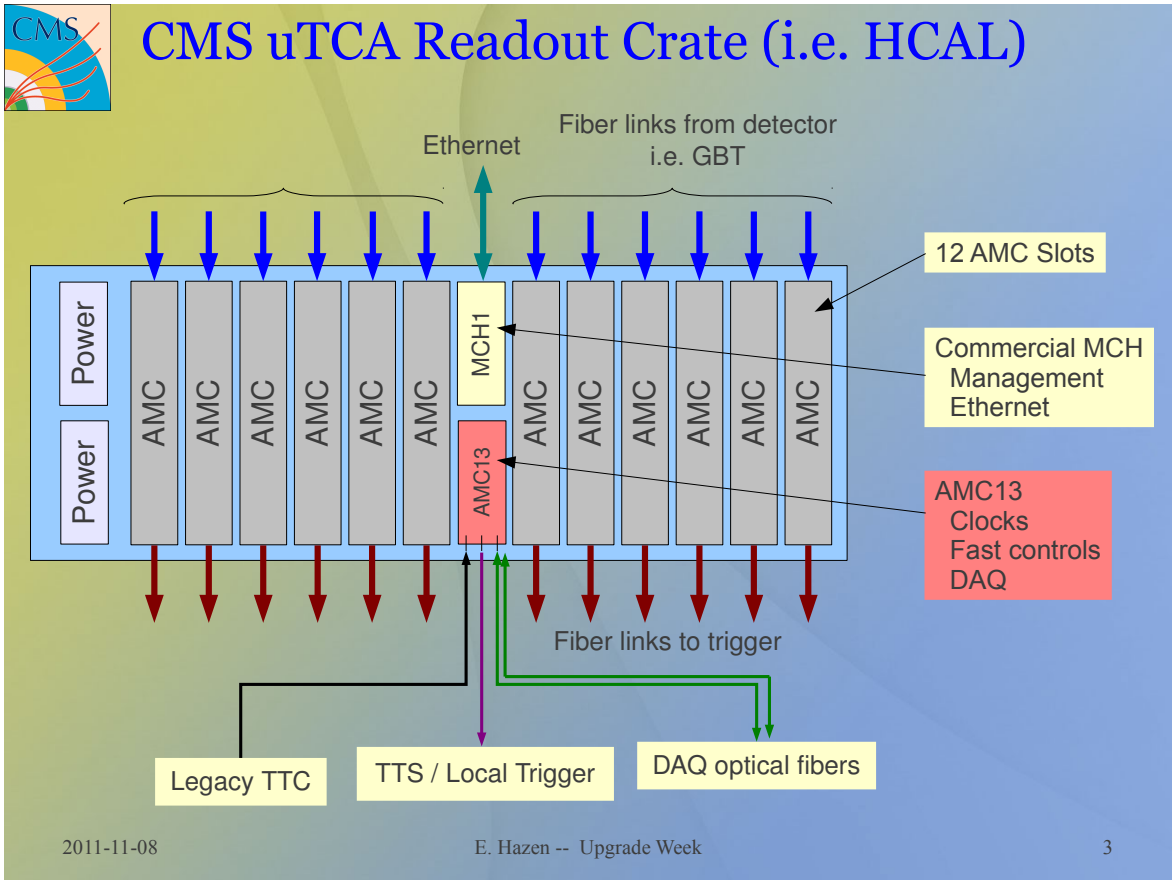
- Same architecture but evolving implementation
  - Common hardware platform
  - Possibly sub-detector specific FW plug-in
- kind of Mezzanine card in a server
  - PCIe gen4, other system IO-bus, ..

- Additional requirement
  - In L0/L1 case, need to bifurcate Trigger data for tracker
    - about 10% of data due to ROI

CMS Trigger DAQ

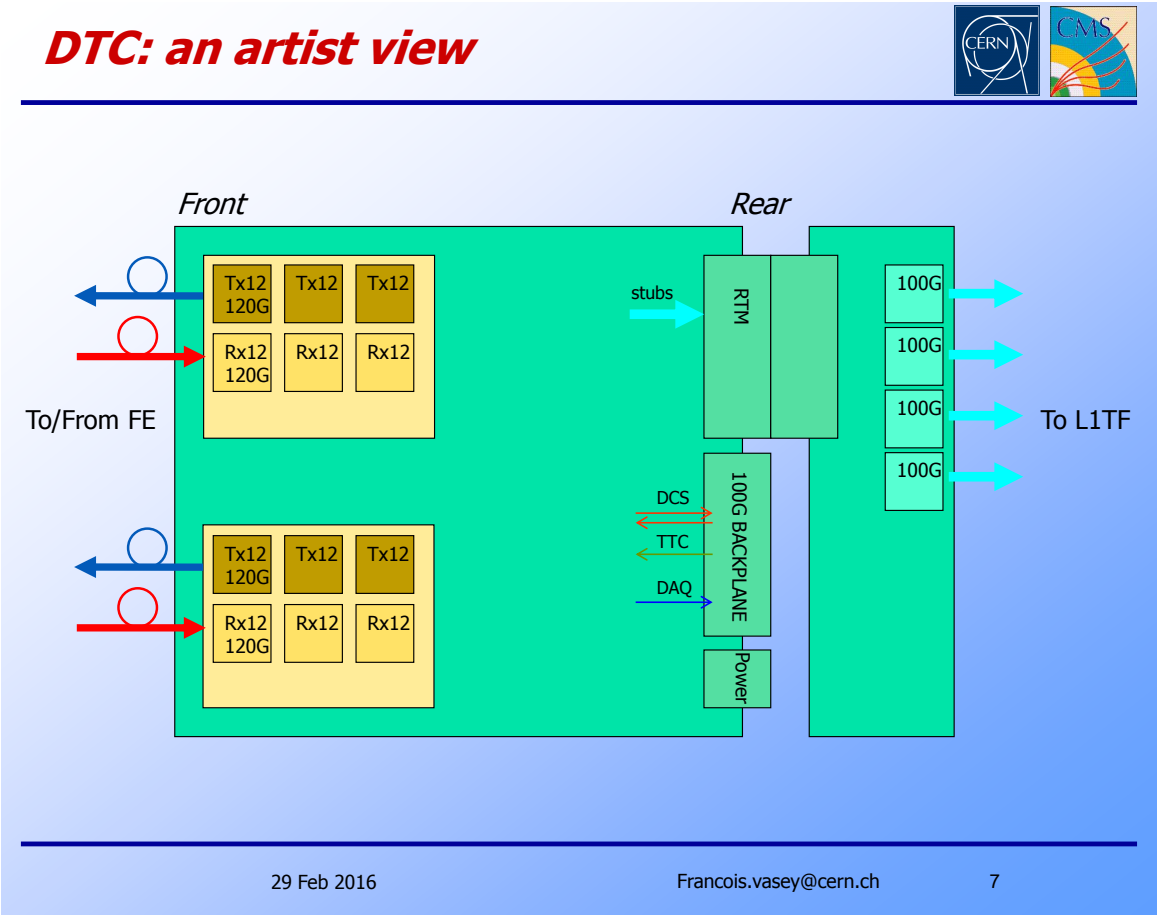


CMS phase-I upgrade



- Bifurcate stream coming from FE to Trigger
- Common Control+Timing HUB and DAQ concentrator (AMC13)
- In L1 upgrade re-use of various AMC cards (plug-in FW)

CMS Phase-II read-out example (TK)



- Note: BW for Trigger/DAQ is ~80/20 %

CMS ideas for DAQ - RU

- Synchronous part done in xTCA
- Asynchronous part of DAQ, some options
  - Point-to-point link to intermediate cDAQ “Hub” card
    - Mezzanine on leaf card
    - From leaf card via xTCA backplane to other slot
      - when concentration useful
  - Link and Protocol
    - Custom or Ethernet L2 or TCP when sufficient memory resources
    - Back pressure
  - cDAQ RU (preferably on surface)
    - PC with a commercial NIC (Ethernet L2), or custom card
    - Concentrator, buffer
    - Protocol converter to EVB network (Ethernet or HPC)

EVB

EVB design / implementation choices

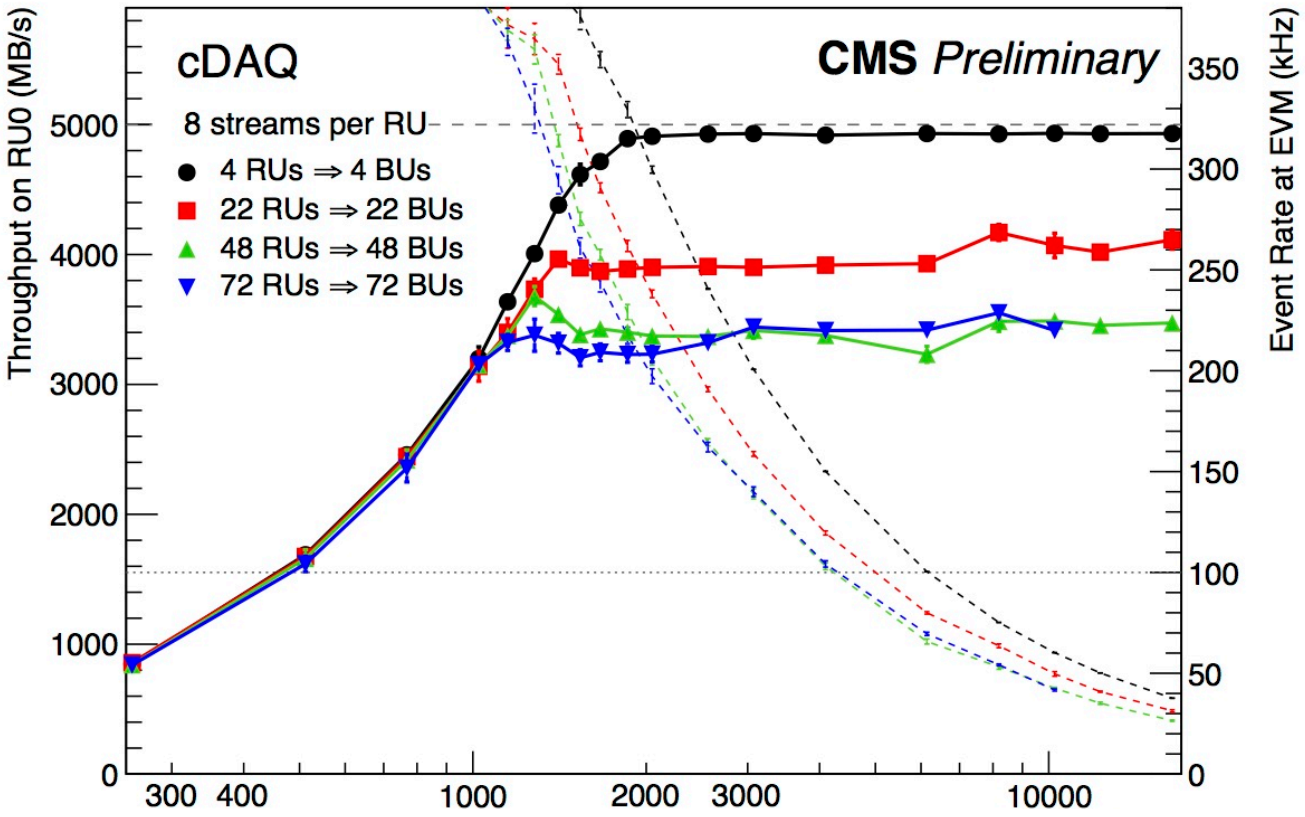
- ROI based or full event building
  - Large implications for EVB – HLT interface and HLT framework
- Lossy (incomplete events) or Lossless EVB
  - All experiments are or move to Lossless
- Network technology
- Protocols
- Event Manager to control EVB process and optionally throttle trigger
- Effective throughput vs bi-section bandwidth
  - "folded" EVB to use bi-directional links
  - Traffic shaping ?
    - Studied by Alice

EVB network technology

- Ethernet
  - Switches with or without “deep buffers”
  - L2 or TCP/IP ?
- HPC: Infiniband, Omnipath, ..
  - Trend for closer integration with CPU (Omnipath)
- 100 Gbps now (4x25 Gbps lanes)
  - Likely 200 Gbps (4x50 Gbps lanes) at LS3

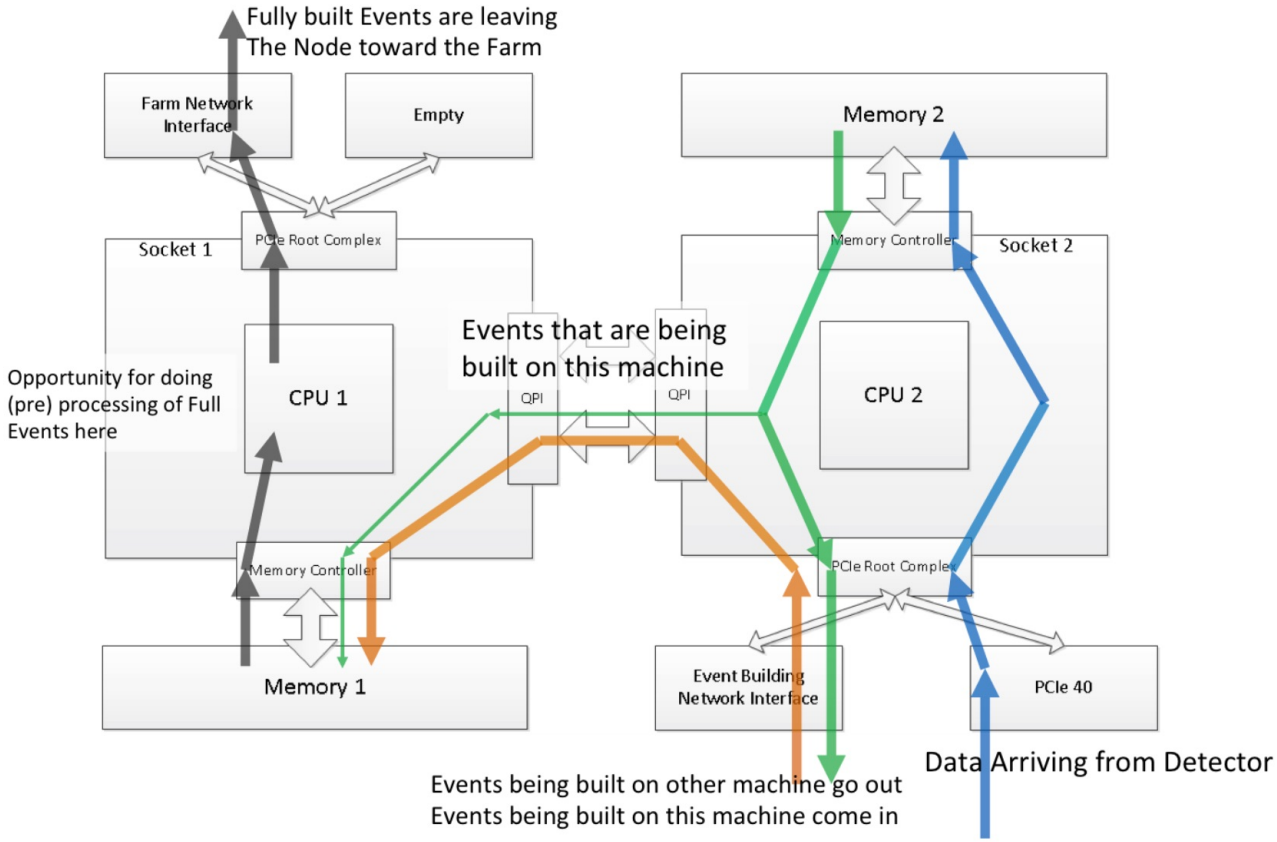
- Interfacing to FPGA based Read-Out Unit
  - Ethernet:
    - Layer2 ok, but transmission unreliable
    - (reduced) TCP/IP possible in FPGA, but needs memory for buffering
  - HPC: difficult

EVB scaling and switch technology



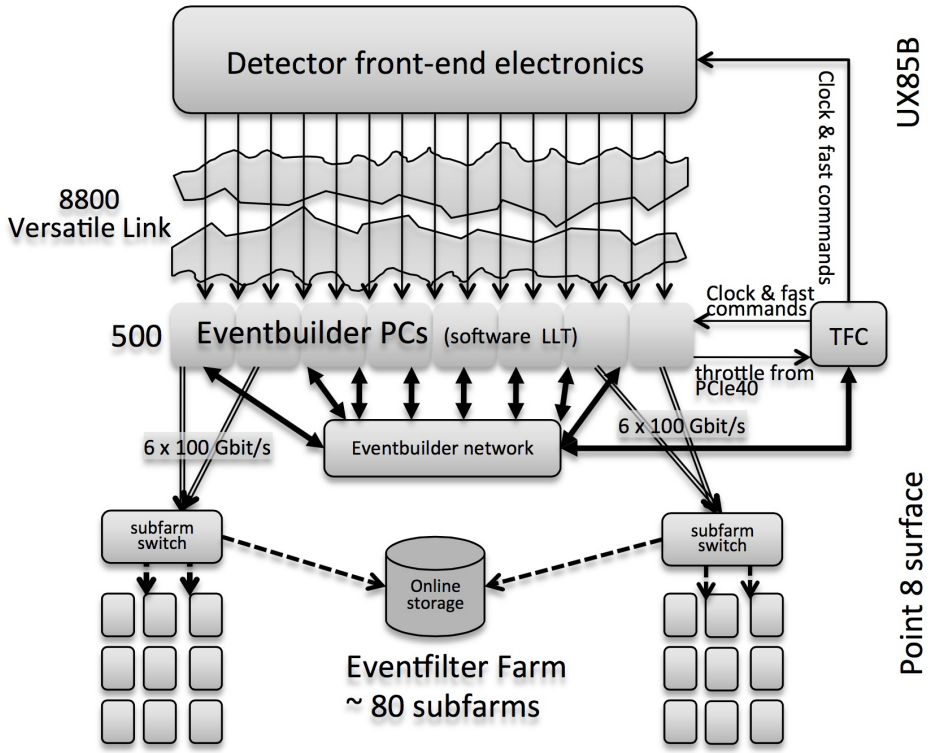
- CMS today Infiniband FDR (56 Gbps) EVB (switches w/o buffers)
- Scaling NxN: Efficiency is about 50% of line BW

Optimization: use of bi-directional links (LHCb)



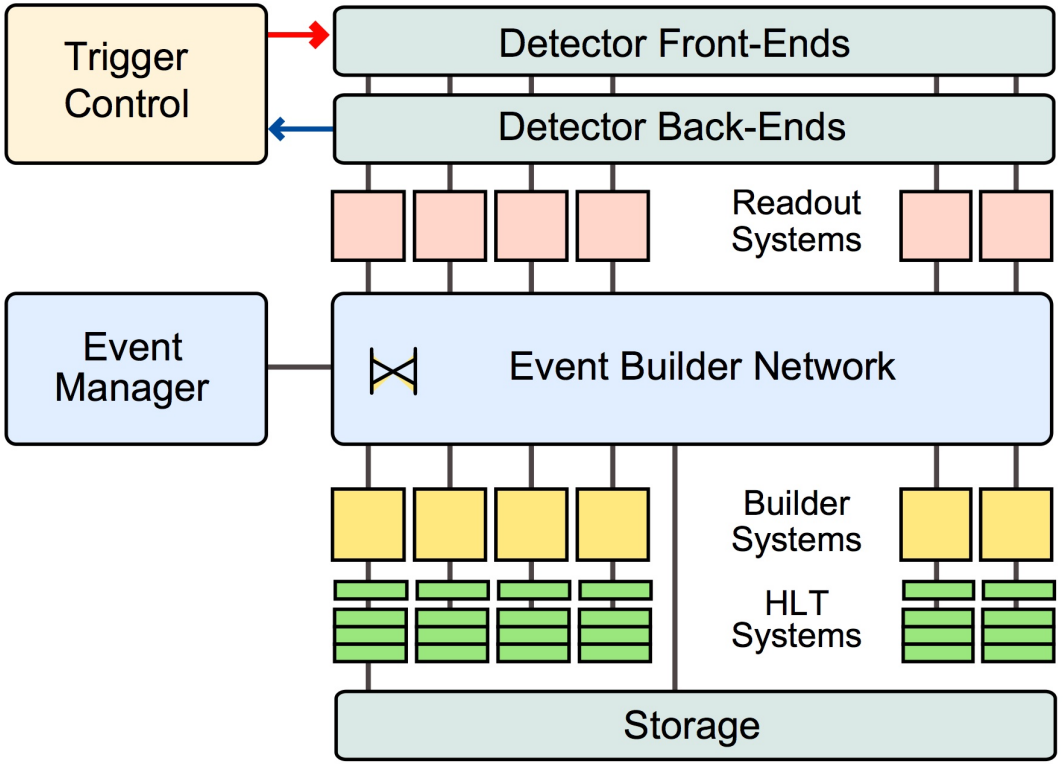
- Same event-builder node is used for read-out unit and builder unit

LHCb EVB



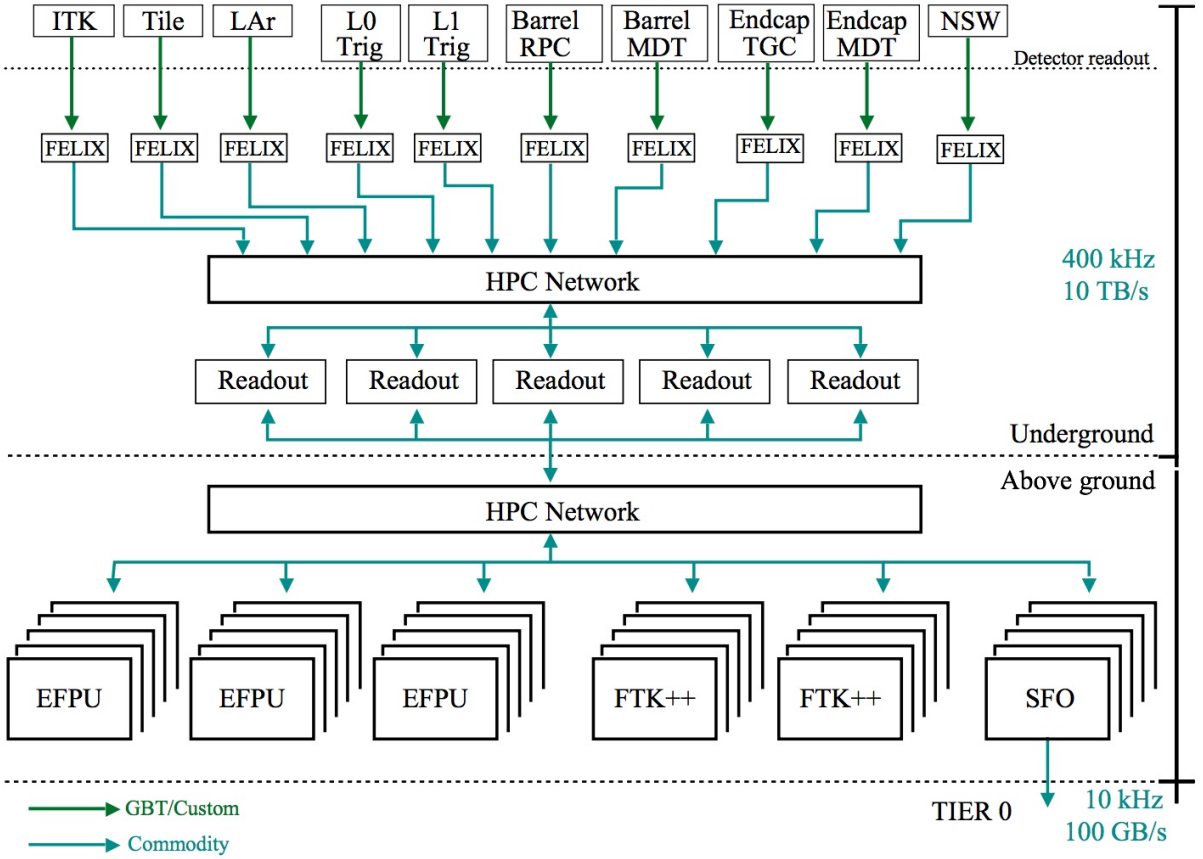
- 4 TB/s throughput
- Full event building, event assignment distributed by TTC system
- Bi-directional event building
- Assume 500 servers, 100 Gbps links, ~50% efficiency

CMS EVB baseline



- 4 TB/s throughput
- Full event building
- Assume 400 servers, 200 Gbps links, ~30% efficiency

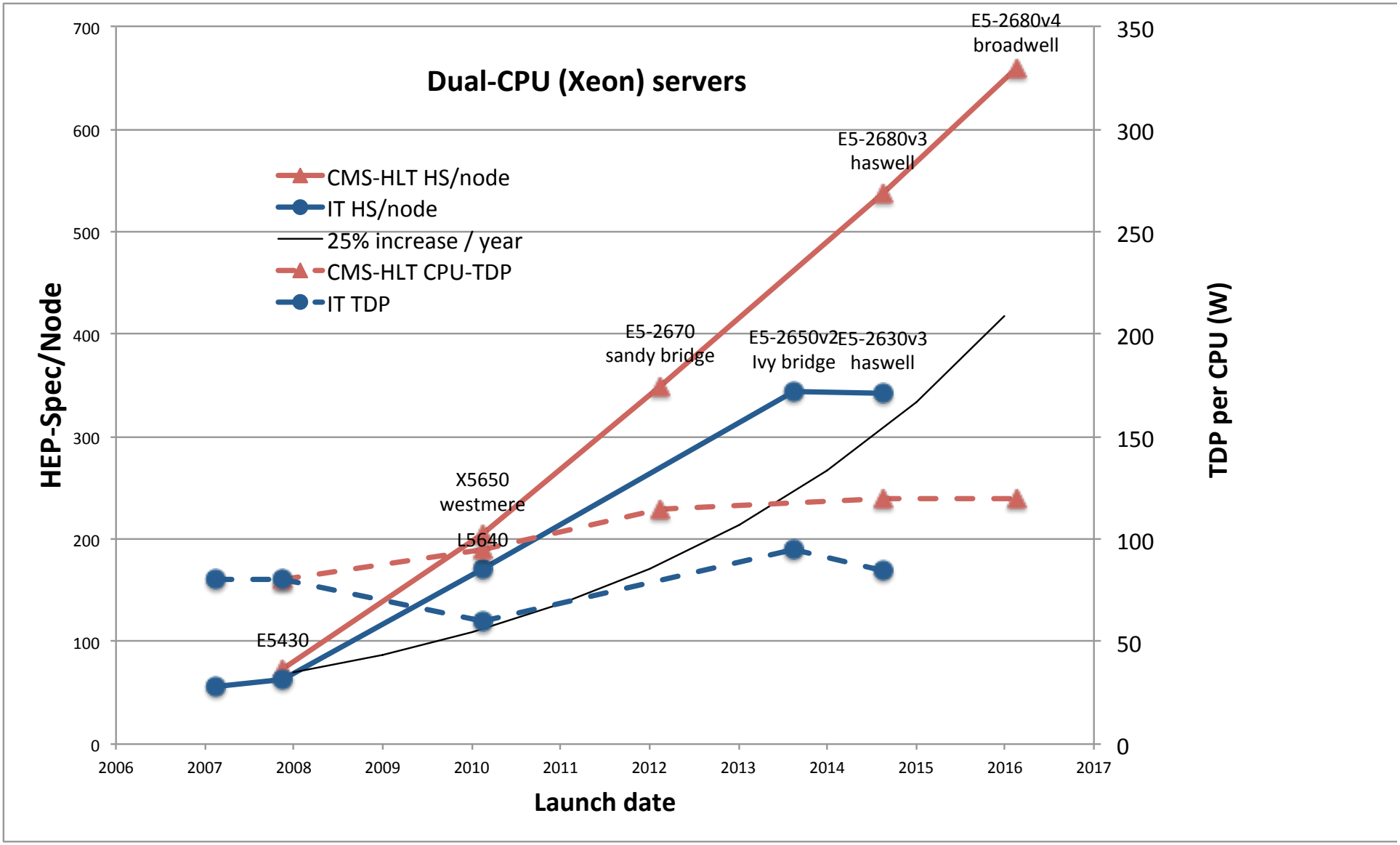
Atlas EVB/HLT



- ROI based event building, on demand by HLT processes
- Object store with large capacity connected to switch

HLT facility

Dual-Xeon server evolution



Extrapolation of HLT farm needs (Atlas/CMS)

- Atlas (Phase-II scoping doc – Sep 2015)
  - Estimate
    - Use L1 TT at 1 MHz, FTK++ at 100 kHz, CPU farm input 400 kHz
  - Result
    - Need ~11 MHS06, split ~50/50 in to FTK++ / CPU
- CMS (TP and scoping doc)
  - Guestimate based on
    - Current detector, PU dependence, possible gain with using L1 Track trigger
  - Result:
    - PU = 140 / 200 and HLT input 500 / 750 kHz: need ~ 5.0 / 11.0 MHS06
- To fix ideas:
  - Need 11 MHS06 in 2026 (Atlas/CMS has ~ 0.7/0.5 MHS now)
    - Ignore that likely LHC-HL will not start with PU=200
  - Compare LHCb: need 3.3 MHS06 in 2021

Extrapolation for HLT farm (dual Xeon based)

- To fix ideas: Need ~11 MHS for HLT in 2027 (Atlas/CMS)
- Observed (~ 2007 – 2015) dual-Xeon servers
  - Roughly constant Power, constant cost per server (with CERN IT "sweet-spot")
  - Start point Q1-2016: server with 0.66 kHS and 0.35 kW and X kCHF

Assumed perf. increase per year of servers	25% / year [WLCG 2014]	12.5 % / year
11 years	12	3.7
#servers in Q1-27	1431	4562
Total power Q1-27	0.5 MW	1.6 MW

- Comments
  - Ignore installed base from run-3
  - Current Atlas and CMDS datacenter ~1 MW cooling looks insufficient
    - Likewise Alice, LHCb specify 2 MW cooling (containerized data center)

Processors for HLT farm

- Dual-Xeon
- Non x86 under consideration
  - ARM etc
- Commercial specialized (co-)processors under consideration
  - Multi-core (Xeon-Phi etc)
  - Co-processing with commercial GPU, FPGA boards
  - Alice has experience with GPU, but Alice's use case different from pp
- Custom co-processors
  - Co-processing with custom HW (FTK++ in Atlas with Ass. Memory)

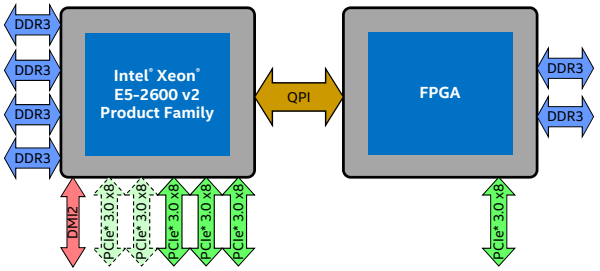
Specialized (co-) processors

- CMS experience, so far
  - CMS “standard” code, alternative platforms not competitive (price/perf) compared to Xeon (See http://arxiv.org/pdf/1510.03676v1.pdf)
- Very active area of research
  - Eg “Connecting the DOTS 2016” for tracking
- Concern
  - (fine grain) parallelization and vectorization
    - Requires extensive re-engineering of the code
  - Portability and long-term maintainability
- HLT farm versus worldwide Tier-n centers
  - HLT farm is under control of experiment, so easier to deploy alternative platforms

An industry trend: CPU+FPGA acceleration

IVB+FPGA Software Development Platform

Software Development for Accelerating Workloads using Xeon and coherently attached FPGA in-socket

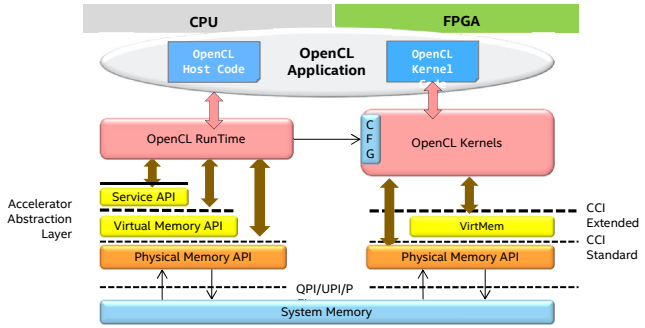


Processor	Intel® Xeon® E5-26xx v2 Processor
FPGA Module	Altera Stratix V
QPI Speed	6.4 GT/s full width (target 8.0 GT/s at full width)
Memory to FPGA Module	2 channels of DDR3 (up to 64 GB)
Expansion connector to FPGA Module	PCIe 3.0 x8 lanes - maybe used for direct I/O e.g. Ethernet
Features	Configuration Agent, Caching Agent,, (optional) Memory Controller
Software	Accelerator Abstraction Layer (AAL) runtime, drivers, sample applications

Heterogeneous architecture with homogenous platform support

intel | 14

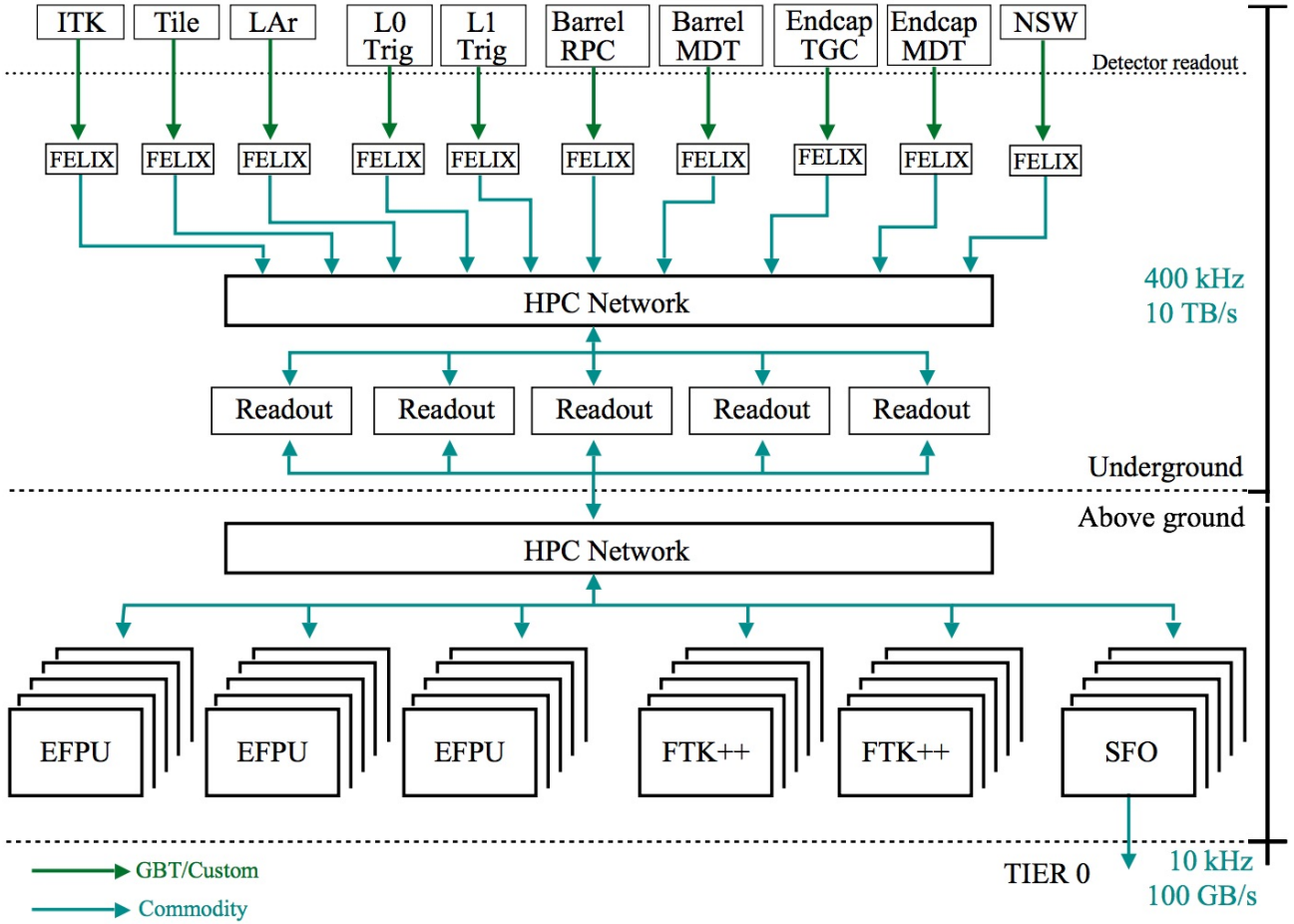
Programming Interfaces : OpenCL



Unified application code abstracted from the hardware environment
Portable across generations and families of CPUs and FPGAs

intel | 16

Atlas HLT CPU / custom FTK++



- HLT process (ROI based) can request processing of event by FTK++ processors via proxy connected to HPC network

STORAGE

Storage

- Alice
  - 80 GB/s, capacity 70 PB
  - Considering Cluster File System (GPS,Lustre) , “Object Store”, EOS, ..
- LHCb
  - 2 GB/s
- Atlas
  - 50 GB/s
  - Use the “Object Store” of the EVB-HLT
- CMS
  - 40 GB/s
  - Could use Cluster File System as now, or something else

CONCLUSION

Areas of R&D for DAQ

- Servers for IO EVB nodes
  - PCIe bus
  - Zero copy, Memory effective throughput
- EVB and HLT architecture
  - Full event building versus ROI
- Network fabrics for Event Collection, building, distribution
  - Ethernet or HPC
  - Traffic shaping ?
- Use of high level languages for FPGA programming
- Large in memory Object store, Storage systems
- Use of SW technologies
  - Eg data analytics tools
- Data center

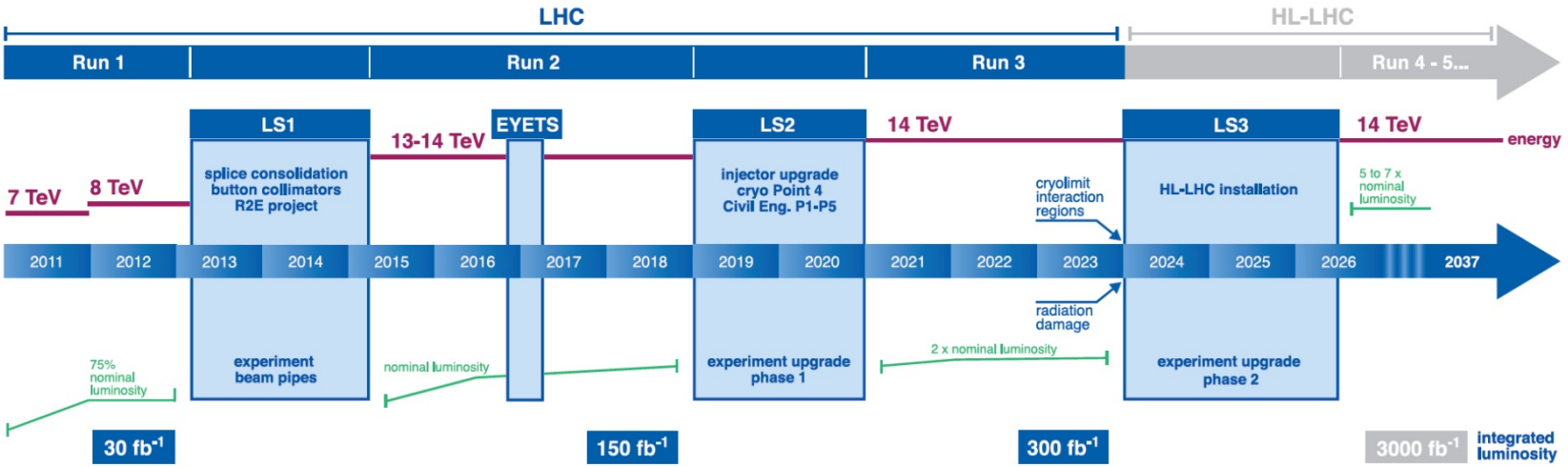
Final Remark

- People effort for
  - Development of SW for control, configuration, monitoring
  - Integration and maintenance
- Very significant compared to the effort for
  - DAQ data flow SW in strict sense
- Hence
  - There is a gain from uniformity by using common HW and SW components

- Trend using common (commercial?) HW, FW and SW building blocks with “plug-ins” for sub-detector specific functionality

EXTRA MATERIAL

LHC / HL-LHC Plan



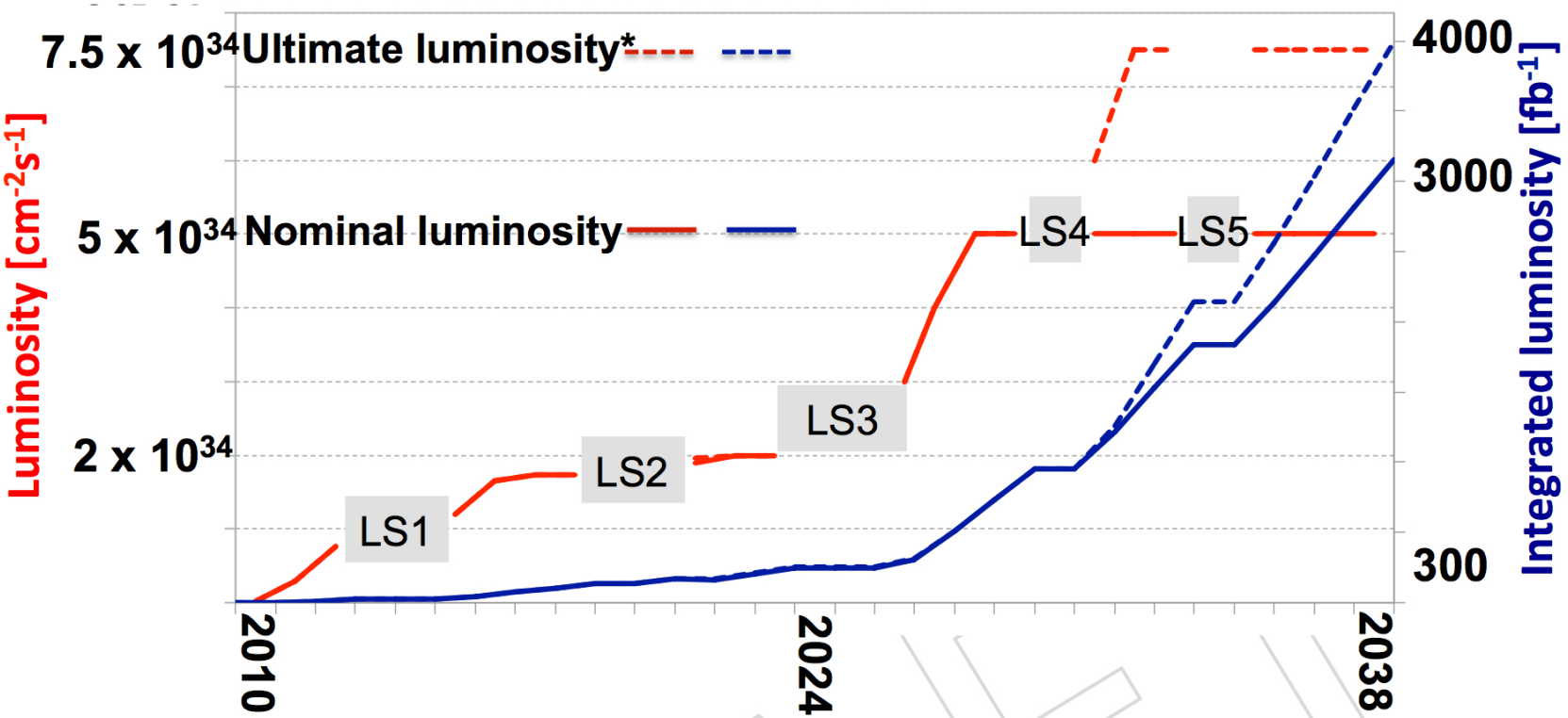
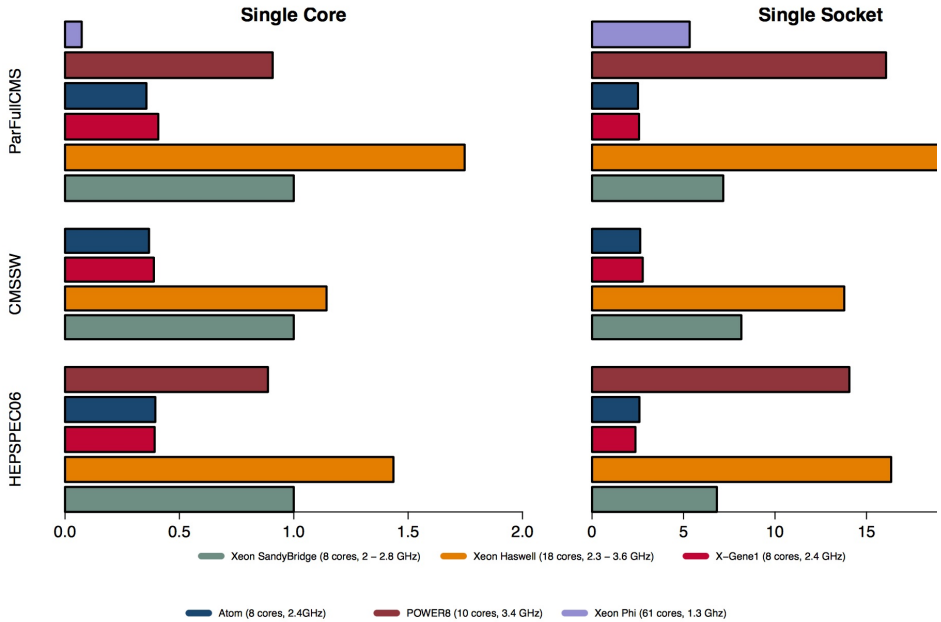
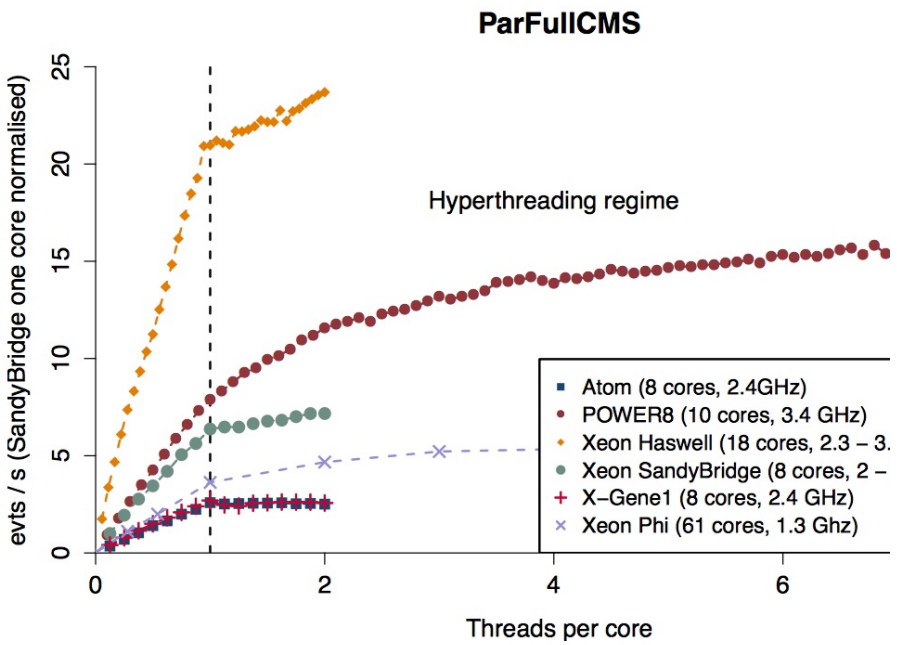


Figure 7: Projected LHC performance through 2040, showing present schedule for long shut-downs of LHC and projected luminosities.

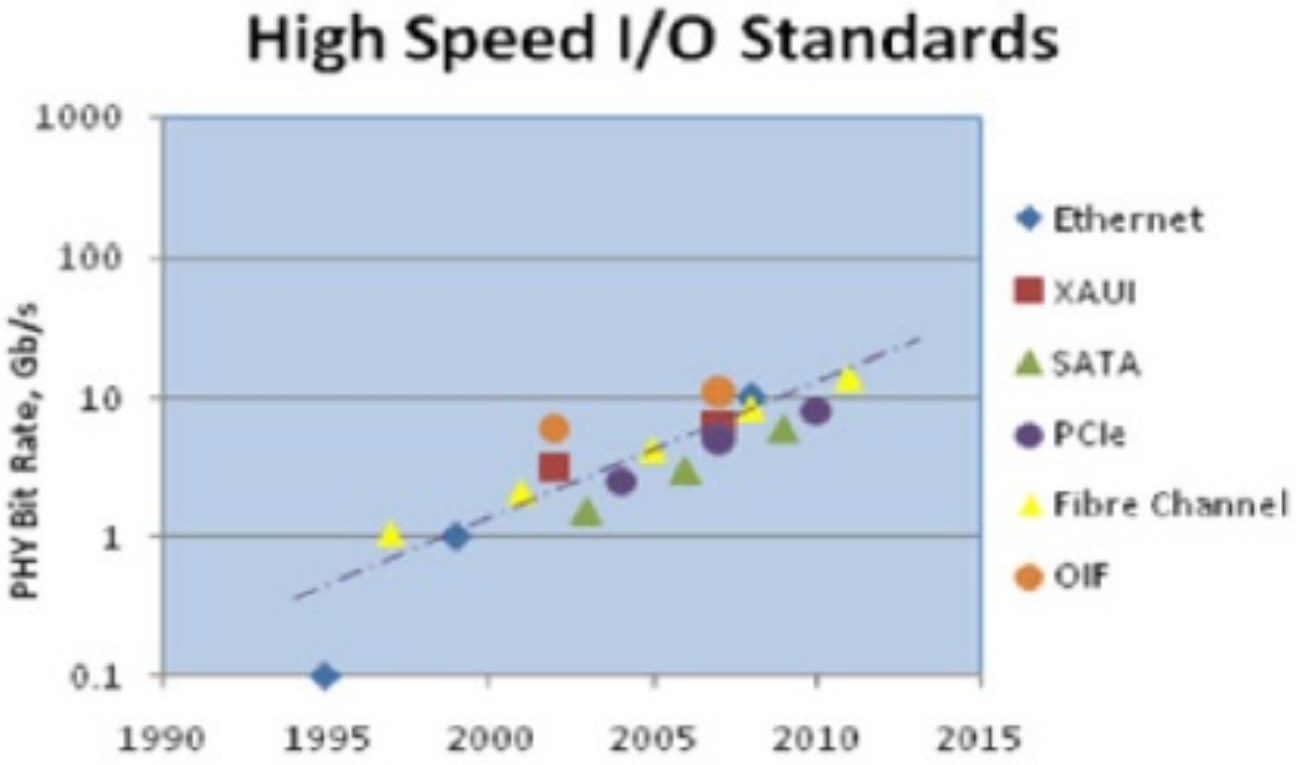
CMS standard offline code on specialized processors





- Not competitive compared to dual-xeon (2015)
  - See http://arxiv.org/pdf/1510.03676v1.pdf

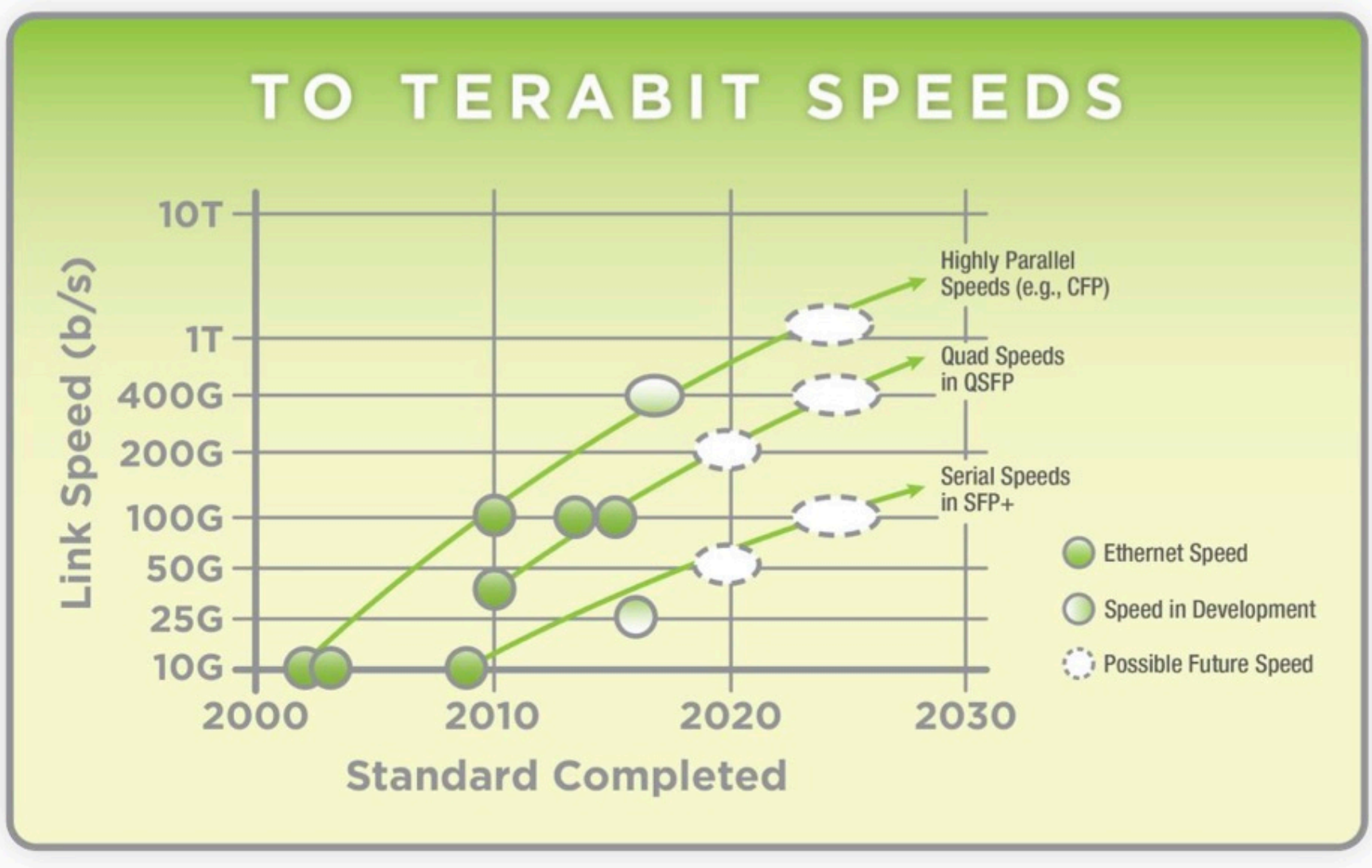
Links and Serialisers



per "lane"

- FPGA xilinx-7 (2013) 10 Gbps
- PCIe gen4 16 Gpbs (2017-18)
- IB EDR 25 Gbps (2015)
- Future HPC Fabric lanes
  - 25 Gbps (2015-2016) , 32 Gbps (2017-18), 50-56 Gbps (19-20)

Ethernet


TO TERABIT SPEEDS
Link Speed (b/s)
10T
1T
400G
200G
100G
50G
25G
10G
2000
2010
2020
2030
Standard Completed
Highly Parallel Speeds (e.g., CFP)
Quad Speeds in QSFP
Serial Speeds in SFP+
Ethernet Speed
Speed in Development
Possible Future Speed