The Phase-2 Upgrade of the CMS Data Acquisition

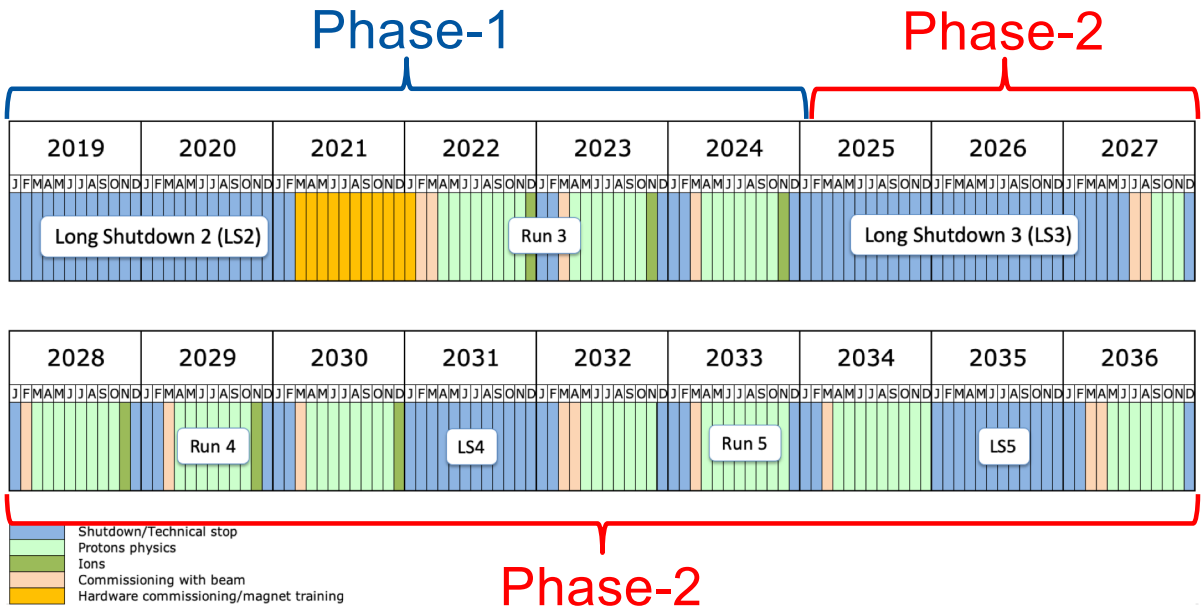
E. Meschi – CERN EP/CMD
for the CMS Data Acquisition Project

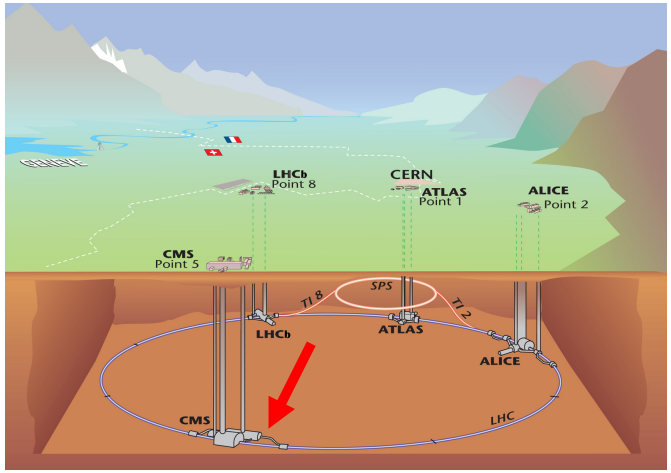
The main purpose of the data acquisition system (DAQ) is to provide the **data pathway**, **aggregation**, and **time decoupling** between the **synchronous detector readout** and data reduction, the **asynchronous selection** of interesting events in the HLT, their **local storage** at the experiment site, and the **transfer** to offline computing for reconstruction, permanent storage and analysis.

Historically, the DAQ requirements of HEP experiments **have been met at a cost of order 7-10%** of the total cost of the experiment / upgrade.

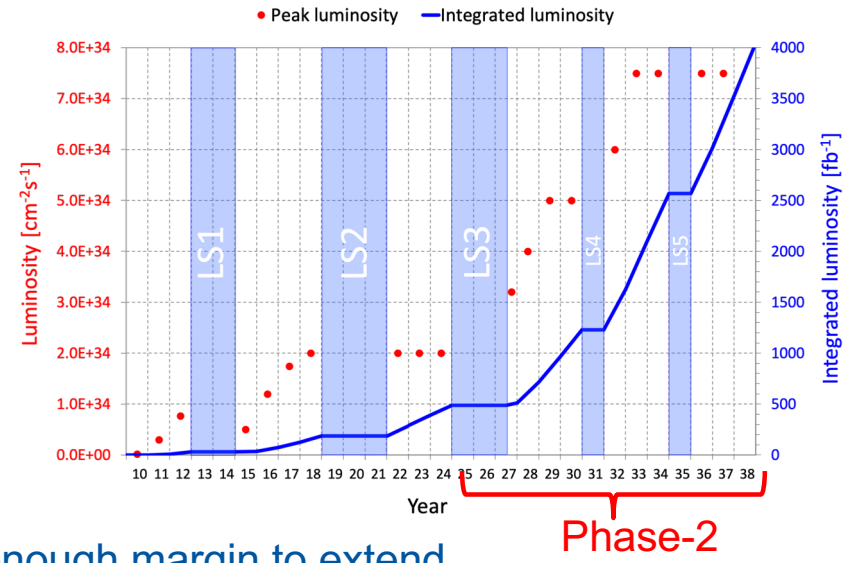
vCHEP2021 – 21 May 2021

High Luminosity LHC





	$\mathcal{L}$	$\langle$ PU $\rangle$	Vertex Density	$\int \mathcal{L}$ / year
Baseline	$5 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$	140	0.8 / mm	$250\,\mathrm{fb}^{-1}$
Ultimate	$7.5 \cdot 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$	200	1.2 / mm	$> 300\,\mathrm{fb}^{-1}$



Target (L1 and HLT): same physics reach as Phase1 + enough margin to extend reach (e.g. extended $|\eta|$ range, VBS, VBF, LLP, displaced muons, light mesons…)

CMS HL-LHC Upgrade

Technical proposal CERN-LHCC-2015-010 https://cds.cern.ch/record/2020886
Scope Document CERN-LHCC-2015-019 https://cds.cern.ch/record/2055167

L1-Trigger/HLT/DAQ

https://cds.cern.ch/record/2283192
https://cds.cern.ch/record/2283193

- Tracks in L1-Trigger at 40 MHz
- PFlow-like selection 750 kHz output
- HLT output 7.5 kHz

Barrel Calorimeters

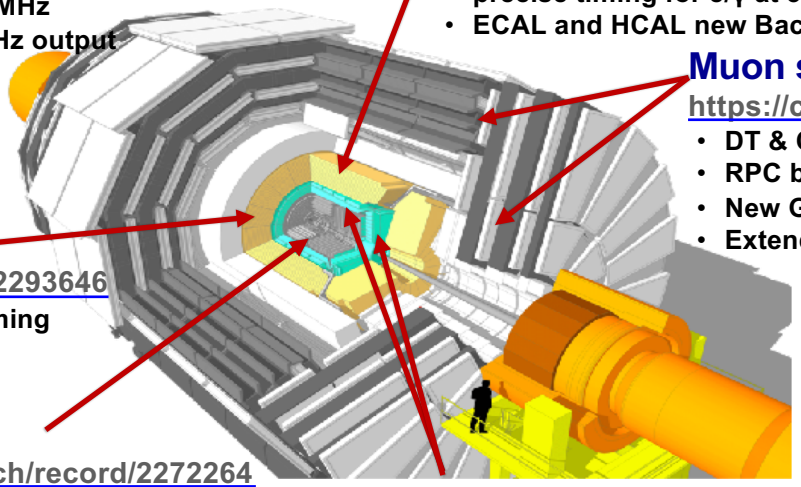
https://cds.cern.ch/record/2283187

- ECAL crystal granularity readout at 40 MHz with precise timing for e/γ at 30 GeV
- ECAL and HCAL new Back-End boards

Muon systems

https://cds.cern.ch/record/2283189

- DT & CSC new FE/BE readout
- RPC back-end electronics
- New GEM/RPC $1.6 < \eta < 2.4$
- Extended coverage to $\eta \simeq 3$



Calorimeter Endcap

https://cds.cern.ch/record/2293646

- 3D showers and precise timing
- Si, Scint+SiPM in Pb/W-SS

Tracker https://cds.cern.ch/record/2272264

- Si-Strip and Pixels increased granularity
- Design for tracking in L1-Trigger
- Extended coverage to $\eta \simeq 3.8$

MIP Timing Detector

https://cds.cern.ch/record/2296612

Precision timing with:

- Barrel layer: Crystals + SiPMs
- Endcap layer: Low Gain Avalanche Diodes

Beam Radiation Instr. and Luminosity, and Common Systems and Infrastructure

https://cds.cern.ch/record/002706512

New paradigms (design/technology) for an HEP experiment to fully exploit HL-LHC luminosity

~50000
FE optical links

ATCA modular electronics

FPGAs with ~100 High-speed serial transceivers

L1 Trigger and Readout

CMS detector	LHC Phase-1	HL-LHC Phase-2	
Peak ⟨PU⟩	60	140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size at HLT input	2.0 MB [a]	7.8 MB	9.9 MB

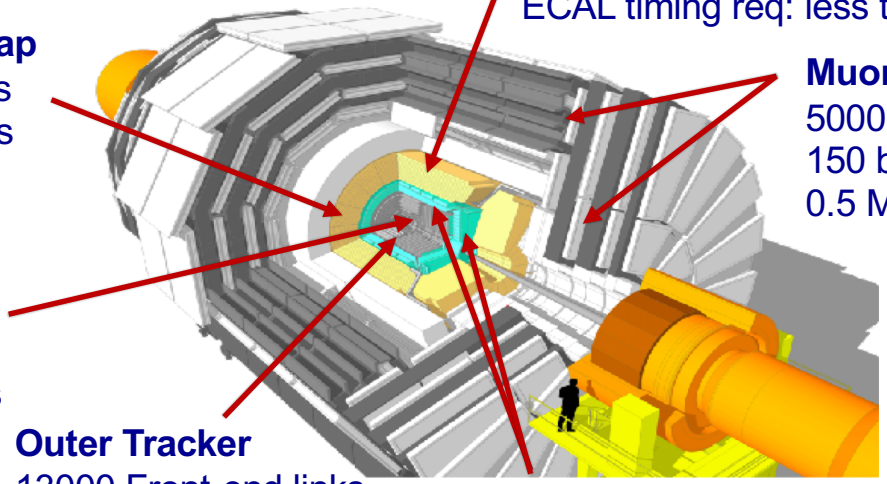
Beam Radiation Instr. and Luminosity, and Common Systems and Infrastructure
Special handling of triggers
Dedicated DAQ

Barrel Calorimeters
10000+ front-end links
144 back-end boards
2.5 MB / evt
ECAL timing req: less than 10 ps RMs jitter

Calorimeter Endcap
8000 front-end links
84 back-end boards
3 MB / evt
126 TPG boards

Muon systems
5000 front-end links
150 back-end boards
0.5 MB / evt

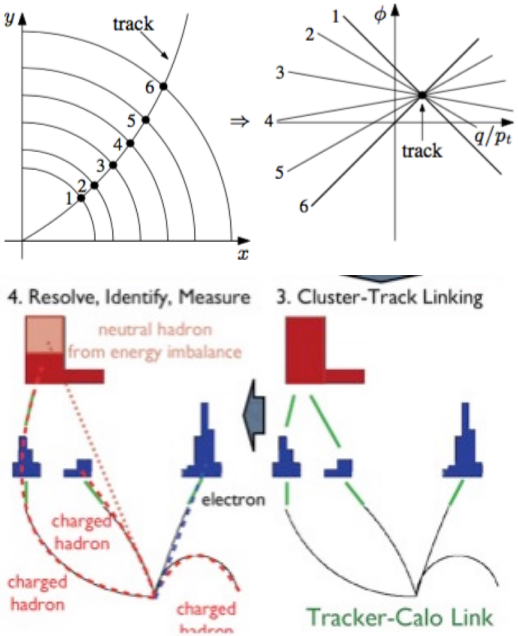
Inner Tracker
2160 front-end links
28 back-end boards
1.4 MB / evt



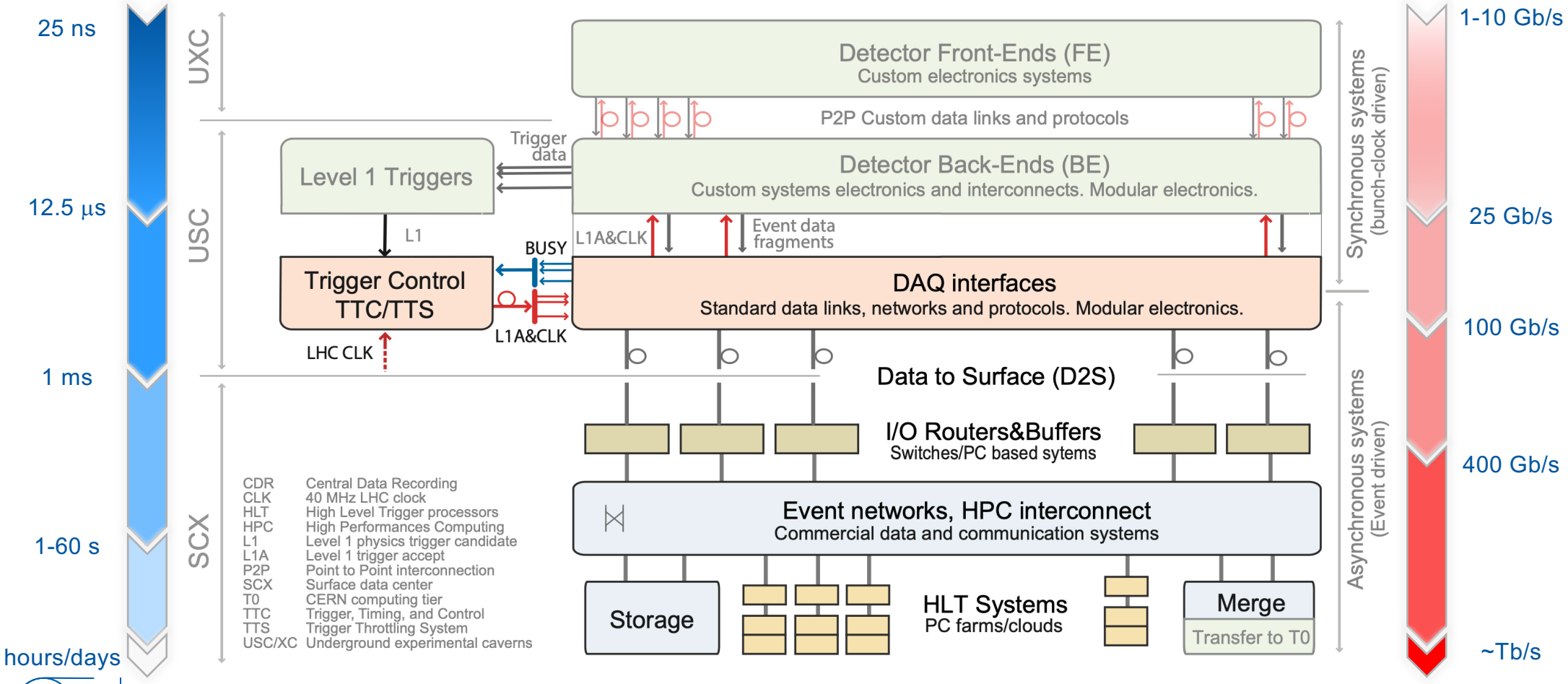
Outer Tracker
13000 Front-end links
216 back-end boards
1.15 MB / evt
180 TF boards

MIP Timing Detector
2500 front-end links
22 back-end boards
0.7 MB / evt
MTD timing req.: less than 15 ps jitter

L1-Trigger
280 back-end boards
0.3 MB / evt



DAQ and HLT: Conceptual View



Technology evolution and DAQ

Adopting new technologies within the **same general conceptual scheme**

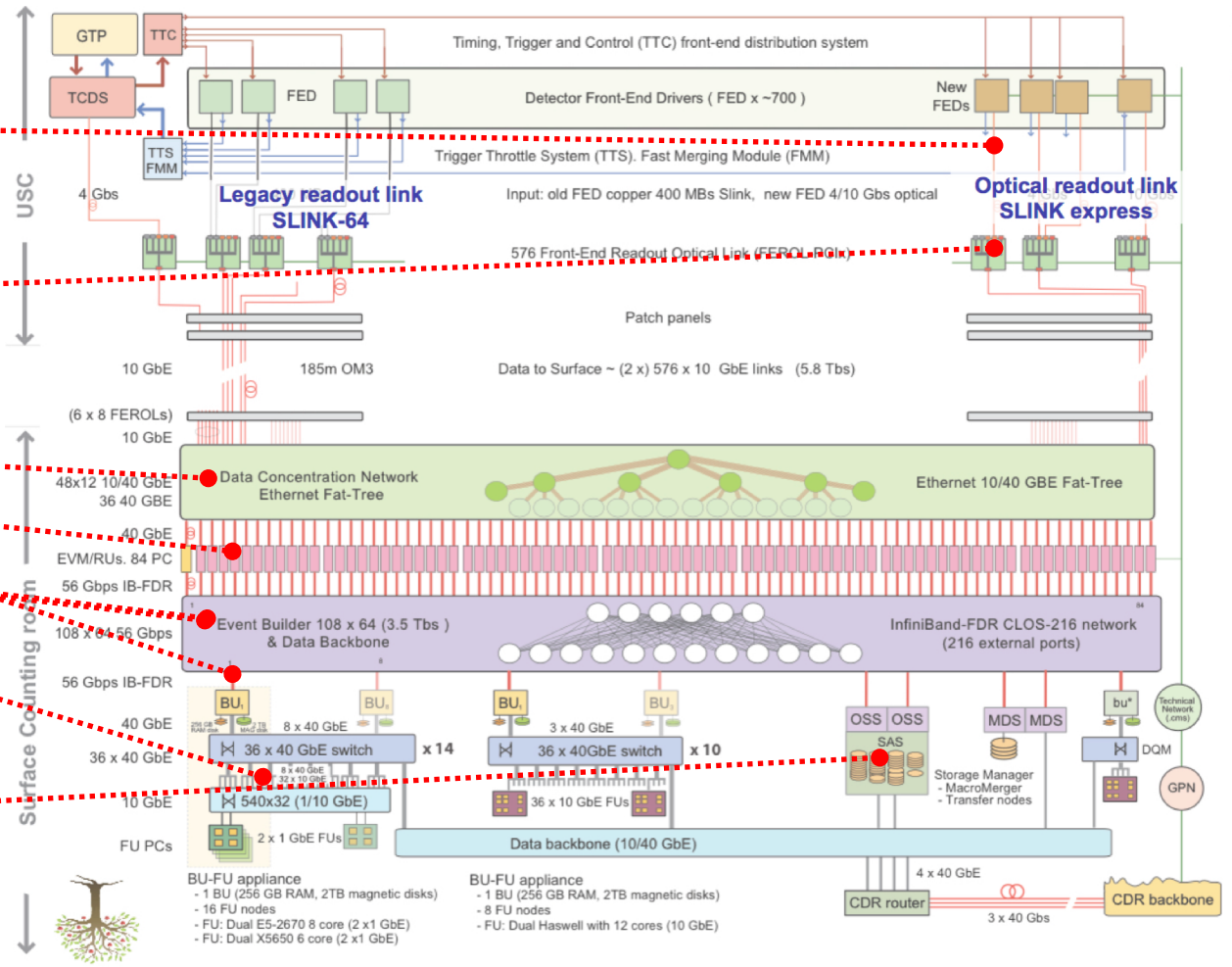
- Modular electronics: from parallel buses to telecom-grade **serial backplanes** and **(optical) point-to-point links**
- FPGA evolution: many **high-speed transceivers** on a single chip
- Networking and cluster interconnects: from niche market for HPC to **hyperscalers** domain
  - Progressive **reduction** of alternative “standards”
  - Continued role (and convergence) of **Ethernet**
- Multi-core architectures, evolution of memory bandwidth, parallelism
  - Combine **more functions** in the same (commodity) hardware
  - Co-processors: **heterogeneous** computing
- **Cloud storage**

Many advances can be tested in a “refresh” and **scaled up** in the next iteration

DAQ technology evolution: Run 2

Phase-1 upgrade

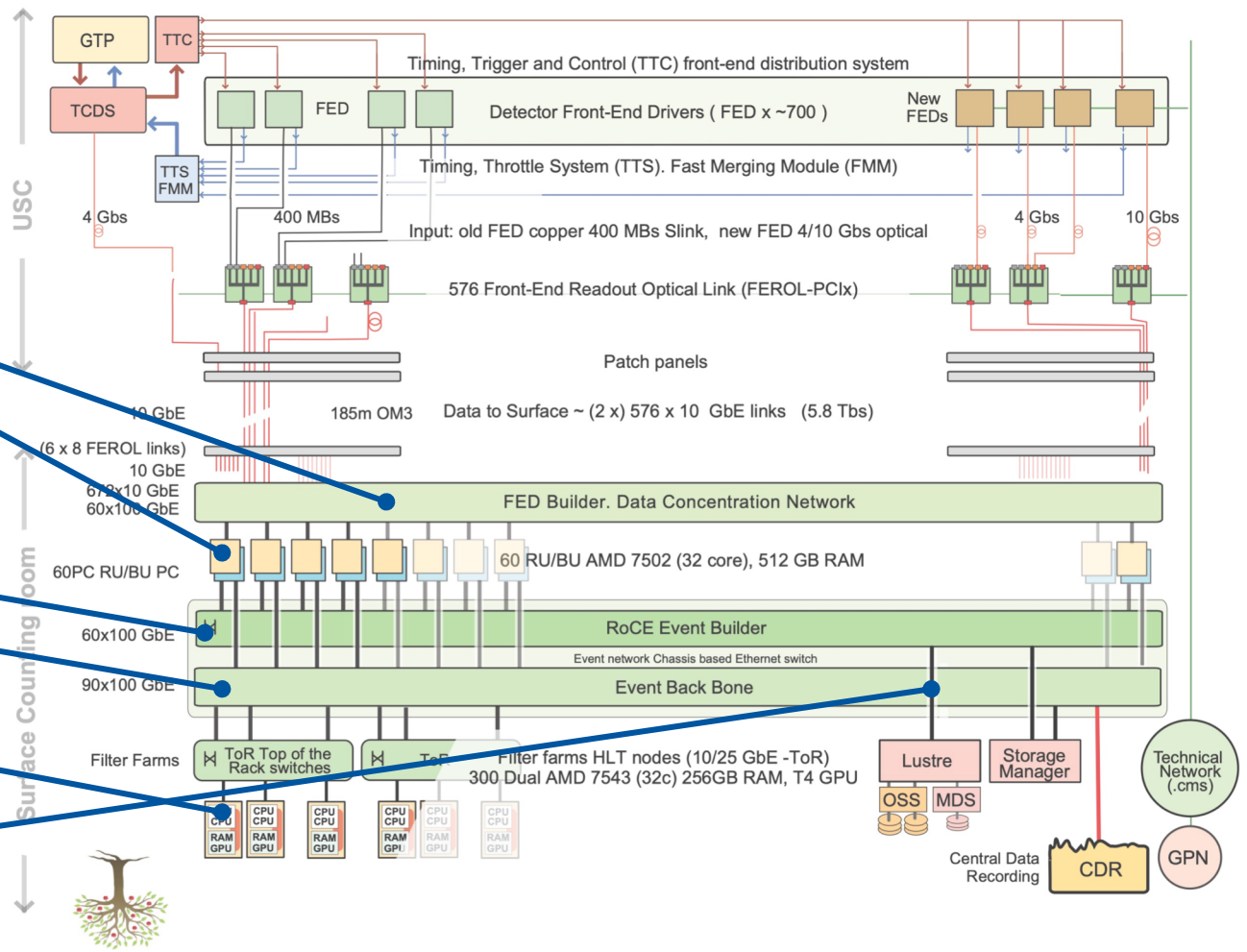
- **Optical** readout links from MicroTCA back-ends
- "Simplified" **TCP/IP in FPGA @10Gb/s**
- 10/40 GbE **"fat-tree"** Data Concentration Network
- 100x100 **"RU-BU" evb**
- **InfiniBand** CLOS FDR switch fabric
- HLT "appliance" w/**10GbE** local switch
- **Cluster** (parallel) **filesystem** based on **Lustre**



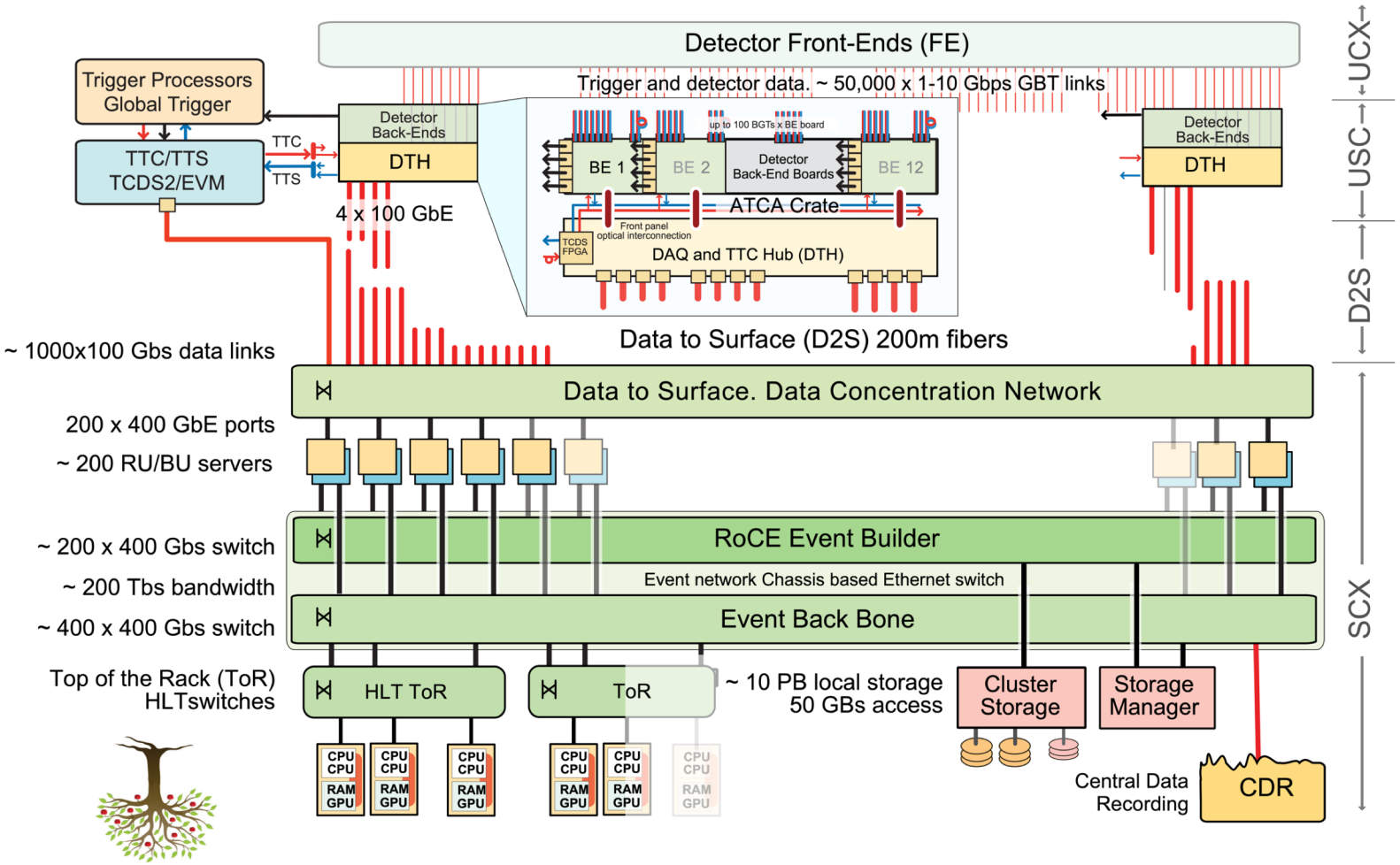
DAQ technology evolution: Run 3

Technology “refresh”

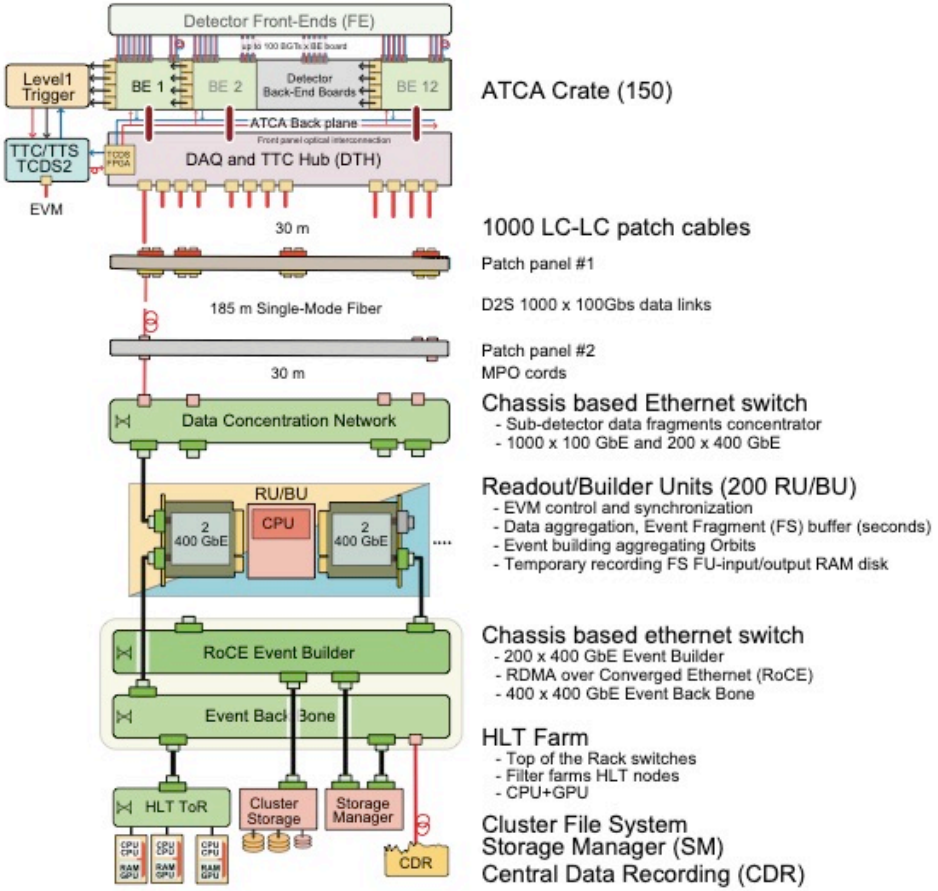
- **10/100 GbE monolithic** Data Concentration Network
- 50-node **“folded” event builder**
- Monolithic **RoCE** 100GbE event builder fabric
- **100 GbE event** backbone
- HLT with **GPU** offload
- Cluster file system on same RoCE network as event builder



Phase-2 DAQ baseline



DAQ technology: Phase-2



- Uniform (ATCA, Xilinx K/V-UP) backend, high-speed serial optical links
- **Precision clock** for timing detectors
- **Unified readout** board combined with trigger and timing distribution and (fast) control
- **100GbE D2S** links running TCP/IP
- 100 ➔ **400 GbE** Data Concentrator
- “Folded” event builder
- **RoCE-based** 400 GbE EvB switch fabric

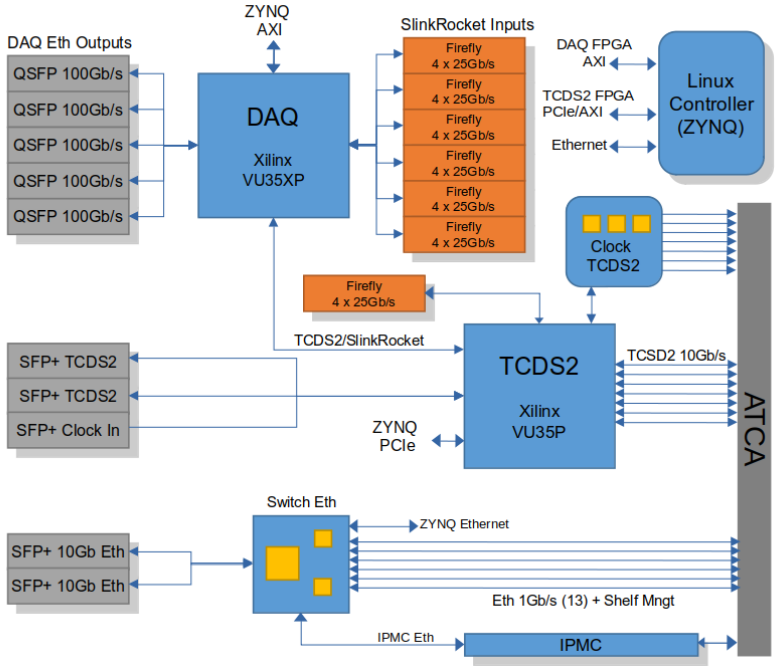
- **Uniform** event backbone
- **Heterogeneous** HLT
- Cluster filesystem storage

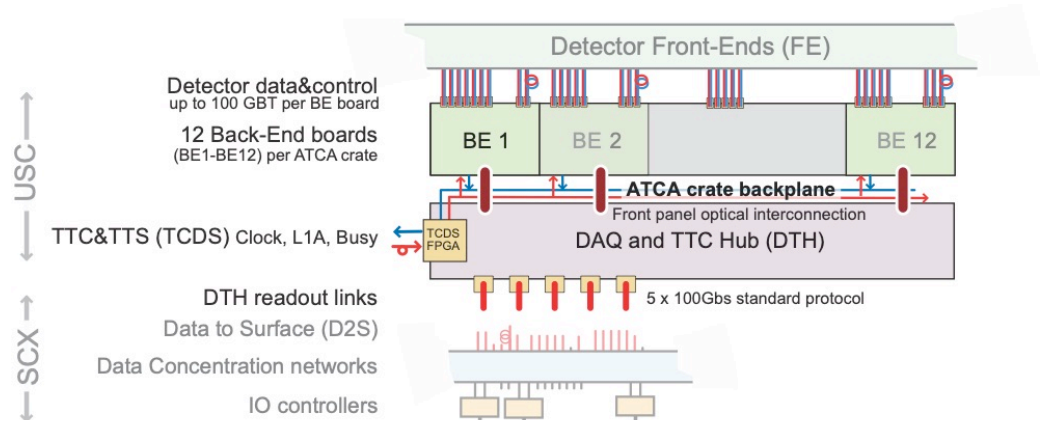
Readout and D2S

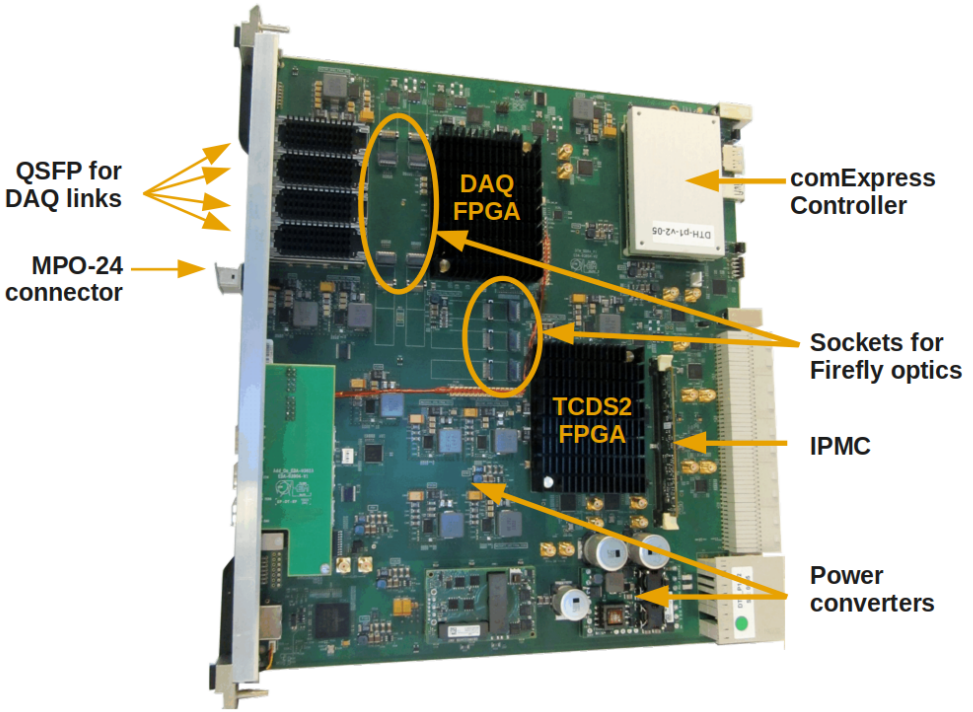
Subdetector	Back-end platform	Back-end FPGA	BE boards	BE crates	Front-end lpGBT links[a]	Subevt size (MB) $\langle PU \rangle = 140$	Subevt size (MB) PU=200	Thru (Tb/s) PU=140 500 kHz	Thru (Tb/s) $\langle PU \rangle = 200$ 750 kHz	Av thru per BE board (Gb/s)	Av thru per BE crate (Gb/s)
Outer Tracker - PS.	Serenity	1xVU13P	108	9	6500	0.50	0.72	2.01	4.31	40	479
Outer Tracker - 2S.	Serenity	1xVU13P	108	9	6500	0.30	0.43	1.21	2.58	24	287
OT Track Finder TPG	Apollo	2xVU13P	180	18		0.01	0.01	0.04	0.06	3	1
Inner Tracker	Apollo	2xVU13P	28	4	1260	1.01	1.44	4.03	8.64	309	2160
MIP Timing Det. - BTL	Serenity	2xKU15P	8	2	864	0.17	0.24	0.67	1.44	180	720
MIP Timing Det. - ETL	Serenity	2xKU15P	14	2	1600	0.31	0.44	1.23	2.64	189	1320
ECAL Barrel	BCP	1xVU13P	108	12	10000	2.11	2.11	8.44	12.67	117	1056
HCAL Barrel	BCP	1xVU13P	18	2	other	0.24	0.24	0.96	1.44	80	720
HCAL HO	BCP	1xVU13P	9	1	legacy	0.03	0.03	0.12	0.18	20	180
HCAL HF	BCP	1xVU13P	9	1	other	0.06	0.06	0.24	0.36	40	360
CALO Endcap	Serenity	2xKU15P	84	12	8000	2.10	3.00	8.40	18.00	214	1500
CALO Endcap TPG S1	Serenity	2xVU7P	72	8	9000	0.11	0.15	0.42	0.90	13	113
CALO Endcap TPG S2	Serenity	2xVU9P	54	6		0.04	0.05	0.14	0.30	6	50
Muon DT	BMT	1xVU13P	84	8	2400	0.11	0.15	0.42	0.90	11	113
Muon CSC	APX	1xVU13P	10	2	other	0.33	0.47	1.32	2.82	282	1410
Muon GEM - GE1/1	APX	1xVU13P	8	1	GBTX	0.002	0.003	0.01	0.02	2	18
Muon GEM - GE2/1	APX	1xVU13P	8	1	GBTX	0.001	0.002	0.01	0.01	1	9
Muon GEM - ME0	APX	1xVU13P	18	2	1728	0.08	0.12	0.34	0.72	40	360
Muon RPC	Serenity	UltraScale+	18	3	other	0.01	0.01	0.03	0.07	4	22
Level-1 Trigger	various	UltraScale+	280	28		0.26	0.26	0.78	1.68	6	60
Total.			>1226	>130	50k	7.77	9.94	31.1	60.3		

DAQ and TCDS Hub

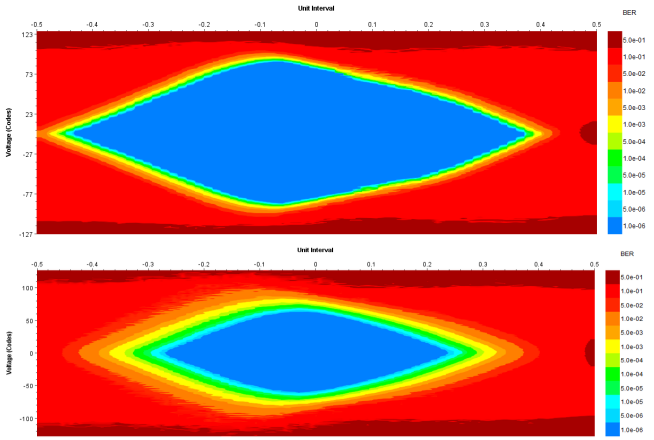
FPGA with HBM buffer
Fly-over optical engines
100G TCP/IP







DTH tests and development



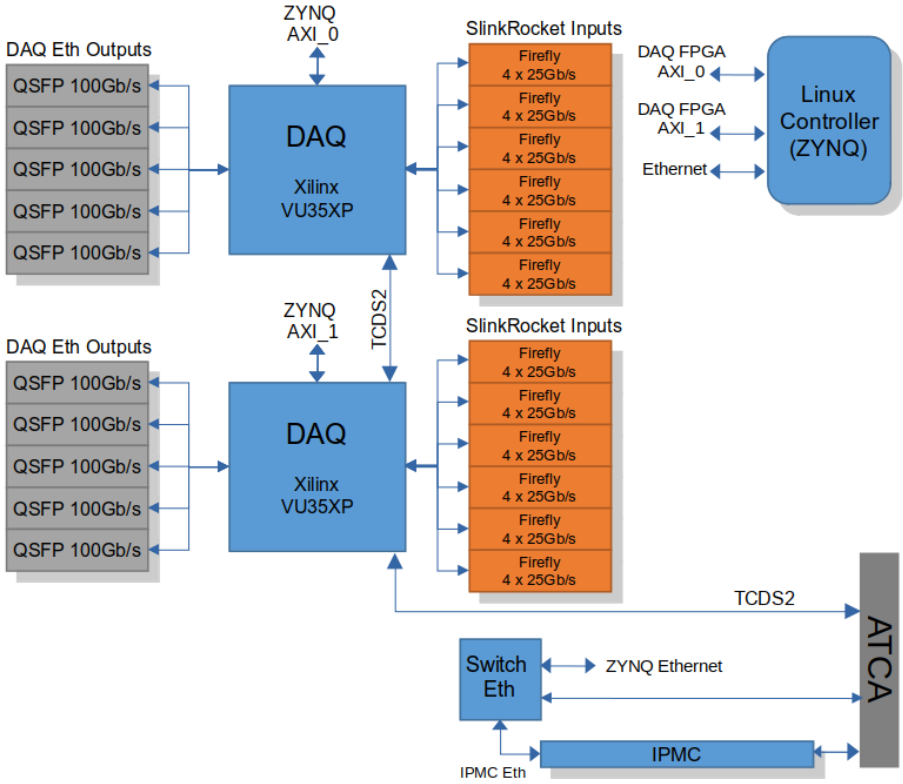
FireFly

QSFP

Prototype	DAQ FPGA	TCDS FPGA	Buffer	DAQ b/w	Eth switch	On-board Controller	Year
DTH-P1v1	KU15P	KU15P	HMC not used	tested 100 Gb/s	No	ComExpress	2019
DTH-P1v2	KU15P	KU15P	DDR not used	tested 400 Gb/s	No	ComExpress	2020
Switch	N/A	N/A	N/A	N/A	MicroSemi VSC7444	N/A	2020
DTH-400-P2	VU35P	VU35P	HBM	400 Gb/s	MicroSemi VSC7444	Trenz (Zynq)	(2022)
DAQ-800	2xVU35P	N/A	HBM	800 Gb/s	N/A	Trenz (Zynq)	(2023)

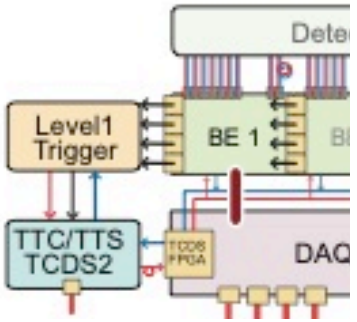
DAQ800

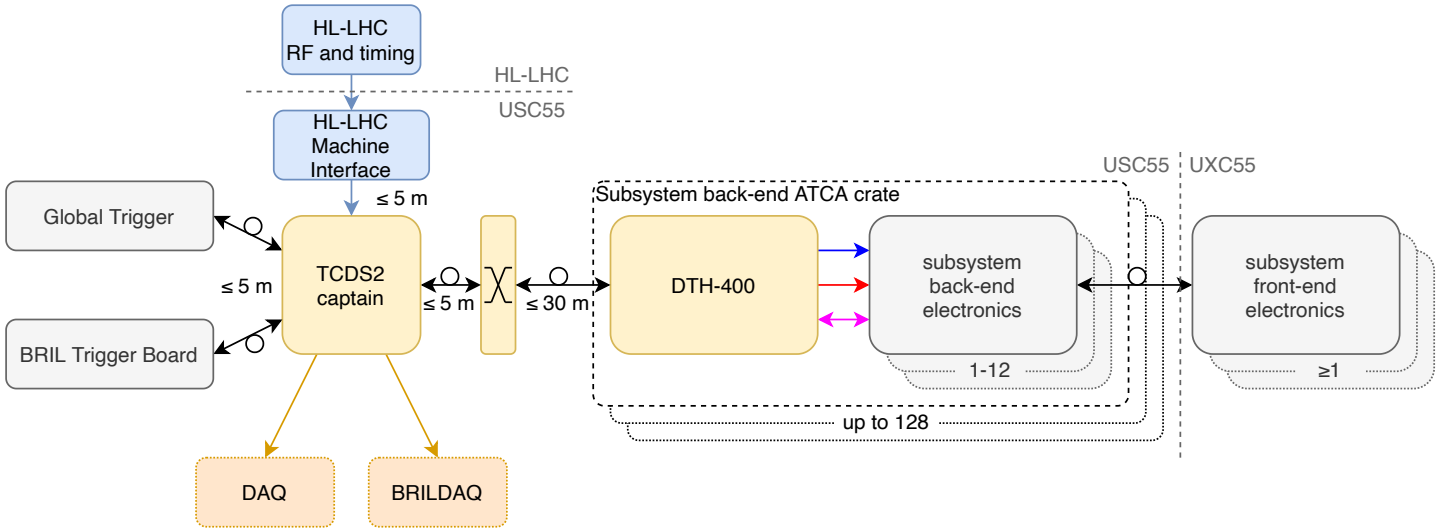
A second type of readout board is needed for systems with large throughput per crate



Trigger and Timing distribution

Trigger and Timing Control and Distribution



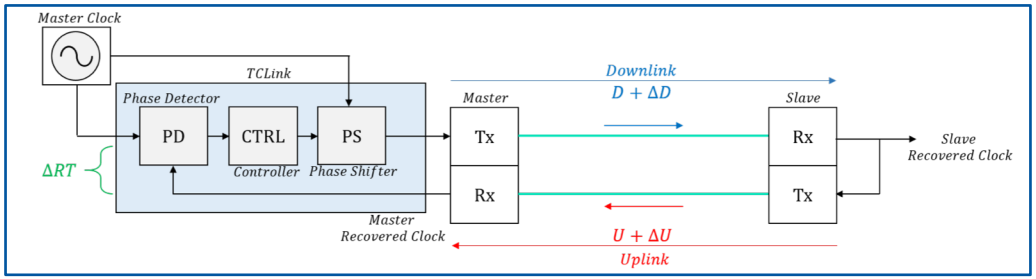


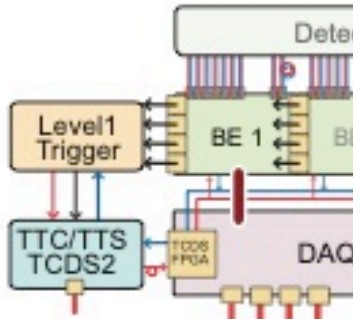
- Trigger/Timing data streams over high speed serial optical links (10 Gb/s)
- High-quality, phase-stabilised clock distribution
  - Random jitter well within timing detectors requirements
- Synchronisation of CMS to HL-LHC orbit
- Distribution of Level-1 physics triggers (with type) and additional L1As for calibration, testing etc.
- Generation and distribution of timing and synchronisation commands with payload
- Collection of back-end and DAQ buffer status and synchronous trigger throttling
- Improved partitioning capability for commissioning, calibration, etc.

TCDS2 tests

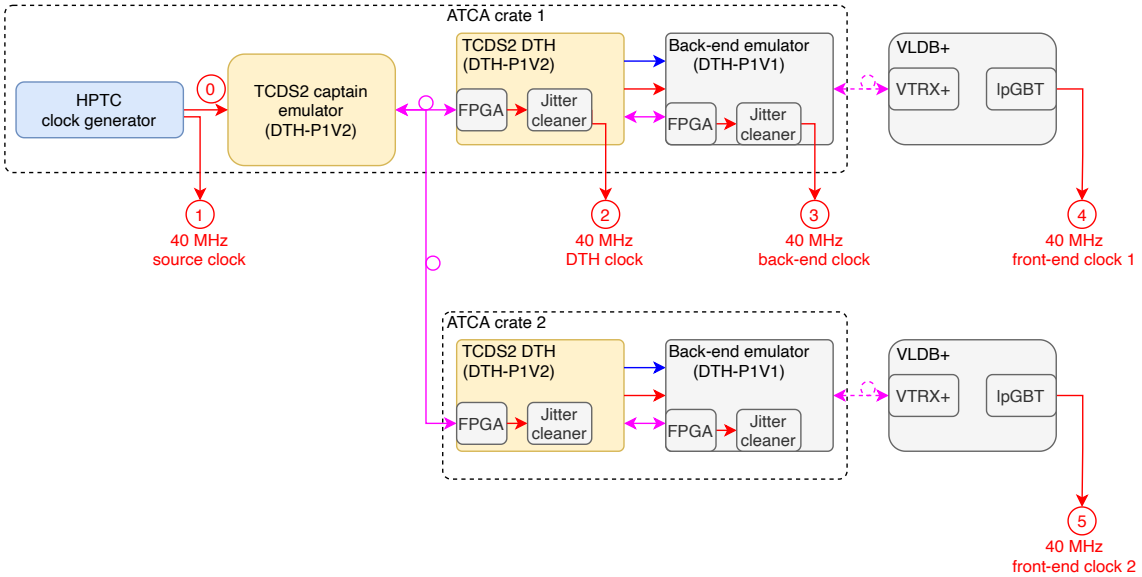
Time-compensated link

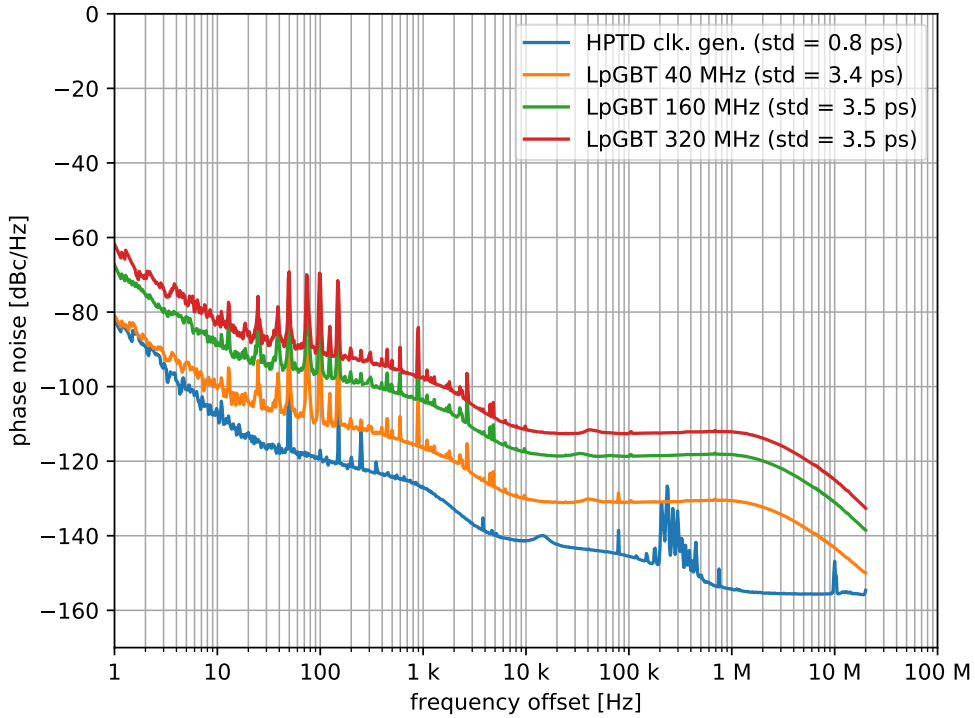
https://indico.cern.ch/event/799025/contributions/3486290/





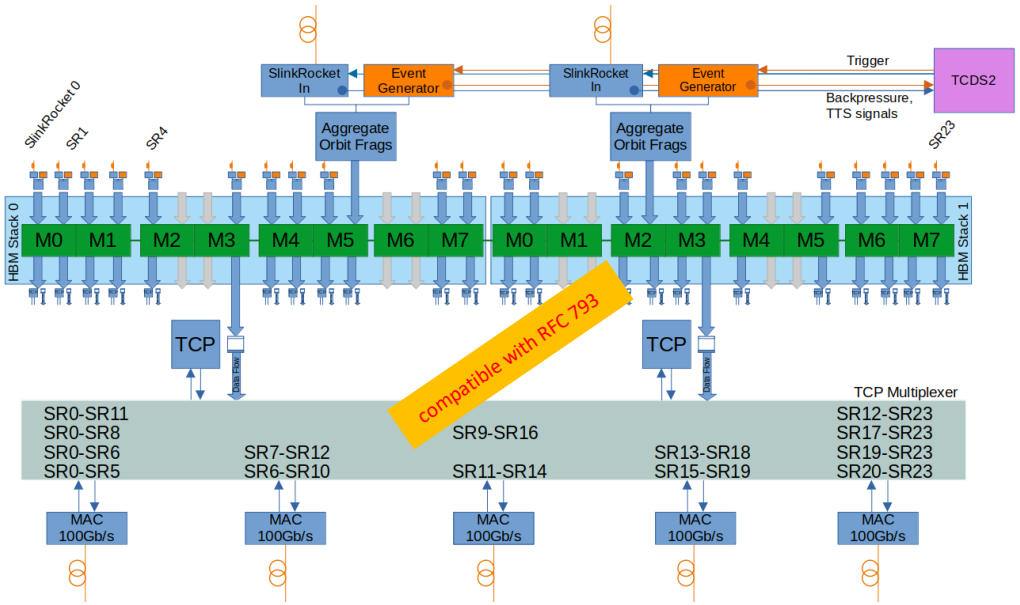
Clock quality tests



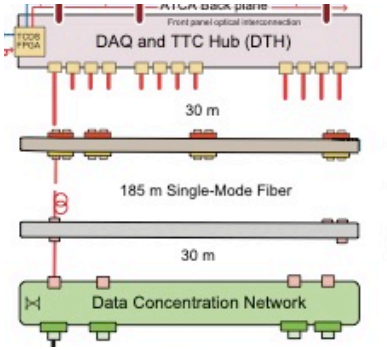


Data Concentration and Event Building

D2S and data concentration



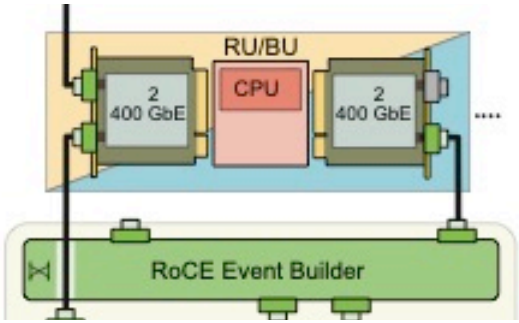
- DTH aggregates fragments from one data source and one LHC orbit (88.9 μs)
- Each source is routed to one TCP stream
- CWDM4 is used to cover the ~200 m to the surface
- Multiple D2S links are aggregated in the data concentration network into 400 Gb/s ports



Operate interconnects at a working point where the line speed is optimally used

Subdetector	BE crates	boards per crate	FF speed (Gb/s)	FF per board	FF per subdet	DTH-400 per crate	DAQ-800 per crate	DTH-400 per subdet	DAQ-800 per subdet	D2S channels per crate	D2S channels per subdet	Avg load D2S channel
Outer Tracker - PS	9	12	25	4	432	1	1	9	9	9	81	0.53
Outer Tracker - 2S	9	12	25	4	432	1	1	9	9	6	54	0.48
Track trigger	18	10	25	1	180	1	-	18	-	1	18	0.03
Inner Tracker	4	7	25	16	448	1	3	4	12	28	112	0.77
MIP Timing Det. - BTL	2	4	25	16	128	1	2	2	4	12	24	0.60
MIP Timing Det. - ETL	2	7	25	16	224	1	3	2	6	28	56	0.83
ECAL Barrel	12	9	25	8	864	1	2	12	24	14	168	0.75
HCAL Barrel	2	9	25	4	72	1	1	2	2	9	18	0.80
HCAL HO	1	9	25	2	18	1	-	1	-	2	2	0.90
HCAL HF	1	9	25	2	18	1	-	1	-	4	4	0.90
CALO Endcap	12	7	16	24	2016	1	3	12	36	21	252	0.71
CALO Endcap TPG S1	8	9	16	2	144	1	-	8	0	2	16	0.56
CALO Endcap TPG S2	6	9	16	2	108	1	-	6	0	2	12	0.50
Muon DT	8	10,12	25	1	84	1	-	8	-	2	16	0.56
Muon CSC	2	5	25	16	160	1	2	2	4	20	40	0.71
Muon GEM - GE1/1	1	8	25	1	8	1	-	1	-	1	1	0.18
Muon GEM - GE2/1	1	8	25	1	8	1	-	1	-	1	1	0.09
Muon GEM - ME0	2	9	25	4	72	1	1	2	2	6	12	0.60
Muon RPC	3	6	25	1	18	1		3		1	3	0.22
Level-1	28	10	25	1	560	1	-	28	-	2	56	0.30
TCDS2	1		25	1	1	1	-	1	-	1	1	0.10
Systems w/ TCDS2 only	4		-	-	-	1		4				
Contingency								6	2			
Total	136				5994			142	110		947	

Event Builder

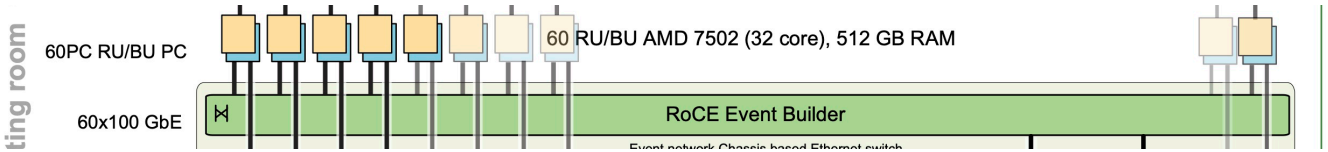


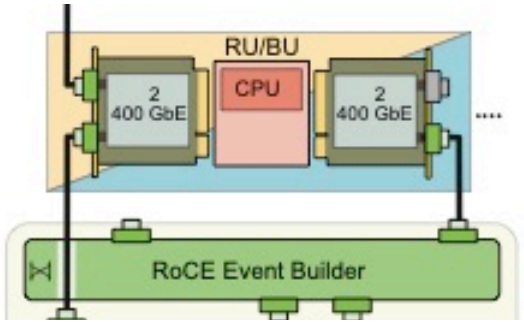
- **RDMA** enables full exploitation of the physical link available bandwidth
  - Run-2 event builder was based on **InfiniBand FDR**
  - **low latency** and **reduced CPU load** at the endpoints
- **RoCE** (RDMA over converged Ethernet)
  - Ethernet switch fabric less expensive and easier to manage
- **Run-3** CMS event builder:
  - single-socket AMD EPYC 7502P processor
  - Mellanox ConnectX-6 **100 GbE** NICs
  - Juniper Networks QFX100016 deep buffered switch

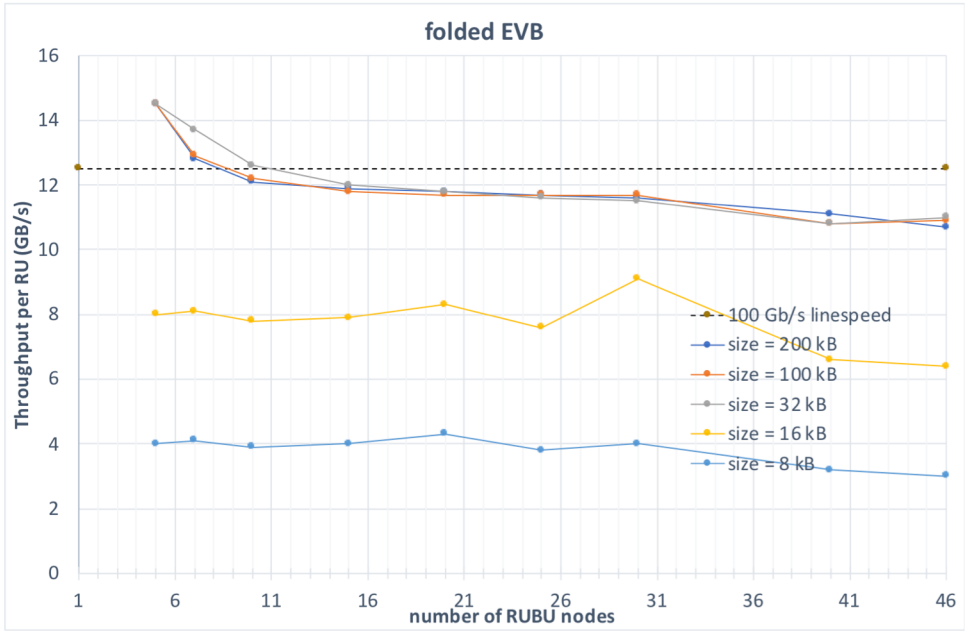
Phase-2
- Ethernet 200 and 400 Gb/s already available on **backbone switches**
- RoCE-based 400 Gb/s will require **corresponding NICs** and servers with **PCIe Gen5** and sufficient memory b/w (**800 Gb/s concurrent I/O**)

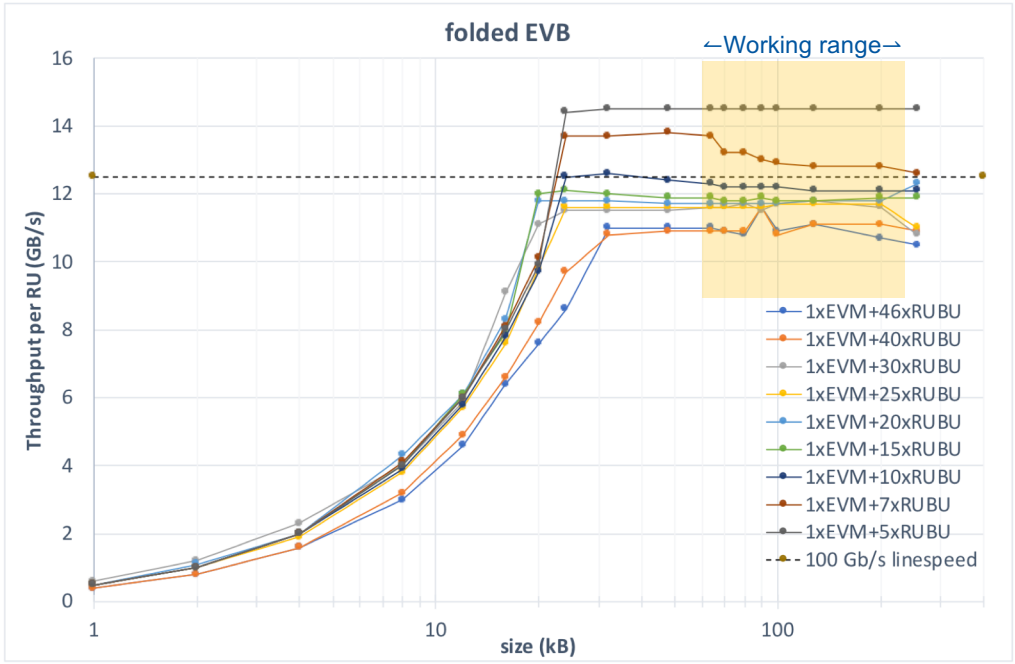
Event Builder

Scaling tests using Run-3 system







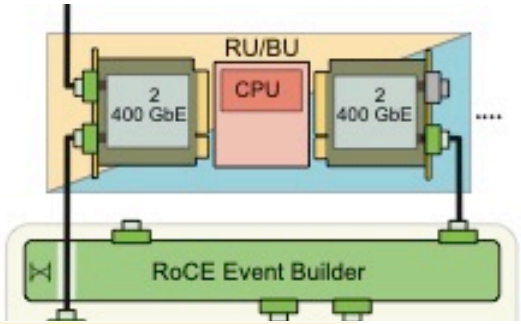


Throughput scales with number of nodes

Close to line speed within working range

Same effective link utilization for 400 GbE needs to be demonstrated

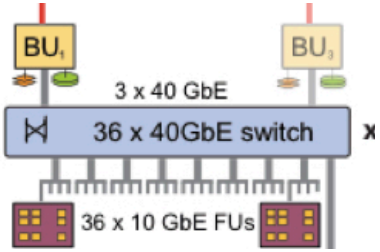
Event Builder



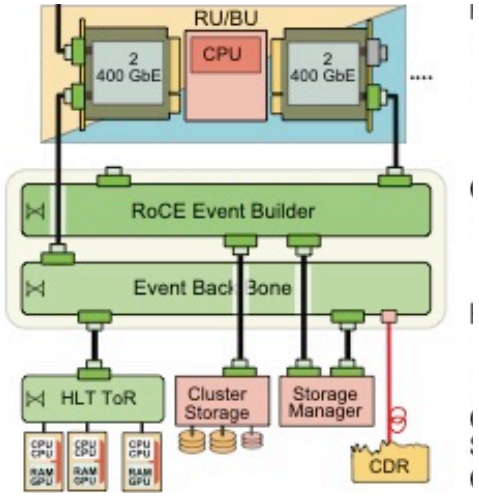
	Run-2	Run-3	Phase-2
L1 accept rate (maximum)	100 kHz	100 kHz	750 kHz
Event Size	2.0 MB [a]	2.0 MB [a]	9.9 MB
Event Network throughput	1.6 Tb/s	1.6 Tb/s	60 Tb/s
D2S modules	FEROL	FEROL	DTH-400/DAQ-800
D2S module Ethernet ports	10 GbE	10 GbE	100 GbE
number of D2S source ports	650	650	950
D2S network, EVB ports	40 GbE	100 GbE	400 GbE
EVB network	Infiniband FDR	100 GbE RoCE	400 GbE RoCE
architecture configuration	non-folded	folded	folded
number of EVB nodes	120	48	200
aggregation	super-fragment	super-fragment	orbit

Online Selection

HLT I/O and orchestration



- Data distribution, HLT execution steering and monitoring **entirely data-driven and file-based**
  - Requires **minimal additions** (I/O modules) w.r.t. offline framework
- **nfs over TCP/IP** is used to access data
- monitoring using **elasticsearch** as back-end
- Run-2/3 system uses **RAMdisk as buffer** for HLT



Input:

- Event (or orbit) file consists of adjacent segments mapping individual data source (front-end data, FED)
- Buffer requirements for Phase-2 scale with rate and event size: **~250 and ~450 TB** respectively for PU140 and 200
- Cost/performance of **other solutions** than RAM currently unsuitable (e.g. SSD endurance) but easily adaptable if progress happens

Output:

- Serialized root objects – self-contained events
- **Easily concatenable** for aggregation / storage
  - Concatenation in subsequent steps

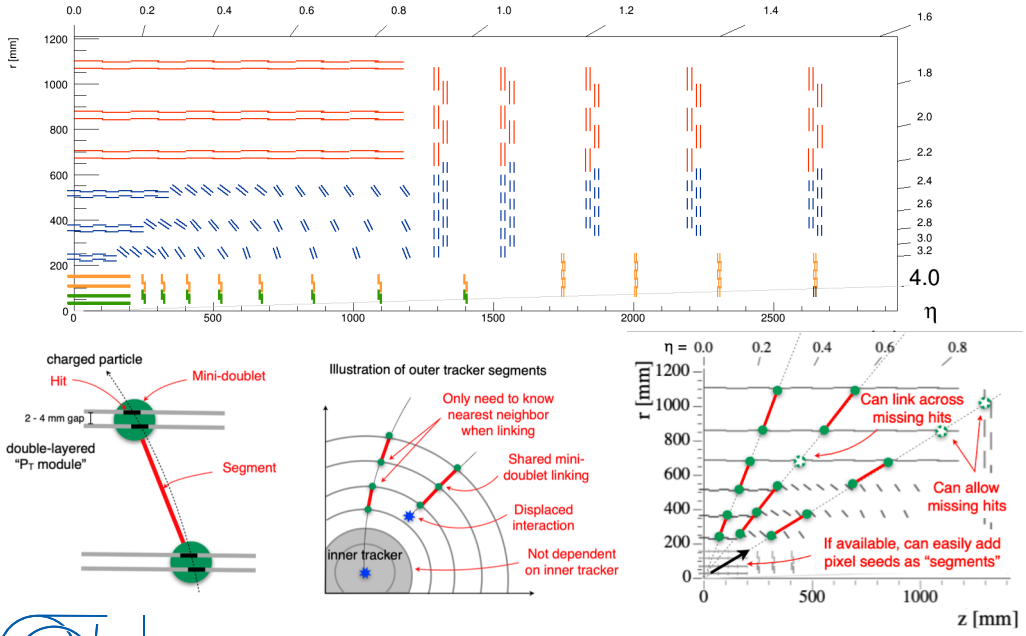
Phase-2:

- NFS throughput on Run-3 (100Gb/s) system **over 90%** of line-speed
- Scaling to be verified on the next-generation high-speed network fabric
- Alternatives could use RoCE or revert to non-file-based access if necessary

High Level Trigger

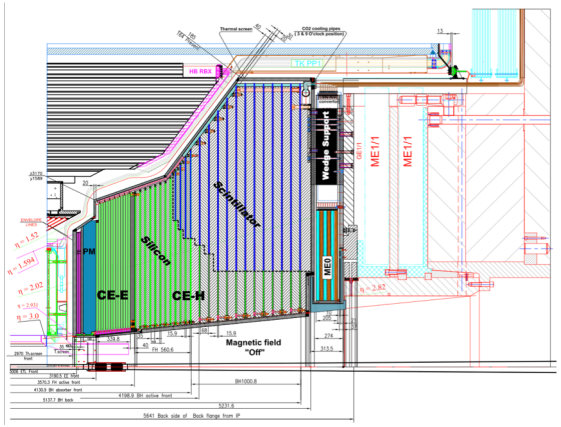
Key design choice from inception: **use same framework as offline** reconstruction

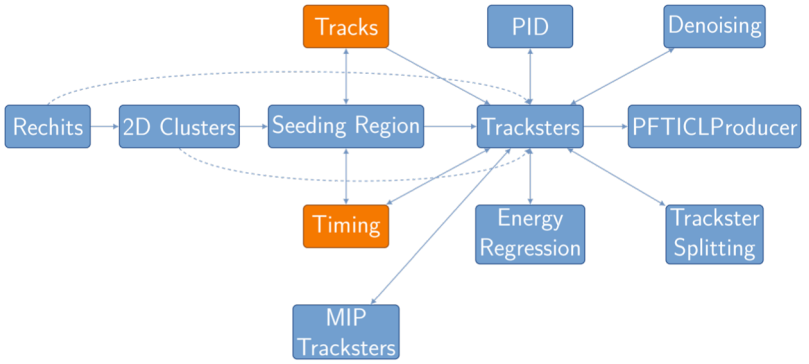
- Flexibility
- Code reuse
- Swift migration of offline-developed algos



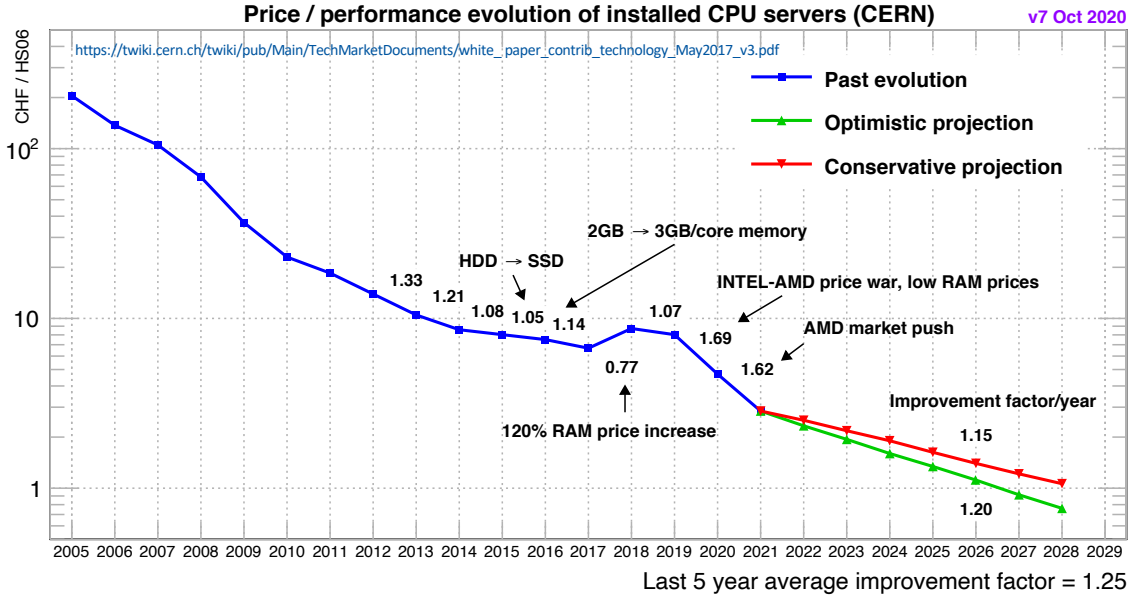
Optimization for selection vs. reconstruction Efficiency vs. rejection target (factor 100 vs. L1)
Optimal use of **limited** computing resources and **Heterogeneous platforms**

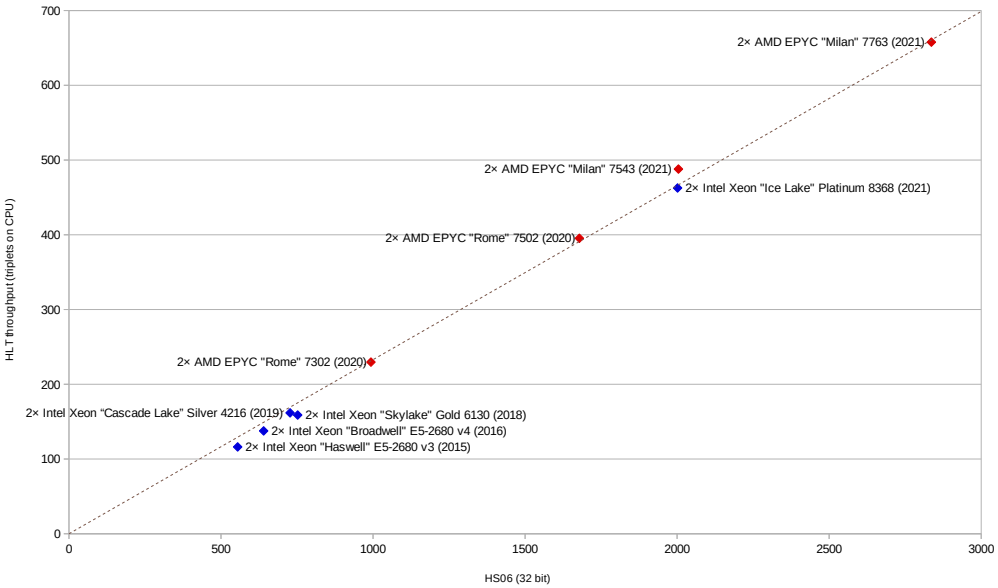
- **Parallelisation**
- **Data structures**





Phase-2 HLT Computing





HS06 benchmark maps well the CMS HLT workflow

A mildly optimistic **price/performance improvement for CPU is 20 %** per year

Complexity leap of HL-LHC detector exceeds performance increase per unit cost

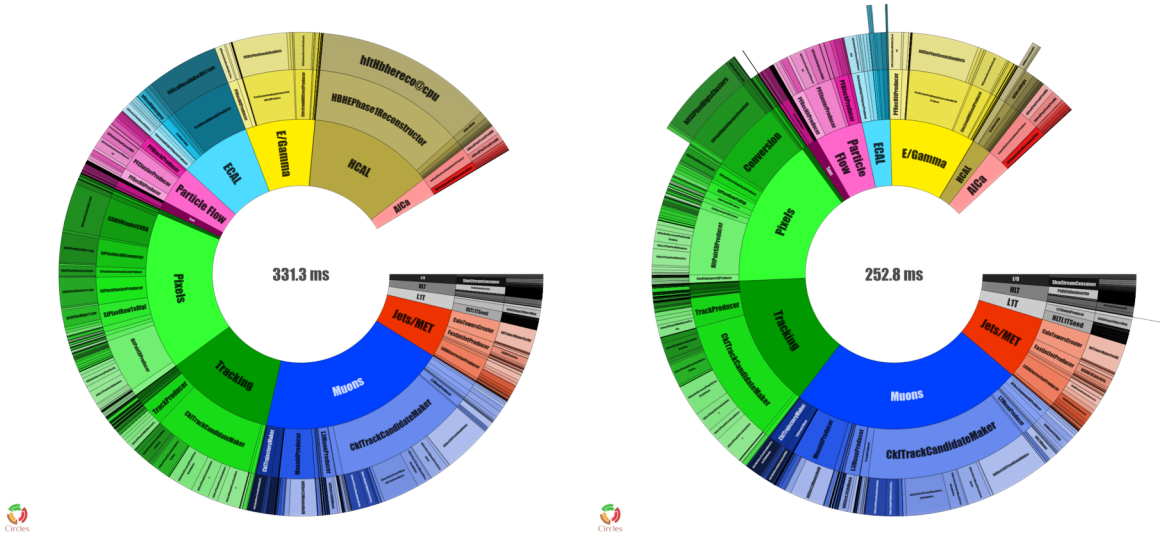
Missing factor 10 ➔ new strategy required

GPUs in HLT: Run-3

Require framework extension: *external worker*, offloading asynchronous work outside framework scheduler:

- *acquire()* reads input data
  - launch asynchronous execution
  - scheduler **can run other tasks in the thread** that called *acquire()*
- completion of asynchronous work notified to framework and result put back into event (*produce()*)

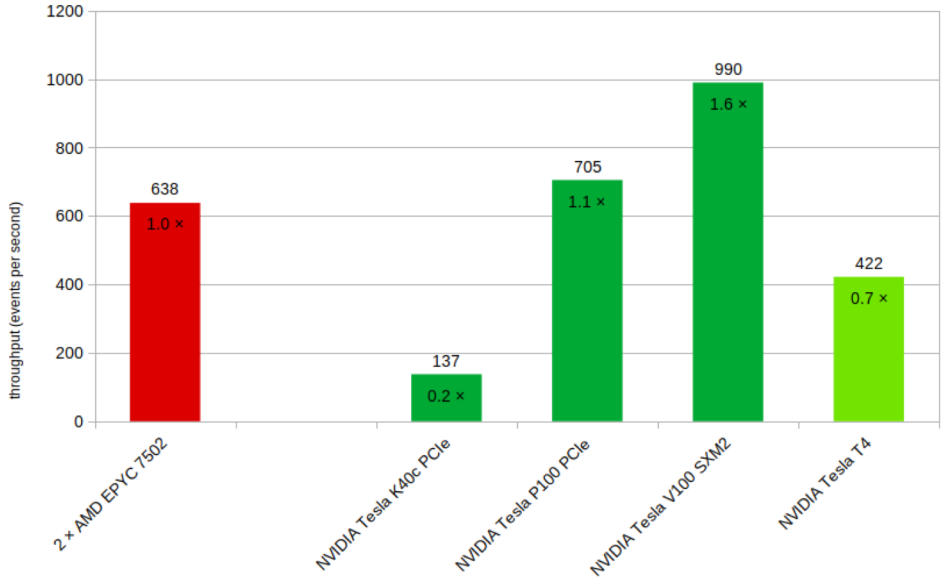
alternative implementations of algorithm (e.g. CPU and GPU) can be chosen at runtime based on environment



About 30 % of Run-3 workflow can be offloaded today to GPU

Deploying heterogeneous HLT

Equivalent throughput capacity of GPU w.r.t. reference CPU



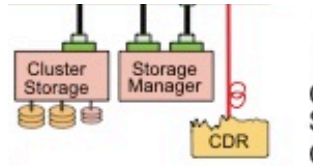
"Work in Progress"

Effective HS06 cost when including offload

Scenario	Year		CPU CHF/HS06	CPU fraction	GPU CHF/HS06	GPU fraction	effective CHF/HS06
HLT TDR	2020		4.2		1.5		
Run-4	2028	20%/y	1.0	50%	0.34	50%	0.66
		15%/y	1.4	50%	0.48	50%	0.93
Run-5	2032	20%/y	0.5	20%	0.16	80%	0.23
		15%/y	0.8	20%	0.27	80%	0.38

Storage and transfer

Storage and Transfer



Storage:

- **Storage appliance** for robustness / availability / bandwidth
- **Parallel filesystem (Lustre)** to aggregate data at scale (hundreds of data sources / destinations)
- Run-2 ➔ Run-3 system update
  - InfiniBand FDR ➔ 100 GbE RoCE
  - 12 GB/s (R+W) ➔ **36 GB/s** (this beautiful (?) piece of equipment)
  - 0.75 PB ➔ 1.3 PB
  - Total capacity (one day of data taking) easily met due to b/w reqs
  - **HDD/SSD admixture** (SSD for metadata)
- Phase-2 minimal requirements not so distant
  - New storage technologies with better endurance and price/perf ?
  - 400 GbE (RoCE or…)

Transfer:

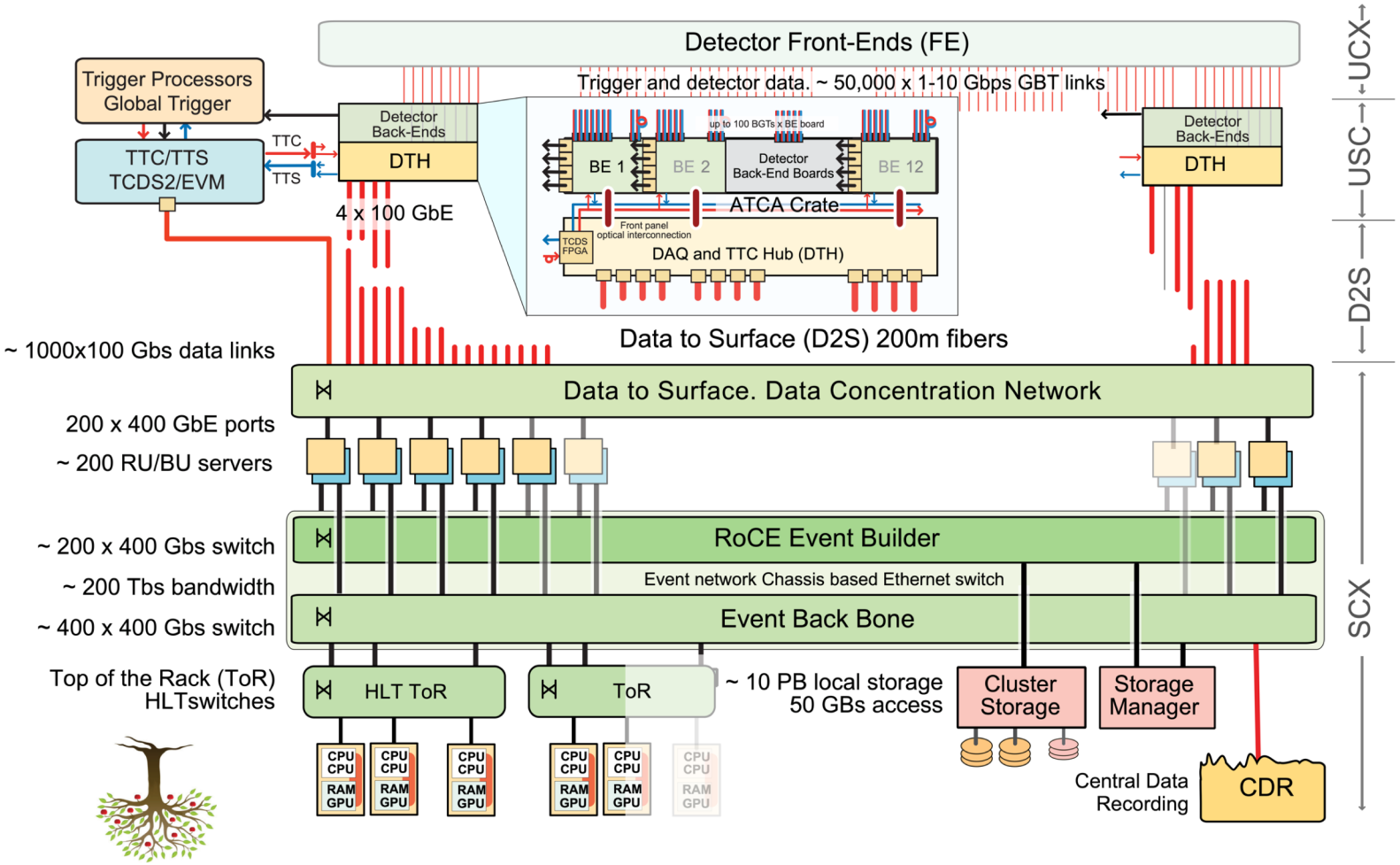
- Long-distance links Run-2 ➔ Run-3: 4 x 40 ➔ 4 x 100 Gb/s (redundant, also used for “HLT cloud”)
- Phase-2 requirements are met with 4 x 400 Gb/s link

Summary

Phase-2 DAQ Parameters

CMS detector	LHC Phase-1	HL-LHC Phase-2	
Peak ⟨PU⟩	60	140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size at HLT input	2.0 MB [a]	7.8 MB	9.9 MB
Event Network throughput	1.6 Tb/s	31 Tb/s	60 Tb/s
Event Network buffer (60 s)	12 TB	234 TB	445 TB
HLT accept rate	1 kHz	5 kHz	7.5 kHz
HLT computing power [b]	0.7 MHS06	17 MHS06	37 MHS06
Event Size at HLT output [c]	1.4 MB	5.5 MB	6.9 MB
Storage throughput [d]	2 GB/s	31 GB/s	61 GB/s
Storage throughput (Heavy-Ion)	12 GB/s	61 GB/s	61 GB/s
Storage capacity needed (1 day [e])	0.2 PB	2.0 PB	3.9 PB

Phase-2 DAQ baseline



Some of the points not touched

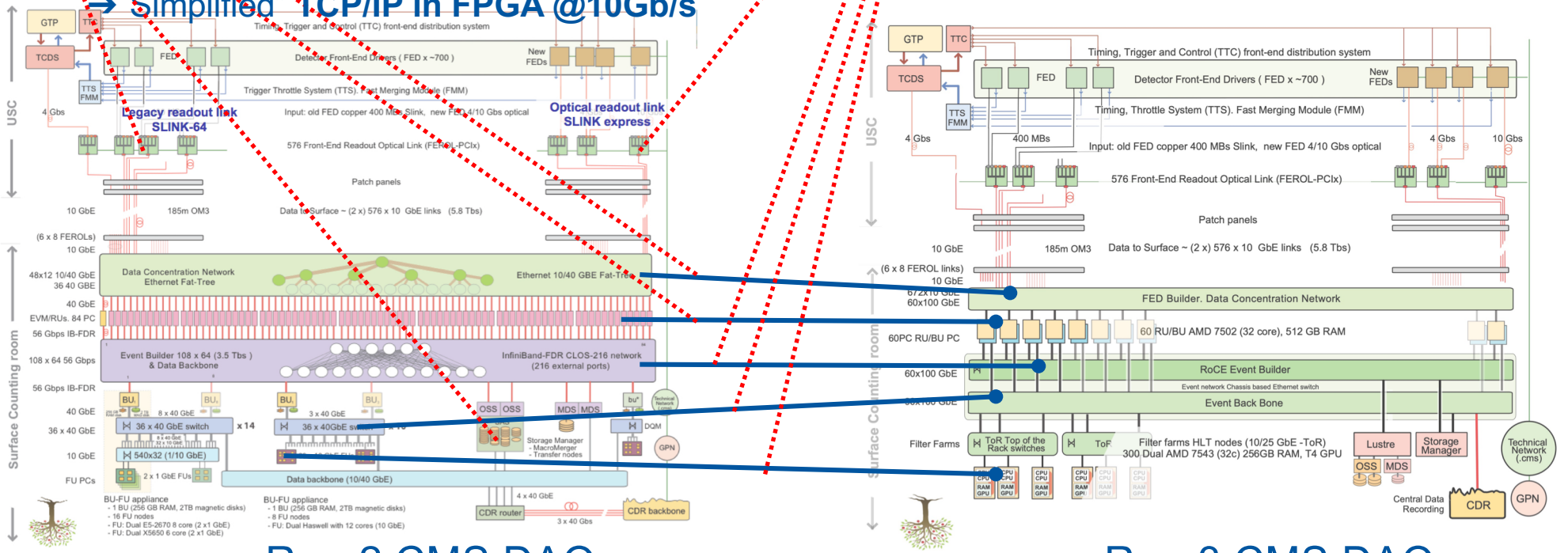
- Run Control and DAQ configuration management
- HLT sw and menu configuration
- Online software: DAQ as a distributed system
- Monitoring, databases, GUIs
- Detector control system
- Technical and IT infrastructure (online cluster, network management, sw deployment, configuration management)

backup

DAQ technology evolution: Run 2 and 3

- 10/40 GbE **“fat-tree”** DCN ➔ 10/100 GE
- monolithic
- ~100x100 **“RU-BU” evb** ➔ ~50 node **“folded”**
- **Cluster** (parallel) **filesystem** based on **Lustre**
- ➔ “Simplified” **TCP/IP in FPGA @10Gb/s**

- ➔ **Optical** readout links from MicroTCA back-ends
- **InfiniBand** CLOS FDR ➔ Monolithic **RoCE** 100GbE
- “appliance” **10GbE** local switch ➔ **100 GbE** backbone
- ➔ HLT with **GPU** offload



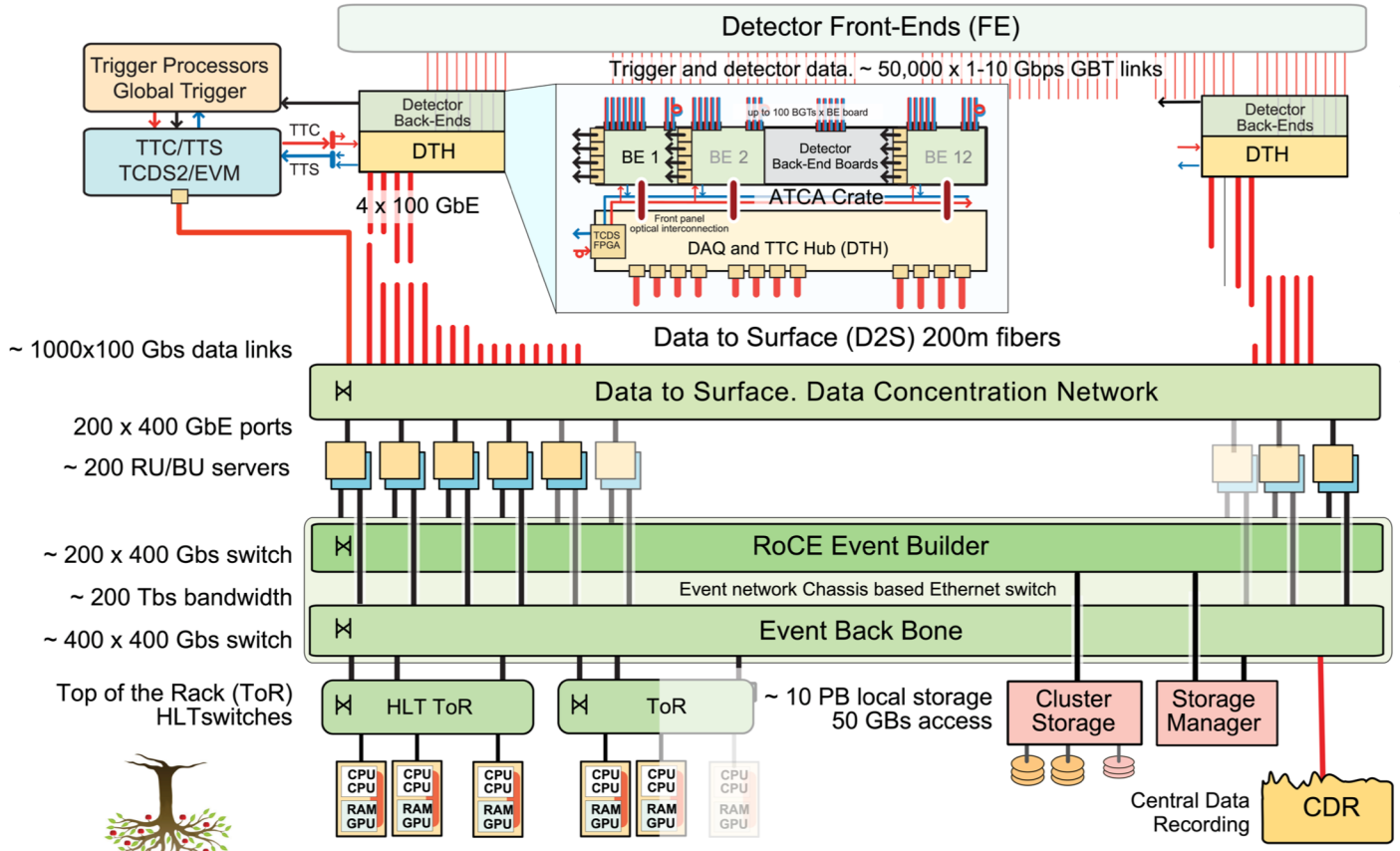
Run-2 CMS DAQ

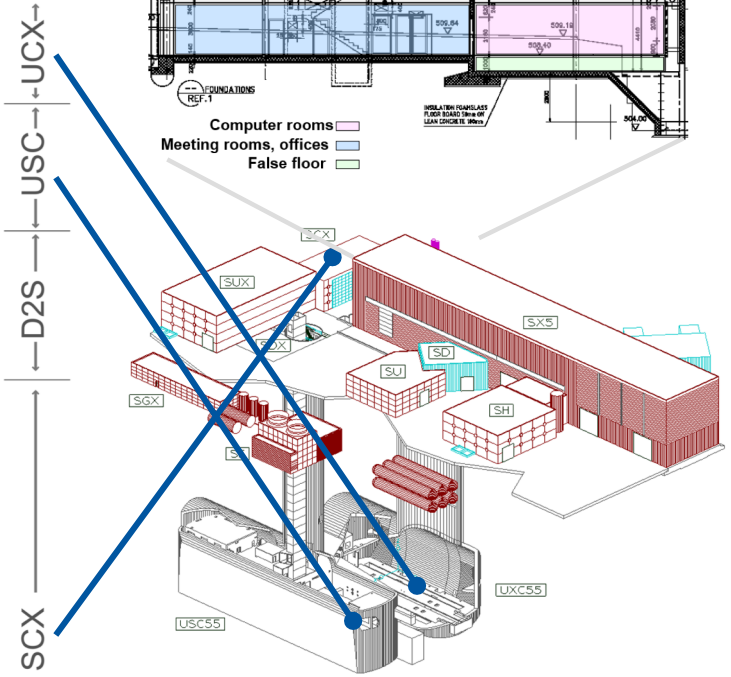
Run-3 CMS DAQ

DAQ and HLT: summary of components

Component	Technology	Estimated quantity
DTH-400 and DAQ-800 boards [a]	ATCA custom board	250 boards
TCDS2 custom boards	ATCA custom board	16 boards
DAQ D2S links	100-GBASE-CWDM4 [b]	1000 links
Data Concentrator Network	Chassis-based [c] switch	1200 ports
Event Builder Nodes [d]	Rack-mount 2U server	200 servers
Event Builder Network	Chassis-based 400 Gb/s switch	200 ports
Event Backbone Network	Chassis-based 400 Gb/s switch	200 ports
ToR switch	Rack-mount [e] switch	42 ToR switch (approx. 5×50 ports)
HLT servers [f]	Rack-mount 1U(2U) server with 2(6) GPU	1600(840) servers
Storage System	Network-attached storage appliance	61 GB/s bandwidth 3.9 PB total storage

Infrastructure and Data Center





Functions	Number of racks		Total Power (kW)	
	2028	**2032**	**2028**	**2032**
Event builder	11		160	
STS	3		30	
DQM	2		40	
HLT	40	42	1830	2120
Core services	15		150	
Validation systems	13		100	
Miscellaneous	10		60	
Starpoints	8		40	
Cloud	10	10	200	400
Total	**112**	**114**	**2610**	**3100**