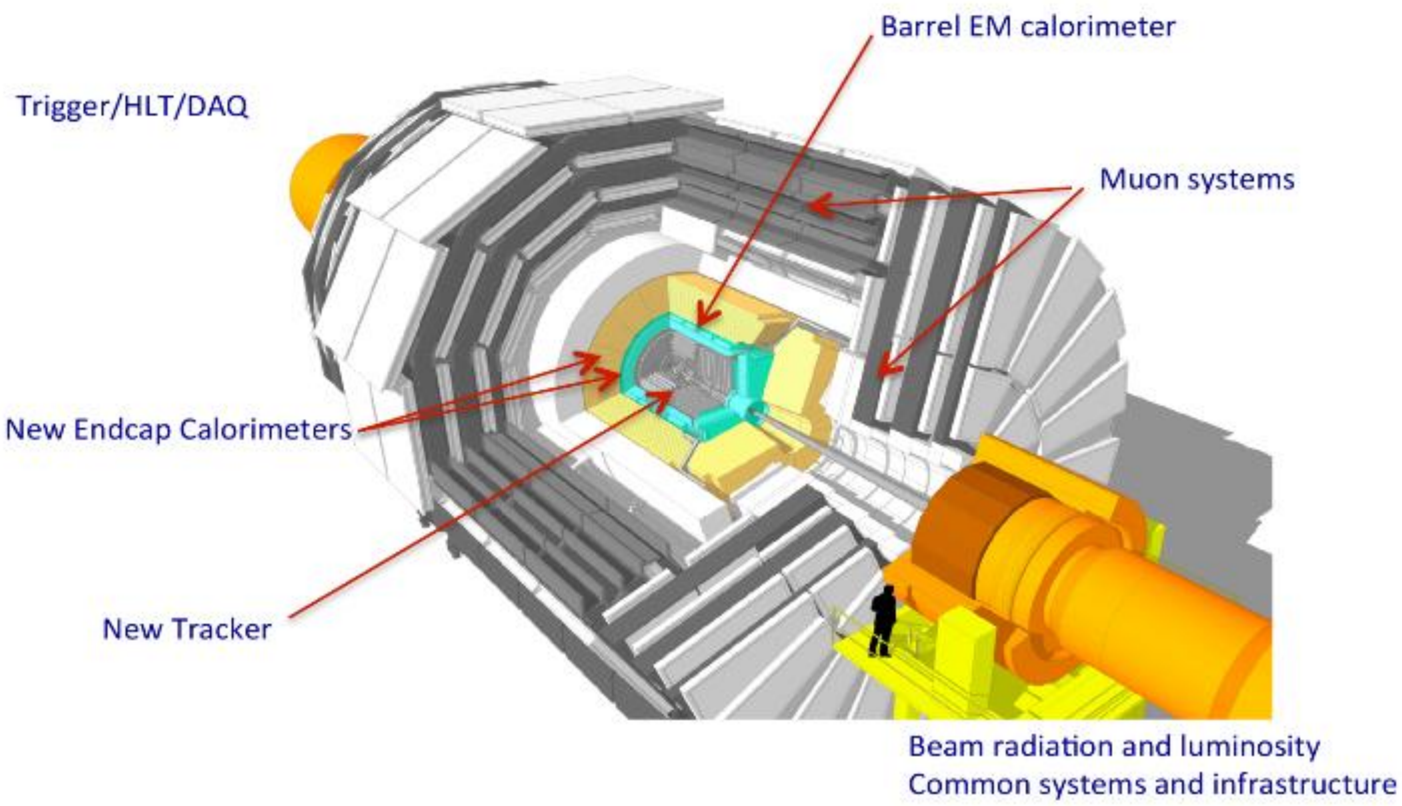
Orbit Builder for CMS Phase-2 at CERN

Presenter: Rafał Dominik Krawczyk, on behalf of the CERN CMS DAQ group
Primary authors: Rafał Dominik Krawczyk-Rice University, Andrea Petrucci-UCSD
Email: rafal.dominik.krawczyk@cern.ch

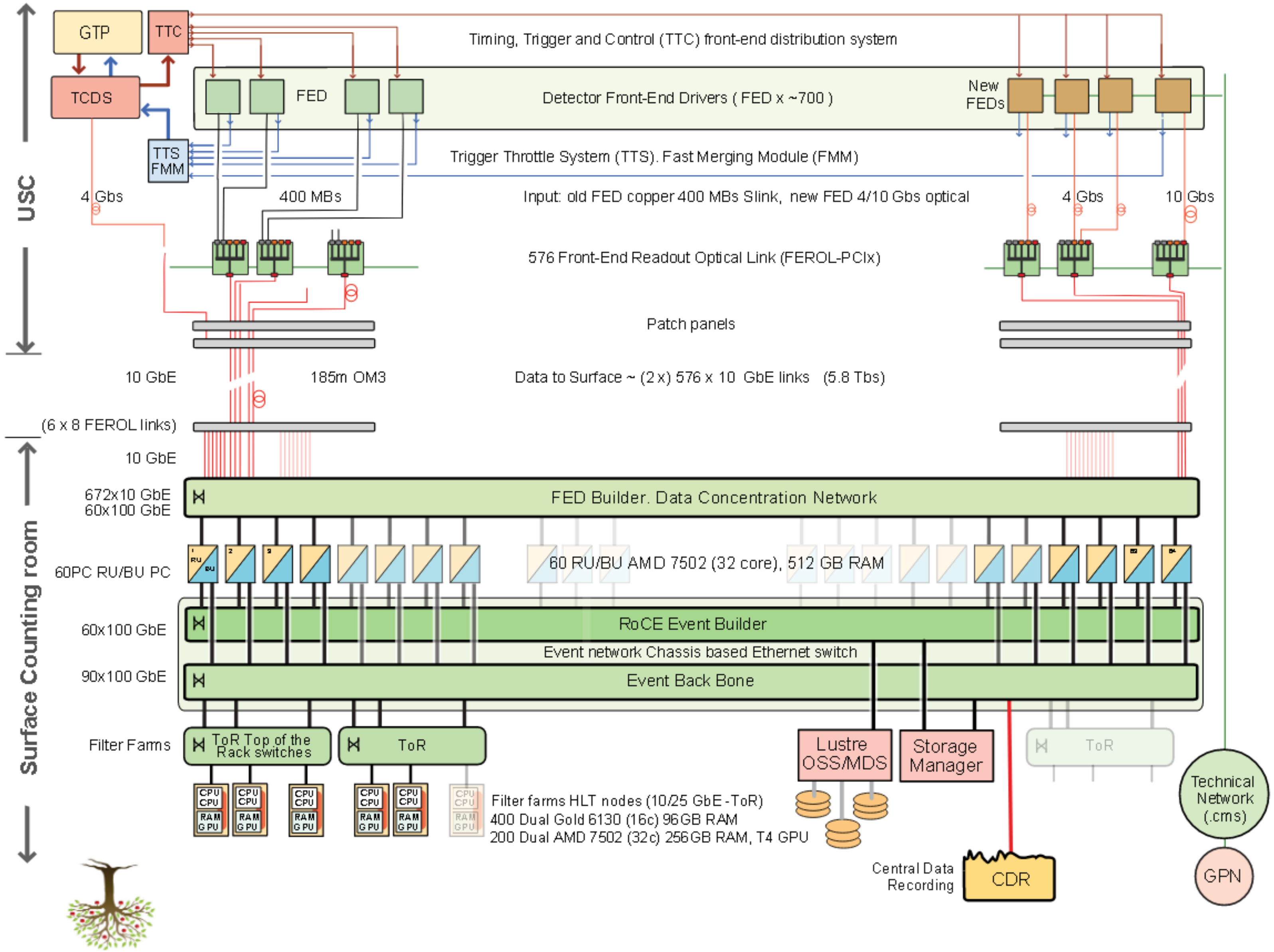
26th International Computing in High Energy & Nuclear Physics Conference
9 May 2023

CMS in Phase-2

- One of four main LHC experiments
- Upgrade for HL-LHC luminosity increase
  - Run 3 (now) 2×10^{34} cm^{-2} s^{-1}
  - Run 4 (2029) 5×10^{34} cm^{-2} s^{-1}
  - Run 5 (2035) 7.5×10^{34} cm^{-2} s^{-1}
- Phase-2 DAQ:
  - Event size → from 2 MB to 8.4 MB
  - L1 Trigger acceptance rate → from 100 kHz to 750 kHz
  - HLT accept rate → 1 kHz to 7.5 kHz
  - Ready in 2025 for Run 4 commissioning

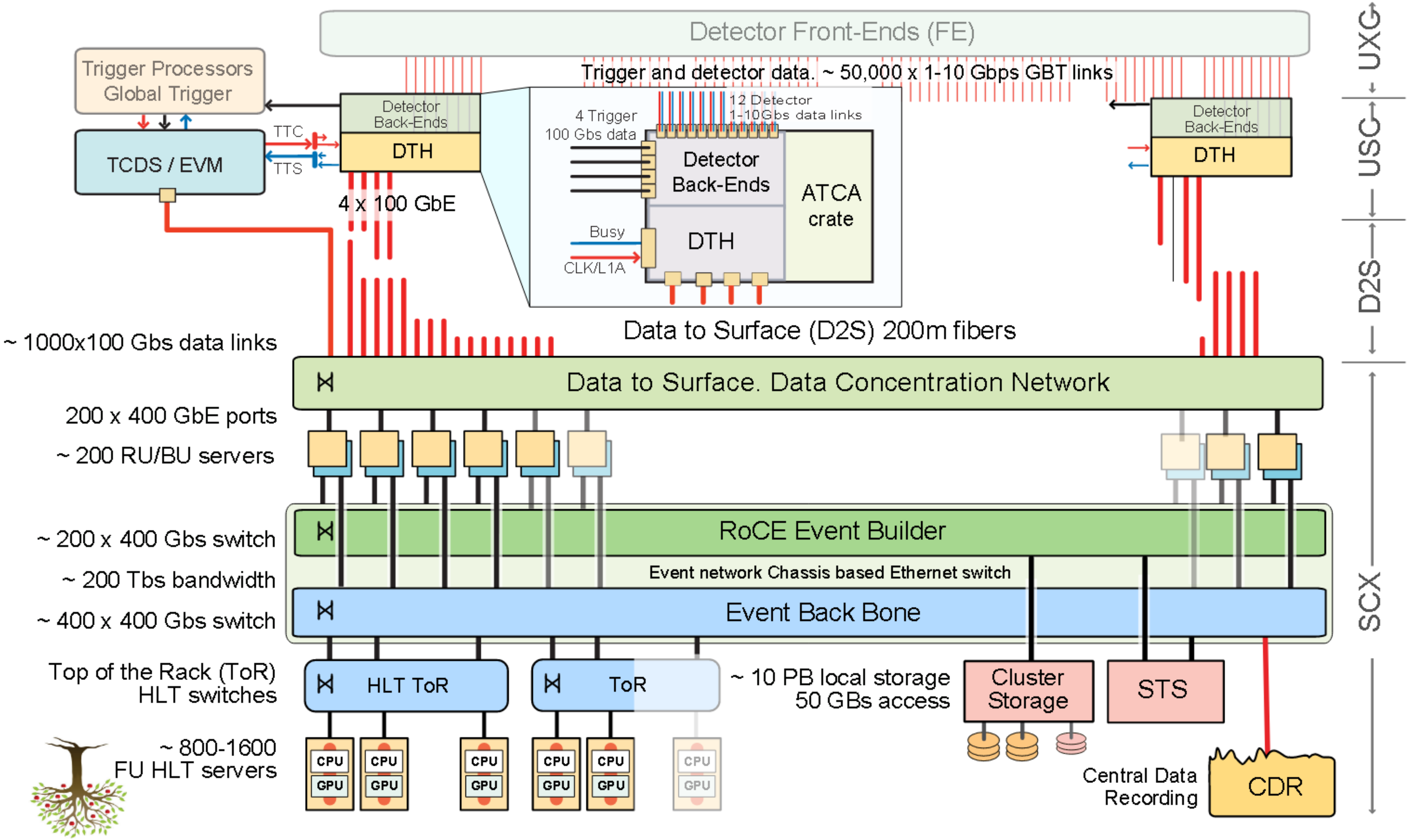


Current DAQ architecture



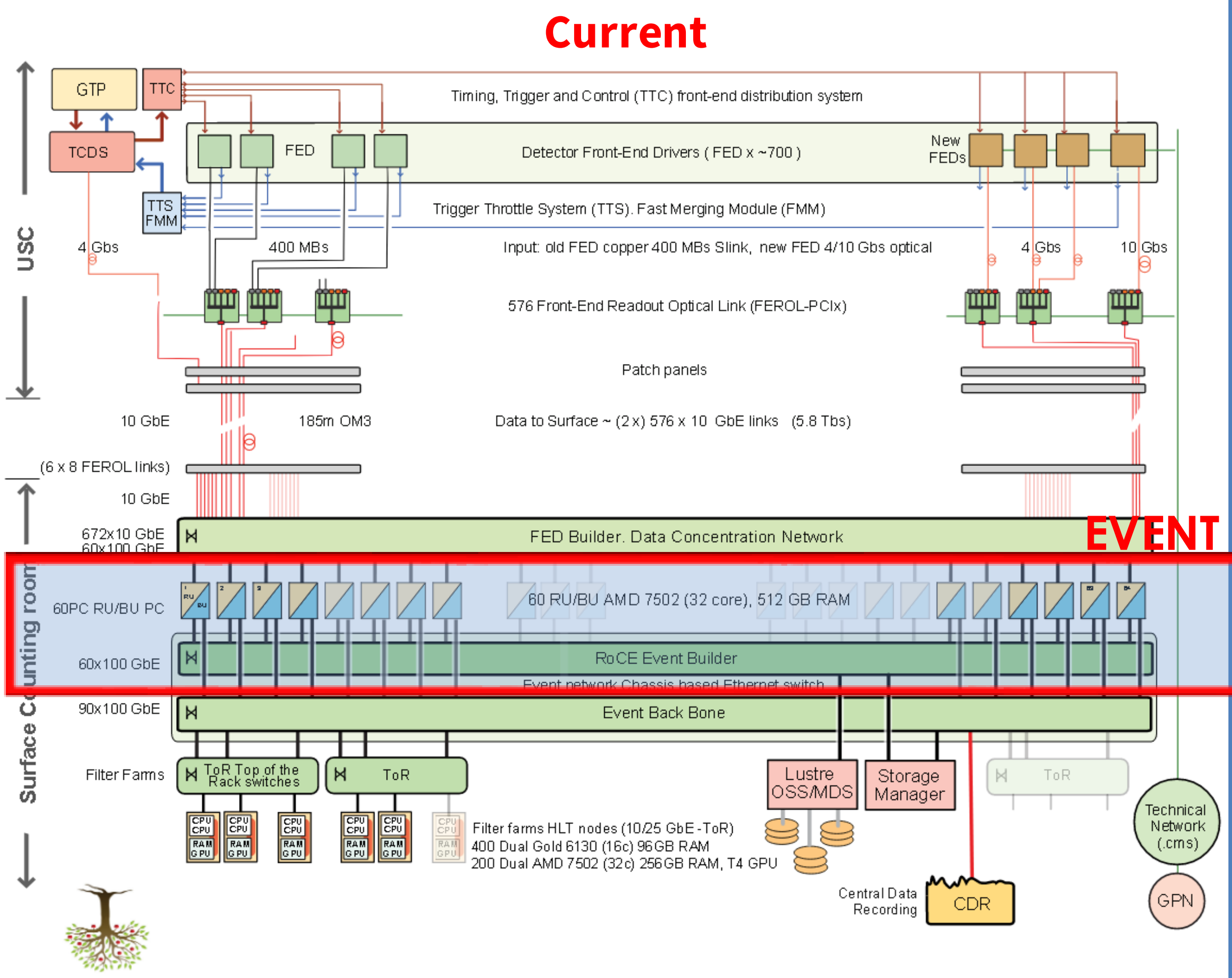
Source: CMS PAPER PRF-21-001

Phase-2 DAQ architecture

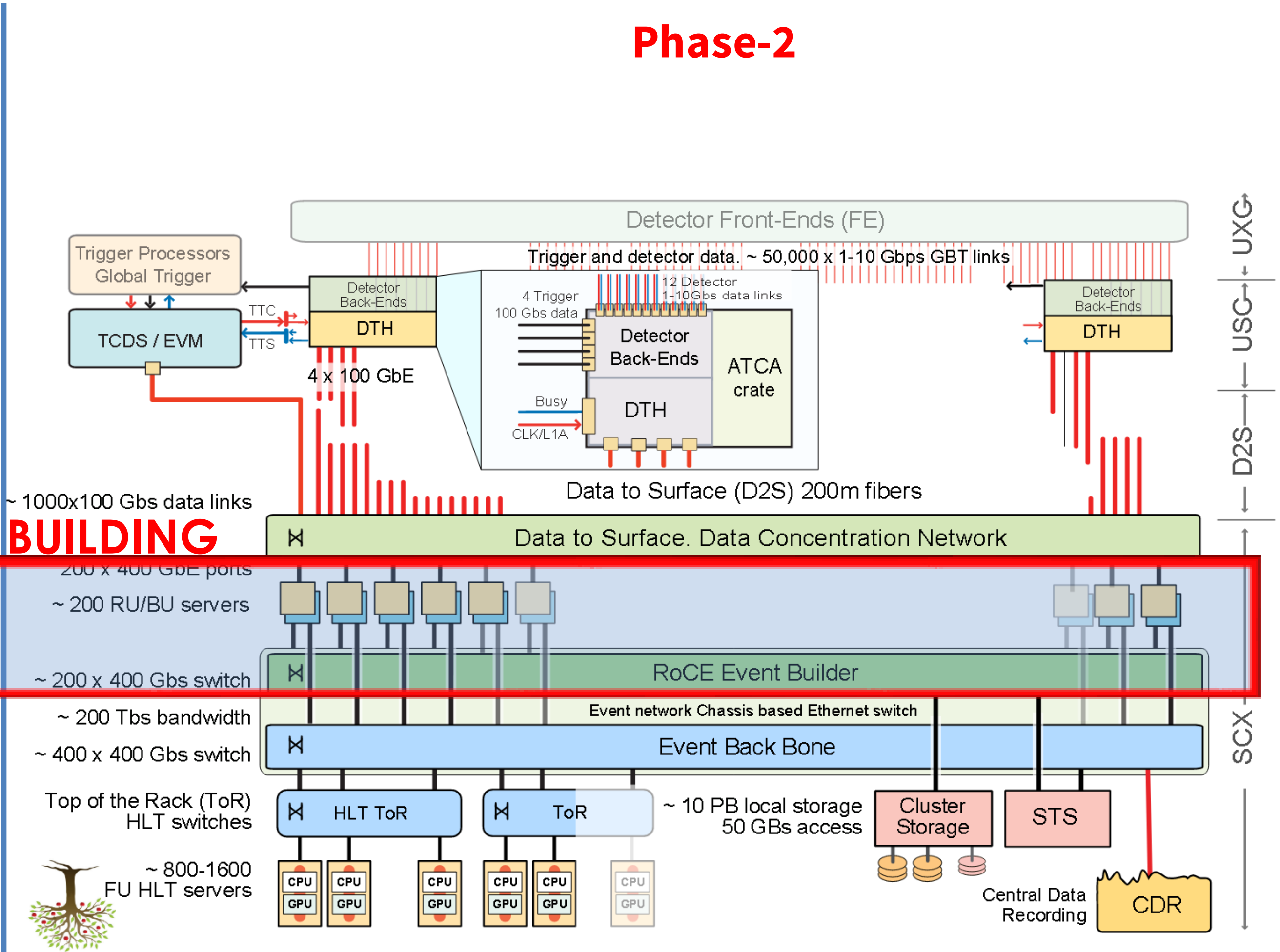


Source: CMS-TDR-022

Current versus Phase-2 DAQ architecture

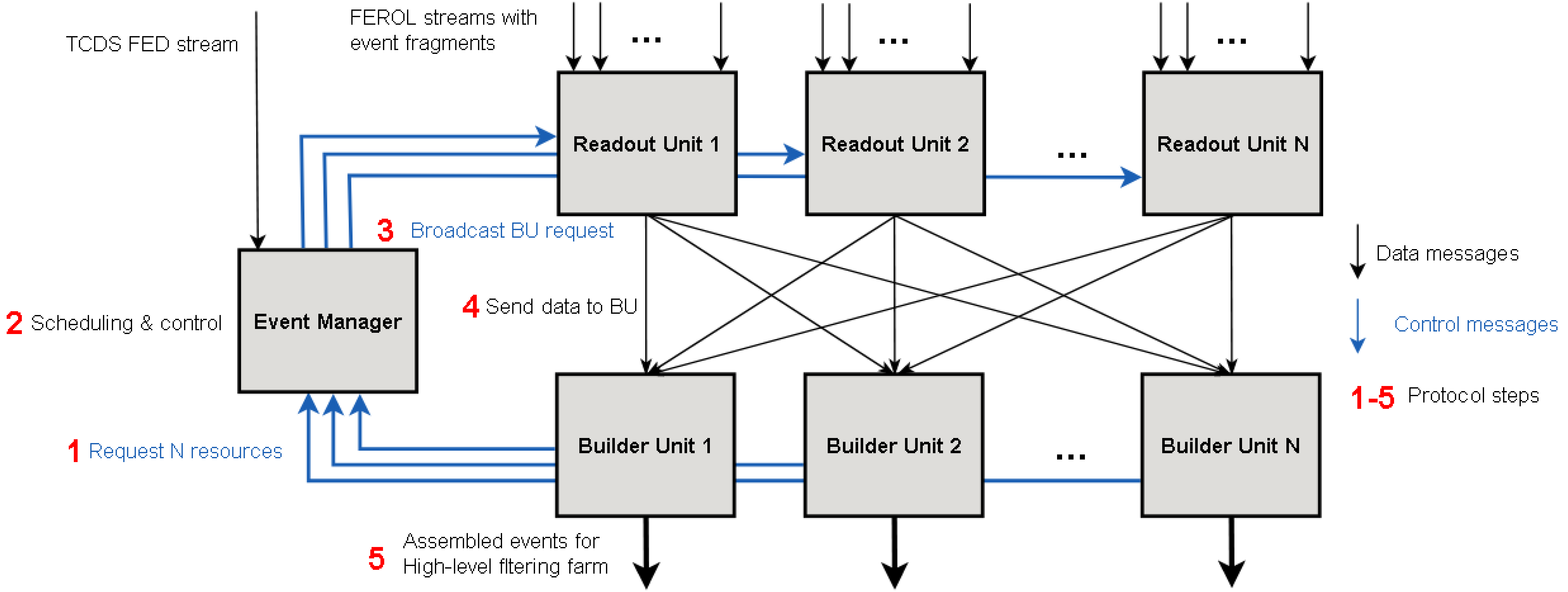


Source: CMS PAPER PRF-21-001



Source: CMS-TDR-022

CMS event building



Primary objective → assembling events from their scattered fragments

Phase-2 DAQ:

- Event size → from 2 MB to 8.4 MB
- L1 Trigger acceptance rate → from 100 kHz to 750 kHz
- HLT accept rate → 1 kHz to 7.5 kHz

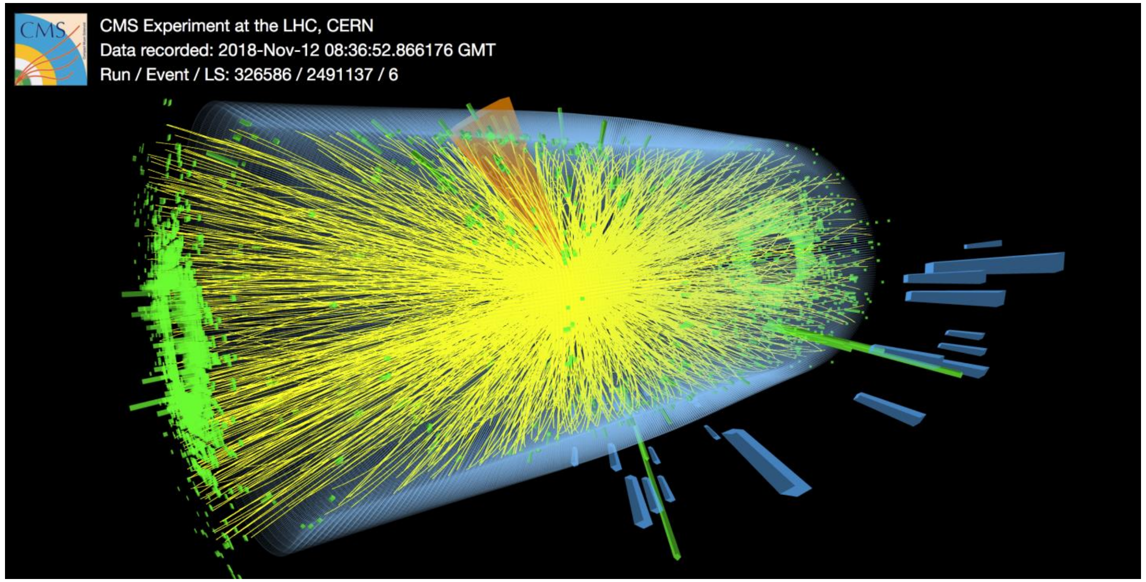
Challenge for Phase-2 → increased workloads:

- Total builder network traffic → from 1.6 Tb/s to 51 Tb/s
- Total servers from ~60 to ~200 servers
- High-performance software, quasi-real-time lossless data taking

Phase-2 event versus orbit

Events in DAQ

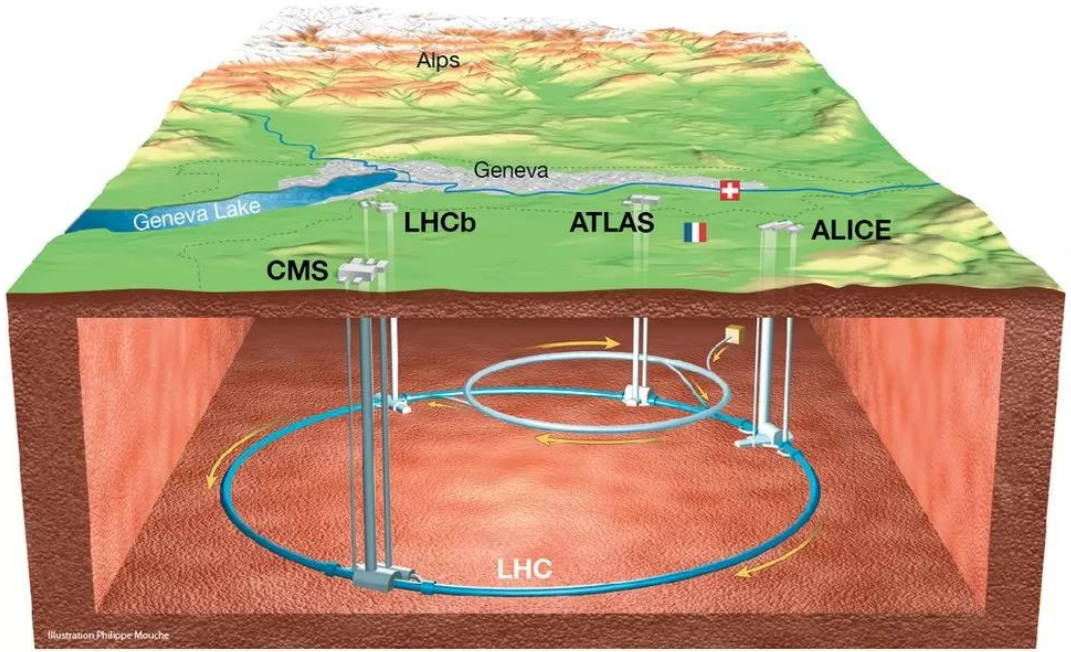
- Corresponds to a **collision** selected by L1 trigger
- **Full event** size → up to 8.4 MB
- Event rate → up to 750 kHz



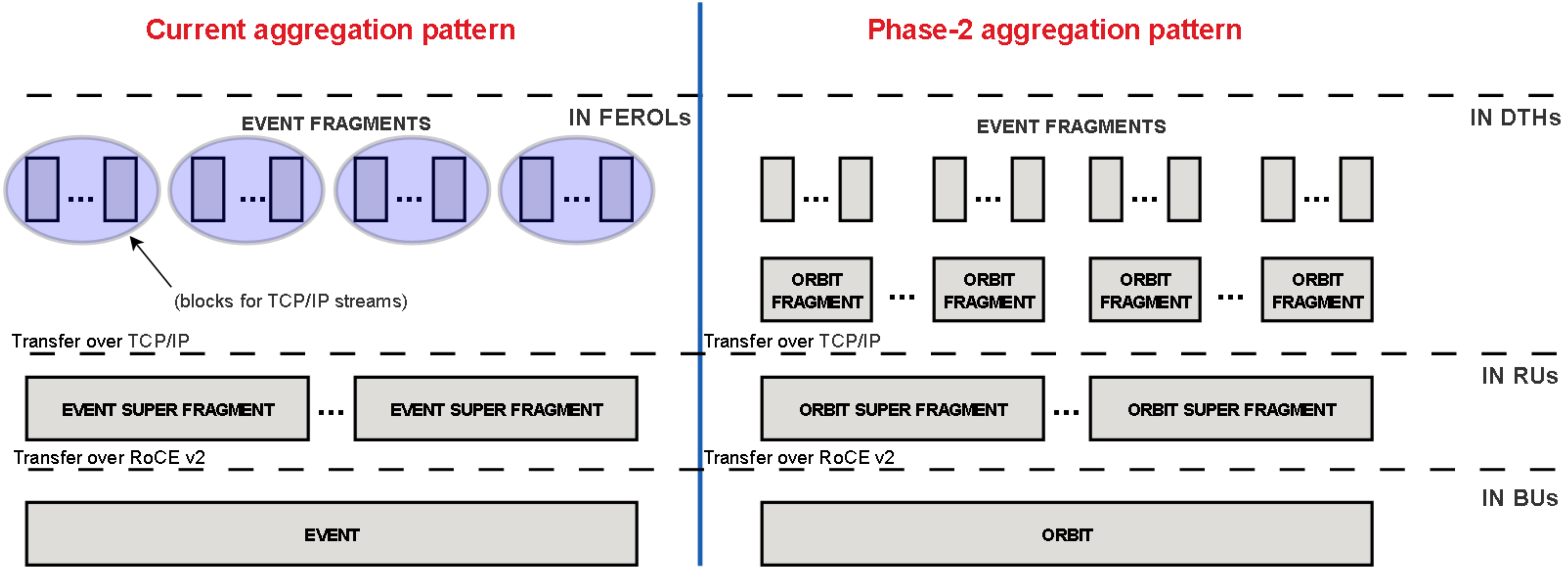
CMS-PHO-EVENTS-2018-010-19

Orbits in DAQ

- A collection of events during one **LHC orbit**
- **Orbit fragment** size→ 50-250 kB
- Orbit rate → 11.2 kHz
- 67 events per orbit on average



Phase-2 orbit builder data aggregation

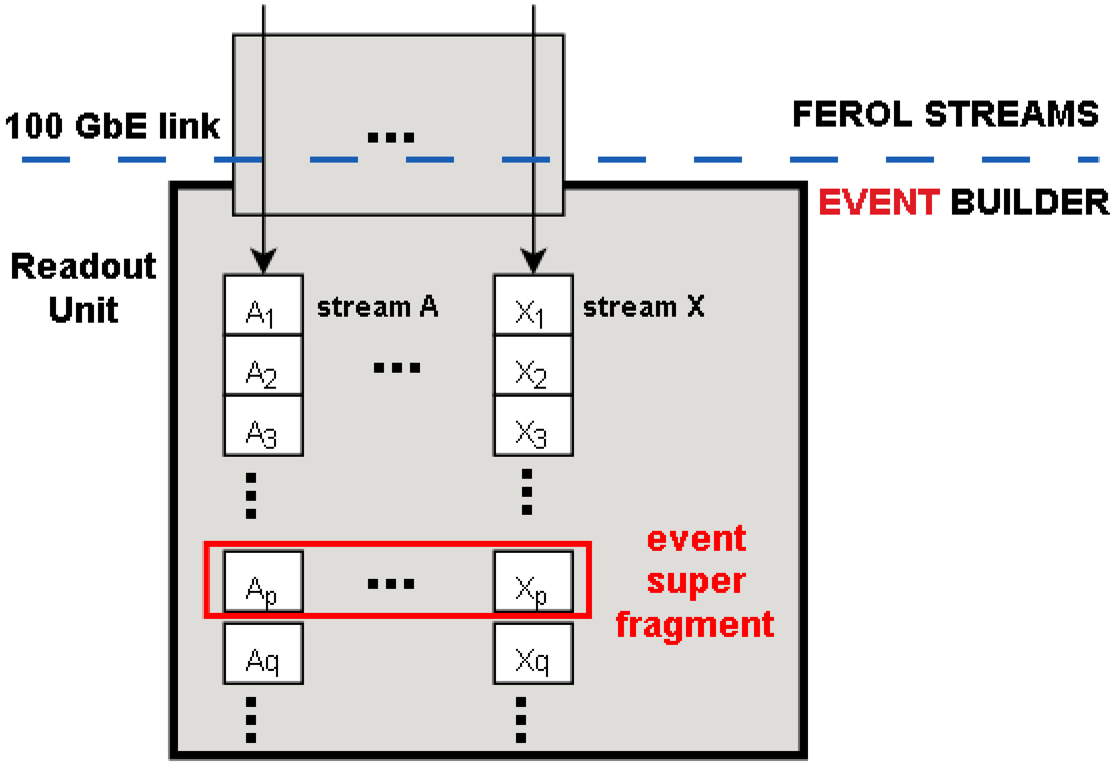


Why selected orbits for Phase-2:

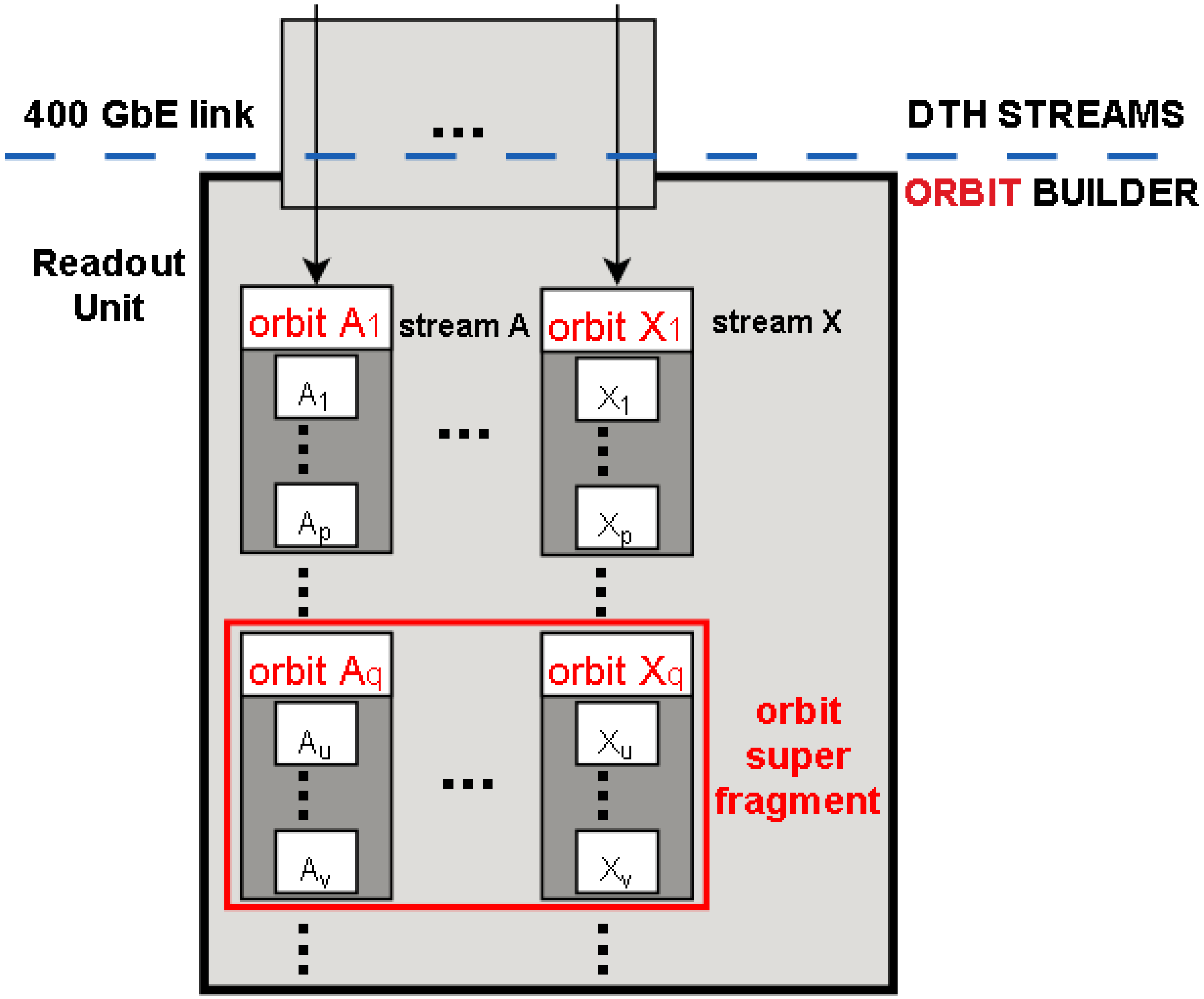
- More data per transmission to RU
- More data per RU-BU transmissions
- Less control messages in the event builder

Event builder versus orbit builder

Current aggregation by event



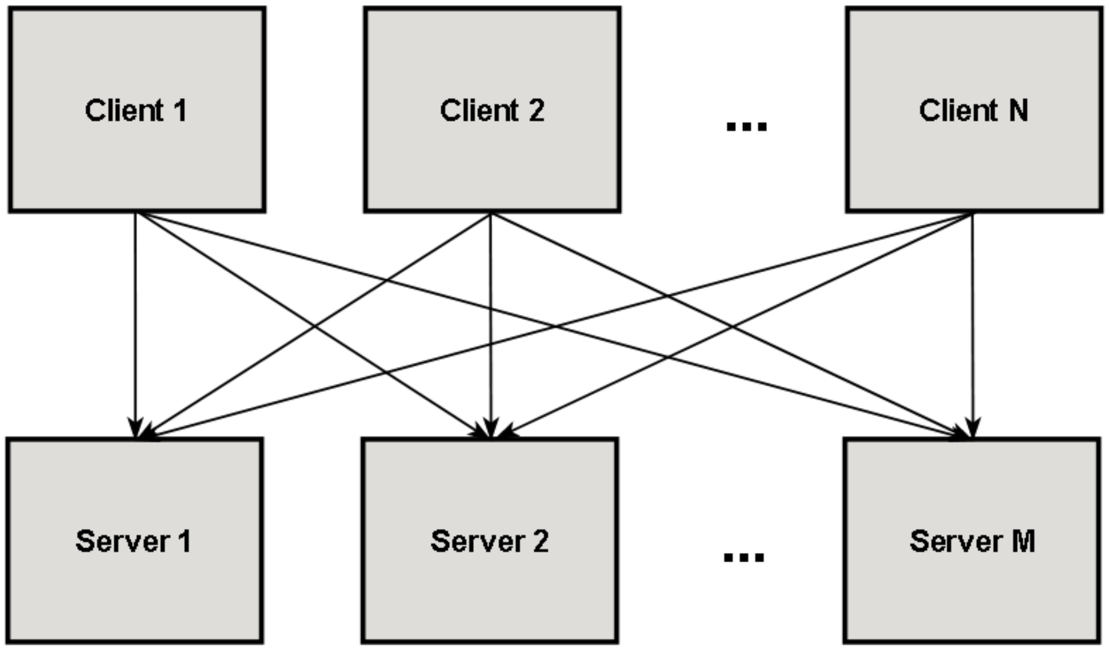
Phase-2 aggregation by orbit

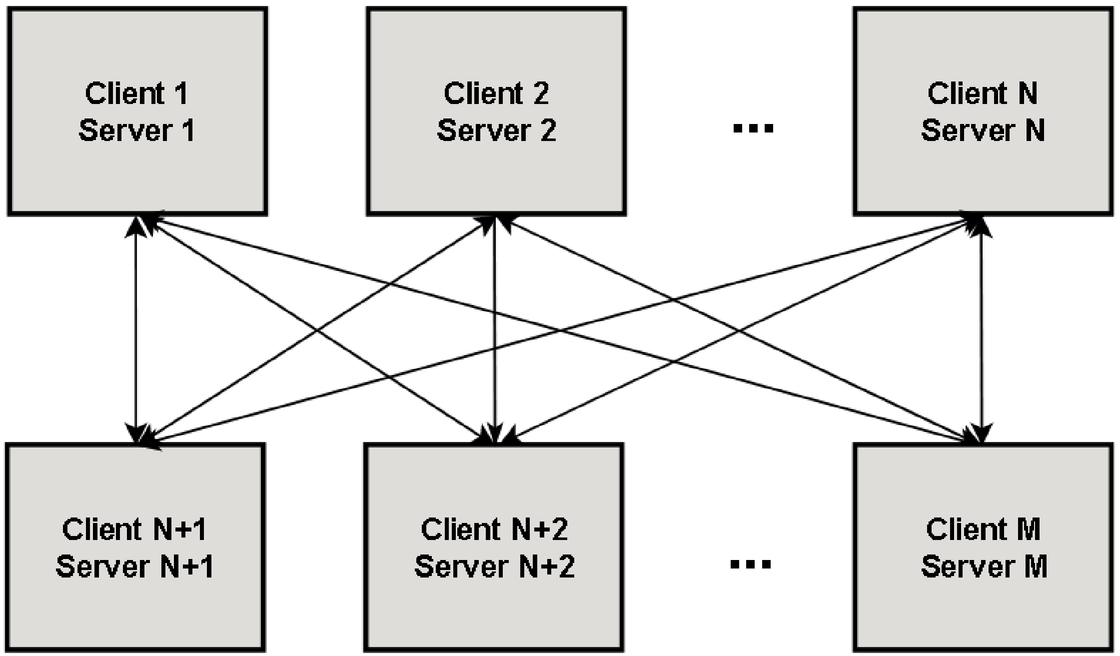


Orbit builder software study

Developed the **pipestream** C++benchmark based on the **XDAQ 2nd generation** online software

- **Emulates Event Builder network traffic**
- **REST** and **finite-state machine** for the runtime control
- High-performance library supporting RDMA over Converged Ethernet (RoCE)
- **YAML** for bootstrap configuration
- See the related CHEP talk → "*Towards a container-based architecture for CMS data acquisition*" by Dainius Šimelevičius
- Runs **standalone** or in **Kubernetes**
- Scheduled data sending over network between different nodes from clients to servers
- Throughput of clients and servers periodically probed through REST



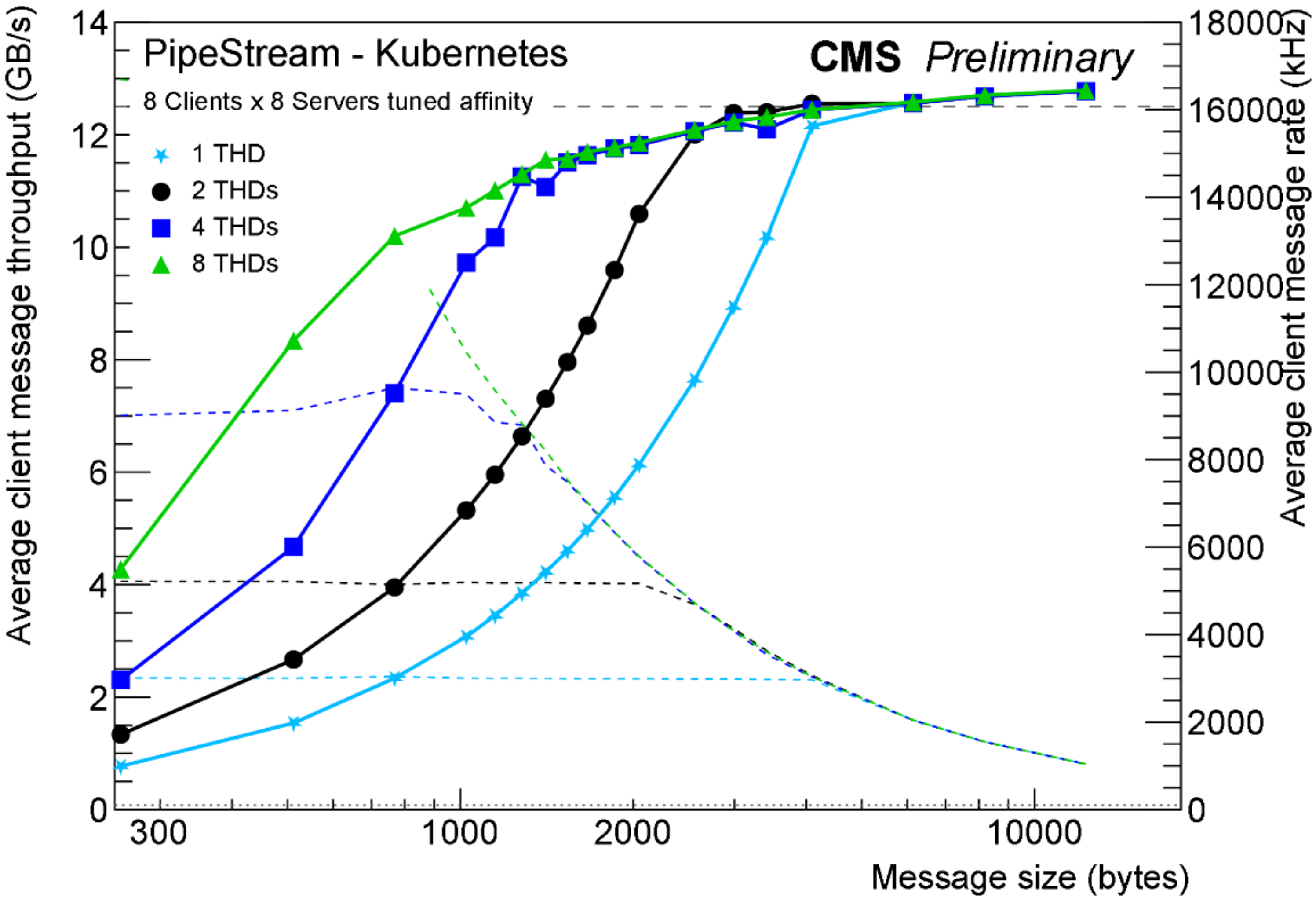


Performance tests

- Tuned parameters:
    - maximal message size
    - buffer size per connection
    - burst size
    - threads number and affinity
    - memory affinity
- Used the existing DAQ Run 3 infrastructure with 100 Gigabit Ethernet
- Measured nodes performance for the all-to-all, CMS event building-like traffic
- One orchestrator and 14 test nodes

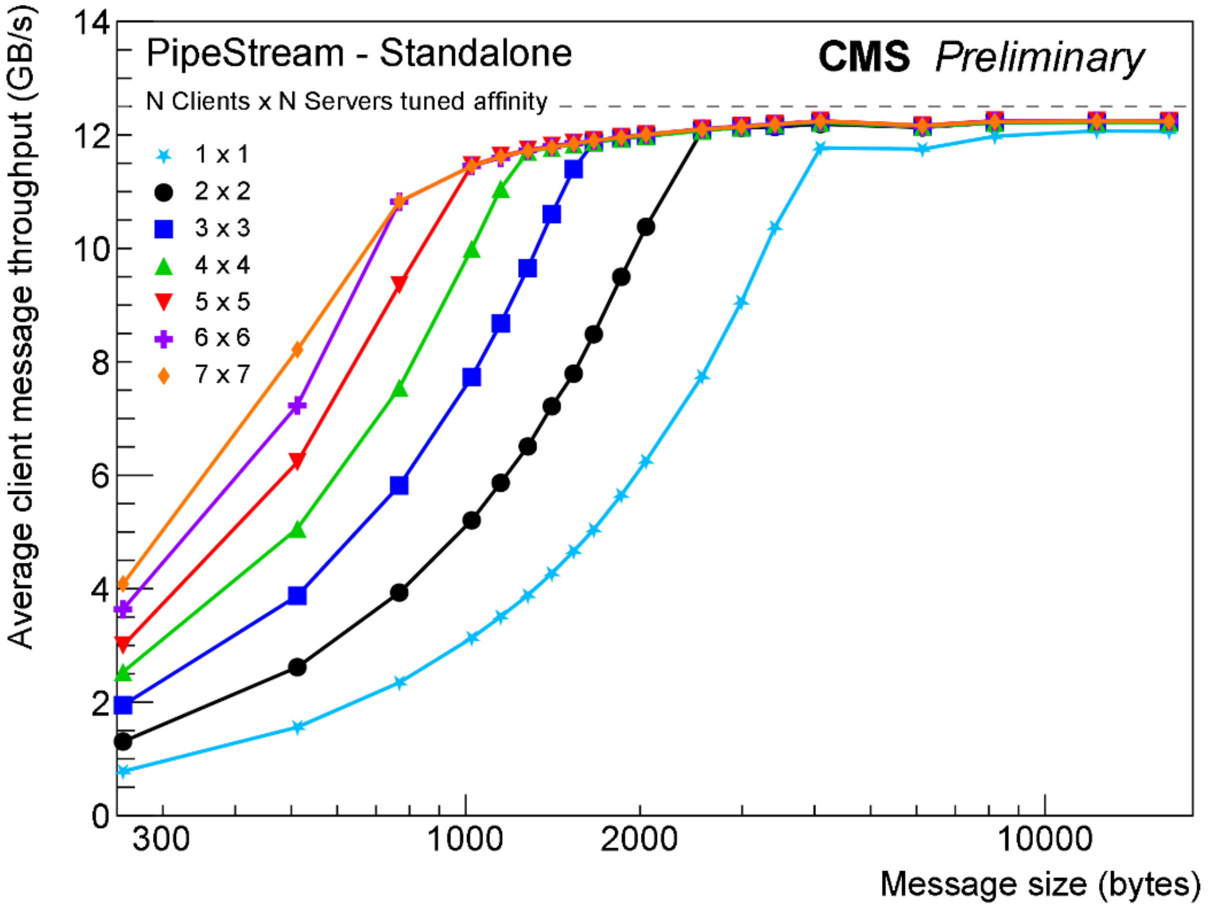
Message rate and throughput over small system

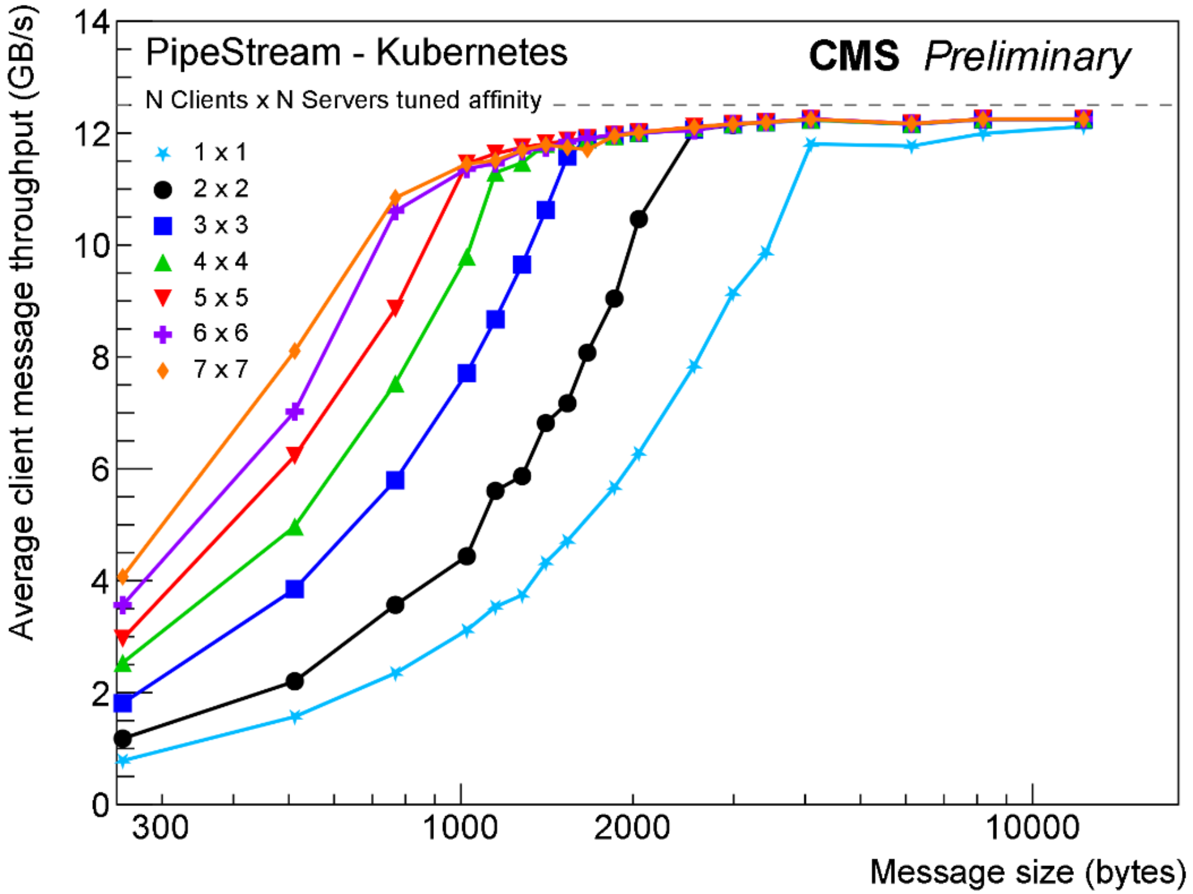
- A **folded** configuration with client and servers sharing nodes
- Checked performance for small message sizes
- Measured message rates



Standalone versus Kubernetes

- A configuration with client and servers on separate node
- Checked performance for CMS-like message sizes
- No performance penalty in K8s





Summary

- Proof of concept → XDAQ 2nd generation framework for the CMS Phase-2 event building use-case
- Initial results good enough to proceed with the development
- Next step → developing into a fully-functional event builder with the presented software platforms

Thank you

Primary authors:
- **Krawczyk, Rafał Dominik (RICE)**
- Petrucci, Andrea (UCSD)

Co-Authors:
Amoiridis, Vassileios (CERN); Behrens, Ulf (RICE); Bocci, Andrea (CERN); Branson, James (UCSD); Brummer, Philipp (CERN); Cano, Eric (CERN); Cittolin, Sergio (UCSD); Da Silva Almeida Da Quintanilha, Joao (CERN); Darlea, Georgiana-Lavinia (MIT); Deldicque, Christian (CERN); Dobson, Marc (CERN); Gigi, Dominique (CERN); Glege, Frank (CERN); Gomez-Ceballos, Guillelmo (MIT); Gutic, Neven (CERN); Hegeman, Jeroen (CERN); Izquierdo Moreno, Guillermo (CERN); Kartalas, Miltiadis (CERN); (RICE); Li, Wei (RICE); Long, Kenneth (MIT); Meijers, Frans (CERN); Meschi, Emilio (CERN); Morovic, Srecko (UCSD); Orsini, Luciano (CERN); Paus, Christoph (MIT); Pieri, Marco (UCSD); Rabady, Dinyar Sebastian (CERN); Racz, Attila (CERN); Sakulin, Hannes (CERN); Schwick, Christoph (CERN); Simelevicius, Dainius (Vilnius University); Vazquez Velez, Cristina (CERN); Zejdl, Petr (CERN); Zhang, Yousen (RICE); Zogatova, Dominika (CERN)

Supplementary slides

Performance tests nodes

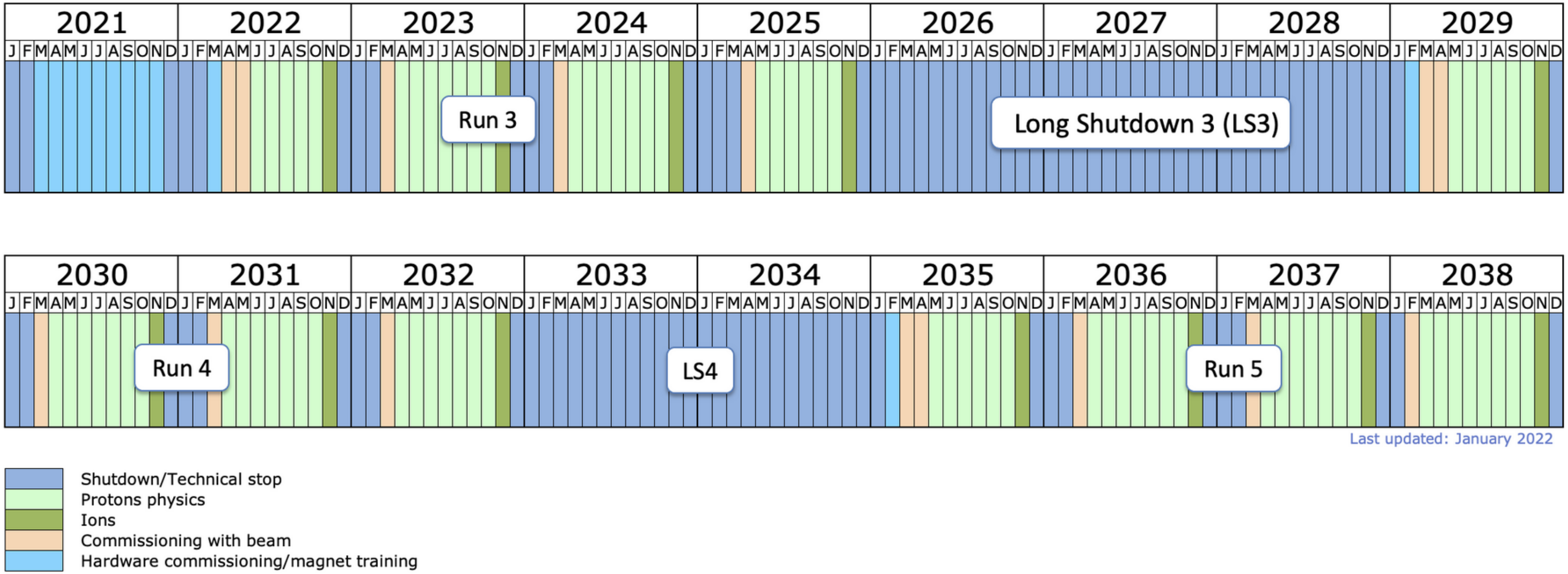
Worker Nodes

CPUs	2 x Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz
RAM	256 GiB DDR4, 2666 MT/s
NICs	Mellanox Connect X-6 in Ethernet mode

Test Network

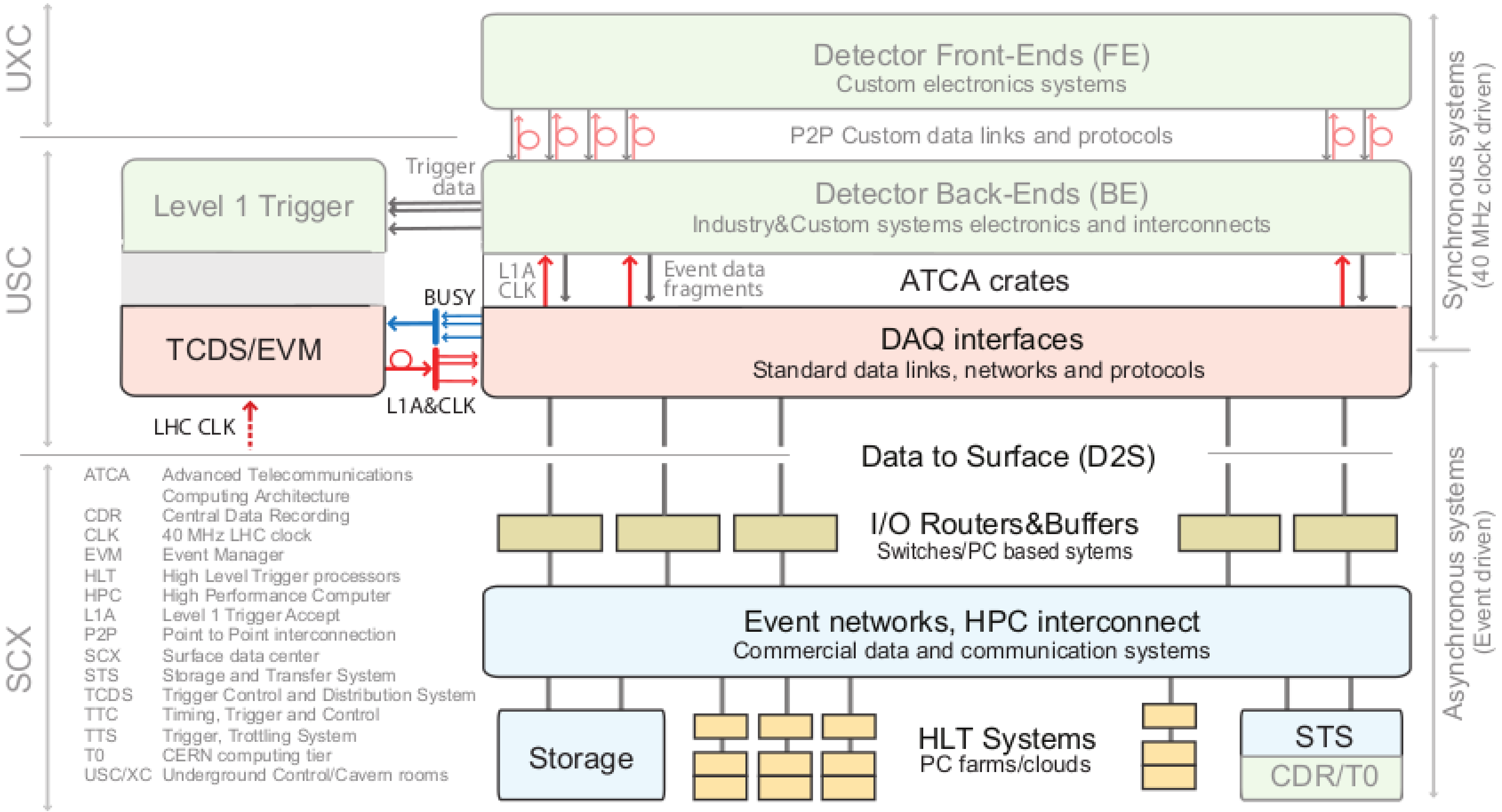
ports	14 x 100 Gbps
Switch	Juniper QFX10000-30C line card (100Gbps)
Chassis	QFX10008

LHC & detectors schedule



Source CERN :Longer term LHC schedule

Conceptual design of Phase-2 CMS DAQ



Source: CMS-TDR-022

Phase-2 tables & figures

CMS detector	LHC	HL-LHC	
	Phase-1	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]	6.1 MB	8.4 MB
Event Network throughput	1.6 Tb/s	24 Tb/s	51 Tb/s
Event Network buffer (60 s)	12 TB	182 TB	379 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	4.3 MB	5.9 MB
Storage throughput [d]	2 GB/s	24 GB/s	51 GB/s
Storage throughput (Heavy-Ion)	12 GB/s	51 GB/s	51 GB/s
Storage capacity needed (1 day [e])	0.2 PB	1.6 PB	3.3 PB

[a]Design value.
[b]Does not include Data Quality Monitoring.
[c]Actual compression factor for Phase-1. For Phase-2 same factor is assumed, see Section 6.2.11.
[d]The storage throughput is defined as the effective throughput with concurrent recording and transfer. The throughput required is determined by the HLT output event size and the additional output streams, see Section 6.2.11.
[e]Assuming an LHC duty cycle, i.e. the fraction of time spent in stable colliding beams, of 75 %.

Source: CMS-TDR-022

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]	900 links
Data Concentrator Network	Chassis-based [c] switch	1100 ports [d]
Event Builder Nodes [e]	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	400 ports
ToR switch	Rack-mount[f] switch	42 ToR switch (approx. 5×50 ports)
HLT servers [g]	Rack-mount 1U(2U) server with 2(6) GPU	1600(840) servers
Storage System	Network-attached storage appliance	102 GB/s bandwidth [h] 3.3 PB total storage

[a]DTH-400 boards with DAQ and TCDS functionality and DAQ-800 boards with DAQ functionality only.
[b]Transceiver, optical module and single-mode optical fibers linking USC to SCX.
[c]Switch with 100 Gb/s and 400 Gb/s line cards.
[d]900 ports 100 Gb/s and 200 ports 400 Gb/s.
[e]A server capable of ≈ 1 Tb/s concurrent input and output is assumed (requires PCIe Gen5).
[f]400 Gb/s uplinks from Event Backbone and 100 Gb/s downlinks to HLT servers.
[g]The values in parentheses are for Run-5.
[h]providing 51 GB/s throughput (read+write).

Source: CMS-TDR-022

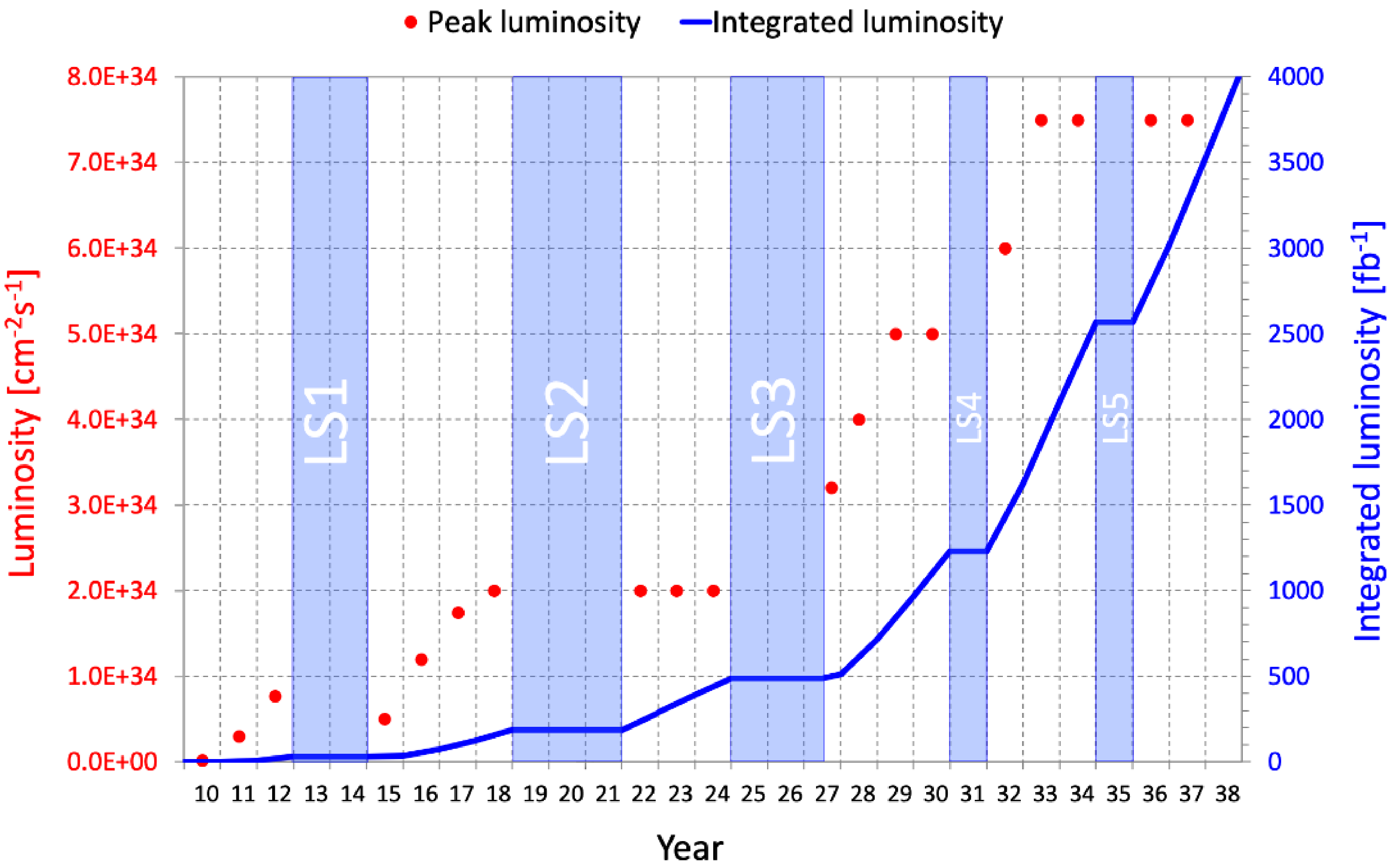
Run 1-3 table

	Run 1	Run 2	Run 3
Event building rate pp	100 kHz	100 kHz	100 kHz
Event size pp [a]	1 MB	2 MB	2 MB
Read-out links S-LINK64 (copper) 400 MB/s[d]	636[b]	575[b] .. 532[c]	528[b]
Read-out links optical[i] 6 Gb/s [d]	-	55[c,e]	55[b,e]
Read-out links optical[i] 10 Gb/s	-	60[b] .. 167[c]	176[b]
FED Builder network technology	Myrinet	Ethernet	Ethernet
FED Builder network speed	2 rails of 2.5 Gb/s	10 & 40 Gb/s	10 & 100 Gb/s
Event builder # of readout units	640	108[c]	50[f]
Event Builder network technology	Ethernet	Infiniband	Ethernet RoCE v2[j]
Event Builder link speed	1-3 rails of 1 Gb/s	56 Gb/s	100 Gb/s
Event Builder parallel slices	8	1	1
Event Builder network throughput	1.0 Tb/s	1.6 Tb/s	1.6 Tb/s
Event Builder # of builder units	1260[g]	73[c]	50[f]
BU RAM disk buffer	none	16 TB	15 TB
HLT # of filter unit motherboards	720[b,g] .. 1260[c,g]	900[b] .. 1084[c]	200[k]
HLT # cores	5.8k[b] .. 13k[c]	16k[b] .. 31k[c]	26k[h,k]
HLT computing power (MHS06)	0.05[b] .. 0.20[c]	0.34[b] .. 0.72[c]	0.65[h]
HLT # of NVIDIA T4 GPUs	-	-	400[k]
Storage system technology	16 SAN[l] systems	1 cluster file system	1 cluster file system
Storage system bandwidth write + read	2 GB/s	9 GB/s	30 GB/s
Storage system capacity	300 TB	500 TB	1.2 PB
Transfer System to Tier-0 speed	2×10 Gb/s	4× 40 Gb/s	4× 100 Gb/s

[a]design value, [b]at the beginning of the run, [c]at the end of the run, [d]main data-taking configuration - excluding links from partition managers used for partitioned running, [e]54 links from mezzanines with optical SlinkExpress, [f]readout and builder unit running on same server ("folded event builder"), [g]filter and builder units running on same server, [h]not including GPU compute power, [i]SlinkExpress, [j]Remote DMA over Converged Ethernet, [k]ordered at the time of writing, [l]Storage-area network

Source: CMS PAPER PRF-21-001

Runs luminosity timeline



Source: CMS-TDR-022