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) $2 \times 10^{34}$ cm$^{-2}$ s$^{-1}$
  - Run 4 (2029) $5 \times 10^{34}$ cm$^{-2}$ s$^{-1}$
  - Run 5 (2035) $7.5 \times 10^{34}$ 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

- **Trigger Processors Global Trigger**
- **TCDS / EVM**
- **Detector Back-Ends**
  - **DTH**
- **Data to Surface (D2S) 200m fibers**
  - ~1000x100 Gbs data links
- **Data to Surface, Data Concentration Network**
  - 200 x 400 GbE ports
  - ~200 RU/BU servers
- **RoCE Event Builder**
  - Event network Chassis based Ethernet switch
  - ~200 x 400 Gbs switch
  - ~200 Tbs bandwidth
  - ~400 x 400 Gbs switch
- **Top of the Rack (ToR) HLT switches**
  - ~800-1600 FU HLT servers
- **Event Back Bone**
  - HLT ToR
  - ToR
  - ~10 PB local storage 50 Gb/s access
- **Cluster Storage**
- **STS**
- **Central Data Recording**
- **Source: CMS-TDR-022**
Current versus Phase-2 DAQ architecture

**Current**
- Timing, Trigger and Control (TTC) front-end/distribution system
- Detector Front-End Drivers (FEDs)
- Trigger Throttle System (TTS) Fast-Peaking Module (FPM)
- Input: 36 FEDs per 400 Mb/s links, now FED 410 Gb/s optical
- 576 Front End Readout Optical Link (FEROL-FCx)
- Data to Surface ~2x576 x 10 GbE links (5.8 To) ~1000x100 GbE data links

**Phase-2**
- Detector Front-Ends (FE)
- Trigger and detector data: ~50,000 x 1-10 Gbps GBT links

**Event Building**
- Event Back Bone
- ~200 RUBU servers
- ~200 x 400 Gbps switch
- ~400 x 400 Gbps switch
- ~800-1000 FU HLT servers

**Source:** CMS PAPER PRF-21-001

**Source:** CMS-TDR-022
CMS event building

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

Primary objective → assembling events from their scattered fragments
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

**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

**Current aggregation pattern**
- EVENT FRAGMENTS
- IN FEROLs
- Transfer over TCP/IP
- EVENT SUPER FRAGMENT
- Transfer over RoCE v2
- EVENT

**Phase-2 aggregation pattern**
- EVENT FRAGMENTS
- Transfer over TCP/IP
- ORBIT FRAGMENT
- Transfer over RoCE v2
- ORBIT SUPER FRAGMENT
- ORBIT

**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 2\textsuperscript{nd} 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); Delcidque, 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); 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

<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPUs</td>
<td>2 x Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz</td>
</tr>
<tr>
<td>RAM</td>
<td>256 GiB DDR4, 2666 MT/s</td>
</tr>
<tr>
<td>NICs</td>
<td>Mellanox Connect X-6 in Ethernet mode</td>
</tr>
</tbody>
</table>

## Test Network

<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>ports</td>
<td>14 x 100 Gbps</td>
</tr>
<tr>
<td>Switch</td>
<td>Juniper QFX10000-30C line card (100Gbps)</td>
</tr>
<tr>
<td>Chassis</td>
<td>QFX10008</td>
</tr>
</tbody>
</table>
### LHC & detectors schedule

<table>
<thead>
<tr>
<th>Year</th>
<th>2021</th>
<th>2022</th>
<th>2023</th>
<th>2024</th>
<th>2025</th>
<th>2026</th>
<th>2027</th>
<th>2028</th>
<th>2029</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 3</td>
<td>Run 4</td>
<td>Long Shutdown 3 (LS3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>2030</th>
<th>2031</th>
<th>2032</th>
<th>2033</th>
<th>2034</th>
<th>2035</th>
<th>2036</th>
<th>2037</th>
<th>2038</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run 5</td>
<td>LS4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Legend:**
- Dark blue: Shutdown/Technical stop
- Medium blue: Protons physics
- Light green: Ions
- Light yellow: Commissioning with beam
- Light blue: Hardware commissioning/magnet training

Last updated: January 2022

Source: CERN: Longer term LHC schedule
Conceptual design of Phase-2 CMS DAQ

Source: CMS-TDR-022
Phase-2 tables & figures

<table>
<thead>
<tr>
<th>Component</th>
<th>Technology</th>
<th>Estimated quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTH-400 and DAQ-800 boards</td>
<td>ATCA custom board</td>
<td>250 boards</td>
</tr>
<tr>
<td>TCDS2 custom boards</td>
<td>ATCA custom board</td>
<td>16 boards</td>
</tr>
<tr>
<td>DAQ D2S links</td>
<td>100-GBASE-CWDM4</td>
<td>900 links</td>
</tr>
<tr>
<td>Data Concentrator Network</td>
<td>Chassis-based switch</td>
<td>1100 ports</td>
</tr>
<tr>
<td>Event Builder Nodes</td>
<td>Rack-mount 2U server</td>
<td>200 servers</td>
</tr>
<tr>
<td>Event Builder Network</td>
<td>Chassis-based 400 Gb/s switch</td>
<td>200 ports</td>
</tr>
<tr>
<td>Event Backbone Network</td>
<td>Chassis-based 400 Gb/s switch</td>
<td>400 ports</td>
</tr>
<tr>
<td>ToR switch</td>
<td>Rack-mount switch</td>
<td>42 ToR switch (approx. 5 x 50 ports)</td>
</tr>
<tr>
<td>HLT servers</td>
<td>Rack-mount 1U(2U) server with 2(6) GPU</td>
<td>1600 (840) servers</td>
</tr>
<tr>
<td>Storage System</td>
<td>Network-attached storage appliance</td>
<td>102 GB/s bandwidth</td>
</tr>
<tr>
<td>Storage capacity needed (1 day)</td>
<td>3.3 PB total storage</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CMS detector</th>
<th>LHC Phase-1</th>
<th>LHC Phase-2</th>
<th>HL-LHC Phase-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak (PU)</td>
<td>60</td>
<td>140</td>
<td>200</td>
</tr>
<tr>
<td>L1 accept rate (maximum)</td>
<td>100 kHz</td>
<td>500 kHz</td>
<td>750 kHz</td>
</tr>
<tr>
<td>Event Size at HLT input</td>
<td>2.0 MB</td>
<td>6.1 MB</td>
<td>8.4 MB</td>
</tr>
<tr>
<td>Event Network throughput</td>
<td>1.6 Tb/s</td>
<td>24 Tb/s</td>
<td>51 Tb/s</td>
</tr>
<tr>
<td>Event Network buffer (60 s)</td>
<td>12 TB</td>
<td>182 TB</td>
<td>379 TB</td>
</tr>
<tr>
<td>HLT accept rate</td>
<td>1 kHz</td>
<td>5 kHz</td>
<td>7.5 kHz</td>
</tr>
<tr>
<td>HLT computing power</td>
<td>0.7 MHS06</td>
<td>17 MHS06</td>
<td>37 MHS06</td>
</tr>
<tr>
<td>Event Size at HLT output</td>
<td>1.4 MB</td>
<td>4.3 MB</td>
<td>5.9 MB</td>
</tr>
<tr>
<td>Storage throughput</td>
<td>2 GB/s</td>
<td>24 GB/s</td>
<td>51 GB/s</td>
</tr>
<tr>
<td>Storage throughput (Heavy-Ion)</td>
<td>12 GB/s</td>
<td>51 GB/s</td>
<td>51 GB/s</td>
</tr>
<tr>
<td>Storage capacity needed (1 day)</td>
<td>0.2 PB</td>
<td>1.6 PB</td>
<td>3.3 PB</td>
</tr>
</tbody>
</table>

\[a\] 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 throughput is defined as the effective throughput with concurrent recording and transfer. The throughput 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
### Run 1-3 table

<table>
<thead>
<tr>
<th>Event building rate pp</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event size pp</td>
<td>100 kHz</td>
<td>100 kHz</td>
<td>100 kHz</td>
</tr>
<tr>
<td>Read-out links S-LINK64 (copper) 400 MB/s</td>
<td>636 MB</td>
<td>575 MB..532 MB</td>
<td>529 MB</td>
</tr>
<tr>
<td>Read-out links optical 6 Gb/s</td>
<td>-</td>
<td>555 MB..608 MB</td>
<td>555 MB..608 MB</td>
</tr>
<tr>
<td>Read-out links optical 10 Gb/s</td>
<td>-</td>
<td>608 MB..167 MB</td>
<td>176 MB</td>
</tr>
<tr>
<td>FED Builder network technology</td>
<td>Myrinet</td>
<td>Ethernet</td>
<td>Ethernet</td>
</tr>
<tr>
<td>FED Builder network speed</td>
<td>2 rails of 2.5 Gb/s</td>
<td>10 &amp; 40 Gb/s</td>
<td>10 &amp; 100 Gb/s</td>
</tr>
<tr>
<td>Event builder # of readout units</td>
<td>640</td>
<td>108 MB</td>
<td>50</td>
</tr>
<tr>
<td>Event Builder network technology</td>
<td>Ethernet</td>
<td>Infiniband</td>
<td>Ethernet RoCE v2</td>
</tr>
<tr>
<td>Event Builder link speed</td>
<td>1-3 rails of 1 Gb/s</td>
<td>56 Gb/s</td>
<td>100 Gb/s</td>
</tr>
<tr>
<td>Event Builder parallel slices</td>
<td>8</td>
<td>1.6 Tb/s</td>
<td>1</td>
</tr>
<tr>
<td>Event Builder network throughput</td>
<td>1260 MB</td>
<td>73 MB</td>
<td>1</td>
</tr>
<tr>
<td>BU RAM disk buffer</td>
<td>none</td>
<td>26 Gb/k</td>
<td>15 TB</td>
</tr>
<tr>
<td>HLT # of filter unit motherboards</td>
<td>720 MB..1260 MB</td>
<td>900 MB..1084 MB</td>
<td>200 MB</td>
</tr>
<tr>
<td>HLT # cores</td>
<td>5.8 MB..13 MB</td>
<td>16 Gb/k..31 Gb/k</td>
<td>26 Gb/k</td>
</tr>
<tr>
<td>HLT computing power (MHS06)</td>
<td>0.05 MB..0.20 MB</td>
<td>0.34 MB..0.72 MB</td>
<td>0.65 MB</td>
</tr>
<tr>
<td>HLT # of NVIDIA T4 GPUs</td>
<td>-</td>
<td>-</td>
<td>400 MB</td>
</tr>
<tr>
<td>Storage system technology</td>
<td>16 SAN</td>
<td>1 cluster file system</td>
<td>1 cluster file system</td>
</tr>
<tr>
<td>Storage system bandwidth write + read</td>
<td>2 GB/s</td>
<td>9 GB/s</td>
<td>30 GB/s</td>
</tr>
<tr>
<td>Storage system capacity</td>
<td>300 TB</td>
<td>500 TB</td>
<td>1.2 PB</td>
</tr>
<tr>
<td>Transfer System to Tier-0 speed</td>
<td>2 x 10 Gb/s</td>
<td>4 x 40 Gb/s</td>
<td>4 x 100 Gb/s</td>
</tr>
</tbody>
</table>

*"design value, "at the beginning of the run, "at the end of the run, "main data-taking configuration - excluding links from partition managers used for partitioned running, "54 links from mezzanines with optical SlinkExpress, /readout and builder unit running on same server ("folded event builder"), /filter and builder units running on same server, /not including GPU compute power, /SlinkExpress, /Remote DMA over Converged Ethernet, /ordered at the time of writing, /storage-area network

Source: CMS PAPER PRF-21-001
Runs luminosity timeline

Source: CMS-TDR-022