One year of LHCb triggerless DAQ: achievements and lessons learned

26th International Conference on Computing in High Energy & Nuclear Physics
08/05/2023

Flavio Pisani for the LHCb Online team
CERN
The low-level trigger saturates in hadronic channels
The instantaneous luminosity in Run 3 will go up to $2 \times 10^{33} \text{ cm}^{-2} \text{ s}^{-1}$

A substantial upgrade is needed to take advantage of the increased luminosity
Online DAQ system overview

Front-End electronics

DAQ cards

Event Builder + HLT1

Buffer Storage

HLT2

32 Tbps

1 Tbps

~40 PB disk storage

164 servers: DAQ + event builder + event filter first pass (on GPGPUs)

~2000 full-duplex control links (Versatile Link)

~11000 half-duplex DAQ links (Versatile Link / GBT @ 4.8G, 300 m)

F. Pisani

One year of LHCb triggerless DAQ: achievements and lessons learned

26th CHEP Conference
The Event Building process in a nutshell

- Every event is divided into multiple fragments
- Every **Readout Unit (RU)** receives a fragment of the event
- Every **Builder Unit (BU)** has to gather all the fragments of the event
- Modular software architecture built in C++
- RU: it reads the data from the DAQ card and sends it over the EB network
- BU: it reads the data the EB network and it writes the built data into the HLT1 input buffer
- The scheduling synchronization is achieved using an in-band data barrier
- Dedicated low-level communication library
- Buffer-isolated critical sections to minimise slowdowns and deadtime
The processing of $N$ events is divided into $N$ phases ($N$ is the number of EB nodes)

- In every phase one RU sends data to one BU, and every BU receives data from one RU
- During phase $n$ RU $x$ sends data to BU $(x + n)\%N$
- All the units switch synchronously from phase $n$ to phase $n + 1$

**Congestion-free traffic on “selected networks” (e.g. fat-tree networks)**
The large amount of system memory available makes the system more robust against latency spikes.

The DMA/RDMA architecture reduces the memory throughput required.
• Buffers and discard policies allow reduce as much as possible the backpressure propagation
• In case of HW issues on a specific node it is possible to move only the RU/BU functionality to another different server
System scalability

Event Builder Scalability Test

- **Ideal**
- **Throughput**

Throughput [Tb/s]

Number of EB servers

Event rate [MHz]

Number of EB servers

**Experiment requirement**

**Event rate**
Pros and Cons after the first year

✓ We removed the physics inefficiencies introduced by the HW trigger
✓ The system is resilient against network latency spikes
✓ We used off-the-shelf components to reduce cost
✓ The converged architecture significantly reduces costs

✘ The HW trigger provides non-physics functionality
✘ Special tuning to optimise the PCIe communications
✘ Highly converged architectures are less flexible and reliable
Conclusions

- The LHCb experiment has been upgraded to perform a full read-out at the bunch-crossing rate
- The design uses as much as possible off-the-shelf parts
- The system has been optimised to be less sensitive to latency spikes
- The DAQ system can sustain the load of 32 Tb/s
- The system has been successfully used for the first part of Run3
THANK YOU FOR YOUR ATTENTION
BACKUP
It exists!

- 163 EB servers
- 24 racks: 18 EB, 2 control, 4 storage
- 28 40-port IB HDR switches: 18 leaf and 10 spine
Network architecture

![Diagram of network architecture with 32 Tbps and 1 Tbps connections, 164 EB servers, 16 storage servers, 200G IB, 100GbE, and 10GbE connections.](image)
The PCIe40: a single custom-made FPGA board for DAQ and Control

- Based on Intel Arria10
- 48x10G capable transceiver on 8xMPO for up to 48 full-duplex Versatile Links
- 2 dedicated 10G SFP+ for timing distribution
- 2x8 Gen3 PCIe
- Efficient and accurate software trigger that can perform online selection with offline-like quality
- One card multiple FW personalities:
  - Readout Supervisor (SODIN)
  - Interface Board (SOL40)
  - DAQ card (TELL40)
Versatile link / GBT

Credit: P. Moreira, S. Baron (CERN)
Barrier synchronization

Distributed tree barrier
EB server data flow
EB hardware layout
### Event size model

<table>
<thead>
<tr>
<th>Sub-detector</th>
<th>fragment size [B]</th>
<th>#tell40 streams</th>
<th>event size [B]</th>
<th>event fraction</th>
<th>MEP size [GB]</th>
<th>MFP size [MB]</th>
<th>RU send size [MB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velo</td>
<td>156</td>
<td>104</td>
<td>16250</td>
<td>0.13</td>
<td>0.49</td>
<td>4.69</td>
<td>14.06</td>
</tr>
<tr>
<td>UT</td>
<td>100</td>
<td>200</td>
<td>20000</td>
<td>0.16</td>
<td>0.60</td>
<td>3.00</td>
<td>9.00</td>
</tr>
<tr>
<td>SCIFI</td>
<td>100</td>
<td>288</td>
<td>28800</td>
<td>0.23</td>
<td>0.86</td>
<td>3.00</td>
<td>9.00</td>
</tr>
<tr>
<td>Rich 1</td>
<td>166</td>
<td>132</td>
<td>22000</td>
<td>0.18</td>
<td>0.66</td>
<td>5.00</td>
<td>15.00</td>
</tr>
<tr>
<td>Rich 2</td>
<td>166</td>
<td>72</td>
<td>12000</td>
<td>0.10</td>
<td>0.36</td>
<td>5.00</td>
<td>15.00</td>
</tr>
<tr>
<td>Calo</td>
<td>156</td>
<td>104</td>
<td>16250</td>
<td>0.13</td>
<td>0.49</td>
<td>4.69</td>
<td>14.06</td>
</tr>
<tr>
<td>Muon</td>
<td>156</td>
<td>56</td>
<td>8750</td>
<td>0.07</td>
<td>0.26</td>
<td>4.69</td>
<td>14.06</td>
</tr>
<tr>
<td>Total</td>
<td>1000</td>
<td>956</td>
<td>124050</td>
<td>1</td>
<td>3.72</td>
<td>30.06</td>
<td>90.19</td>
</tr>
</tbody>
</table>
Multiple Fragment Packet (MFP)

MFP header

```
<table>
<thead>
<tr>
<th>Time Slot</th>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0xCE</td>
<td>PSIZE</td>
</tr>
<tr>
<td>1</td>
<td>0x40</td>
<td>NFRAGS</td>
</tr>
<tr>
<td>2</td>
<td>SRCID</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>EVID</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ALIGN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FVERSION</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FTYPE 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FTYPE 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FTYPE 3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FTYPE 4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FTYPE N</td>
<td>padding to 32 bits</td>
</tr>
<tr>
<td></td>
<td>FSIZE 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSIZE 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FSIZE N</td>
<td>padding to $2^{\text{ALIGN}}$</td>
</tr>
</tbody>
</table>
```

MFP

```
<table>
<thead>
<tr>
<th>Time Slot</th>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>MFP header</td>
<td>padding to $2^{\text{ALIGN}}$</td>
</tr>
<tr>
<td>1</td>
<td>Fragment ($event = \text{EVID} + 0, \text{source} = \text{SRCID}$)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Fragment ($event = \text{EVID} + 1, \text{source} = \text{SRCID}$)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Fragment ($event = \text{EVID} + \text{NFRAGS} - 1, \text{source} = \text{SRCID}$)</td>
<td></td>
</tr>
</tbody>
</table>
```

F. Pisani  
One year of LHCb triggerless DAQ: achievements and lessons learned  
26th CHEP Conference
Multiple Event Packet (MEP)

**MEP header**

<table>
<thead>
<tr>
<th>0xFA</th>
<th>0xCE</th>
<th>NMFPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>PWORDS</td>
</tr>
<tr>
<td></td>
<td>SRCID 1</td>
<td>SRCID 2</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>SRCID N</td>
<td></td>
<td>padding to 32 bits</td>
</tr>
<tr>
<td>OFFSET 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFSET 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OFFSET N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MEP**

```
0 1 2 3
```

- **Header**
  - MEP header
    - padding as needed
  - MFP \( (source = \text{SRCID} 1) \)
    - padding
  - MFP \( (source = \text{SRCID} 2) \)
    - padding as needed
  - ... 
  - MFP \( (source = \text{SRCID} N) \)
    - padding as needed

- **Array of MFPs**
  - Array of source IDs
  - Array of MFP offsets