Trigger & DAQ at the LHC: filtering data from 50 TB/s to 1 GB/s

CERN openlab
summer student lectures

Flavio Pisani
CERN EP/LBC
Intro & Acknowledgments

• The circle of slides:
  ○ Much of the material for this talk was originally prepared by N. Neufeld and T. Colombo for the 2019 edition
  ○ They in turn acknowledges that it much of his presentation was naturally based on the work of other colleagues
  ○ A missing indication of origin of a figure does not imply that it is originally mine

• Trigger and DAQ are vast subjects covering a lot of physics and engineering:
  ○ This talk is meant to give a high-level overview
  ○ Some inevitable gross simplifications are to be expected
  ○ Many topics are left out altogether
The Large Hadron Collider
The Large Hadron Collider

- 27 km
- Vacuum at $10^{-13}$ atm
- $\geq 9600$ magnets
- Magnetic field $\sim 8$ T
- Magnets at $-271.3^\circ$ C
- Energy in the beam corresponds to a car at the speed of sound
- 4 large experiments
- Cost: 5 billion CHF
Colliding beams

Proton-Proton: 2835 bunch/beam
Protons/bunch: $10^{11}$
Beam energy: 7 TeV (7 x $10^{12}$ eV)
Luminosity: $10^{34}$ cm$^2$ s$^{-1}$
Crossing rate: 40 MHz
Collisions: $\approx 10^7$ - $10^9$ Hz

Selection of 1 in $10,000,000,000,000$
The LHC experiments

- **ALICE** – “A Large Ion Collider Experiment”
  - L x W x H: 26 x 16 x 16 m
  - Weight: 10000 t
  - 35 countries, 118 Institutes
  - Material costs: 110 MCHF
- **LHCb** – “LHC beauty”
  - L x W x H: 21 x 13 x 10 m
  - Weight: 5600 t
  - 15 countries, 52 institutes
  - Material costs: 75 MCHF
- **ATLAS** – “A Toroidal LHC ApparatuS”
  - L x W x H: 46 x 25 x 25 m
  - Weight: 7000 t
  - 38 countries, 174 institutes
  - Material costs: 540 MCHF
- **CMS** – “Compact Muon Solenoid”
  - L x W x H: 22 x 15 x 15 m
  - Weight: 12500 t
  - 40 countries, 172 institutes
  - Material costs: 500 MCHF
Lead-lead collision @ ALICE
Higgs boson @ ATLAS

Run: 204153
Event: 35369265
2012-05-30 20:31:28 CEST
Another Higgs boson @ CMS
Rare B meson decay @ LHCb

$B^0_s \rightarrow \mu^+\mu^-$

All reconstructed tracks

14 mm

Only wall reconstructed tracks with $p_t > 500$ MeV
High-energy physics: the art of finding needles in haystacks

- Contemporary high-energy physics focuses on rare processes
- The vast majority of the collisions is “boring”
- Interesting physics is ≥ 9 orders of magnitude rarer: one in a billion or more!
- Need to efficiently identify these interesting processes before reading out and storing the data

![Graph showing cross-sections of various particle processes at different energies](image)

- W & Z bosons: \(1 \text{ kHz}\) Discovered: 1983 @ SPPS
- t quark: 10 Hz Discovered: 1995 @ Tevatron
- Higgs boson 0.1 Hz Discovered: 2012 @ LHC

The haystack (hard collisions): 1 GHz
Finding the needle

This is what we’re looking for: Higgs boson decaying in four easily identifiable muons.

The LHC produces a few of these per day
Finding the needle

This is where it hides: tens of other hard collisions producing 1000s of particles

The LHC makes 40 million of these per second!
LHC experiments: ≥ 10 million sensors

- Accurately measuring momentum (p) and energy (E) of the thousands of particles produced by the collisions requires **fine-grained** particle detectors
- High granularity ⇒ big data: ~1 MB per bunch crossing
- Theoretical output: 40 MHz x 1 MB = 40 TB/s
40 TB/s, visualised

- 40 TB/s ⇒ **Forty 10 TB hard drives filled in ten seconds**
  - Saving all this data for later analysis is impossible
  - Even if we had infinite money, we don’t have infinite space (~5.4 m³/h)
- Real-time data selection is necessary
  - In HEP-speak: **triggering**
  - Acquire and save only interesting bunch crossings ("events")
Data selection

- Particle beams cross every 25 ns (40 MHz)
  - ~25 “hard” collisions per bunch crossing
  - Up to $10^9$ collisions per second

- Typical trigger architecture
  - Two or more consecutive steps
  - Step 1 (“Low-level Trigger”):
    - Custom hardware
    - 40 MHz → 100-1000 kHz
  - Step 2 (“High-level trigger”)
    - Software
    - 100-1000 kHz → 1-10 kHz
Step 1: Low-level Trigger
Low-level Trigger

● Several technological limitations make impractical to read out tens of millions of sensors at 40 MHz
● ATLAS & CMS would output ≥ 40 TB/s
   ○ This would require a lot of fast, low-power, low-cost links, but:
     ■ Radiation-hard optical links are slow (~5 Gb/s today, ~10 Gb/s tomorrow)
     ■ Fibres and cables “steal” space from the detector itself
● Remember: most of these millions of events per second are totally uninteresting: one Higgs event every 10 seconds
● The first data-selection step (Level-1 trigger) must somehow select the more interesting events without reading out all the sensors
Selection based on a subset of sensors

- Use data from fast sensors for Low-level decision
- Try to identify high-momentum and high-energy particles
- The remaining sensors have to buffer the data in their pipelines until a decision is made
- The Low-level trigger has hard latency constraints
- Typical latency: a few μs
- This requires custom electronics
- When the Low-level trigger accepts an event, it triggers the readout of the detector (hence the name)
Fast local algorithms

Calorimeters:
- Cluster finding
- Energy deposition evaluation
- Coarse-grained

Muon systems:
- Track finding
- Momentum evaluation
- Dedicated fast sensors
Custom electronics

- Sophisticated hardware (FPGAs, ASICs)
- Hundreds of custom-built boards process small pieces of the collision at enormous speeds
- They give a crude but effective decision, based on simple criteria
Summary

- The Level 1 Triggers are implemented in custom electronics:
  - difficult / expensive to upgrade or change
  - maintenance by experts only
- Decision time: ~ a small number of microseconds
  - The Level 1 Triggers are **hard real-time** systems
- They use “simple” hardware-friendly algorithms
  - Particle identification (high-energy particles & jets)
  - Local pattern recognition based on coarse-grained data
- Working with partial information and with drastic simplifications has a price: potentially interesting and valuable events are lost
- Future directions:
  - Eliminate / reduce hardware Level-1 (ALICE, LHCb)
  - Substantially upgrade Level-1 (ATLAS, CMS)
Step 2: 
"High-level" trigger (HLT)
HLT: Full “reconstruction” of the collisions

- Pack the knowledge of thousands of physicists and decades of research into a huge sophisticated algorithm
  - **Reconstruct all charged particle trajectories**
  - Find segments, connect them, re-fit to physical trajectory
  - Associate the particles with the correct p-p hard collision
  - Multiple interactions in each crossing
- Measure all of the energy depositions in the calorimeters with fine granularity
- Associate tracks and energy depositions
Data decoding

- What we get from the detector:
  - Sensor 1244 has measured a signal of 120 at time 1333096259.344245

- What we need to reconstruct what happened:
  - Signal at position \( x = 1.2 \text{ cm}, y = 4.5 \text{ cm}, z = 3.2 \text{ cm} \), deposited energy of 100 keV

- Requires precise information about location of the detector element (alignment) and of its signal sensitivity (calibration)
Track finding: pattern recognition

- Start with a collection of “hits” (footprints) on the various layers of the detector
Track finding: pattern recognition

- Reconstruct the trajectory of the particle that left the footprints
- Deviation on a curved trajectory in a normal plane around the magnetic field
- Radius of curvature allows to determine the particle’s momentum
Track finding: pattern recognition

- Can become very complicated
- Lots of tracks, rings, curved / spiral trajectories, spurious measurements and various other imperfections
- Problem scales like $N^L$
  - $N$ is the number of particles
  - $L$ is the number of layers
Vertex finding

Primary vertex: Reconstruct where the collision occurred

Secondary vertex: Detect particle decays

\[ \Xi^- \rightarrow \Lambda^0 + \pi^- \rightarrow p + \pi^- + \pi^- \]
Mass: 1.322 GeV/c²
\( p_T \): 1.459 GeV/c
Decay length: 6.85 cm

Run 104892, raw data chunk 09000104892020.130, event is chunk 0
Software on commodity hardware

- Commodity x86 servers
- ~2000 per experiment
- Huge existing codebases
- A few millions of lines of code
- Mostly single-threaded
- Exploit the inherent parallelism of the data
- Current figures:
  - Input: 1000-100 kHz
  - Output: 10-1 kHz
  - Processing time per event: 10-100 ms
Summary: reconstruction steps

- Data decoding
  - Associate geometrical information and detector calibration to the data
- Clusterisation
  - Particles frequently create signal in adjacent sensors
  - These signals are combined into “clusters”
- The clusters are combined into “tracks” (pattern recognition)
  - Clusters of hits associated to the same particle are grouped together
- Tracks are combined to reconstruct vertices
- Events are selected according to the reconstructed physics objects
  - e.g. 4 muons with a transverse momentum > 20 GeV
Data Acquisition & Event Building,
or: how do we get the data to the HLT?
DAQ architecture

- Detector
- Readout Units
- DAQ network
- Compute Units

Custom point-to-point radiation-hard link from the detector front-end

DAQ links commodity LAN (Ethernet, InfiniBand)

Links into compute units lower variant of DAQ links (HLT is CPU limited)

10000 x

~ 1000 x

~ 3000 x
Data flow

- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering (High Level Trigger – HLT)
Data flow

- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering (High Level Trigger – HLT)
Data flow

- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering (High Level Trigger – HLT)
Data flow

- Every Readout Unit has a piece of the collision data
- All pieces must be brought together into a single Compute unit
- The Compute Unit runs the software filtering (High Level Trigger – HLT)
Event Building in a nutshell

- RU (Readout Unit): it reads the data out from the detector
- BU (Builder Unit): it receives the full event data
DAQ networks

- The DAQ network aggregates data from different Readout Units ⇒ Many-to-one communication
- Data transfers are driven by the availability of the data from the detector ⇒ Synchronous, bursty traffic
- When many sources send synchronous microbursts of data to a destination ⇒ The network buffers are overflowed ⇒ Congestion / packet loss ⇒ Must be kept under control, otherwise:
  - “Catastrophic throughput collapse”
DAQ networks

- The DAQ network aggregates data from different Readout Units
  ⇒ Many-to-one communication
- Data transfers are driven by the availability of the data from the detector
  ⇒ Synchronous, bursty traffic
- When many sources send synchronous microbursts of data to a destination
  ⇒ The network buffers are overflown
  ⇒ Congestion / packet loss
  ⇒ Must be kept under control, otherwise:
- “Catastrophic throughput collapse”
Traffic shaping to the rescue: linear shift

- 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 \( n + 1 \)
Putting it all together
ATLAS Trigger & DAQ in Run2
CMS Trigger & DAQ in Run2

- **Level 1 Trigger**
- **Event Manager**
- **Detector Front-Ends**
  - **Builder Network**: 100 GB/s
- **Readout Systems**
- **Control and Monitor**
- **Filter Systems**
- **Computing Services**

- 40 MHz
- 100 kHz
- $10^2$ Hz
# Trigger & DAQ Run 2 datasheets

<table>
<thead>
<tr>
<th></th>
<th>Level 1 Rate [Hz]</th>
<th>Event size [B]</th>
<th>Readout bandwidth [GB/s]</th>
<th>HTL Out data rate [GB/s]</th>
<th>HTL Out event rate [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALICE</td>
<td>Pb-Pb 100 p-p 10^3</td>
<td>5x10^7, 2x10^6</td>
<td>25</td>
<td>1250, 200</td>
<td>10^2, 10^2</td>
</tr>
<tr>
<td>ATLAS</td>
<td>10^5</td>
<td>1.5x10^6</td>
<td>50</td>
<td>1000</td>
<td>10^2</td>
</tr>
<tr>
<td>CMS</td>
<td>10^5</td>
<td>10^6</td>
<td>100</td>
<td>1000</td>
<td>10^2</td>
</tr>
<tr>
<td>LHCb</td>
<td>10^6</td>
<td>5x10^4</td>
<td>50</td>
<td>700</td>
<td>1.2x10^4</td>
</tr>
</tbody>
</table>
The future of DAQ: Run3 and Beyond
The LHC long-term planning

<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></td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
</tr>
<tr>
<td>Run 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Run 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Long Shutdown 3 (LS3)</td>
<td></td>
<td></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></td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
</tr>
<tr>
<td>Run 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Run 5</td>
</tr>
<tr>
<td>LS4</td>
<td></td>
<td></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>2039</th>
<th>2040</th>
<th>2041</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>JFMAM</td>
<td>JFMAM</td>
<td>JFMAM</td>
</tr>
<tr>
<td>LS5</td>
<td></td>
<td></td>
<td>Run 6</td>
</tr>
</tbody>
</table>

Last update: April 2023
## Experiment upgrade plans

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Event-size [kB]</th>
<th>Rate of events into HLT [kHz]</th>
<th>HLT bandwidth [Gb/s]</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALICE</td>
<td>20000</td>
<td>50</td>
<td>8000</td>
<td>2022</td>
</tr>
<tr>
<td>ATLAS</td>
<td>4000</td>
<td>200</td>
<td>6400</td>
<td>2026</td>
</tr>
<tr>
<td>CMS</td>
<td>4000</td>
<td>1000</td>
<td>32000</td>
<td>2026</td>
</tr>
<tr>
<td>LHCb</td>
<td>100</td>
<td>40000</td>
<td>32000</td>
<td>2022</td>
</tr>
</tbody>
</table>
Why can LHCb read out all the collisions?

- Spectrometer geometry (fibres/cables are not "in the way")
- Relatively low radiation levels permit to relax the constraint on the FPGAs used for "middle" layer processing
- Smaller total event-size ~100 kB
LHCb DAQ system for Run3

- VeLo
- RICH
- UT
- SciFi
- Calo
- Muon

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

~2000 full-duplex control links (Versatile Link)

32 Tbps

32 Tbps

1 Tbps

1 Tbps

~160 servers: DAQ + event builder + event filter first pass (on GPGPUs)

~40 PB disk storage

Event filter second pass (~4000 servers)
• Data is not separated into events anymore
• Each sub-detector produces a continuous stream of data
• The data stream is chopped in time frames of a specific length
Streaming/continuous readout

- Each data stream is individually read out
- Partial time frames are assembled into full time frames
- Subsequent data processing correlates data within the frame
Summary

The LHC experiments need to reduce
● ~50 TB/s to ~1 GB/s (or rather, 25 PB/year)
● This is achieved with massive use of FPGAs, custom ASICs, x86 CPUs and GPGPUs
● Large commodity local area networks are used to distribute data among the individual computing elements
● The future will see massive increase of required programmable computing power and required networking bandwidth, much more data will be moved off detector
Thank you for your attention