ALICE trigger and computing upgrade: design, technology and performance

B. von Haller for the ALICE Collaboration

CERN

6.10.2016
Outline

- ALICE trigger & computing upgrade in a nutshell
- Triggers, Heart Beats and Timeframes
- Hardware and computing model
  - Nodes and network
  - Hardware acceleration
  - O2 Facility and Tiers
- Software
  - Paradigms
  - Development processes and tools
- Project organisation
- Schedule
ALICE Trigger/DAQ Upgrade in a nutshell

Requirements

1. After LS2, LHC min bias PbPb at 50 kHz
   ▶ ~100 x more data than during Run 1
   ➔ Too much data to be stored
2. Physics topics addressed by ALICE upgrade
   ▶ Rare processes
   ▶ Very small signal over background ratio
   ▶ Needs large statistics of reconstructed events
   ▶ 13nb\(^{-1}\) for PbPb
   ➔ Triggering techniques very inefficient if not impossible in most cases
3. TPC inherent rate (drift time ~100 µs) < 50 kHz
   ➔ Support for continuous read-out (TPC), as well as triggered read-out

New computing system

• Read-out the data of all interactions
• Compress these data intelligently by reconstructing them online
• One common online-offline computing system: O\(^2\)
   ➔ Paradigm shift compared to approach for Run 1 and 2
Asynchronous (few hours) event reconstruction with final calibration

Compressed Sub-Timeframes
Base Line correction, zero suppr. Readout
Data aggregation Local data processing

Continuous and triggered streams of raw data

3.4 TB/s

Hi:50kHz pp/pA:200kHz

First Level Processors (FLP)

Data aggregation Synchronous global data processing

Compressed Timeframes

500 GB/s

Hi:50kHz pp/pA:200kHz

Event Processing Nodes (EPN)

Data storage (60 PB)
1 year of compressed data
Write 170 GB/s, Read 270 GB/s

Compressed Timeframes

90 GB/s

Hi:50kHz pp/pA:200kHz

Tier 0, Tiers 1 and Analysis facilities

Asynchronous (few hours) event reconstruction with final calibration

Reconstructed events

20 GB/s
Continuous & triggered read-out

- GBT
- TTC
- TTC PON

O^2: Online and Offline Computing System
FLP: First level processor
DCS: Detector Control System
TTS: Trigger and Timing Distribution System
CTP: Central Trigger Processor
GBT: Gigabit Transceiver
FTL: Fast Trigger Links
TTC: Timing Trigger Control

Common Readout Unit (CRU)

FIT ZDC ACO TOF EMC PHS

GBT

→data & ←configuration

FIT

←trigger (ITS/MFT/TRD)

TTC

BTC

→data & ←configuration

TTC PON

O^2 & DCS

FIT ZDC ACO TOF EMC PHS

ACO CPV EMC HMP PHS

←trigger

DDL1 or 2

→data & ←configuration

CRORC

O^2 & DCS

FIT ZDC ACO TOF EMC PHS

ACO CPV EMC HMP PHS

←trigger

DDL1 or 2

→data & ←configuration
Trigger system

Run 1 and 2:
- TTC (Timing Trigger Control) used for trigger distribution (unidirectional)
- Busy/throttling transported electrically on a separate link

Run 3:
- TTC PON for detectors with continuous readout
  - Broadcast and continuous serial stream
  - LHC Bunch Clock synchronous
  - Send clock
  - Time Division Multiplexing
  - Shared bandwidth, lower line rate
  - LHC Bunch Clock synchronous
  - Send busy/throttling

PON: Passive Optical Network
Commercial system adapted for the CERN requirements
Triggers, Heart Beats, Timeframe

**Heart Beat (HB)**
Issued in continuous & triggered modes to all detectors

**Physics trigger**
Can be sent to upgraded detectors will be sent to non-upgraded detectors

---

**Read-out: Continuous and triggered**

**Heart Beat Frames (HBF):** data stream delimited by two HBs

**Sub-Time Frame (STF) in FLP 0:** grouping of (~256) consecutive HBFs from one FLP

**Time Frame (TF):** grouping of all STFs from all FLPs for the same time period from triggered or continuously read out detectors

---

**Typical values:**
- HB: 1 per orbit, 89.4 µs: ~10 kHz
- TF: 1 every ~20 ms: ~50 Hz
- \( \rightarrow 1 \text{ TF} = \sim 256 \text{ HBF} \)
System throttling

1. HB trigger/message
2. HBF/Trigger Transmitted?
3a. HBF
3b. HBF/trigger acknowledge/message
4. Delete HBF from buffer
   Discard incoming data
5. Delete HBF from memory

CTP

CTP FLP

CRU

FLP

Readout Links

HB Map
1 1 1 0 1
1 0 1 1 1

HB Map
1 1 1 0 1
1 1 1 1
O2 Hardware facility

- **Detectors**
  - 270 FLPs
  - First Level Processors (FLPs)
- **Switching Network**
  - Input: 270 ports
  - Output: 1500 ports
  - 500 GB/s
- **Event Processing Nodes (EPNs)**
  - 1500 EPNs
- **Storage Network**
  - Input: 1500 ports
  - Output: 34 ports
  - 90 GB/s
- **Storage Arrays**
  - 68 Storage Arrays
  - 60 PB
- **34 Storage Servers**
  - RD and WR 440 GB/s
- **Input: 9000 Read-out Links**
  - 3.4 TB/s
# Readout & FPGA Hardware acceleration

<table>
<thead>
<tr>
<th>RORC 1</th>
<th>C-RORC</th>
<th>CRU</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="RORC 1 Card" /></td>
<td><img src="image2" alt="C-RORC Card" /></td>
<td><img src="image3" alt="CRU Card" /></td>
</tr>
<tr>
<td>2 ch @ 2 Gb/s PCIe gen.1 x4 (1 GB/s)</td>
<td>12 ch @ up to 6 Gb/s PCIe gen.2 x 8 (4 GB/s)</td>
<td>24 ch @ 5 Gb/s PCIe gen.3 X 16 (16 GB/s)</td>
</tr>
<tr>
<td>Custom DDL protocol</td>
<td>Custom DDL protocol (same protocol but faster)</td>
<td>GBT</td>
</tr>
<tr>
<td>Protocol handling TPC Cluster Finder</td>
<td>Protocol handling TPC Cluster Finder</td>
<td>Protocol handling TPC Cluster Finder</td>
</tr>
<tr>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 3</td>
</tr>
<tr>
<td>LS1</td>
<td>LS 2</td>
<td>Run 3</td>
</tr>
<tr>
<td>Run 1</td>
<td>Run 2</td>
<td>Run 3</td>
</tr>
</tbody>
</table>
Hardware acceleration (FPGA)

Performance of the FPGA-based FastClusterFinder algorithm for DDL1 (Run1) and DDL2 (Run2) compared to the software implementation on a recent server PC.
Hardware acceleration (GPU)

- For TPC track finding on EPNs (as today’s HLT)
- Possibly more use cases depending on R&D

Tracking time of HLT TPC Cellular Automata tracker on Nehalem CPU (6Cores) and NVIDIA Fermi GPU.
Network performance tests

Comparison of Ethernet, IP over InfiniBand and IP over Omni-Path

40 GbE: 40 Gigabit Ethernet
OPA: Intel® Omni-Path
IB: InfiniBand
O2: ALICE Online-Offline framework
Storage

Client File Systems performance tests

ADM: Administration server
MD: Metadata server
DS: Data server
CLI: Client
GE: Gigabit Ethernet
IB: InfiniBand
GPFS: General Parallel File System
Lustre: open-source parallel file system
Ceph: distributed object store
RADOS: Reliable Autonomic Distributed Object Store
O² Farm

- ~100k CPUs, ~5k GPUs, ~500 FPGAs
- FLPs at P2 in existing CR1
- EPNs and storage need a new dedicated room
  - Space and weight limitations
- Two scenarios
  - CR0
    - Container(s)
    - Call for tender (common with LHCb and neutrino platform)
  - Common Data Center in Prevessin
    - An alternative to the CR0 at P2 has been proposed by CERN
    - New common data center in Prevessin
    - Solution based on the GSI Green Cube
    - Being studied
Computing Model

Glossary
- RAW: raw data
- CTF: Compressed Time Frame
- ESD: Event Summary Data
- AOD: Analysis Data Object
- O2: Online-Offline facility
- T0, T1, T2: Grid Tier 0, 1, 2
- AF: Analysis Facility
- MC: Monte-Carlo
- HISTO: Subset of AOD specific for a given analysis
Software Design

- **Message-based multi-processing**
  - Ease of development
  - Ease to scale horizontally
  - Possibility to extend with different hardware
  - Multi-threading possible within processes
- **ALFA : ALICE-FAIR concurrency framework**
  - Data transport layer
  - ZeroMQ
  - Multi-process
  - Steady development
- **AliceO2**
  - Prototyping
  - Development started
Software

Development processes

▶ Especially important
  ▶ Most software will run online
  ▶ Large code base
  ▶ Many people involved, from different backgrounds
    ▶ That is also a blessing as we have a large expertise to pick from
O^2 Organization

➤ Today
  ➤ 3 distinct projects: DAQ, HLT, Offline
  ➤ 2 different code base: HLT+Offline and DAQ
  ➤ 2 different reconstructions: HLT and Offline

➤ ALICE O^2
  ➤ 1 project
  ➤ 1 code base
  ➤ 1 reconstruction
O² Organization

- Steering board
  - Project leaders of the DAQ, HLT and Offline
- 13 Computing working groups (CWGs)
  - Made of people from the DAQ, HLT and Offline projects
  - Plus participants from many detectors
- Plenaries
- Successful transition
  - Cross-fertilization of the different cultures
  - There is a strong and growing sense of belonging
O² Schedule

2012-2013

High-Level Design R&D
Trigger TDR
Project organization

2014-2015

Design R&D
Demonstrators
O² TDR
Products selection

2016-2017

Detailed design R&D
Prototyping
Development
Products selection
Prototypes

2018-2019

Detailed design R&D
Prototyping
Development
Products selection
Prototypes
Final components
Deployment
Commissioning

2020-2021

Development
Products selection
Final components
Deployment
Commissioning
Production

Run 1
LS1
Run 2
LS 2
Run 3
ALICE Trigger and computing upgrade is an exciting project with very ambitious requirements
  ▶ 13nb⁻¹ for PbPb
  ▶ >3TB/s continuous and triggered detector input
▶ New trigger and throttling strategy to handle the continuous readout
▶ Major paradigm change with combined offline and online system
▶ Hardware system
  ▶ HW acceleration (FPGAs, GPUs)
  ▶ O² farm with ~100 k CPU cores and ~5000 GPUs
▶ Software development
  ▶ Collaboration with FAIR team in Germany and JPARC experiments in Japan allows a steady progress for the framework
▶ People from different groups and backgrounds work together
  ▶ We all improved by doing so!
▶ On track despite a challenging schedule
O² architecture (1)

- **Synchronous**
  - **Raw data input**
  - **Local processing**
  - **Frame dispatch**
  - **Global processing**
  - **Storage**

**Detector data samples interleaved with synchronized heartbeat triggers**

**FLPs**
- Buffering
- Local aggregation
- Time slicing
- Data Reduction 0
  - e.g. clustering
  - Tagging
- Calibration 0
  - on local data, i.e., partial detector

**EPNs**
- Timeframe building
- Full timeframe
- Detector reconstruction
  - e.g. track finding
- Data Reduction 1
- Calibration 1
  - on full detectors
  - e.g. space charge distortion
- QC

**O²/T0/T1**
- Storage
- Archive
- CTF
- AOD

**Trigger and clock**
O² architecture (2)
O² architecture (3)

Storage
- Reconstruction passes and event extraction
- Asynchronous

Simulation

Analysis
- Analysis Facilities
- AOD
- Analysis
- Storage
  - Histograms, trees

Event extraction
Tagging
Global reconstruction
AOD extraction

Compressed timeframes

Archive
- CTF
- AOD

Event Summary Data
Analysis Object Data

Reconstruction
AOD extraction

ESD, AOD

O²/T0/T1

O²/T0/T1

Global reconstruction
Event extraction
Tagging

CTF

O(1)

Calibration 2

Voltage

QC

O(10)

Reconstruction
Event building
AOD extraction

CTF

Simulation

T2

Analysis

O(1)

Analysis Object Data

ESD, AOD

Compressed timeframes

B. von Haller | O2 Project | 19.05.2015
Physics software design
Processing workflow

All FLPS
Raw data
Local Processing
E.g. Clusterization Calibration

EPN: synchronous
Detector Reconstruction
E.g. TPC & ITS Track finding
Inter-detector matching procedures
Final calibration, 2nd matching

asynchronous
Final matching, PID, Event extraction

Step 0
Step 1
Step 2
Step 3
Step 4

CTF
AOD
Computing requirements for processing

Computing requirements -> Total: ~100,000 CPU cores 5000 GPU chips

<table>
<thead>
<tr>
<th>Detector</th>
<th>Process</th>
<th>Processing requirement</th>
<th>Processing Platform</th>
<th>System reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPC</td>
<td>Calibration</td>
<td>1000</td>
<td>CPU</td>
<td>Intel I7-4600U 2.70 GHz</td>
</tr>
<tr>
<td>TPC</td>
<td>Track seeding, following</td>
<td>5000</td>
<td>GPU</td>
<td>AMD S9000</td>
</tr>
<tr>
<td>TPC</td>
<td>Track merging, fitting</td>
<td>15000</td>
<td>CPU</td>
<td>Intel I7-980X 3.60 GHz</td>
</tr>
<tr>
<td>ITS</td>
<td>Tracking</td>
<td>75000</td>
<td>CPU</td>
<td>Intel I7-2720QM 2.20 GHz</td>
</tr>
<tr>
<td>MCH</td>
<td>Preclustering</td>
<td>200</td>
<td>CPU</td>
<td>Intel I7 2.30 GHz</td>
</tr>
<tr>
<td>MCH</td>
<td>Clustering</td>
<td>5000</td>
<td>CPU</td>
<td>Intel I7 2.20 GHz</td>
</tr>
</tbody>
</table>

Goes together, merging and fitting can run on GPUs too.

Theoretically could run on GPU.

Being ported to GPU, conversion factor unknown.
### Physics programme and data taking scenarios

**Chapter 2**

<table>
<thead>
<tr>
<th>Year</th>
<th>System</th>
<th>$\sqrt{s_{NN}}$ (TeV)</th>
<th>$L_{int}$ (nb$^{-1}$)</th>
<th>$N_{collisions}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>pp</td>
<td>14</td>
<td>0.4</td>
<td>$2.7 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>Pb-Pb</td>
<td>5.5</td>
<td>2.85</td>
<td>$2.3 \cdot 10^{10}$</td>
</tr>
<tr>
<td>2021</td>
<td>pp</td>
<td>14</td>
<td>0.4</td>
<td>$2.7 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>Pb-Pb</td>
<td>5.5</td>
<td>2.85</td>
<td>$2.3 \cdot 10^{10}$</td>
</tr>
<tr>
<td>2022</td>
<td>pp</td>
<td>14</td>
<td>0.4</td>
<td>$2.7 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>pp</td>
<td>5.5</td>
<td>6</td>
<td>$4 \cdot 10^{11}$</td>
</tr>
<tr>
<td>2025</td>
<td>pp</td>
<td>14</td>
<td>0.4</td>
<td>$2.7 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>Pb-Pb</td>
<td>5.5</td>
<td>2.85</td>
<td>$2.3 \cdot 10^{10}$</td>
</tr>
<tr>
<td>2026</td>
<td>pp</td>
<td>14</td>
<td>0.4</td>
<td>$2.7 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>Pb-Pb</td>
<td>5.5</td>
<td>1.4</td>
<td>$1.1 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>p-Pb</td>
<td>8.8</td>
<td>50</td>
<td>$10^{11}$</td>
</tr>
<tr>
<td>2027</td>
<td>pp</td>
<td>14</td>
<td>0.4</td>
<td>$2.7 \cdot 10^{10}$</td>
</tr>
<tr>
<td></td>
<td>Pb-Pb</td>
<td>5.5</td>
<td>2.85</td>
<td>$2.3 \cdot 10^{10}$</td>
</tr>
</tbody>
</table>
### Data types characteristics

- **TF size - Duration of the time window ($t_{TF}$)**
  - Data lost at the edges: $0.1/t_{TF}(ms)$
  - For calibration and reconstruction: 20ms - 100ms
  - Shorter is better for buffering and distribution
  → 20ms (1000 interactions in Pb-Pb at 50kHz)

<table>
<thead>
<tr>
<th>Data type</th>
<th>Size (GB)</th>
<th>Tape copy</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF (Pb-Pb)</td>
<td>10</td>
<td>No</td>
</tr>
<tr>
<td>CTF (Pb-Pb)</td>
<td>1.6</td>
<td>Yes</td>
</tr>
<tr>
<td>ESD</td>
<td>15% of CTF</td>
<td>No</td>
</tr>
<tr>
<td>AOD</td>
<td>10% of CTF</td>
<td>Yes</td>
</tr>
<tr>
<td>MC</td>
<td>100% of CTF</td>
<td>No</td>
</tr>
<tr>
<td>MCAOD</td>
<td>30% of ESD</td>
<td>Yes</td>
</tr>
<tr>
<td>HISTO</td>
<td>1% of ESD</td>
<td>No</td>
</tr>
</tbody>
</table>
Network layout 2: 4 independent EPN subfarms
O2 facility design (3)

Network layout 3: Super-EPNs

1 FLP

40/56 Gb/s

1 FLP

10 X 40/56 Gb/s

25 FLP

10

50

SEPN

2 X 40/56 Gb/s

1 SEPN

50

1 EPN

10 Gb/s

30 EPN

1471 EPN

1500 EPN

250 FLP

226 FLP
O2 facility design (4)

Simulation – Link speed

Left: Network Layout 2:
Link speed on the FLPs and EPNs for a network layout with 4 EPN subfarms for 100 parallel transfers from the FLPs.

Right: Network Layout 3:
Link speed on the FLPs and Super-EPNs (configuration based on an Infiniband network at 56 Gb/s)
O2 facility design (6)

Simulation - system scalability

Latency of the timeframes for different interaction rates using layout 2 (left) and layout 3 (right)

→ Layout 2 is cheaper but scales up to 90kHz only.
## O2 facility – Power and cooling

Table 10.5: Location, rack space, power and cooling needs of the O² facility.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of items</th>
<th>Item height (U)</th>
<th>Location-Total height (U)</th>
<th>Number of Racks</th>
<th>Power per rack (kVA)</th>
<th>Total power (kVA)</th>
<th>Cooling per rack (kW)</th>
<th>Total cooling (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FLP</td>
<td>250</td>
<td>2</td>
<td>CR1 - 35</td>
<td>18</td>
<td>12</td>
<td>216</td>
<td>14</td>
<td>252</td>
</tr>
<tr>
<td>EPN</td>
<td>1500</td>
<td>1</td>
<td>CR0 - 54</td>
<td>34</td>
<td>50</td>
<td>1700</td>
<td>50</td>
<td>1700</td>
</tr>
<tr>
<td>SEPN</td>
<td>50</td>
<td>1</td>
<td>CR0 - 54</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Storage</td>
<td>34</td>
<td>9</td>
<td>CR0 - 54</td>
<td>7</td>
<td>50</td>
<td>350</td>
<td>50</td>
<td>350</td>
</tr>
<tr>
<td>Network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dataflow</td>
<td>10</td>
<td>3</td>
<td>CR1 - 40</td>
<td>2</td>
<td>12</td>
<td>24</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>Dataflow</td>
<td>15</td>
<td>3</td>
<td>CR0 - 54</td>
<td>2</td>
<td>12</td>
<td>24</td>
<td>12</td>
<td>24</td>
</tr>
<tr>
<td>Control</td>
<td>4</td>
<td>1</td>
<td>CR1 - 40</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Control</td>
<td>14</td>
<td>1</td>
<td>CR0 - 54</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Services</td>
<td>110</td>
<td>1</td>
<td>CR1 - 40</td>
<td>4</td>
<td>12</td>
<td>48</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>CR1</td>
<td>25</td>
<td></td>
<td>300</td>
<td></td>
<td>336</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CR0</td>
<td>45</td>
<td></td>
<td>2086</td>
<td></td>
<td>2086</td>
</tr>
<tr>
<td>Grand total</td>
<td></td>
<td></td>
<td></td>
<td>70</td>
<td></td>
<td>2398</td>
<td></td>
<td>2434</td>
</tr>
</tbody>
</table>
Network performance tests

Comparison of Ethernet, IP over InfiniBand and IP over Omni-Path

- 40 GbE (O2)
- 40 GbE (max)
- OPA (O2)
- OPA (max)
- IB (O2)
- IB (max)

Saturating the Sender with Ethernet

40 GbE: 40 Gigabit Ethernet
OPA: Intel® Omni-Path
IB: InfiniBand
O2: ALICE Online-Offline framework