



ALICE

O² Report
Computing Working Group 1
Architecture

O² Data storage: needs and work plan

Version 1.0

S. Chapeland ^a, U. Fuchs ^a, P. Vande Vyvre ^a
^a *CERN*

May 31st, 2016

Abstract

This note summarizes the data storage needs of the future O² system as a preparation for the performance tests of data storage equipment and software. It also includes a work plan for the activities to be performed from now till the deployment of the data storage system.

This note is meant to be circulated to people who are not familiar with the O² project such as potential commercial providers of data storage systems.

Changelog

1.0	2016-05-31	Version circulated out of the project
0.2	2016-05-30	Remove capacity from table 3 and move capacity to 3.2.
0.1	2016-05-25	Initial draft version

Contents

Changelog	2
Contents	2
1 Introduction	3
2 O² system	4
3 Data storage needs	5
3.1 Data types	5
3.2 Data storage access pattern, capacity and throughput needs	6
4 Work plan	7
5 References	8

1 Introduction

ALICE (A Large Ion Collider Experiment) [1] is taking data since 2009. The processing of the ALICE data is performed on the Worldwide LHC Computing Grid) WLCG. The Tier 0 is located at CERN and there are several Tiers 1 and 2 in the countries and institutes member of the ALLICE collaboration.

In 2020, the ALICE experiment at CERN will start collecting data with an upgraded detector and a new Online and Offline computing system called O². The upgrade is based on the Large Hadron Collider (LHC) running conditions which will deliver an interaction rate of 50 kHz with lead beams (Pb-Pb). ALICE will also take data with collisions of proton beams (pp) and p-Pb but the Pb-Pb interactions are those producing the largest amount of data.

The main physics topics addressed by the ALICE upgrade are precise measurements of physics probes characterised by a very small signal-to-background ratio requiring very large statistics. This large background makes the traditional reduction of data volume by the selection of some interactions (the *triggering*) very inefficient, if not impossible.

The Time Projection Chamber (TPC) is the detector producing most of the data volume and it is also the slowest with a drift time of $\approx 100 \mu\text{s}$. In order to keep up with the 50 kHz interaction rate, the TPC will require the implementation of a continuous read-out process to deal with event pile-up and avoid trigger-generated dead time. The resulting data throughput from the detector has been estimated to be greater than 3 TB/s for Pb-Pb events, roughly two orders of magnitude more than during the initial period of data taking.

To minimise the cost and requirements of the computing system for data processing and storage, the O² system is designed for a maximal reduction of the data volume read out from the detector as early as possible during the data-flow. The data of all collisions will be shipped to the O² facility at the anticipated interaction rate of 50 kHz for Pb-Pb collisions or 200 kHz for pp and p-Pb collisions.

The data volume reduction will be achieved in several steps synchronously with data taking. The detector raw data will be *reconstructed* and be replaced by the main characteristics of the particles that passed through such as their momenta, directions, and the primary vertex of the event. Data produced during this stage will be stored temporarily in the O² system.

Taking advantage of the duty factor of the accelerator and the experiment, the second reconstruction stage will be performed asynchronously, using the final calibration in order to reach the required data quality. The O² facility will be sufficient to perform both the synchronous and asynchronous reconstruction during pp data taking. The asynchronous reconstruction will process data archived at the O² facility or parked at the Tier 0 in the CERN Computing Center. However, during Pb-Pb data taking, part of the asynchronous data processing load will be shed to external sites such as Grid Tier 0 and Tier 1s that can provide archiving capability.

2 O² system

The overall dataflow of the O² system is shown in Figure 1. Data produced by the detectors is transferred by custom read-out links to the First-Level Processors (FLPs) using C-RORCs or CRUs link adapters. The CRUs will perform some data compression. Several streams may be aggregated on each FLP and buffered in memory. These nodes achieve a first data reduction and compression by performing the local data processing tasks on data fragments, without needing to have the full detector data set (e.g. TPC cluster finding).

The continuous streams of data samples are split into data frames, using as a reference clock arbitrary heart beat triggers. The frames are accumulated during a time period of the order of 20 ms, to minimise incomplete data at their boundaries for the collisions producing tracks spanning across the frame boundaries. All FLPs produce a Sub-Time Frame (STF) for each time period.

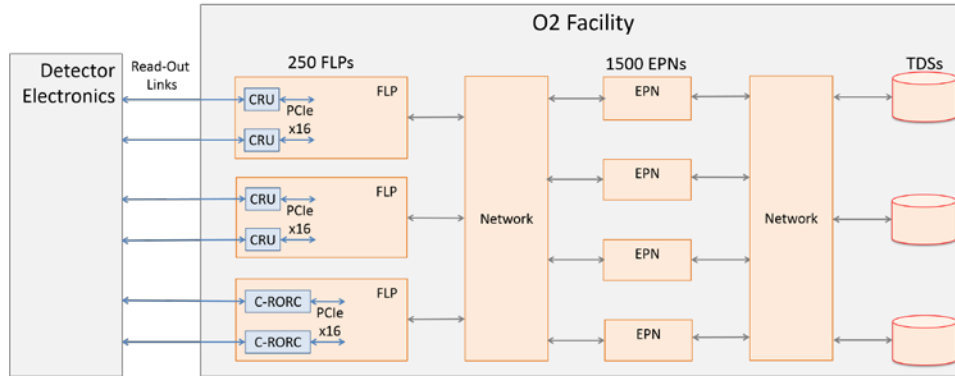


Figure 1 General dataflow of the O² system.

The STFs are then dispatched to the Event Processing Nodes (EPNs) for aggregation. The STFs related to the same time period and from all FLPs are received by the same EPN and aggregated into a Time Frame (TF).

EPNs perform the reconstruction for each detector, further reducing the data volume by an average factor of 8. The fully Compressed TFs (CTFs) are then stored to the Temporary Data Storage (TDS) in the O² facility and a fraction of one third is exported to the Tier 0 / Tier 1 data centres. The CTFs are archived to tape storage in the Tier 0 of the WLCG.

The O² system will access the WLCG for several purposes:

- Export some processing to the Grid during peak periods;
- Perform data processing jobs from the Grid when there is some spare capacity;
- Archive data to tape.

The O² system processing flow is shown in Figure 2.

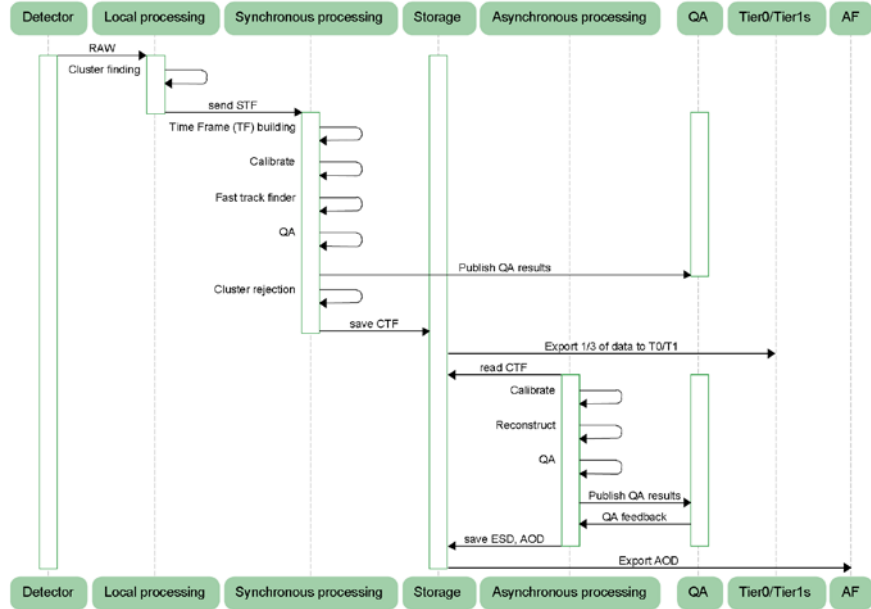


Figure 2 O² processing flow.

3 Data storage needs

3.1 Data types

The data types used by the O² system are shown in Table 1. Three different levels of persistency are defined:

- Transient for data which are used in the servers memories but not recorded to the TDS;
- Temporary for the data stored to the TDS in the O² system until a new version is produced but which are not archived;
- Persistent for the data stored to the TDS in the O² system and later archived to tape in the Grid Tier 0.

Table 1 O² data types.

Acronym	Description	Persistency
RAW	Raw data as it comes from the detector	Transient
STF	Sub-Time Frame containing raw data from a single FLP for a period of time ≈ 20 ms.	Transient
CTF	Compressed Time Frame containing processed raw data of all active detectors for a period of time ≈ 20 ms. Once written, CTF becomes read only data.	Persistent
ESD	Event Summary Data. Auxiliary data to CTF containing the output of the reconstruction process linking the tracks and clusters.	Temporary
MC	Monte Carlo. Simulated energy deposits in sensitive detectors. Removed once the reconstruction of MC data	Transient

	is completed.	
AOD	Analysis Object Data containing the final track parameters in a given vertex and for a given physics event. AODs are collected on dedicated facilities for subsequent analysis.	Persistent
MCAOD	Analysis Object Data for a given simulated physics event. Same as AOD with addition of kinematic information that allows comparison to MC. MCAODs are collected on dedicated facilities for subsequent analysis.	Persistent
HISTO	The subset of AOD information specific for a given analysis. Can be generated during analysis but needs to be offloaded from the Grid.	Temporary

3.2 Data storage access pattern, capacity and throughput needs

The data storage system should provide a flat access from all the EPNs or at least within large groups of EPNs so that the failure of a few nodes would not affect the data accessibility. The access is always done by accessing entire files in one sequence: open/read/close or open/write/close.

The number of physics interaction accumulated during one year, the size of the different data types and the number of versions required simultaneously are shown in Table 2.

This is used to evaluate the total data storage needs in the O² system for a typical year with pp and Pb-Pb data taking and production of ESDs and AODs by the O² system. In addition to the total needs, some reserve is foreseen for the data redundancy (e.g. RAID) and for some spare capacity. All these needs result in a total need of 70 PB of data storage capacity.

Table 2 Relative size of different data types.

Data type	Number of events	Event size (kB)	Number of versions	Total size O2 (PB)	Tape Copy
CTF (pp)	2.7 E+10	50	1	1.4	Yes
CTF (p-Pb)	1.0 E+11	100	1		Yes
CTF (Pb-Pb)	2.3 E+10	1600	1	36.8	Yes
ESD		15% of CTF	1	5.7	No
AOD		10% of CTF	2	7.6	Yes
MC		100% of CTF	1		No
MCAOD		30% of CTF	2		Yes
HISTO		1% of CTF	1	0.4	No
Total				51.9	
Redundancy (20%)				10.4	
Spare capacity (12.5%)				7.8	
Grand Total				70.0	

The data storage capacity and throughput for each data type is summarized in Table 3. All these needs result in a total throughput need of ~450 GB/s (270 for read and 170 for write).

Table 3 Average data throughput for each data type.

	Read rate (file/s)	Write rate (file/s)	Read throughput (GB/s)	Write throughput (GB/s)
CTF	100	50	200	100
ESD	100	100	30	30
AOD	200	200	40	40
Total	400	350	270	170

4 Work plan

The future O² TDS has to be partially deployed beginning of 2019 for the O² system commissioning and for supporting the detector commissioning. Many steps are needed in order to build a realistic Proof of Concept, to characterize it, to dimension the full system and to perform all the administration associated to such a large purchase.

The work plan to achieve these goals includes the tasks indicated in Table 4.

Table 4 Tasks and schedule.

1	Q2 - Q3 2016	Definition of the requirements summarized in this note
2	Q3 - Q4 2016	Identify possible storage solutions: - Compatible with the foreseen infrastructure; - Including the network, the number of servers and clients, the operating system, the topology, the access patterns and the scalability.
3	Q3 - Q4 2016	Identify a “basic building block”. Build a minimal setup representing a scale-down model of the whole system which would constitute a proof of concept.
4	Q4 2016 - Q2 2017	Verify the system performance - Define typical realistic data traffic and access patterns; - Assemble a benchmark suite; - Characterize the performance of the “building block”.
5	Q1 - Q4 2017	Combine multiple “basic building blocks” and make the evidence of the scalability potential.
6	Q1 - Q4 2017	Check with users of existing, larger installations using the selected solution(s).
7	Q4 2017 - Q4 2018	Market survey of potential providers of the selected solution(s) and invitation to tender for the procurement
8	Q1 - Q2 2019	Procure and deploy a fraction of the final system (e.g. 10%)
9	Q3 2019	Perform a data challenge with the deployed fraction before the procurement of the overall system
10	Q1 - Q4 2020	Detector commissioning and data taking with cosmic rays

11	Q1 - Q2 2020	Procure and deploy the remaining part of the final system (e.g. 90%)
12	Q2 2021	Data taking with pp beam

5 References

- [1] ALICE Collaboration, “The ALICE experiment at the CERN LHC”, 2008 JINST 3 S08002, 2008.
- [2] ALICE Collaboration, “Technical Design Report for the Upgrade of the Online–Offline Computing System”, CERN-LHCC-2015-006, 2015.