The ALICE online data storage system

R Divià¹, U Fuchs¹, I Makhlyueva¹³, P Vande Vyvre¹, V Altini², F Carena¹, W Carena¹, S Chapeland¹, V Chibante Barroso¹, F Costa¹, F Roukoutakis¹⁴, K Schossmaier¹, C Soòs¹ and B von Haller¹ (for the ALICE collaboration)

¹ CERN, Physics Department, Geneva, Switzerland.
² INFN, Dipartimento di Fisica dell’Università and Sezione INFN Bary, Italy.

E-mail: Roberto.Divia@cern.ch

Abstract. The ALICE (A Large Ion Collider Experiment) Data Acquisition (DAQ) system has the unprecedented requirement to ensure a very high volume, sustained data stream between the ALICE Detector and the Permanent Data Storage (PDS) system which is used as main data repository for Event processing and Offline Computing. The key component to accomplish this task is the Transient Data Storage System (TDS), a set of data storage elements with its associated hardware and software components, which supports raw data collection, its conversion into a format suitable for subsequent high-level analysis, the storage of the result using highly parallelized architectures, its access via a cluster file system capable of creating high-speed partitions via its affinity feature, and its transfer to the final destination via dedicated data links.

We describe the methods and the components used to validate, test, implement, operate, and monitor the ALICE Online Data Storage system and the way it has been used in the early days of commissioning and operation for the ALICE Detector. We will also introduce the future developments needed from next year, when the ALICE Data Acquisition System will shift its requirements from those associated to the test and commissioning phase to those imposed by long-duration data taking periods alternated by shorter validation and maintenance tasks which will be needed to adequately operate the ALICE Experiment.

1. Introduction

The ALICE experiment at CERN has the unprecedented requirement to be able to store 1.25 GB/s sustained of objectified data on Permanent Data Storage (PDS). This data comes from several detectors being readout via the ALICE Data Acquisition (DAQ) system and under the supervision of the ALICE Experiment Control System (ECS), two packages which have been specifically designed and developed to fulfil the requirements of the ALICE Online project. The PDS domain is, on the other hand, more Offline-oriented and its main goal is to provide CERN-wide and GRID-wide services such as the ALICE Environment for the GRID (AliEn) [1] and the CERN Advanced STORage manager (CASTOR)[2]. To inter-connect Online and Offline domains and to ensure a

³ Now at ITEP Institute for Theoretical and Experimental Physics, Moscow, Russia
⁴ Now at University of Athens, Physics Department, Athens, Greece
reliable and smooth data recording procedure, a dedicated architecture (hardware and software) has been designed, tested and operated: the ALICE online data storage system, based on the Transient Data Storage (TDS). We will herein review this architecture, the history of its development and the first experiences coming from the detectors commissioning and from the ALICE first operation, to end up with a peek at the planning for the near future.

2. Design and validation of the architecture
The transfer of the data from the experiment to the PDS has been one major area of concern for the ALICE collaboration. Events must be formatted, stored, and transferred at very high rates, with a target throughput during heavy-ion runs of 1.25 GB/s sustained over several weeks. During proton runs, the requirements are less challenging in terms of sustained throughput but, due to the much longer duration of these run periods, the amount of accumulated proton-proton data is equivalent to its heavy-ion counterpart. Another issue to be addressed is the liaison between the Online and the Offline domains, where requirements and priorities do not necessarily match. Particular attention has therefore been given to guarantee an appropriate throughput, reliability and flexibility for all of the above interconnections.

Figure 1. ALICE DAQ architecture.

Figure 1 shows the architecture of the ALICE DAQ and its links to other components such as the Central Trigger Processor (CTP), the High Level Trigger (HLT) farm, the Permanent Data Storage (PDS) and the ALICE environment for the GRID (AliEn). What we want to highlight in the above picture is the path followed by the data and the different domains being crossed: the detectors’ domain
(up to the Front-End Readout - FERO), the DAQ domain (hosting the Local Data Concentrators or LDCs, the Global Data Collectors or GDCs, the TDS, the DAQ network and the Storage network), the PDS domain and the AliEn domain. Each of the above domains has its own set of requirements and is under the direct responsibility of a separated team.

Key requirements of the ALICE Online data storage architecture are performance, reliability and support. Performance requirements are rather obvious: experimental data must be moved at the target sustained and peak rates, from the ALICE detectors to the PDS system, over a distance of several kilometres and using whenever possible standard and existing infrastructure. Reliability is needed to maximize the availability of the experiment whenever the LHC is operational: as much data as possible must be stored during the data taking periods without stalling the experiment's detectors or without modifying the triggering profile. Finally, support and operation must be ensured during the whole lifetime of the experiment, in agreement with the evolution of the components (hardware and software), technologies and packages.

At the time of the ALICE technical proposal [3] it was not at all obvious how to accomplish the given requirements using any available and cost-effective solution. However, it was clear that standard technologies were moving fast and in the right direction. The decision was taken to keep evaluating the candidate components, in the specs and on the field, and to follow their evolution as time went by. Network interfaces, mass storage interfaces, mass storage systems (hardware and software) operating systems and their kernels, Cluster Files Systems (CFS): all these items have been continuously evaluated, profiled, and tested during the past years in preparation for the ALICE experiment operation.

If possible, these evaluations peaked in the design, installation and operation of complete, autonomous setups whose components (hardware and software) were already in place for other purposes, e.g. existing data acquisition systems (during downtime of their associated experiment) or Mass Storage Systems (MSSs) from the central CERN services (during shutdown periods of the accelerator facilities). These activities evolved into the so-called "ALICE Data Challenges" (ADCs) [4][5][6], periodic exercises that have been scheduled at regular basis since January 1999 and that provided invaluable information on how to build the ALICE Storage System. The latest ADC, held in late 2006, confirmed once more the feasibility of the chosen architecture while moving a total of 1.8 PB from 18*10^6 files between the ALICE experimental area and the PDS.

The main outcome from evaluations and ADCs is a solid proof of feasibility for the use of commercial solutions to implement an efficient and reliable PDS system, a key concept that has been fully accepted by the ALICE collaboration. From the storage at the experimental area to the transfer of the same onto PDS, no dedicated hardware or firmware is needed. User level packages have been developed to synchronize the recording and migration procedures with the ALICE DAQ and ECS, to coordinate the data streams and to make the best use of the available components.

A key observation made during the ADCs is the infeasibility of moving data into PDS directly within the acquisition process. The latencies introduced by the protocol established between the experimental area and the central CERN PDS would have surely implied blocking the detectors during transient slowdowns of the recording procedure (e.g. when closing/opening a data file) and this cannot be taken lightheartedly. Furthermore, the chosen PDS system (the CERN Advanced STORage manager, or CASTOR) has not been designed to ensure the required features at the level of the individual data stream and therefore cannot ensure the required smoothness of the data recording streams. We eventually took the decision to insert between the DAQ and PDS an intermediate data buffer whose peak and sustained capabilities match the ALICE requirements and whose size is sufficient to store an entire LHC spill of data. This data buffer, known as Transient Data Storage
(TDS), proved to ensure the needed features and to satisfy the base requirements for a reasonable cost. The main role of the TDS is to support two cooperating activities: recording of ALICE experimental data while previously collected data is migrated to the PDS, to be accomplished without significant delays or chronic slowdowns while giving a significant grace period in case of catastrophic failures of the liaison between DAQ and PDS or of the PDS itself.

The last important outcome from the ADCs is the scalability of this type of architecture. We have proven the feasibility of improving linearly the throughput of the data stream (up to the planned design features) simply by adding more hardware. This allows a staged deployment of the computing infrastructure on both ends of the link (experimental area and computer center), to be scheduled between the beginning of the ALICE commissioning phase and the subsequent years of LHC operation. Therefore we were not obliged to install all the TDS and PDS at once but we could rather provide the minimum resources to fulfill the specific requirements for any given running period.

3. Data flow and components

From the moment of the readout from the Front-End Electronics (FEEs) to the final recording stage, data coming from ALICE detectors is moved and stored over several media. The FEE modules, located on or immediately close to the detectors, gather the raw data and ship it via dedicated optical links [7] to the counting room located in the ALICE access shaft [3] where the payload is formatted, the event is built, the stream is objectified and the data is recorded on a local disk. All this must be done on the fly and latencies have to be avoided at all costs to avoid stalling the data flow (an operation that is fully supported by the architecture but that would affect the efficiency of the detector and/or upset the trigger profile). This task is accomplished by the Transient Data Storage Manager (TDSM) package.

On the counting room side of the optical links, data first is readout and stored in memory banks by a set of LDCs. As soon as all the input links local to one LDC have provided their contribution, the sub-event is sent from the LDC to the GDC where the event is built, the data is objectified in ROOT [8] format and stored on hard disk. This is the first time that, after having been transferred from memory to memory and from machine to machine via dedicated optical link, shared memory and Gigabit Ethernet over copper, data is recorded on a permanent magnetic media. Each GDC is assigned a dedicated write volume that belongs to the Transient Data Storage (TDS) and accessed via FibreChannel through a Qlogic QLA2460 interface via Qlogic SANbox 5602 16-port FibreChannel switches. The write volume is driven by an Infortrend A16F (models G2422 and G2430) disk array handled via StorNext 3.1.2 Cluster File System, organized in an architecture particularly optimized for non-simultaneous, fast single-stream writing and multi-stream reading. The disk volume used to store the experimental data is shared neither in write nor in read mode, to avoid any interference with the data taking process due to transient or chronic overload or thrashing of the MSS hardware and software. The GDC keeps writing in its assigned volume until (1) the occupancy of the volume exceeds a given threshold, (2) a request to switch of write volume is raised by the operator (e.g. to request an early analysis of a particular data set), or (3) the run period completes. When any of the above events occurs, the volume changes its status from writing mode into reading mode and becomes candidate for the migration of its content to the Computer Centre.

All transfers between the ALICE experimental area and the Computer Centre are handled by a set of machines named TDSM Movers. These are dedicated hosts equipped with links to the TDS (same hardware as for the GDCs) and to CASTOR (via Gigabit Ethernet over the standard CERN backbone). The TDSM Movers wait for volumes to be migrated and, when a candidate is found, they are given one or more files to move to CASTOR and to register into AliEn. As there are no particular requirements for what concerns transient latencies in the output of the TDSM Movers, it is possible to
have several files being moved at once (either from a single machine or from multiple hosts). Each mover works in cooperation with all the others and can easily maximize the usage of its link to the Computer Centre while reducing the latencies imposed by the Mass Storage System simply by running more simultaneous copies of the transfer process (each copy works on a separate file). It is therefore straightforward to profile the movers’ typical usage as a function of the expected output rate from the ALICE experiment by adjusting the number of parallel moves per TDSM mover and the number of TDSM movers. Files are migrated out of the “emptying” volumes until these are totally emptied, at which point the TDS area becomes again available for writing.

![Diagram of TDSM architecture](image)

**Figure 2. TDSM architecture.**

The architecture of the TDSM system is shown in figure 2. Data flows top to bottom, from the GDCs to the TDS, the File Movers and CASTOR. In the TDS box of this particular example we can see three separate groups of volumes: each group has its own dedicated hardware path between hosts and disks. On the bottom-left of the figure we can see the AliEn spooler machine, which handles the liaison with AliEn. The movers keep an active liaison with the ALICE logbook [9], a key tool to monitor at the level of the DAQ/ECS operator the status of the transfers for each individual run and for the experiment as a whole. On the left side of the figure we can see the TDSM configuration and control database and the TDSM Manager host, both described later in this section.

By using the “affinity” feature of the StorNext file system, it has been possible to give to all machines (managers, writers and readers) a global and unique view of all the disks while retaining the capability to confine a given write or read operation to a well-defined hardware volume. Therefore, it is trivial to dedicate a disk entity (a set of disk volumes working together in RAID) for a particular type of operation (in the case of the ALICE TDS: single-stream writing or multi-stream reading), optimizing the throughput and reducing the number of potential conflict for the usage of its resources. Yet, it was not needed to create individual file system (e.g. one directory per disk volume), an
operation that would have held similar results at the price of a reduced flexibility and increased complexity, for the configuration, installation and operation of the TDS.

Disk volumes, Movers and GDCs are organized in groups. Each group shares the same hardware resources to move data (e.g. the same FibreChannel switch). The TDS can ensure the fastest overall transfers by privileging data transfers within the same group. Data traffic crossing the groups’ barrier is allowed but may overload the inter-groups hardware links. The TDSM gives priority to intra-group transfers whenever this is possible, eventually reverting to inter-group assignments if this cannot be avoided.

The coordination of the usage of the disk volumes, the control of the activity of the TDS Movers and the monitoring of the resources of the TDS are the tasks of the TDS Manager (TDSM) package. The TDSM is based on a distributed, multi-process and multi-host architecture synchronized via a dedicated MySQL database. The TDSM keeps track of the status of disk volumes, serves requests for write buffers issued by the GDCs, handles operator command for TDS control, coordinates the activity of the TDS Movers, registers the status of the migration in the ALICE logbook and synchronizes the liaison with AliEn. The central TDSM database is used to store configuration parameters, to describe the status of the components of the TDSM, to record the history of the system and to distribute common procedures amongst all the actors. A central host, named TDSM Manager, coordinates the activities of the other actors (Movers and GDCs) and handles the liaison with the ALICE trigger system.

For a system as complex as the ALICE TDSM it is vital to be able to monitor the status of the individual components as well as to have a clear picture of the behaviour of the organism as a whole. The ALICE operator can follow at any time the status of the data stream by means of several online monitoring tools:

- by direct access to the TDSM database, to get the status of each component (disk volumes, GDCs, Movers, Manager);
- using history records stored in the TDSM database, to get usage statistics grouped by different criteria (machines, volumes, disk arrays, groups);
- via the ALICE logbook, to get how much data has been recorded and transferred, in general and on a per-run information: how many files have to be moved, how many are been moved, how many have been moved and registered in AliEn;
- by means of the ALICE resource monitoring system, based on Lemon [10], to access information on the setup of the system and to access several key metrics collected on all the machines part of the DAQ and of the TDSM.

4. Alternate trigger profiles
The ALICE Data Acquisition system is capable to sustain data streams of much higher bandwidth than the design value of 1.25 GB/s. Although all the ALICE detectors can be blocked at any time (any element in the DAQ chain is allowed to pause its upstream components via a backpressure-based architecture), this should be avoided as it has two undesirable side effects: the increase of the detector dead time and the altering of the trigger profile associated to the output of the experiment.

Thanks to the high flexibility of the ALICE trigger system, it is possible to steer on the fly the policies used to trigger events. The TDSM uses this feature to keep under control the occupancy of the data disks. Whenever the amount of used resources exceeds a predefined threshold, the Central Trigger Processor (CTP) is requested to switch from the standard trigger mode to a "rare trigger" mode, where only trigger conditions which are less likely to occur are enabled. The effect is a less sustained data throughput which gives a chance to the TDS to free up disk resources (assuming that
the data stream to the CERN Computer Centre runs at its nominal speed) while the achieved trigger profile respects the design objectives of enabling only the rare triggers. The ALICE experiment does, if necessary, toggle between the above two trigger profiles as a function of the occupancy of the TDS, keeping the average data rates around their design values while minimizing the dead time of the detector and without upsetting the quality of the events.

5. Experiences during 2008

The TDS as implemented in 2008 fulfilled all the user requirements. Data has been recorded flawlessly at the experimental area, delivered to the Computer Centre and published into AliEn, serving the requests of several parallel partitions during commissioning runs, cosmics runs and early LHC beam data collection. Many ad-hoc tests have been performed in between and during the production runs to identify hot-spots in the recording process and to test alternate configurations for the various components. Special attention was devoted to the Objectification of the data streams as we had for the first time actual detector data to run them on. Alternate configurations with varying file size and alternate sets of TDSM-related parameters were tested and evaluated during the same period. During the year 2008, two major exercises took place: a cosmics run session (1Q08), where 70 TB have been written in about 1500 runs, and the final ALICE commissioning stage that saw more than 100 TB written in about 6800 runs. No problems were observed in neither of the two periods and all the components operated as planned.

The version of the TDS deployed in 2008 was based upon the results of the past ADCs and its architecture was tailored to ALICE seen as a single, massive Data Acquisition system. However, during the commissioning phase and the 2008 runs it became rather clear that this approach did not fit to the reality, where we witnessed most of the time to a set of several parallel data acquisition systems used either for testing purposes or to implement independent readout partitions. The existing TDSM configuration was therefore adapted to fulfill these modified requirements and a second version of the TDS and its associated software was put in operation during the 2008 winter shutdown. The new architecture is designed in full respect of the observed usage pattern and makes a more efficient use of the available resources, both for parallel and for global runs. First validation tests confirmed the feasibility of the new implementation that has become the production environment for 2009.

As the objectification procedure applied to real detector data when running in real conditions proved to use much more CPU than expected, the need for more processing power (and therefore more GDCs) became rather obvious. This problem was temporarily bypassed by allocating more machines to the global runs, which eventually distributed the load amongst the available hosts, but this could not be used as a final solution as we quickly ran out of hardware resources (especially of links between GDCs and TDS). For the future runs we plan, as described in the next section, to increase the amount of resources (physical and virtual) used to format and to write the experimental data. At the same time we have started a detailed profiling campaign, both at Point 2 and in our development labs, to identify hot-spots within the objectification procedure and to improve its efficiency.

6. Future expansions

The hardware architecture deployed in 2008 did neither allow further expansions nor give enough “hot” redundancy in case of hardware failures. Extra resources had to be planned, to fulfill the staged installation procedure of the ALICE DAQ system and to ensure adequate spare resources for a steady, reliable operation of the system. In the course of 2009, two extra FibreChannel switches SANBox 9000 will be added to the existing setup. Each switch can host up to 8 blades, either with 16 * FC 4 Gb ports or with 4 * FC 10 Gb ports per blade. By equipping the two new switches with 9 “slow” cards, to be used for hosts and equipments, and 2 “fast” cards to allow the inter-switch connectivity, the TDS will get enough resources (particularly CPU power available for the objectification procedure) to ensure a reliable and up-to-specs environment at the time of LHC startup.
Another new requirement raised in 2008 is the need for more GDCs to be allocated for standalone tests or to be assigned to very low-rate runs. These machines need to have full access to the TDS at relatively low throughput (compared to a standard GDC). The allocation of a dedicated host with direct attachment to the TDS was delivering more performance than strictly required, while consuming resources such as FibreChannel ports and user licences. We had to find an alternate data path between GDCs and TDS. The solution came by using StorNext over IP, a flexible and distributed access method to the data volumes and their associated Cluster File System which runs over Ethernet. This new path has been fully validated and went into production already at the end of 2008. Servers linked via Ethernet on one side and via the same interface as for GDCs/Movers on the other could handle requests coming from a set of detached machines, using a distributed and self-balancing architecture for a lower rate total throughput. This approach has been fourthly expanded for the architecture of the TDS to be deployed in 2009.

7. Conclusions
The TDS approach for data recording at the ALICE experiment gave the expected results from operation day one. The experience cumulated during the ADCs proved to be extremely valuable and allowed a seamless integration of all the components of the PDS. The new implementation of the TDS is closely tailored to the usage pattern required to run the ALICE experiment and the enhanced software architecture makes a much better use of the available resources. The scalable architecture of the TDS allows, if necessary, the improvement of the data rates and/or of the number of data producers either by upgrading the existing components or by expanding them. Seen this, we can safely conclude that the ALICE online data storage system is ready for the real challenge, when the LHC will start circulating beams and producing the first collisions at Point 2.

References
[1] Buncic P, Peters A J and Saiz P 2003 The AliEn system, status and perspectives Proc. CHEP 2003 (La Jolla, CA: spires)
[2] Cancio G, Duellmann D and Pace A 2009 Data Management Evolution and Strategy at CERN Proc. CHEP 2009 (Prague, CZ: JPCS)
[3] AA VV 1995 A Large Ion Collider Experiment - Technical Proposal (CERN, CH: LHCC-95-71)
[4] Rademakers F 2000 The ALICE Data Challenges Proc. CHEP 2000 (Padova, I: spires)
[5] Divià R et al. 2003 Challenging the challenge: handling data in the Gigabit/s range Proc. CHEP 2003 (La Jolla, CA: spires)
[6] Carena W, Divià R, Saiz P, Schossmaler K, Vascotto, A and Van de Vyvre P 2001 Online Performance Monitoring of the Third ALICE Data Challenge (CERN, CH: CERN-ALI-2001-046)
[7] Carena W, Csató P, Dénes E, Divià R, Schossmaler K, Soós C, Sulyán J, Vascotto A and Vande Vyvre P 2001 PCI Based Read-out Receiver Card in the ALICE DAQ System (CERN, CH: ALICE-PUB-2001-47)
[8] Brun R 2009 Global Overview of the current ROOT system Proc. CHEP 2009 (Prague, CZ: JPCS)
[9] Chibante Barroso V et al. The ALICE Electronic Logbook Proc. CHEP 2009 (Prague, CZ: JPCS)
[10] Gancio C et al. Current Status of Fabric Management at CERN Proc. CHEP 2004 (Zurich, CH: indico)