The ALICE DAQ: current status and future challenges

M. Arregui [h], W. Carena [h], S. Chapeland [h], P. Csato [d], E. Denes [d], R. Divià [h], B. Eged [e], P. Jovanovic [b], T. Kiss [d], V. Lindenstruth [i], Z. Meggyesi [d], I. Novak [e], F. Rademakers [f], D. Roehrich [a], G. Rubin [h,d], D. Tarjan [e], N. Toth [e], K. Schossmaier [h], B. Skaali [j], C. Soos [e], R. Stock [g], J. Sulyan [d], P. Vande Vyvre [h], A. Vascotto [h], O. Villalobos Baillie [b] and B. Vissy [c] for the ALICE Collaboration,

[a] *Department of Physics, University of Bergen, Norway*

[b] *School of Physics and Space Research, University of Birmingham, United Kingdom*

[c] *ELTE University, Budapest, Hungary*

[d] *KFKI Research Institute for Particle and Nuclear Physics, Hungarian Academy of Sciences, Budapest, Hungary*

[e] *Technical University, Budapest, Hungary*

[f] *GSI, Darmstadt, Germany*

[g] *Institut für Kernphysik, Johann-Wolfgang Gothe Universität, Frankfurt, Germany*

[h] *CERN, Geneva, Switzerland*

[i] *Institut für Hochenergiephysik, Ruprecht-Karls Universität, Heidelberg, Germany*

[j] *Department of Physics, University of Oslo, Norway*

Abstract

The ALICE data acquisition system has been designed to support an aggregate event-building bandwidth of up to 2.5 GByte/s and a storage capability of up to 1.25 GByte/s to mass storage. A general framework called the ALICE Data Acquisition Test Environment (DATE) system has been developed as a basis for prototyping the components of the DAQ. DATE supports a wide spectrum of configurations from simple systems to more complex systems with multiple detectors and multiple event builders. Prototypes of several key components of the ALICE DAQ have been developed and integrated with the DATE system such as the ALICE Detector Data Link, the online data monitoring from ROOT and the interface to the mass storage systems. Combined tests of several of these components are being pursued during

the ALICE Data Challenges. The architecture of the ALICE DAQ system will be presented together with the current status of the different prototypes. The recent addition of a Transition Radiation Detector (TRD) to ALICE has required a revision of the requirements and the architecture of the DAQ. This will allow for a higher level of data selection. These new opportunities and implementation challenges will also be presented.

Key words: (PACS code: 07.05.B, 07.05.H, 07.05.K); DAQ; optical link; event building; filtering; data storage; triggering.

---

1 Introduction

The ALICE Trigger/DAQ system was initially designed for a set of requirements described in the ALICE Technical Proposal [1,2]. The system has to operate with different beam types: pp, p–ion and ion–ion. The original rate and size of events acquired for the extreme case (Pb–Pb) resulted in an aggregated bandwidth of 2.5 GBytes/s in the DAQ system (including the event building) and 1.25 GBytes/s to the mass storage. The system must also be able to combine two different types of physics events: a slow rate of central triggers generating the largest fraction of the total data volume, together with a faster rate of dimuon events.

These requirements and the corresponding Trigger/DAQ architecture were relatively stable and were refined for the preparation of the Technical Design Reports (TDR) [3]. This has led to the overall architecture of the ALICE DAQ, control and trigger systems. The interfaces between these systems and with the sub-detectors are shown in Fig. 1. This paper will cover the DAQ and trigger systems.

The original Trigger/DAQ requirements have since evolved. The first evolution consisted in the addition of buffering to the front-end electronics of the ALICE sub-detectors. The second was an increase by a factor 2 of the estimated data volume of the Time Project Chamber (TPC) that produces more than 90% of the total data volume. The third is the addition of the Transition Radiation Detector (TRD) [4] that will allow us to identify electrons and will provide an additional input signal to the trigger. It allows a new trigger type, the dielectron trigger, and opens to the ALICE DAQ the possibility of adding new online processing (region-of-interest readout, third-level filter, and improved data compression).

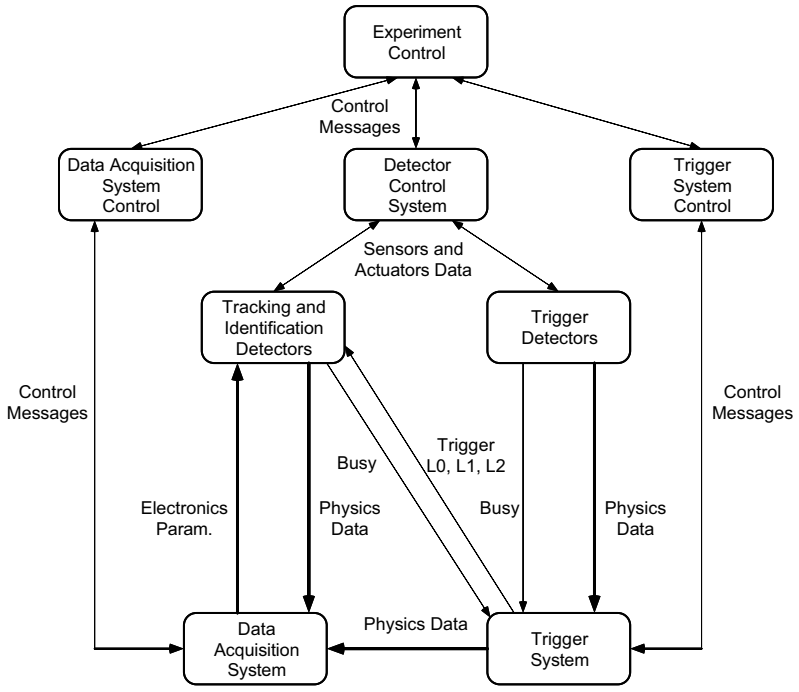

Fig. 1. The overall architecture of the ALICE DAQ, Trigger and control systems.

2 The original requirements

The original Trigger/DAQ architecture was designed to:

- handle widely different running conditions [1];
- balance the capacity to record central collisions which generate large events with the ability to acquire the largest possible fraction of rare events;
- accommodate detectors with vastly differing response time. Some trigger detectors read out for every bunch-crossing, while the slowest one (TPC) has a drift time of 100 μs.

To accommodate these requirements, a 3-level trigger system was designed. The trigger level 0 is needed by some of the front-end electronics which require a strobe 1.2 μs after the time of passage of the particle. This trigger is essentially the signal initiating the conversion process of the sample-and-hold electronics. The trigger level 1 combines input from all the trigger detectors. This trigger comes 2.4 μs after the interaction time. At this time, it is possible to make a refined centrality decision using both the Micro-Channel Plates (MCP) and the Zero-Degree Calorimeter (ZDC) and to have a dimuon trigger.

The DAQ architecture was designed as a two-step process. The data transfer from the front-end electronics to the DAQ system is performed in parallel for

[1] Pb–Pb collisions with 125 ns bunch-crossing intervals, Ca–Ca collisions with 125 ns bunch-crossing intervals and p–p collisions with 25 ns bunch-crossing intervals.

all the sub-detectors over hundreds of optical links. This transfer is initiated by the trigger level 1. The DAQ system then waits for the final trigger decision (trigger level 2) before starting the event building. After the event building, the data are compressed by a factor 2. This results in a data throughput of 1.25 GBytes/s to the mass storage.

3 The updated requirements

The first new requirement was introduced with the addition of buffering in the front-end electronics of the ALICE sub-detectors. This buffering was added to decouple the busy status of each detector from the data transfer to the DAQ. This allows also a better usage of the bandwidth between the detectors and the DAQ and a reduction in the number of optical links needed. The data transfer to the DAQ is also delayed till the trigger level 2 decision.

The second new requirement concerns an increase of the TPC event size. The latest simulations show that the TPC average occupancy will be higher than originally anticipated with a value of the order of 25% instead of 15%. In the meantime the number of samples per channel has been reduced from 1024 to 512 [2]. Each data sample results in a 10 bit word. The total TPC event size is estimated now to 76 MBytes after zero-suppression.

The third new requirement consists of the recent addition of a TRD to the ALICE experiment [4]. This addition induces the following consequences:

- The TRD will bring an additional data volume to an event whose size has already been increased considerably since the original estimation. The updated total data volume is summarised in Table 1.
- With the TRD, ALICE is now equipped with a fast electron detector. This new input will be used as input for the trigger level 1 that has consequently been delayed from 2.4 to 5.5 μs after the interaction. The updated trigger delays and rates are summarised in Table 2.

An estimation of the total bandwidth needed in the DAQ requires an evaluation of the data throughput for each type of trigger. The rates required to accumulate enough statistics in a one-year period of data taking are of the order of few 10^6 events for hadronic physics, at least few 10^7 events for charm

[2] The present physics simulation has shown that the best precision would be given for more than 512 samples per channel. This will most probably still evolve between now and the beginning of data-taking. It will also be adapted once a first experience has been gained with the detector. The TPC electronics is designed such that this number can be adapted at run-time at up to 768 samples. In this paper, we have adopted the value of 512 samples per channel.

Table 1
Event size of the ALICE sub-detectors for Pb–Pb central events

Detector	**Subdetector or option**	**Minimum event size (MByte)**	**Maximum event size (MByte)**
Inner Tracking System (ITS)	Si Pixel	0.140	0.280
	Si Drift	1.500	1.500
	Si Strips	0.160	0.160
Time Projection Chamber (TPC)		56.100	75.900
Transition Radiation Detector (TRD)		8.000	8.000
Time Of Flight (TOF)		0.180	0.180
Photon Spectrometer (PHOS)		0.020	0.020
High Momentum Particle Identification (HMPID)	RICH	0.120	0.120
Dimuon Forward Spectrometer (MUON)	Trigger and Tracking	0.150	0.150
Photon Multiplicity Detector (PMD)		0.030	0.120
Trigger System, Micro-Channel Plates (MCP), Zero-Degree Calorimeter (ZDC)	Central Trigger, Trigger Detectors	0.120	0.120
Total		66.500	86.500

and dielectron physics, and at least 10^9 events for dimuon events. Given a lead–lead run of a few weeks per year, the rates are of the order of few events/s for hadronic physics, few tens for charm and dielectrons, and few hundreds for dimuons.

The first three types of physics require the data from all the detectors while the dimuon needs the data from some detectors (muon arm and a fraction of the inner tracking system). The charm and dielectron physics both need the data from all the detectors. It should be noted that the introduction of the TRD input to the dielectron trigger for the events may introduce a bias for the charm physics. This point is going to be investigated in the collaboration but in this paper we have assumed that the triggers for these two types of

Table 2
Trigger delays and input detectors

Trigger Level	**Input Detectors**	**Delay**	**Action**
0	ZDC	1.2 μs	Front-end electronics strobe
1	MCP, Muon, TRD	5.5 μs	Trigger decision Event number distribution
2		Up to 100 μs	Past-future Protection

physics are disjoint.

The sets of raw data used for different types of physics sometimes overlap. In this case the same data set is used for the different sets of physics. [3] Depending on the type of physics, the set of detectors read out may also vary [4] [5].

Finally, the global data throughput will be reduced by data compression. A general-purpose data compression was included in the original architecture to reduce the data throughput by a factor 2 to the arbitrary limit of 1.25 Gbytes/s. that has been put as a maximum mass storage bandwidth.

The collaboration is now elaborating running scenarios including the various types of physics and triggers. One example is given below. Given the latest estimation of event sizes [5–7], the aggregated data throughput is much higher than the available mass storage bandwidth. The standard data compression (factor 2) will not be sufficient to reduce it to an acceptable level (see Table 3).

Several other ways to reduce this huge data volume were therefore introduced in the architecture:

- a readout reduced to the region-of-interest for the dielectron triggers;
- a level-3 filtering for the dielectron triggers;

[3] Different types of physics use raw data from all detectors collected after a central trigger or a minimum-bias trigger. This is the case for the full events triggered by a central trigger that are used for hadronics, charm and dimuon physics. It is also the case for the full events triggered by minimum-bias events that are used for hadronics and charm physics.

[4] For the dielectron physics, the following set of detectors will be read out: the ITS, the TPC and the TRD (possibly a partial readout), the PHOS, the MUON, the PMD, the trigger (FMD and ZDC), and the trigger system.

[5] For the dimuon physics, the following set of detectors will be read out: the pixels layers of the ITS, the PHOS, the MUON, the PMD, the trigger (FMD and ZDC), and the trigger system.

Table 3
Data throughput for the different types of physics

Physics	**Trigger**	**Detectors**	**Event Size (MByte)**	**Rate (Event/s.)**	**Data Throughput (MByte/s)**	**Comment**
Hadronic	Central	All	≈87.0	2		See footnote (3)
	Min. Bias	All	≈22.0	2		See footnote (3)
Charm	Central	All	≈87.0	20	≈ 870	After compr.
	Min. Bias	All	≈22.0	20	≈ 220	After compr.
Dielectron	Central + Dielectron	See footnote (4)	≈4.6	200	≈ 460	After compr.
	Min. Bias + Dielectron	See footnote (4)	≈1.2	200	≈ 120	After compr.
Dimuon	Central + Dimuon	See footnote (5)	≈0.6	1000	≈ 300	
	Central	See footnote (5)	≈87.0	20		See footnote (3)
Total					≈ 1970	

- a detector-specific data compression system for the TPC data.

4 The architecture

4.1 The software framework

The development of such a large system by several teams requires clear interfaces to be defined between the components constituting the system and the integration of these components to be started early enough. Furthermore, the long development and production periods of the experiment will inevitably lead to several generations of some of the components before and during the data-taking period. This evolution has to be planned from the start.

In addition to the development of the final system, the DAQ group was also

confronted with requests for DAQ systems for the test beams activities.

The idea therefore came to develop a software framework able to surround the development work of the final DAQ system and to support the present detector activities of the experiment: the Data Acquisition and Test Environment (DATE).

At any given time, one production version of the system is released and constitutes the reference framework in which the prototypes are integrated. It is also used for the functional and performance tests of the DAQ system.

The DATE system has been used by the ALICE test beams and the CERN fixed-target experiment NA57. Several major releases of the system have been made [8]. It has been used since 1997 for different systems and has allowed the collection of the order of 10^9 events. The DAQ team of the Compass experiment has used DATE as well and has contributed to its development by porting it to more platforms (PC/Linux and Alpha).

4.1.1 The DATE framework architecture

The DATE framework is a distributed process-oriented system. It was designed to run on Unix platforms connected by an IP-capable network and sharing a common file system such as NFS. It uses the standard Unix system tools available for process synchronisation and data transmission. The system characterises different functions:

- The Local Data Concentrator (LDC) collects event fragments and reassembles them into sub-events. The LDC is also capable of doing local data recording (if used in stand-alone mode) and online sub-event monitoring.
- The Global Data Collector (GDC)puts together all the sub-events pertaining to the same physics event, builds the full events and sends them to the mass storage system. The GDC is capable of online event monitoring.
- The DATE run-control controls and synchronises the processes running in the LDCs and the GDCs. It can run on a LDC, a GDC or another computer.
- The monitoring programs receive data from LDCs or GDCs streams. They can be executed on any LDC, GDC or any other machine accessible via the network.

4.1.2 Detector-specific software

The DATE framework incorporates several APIs that allows the addition of detector-specific software. This is the case for:

- The readout that can be either programmed as C-code routines or defined

with an interactive application.
- The online monitoring program that can be either programmed as a C/C++ program or defined inside the ROOT framework [9,10].

4.2 The trigger system

The trigger protocol consist of signals sent from the central trigger to the detectors: (trigger levels 0, 1 and 2 and event number) and the busy signal sent from the detectors to the central trigger:

- Level 0 provides the earliest signal to strobe detector front-end electronics. The fastest possible copper coaxial cable will be used to achieve the fixed latency of 1.2 μs.
- Level 1 is based on all the available information about the interaction required to make a trigger decision. The trigger level 1 is delivered at a time equal or less than 5.5 μs through the RD12 Trigger, Timing and Control system (TTC) [11].
- Level 2, the final level of trigger, comes up to 100 μs after the interaction has taken place. Because of the long drift time of the TPC and the non-negligible probability of pile-up of several central collisions during this drift time, the final trigger decision can only be made at the end of this drift time. This is the role of the level-2 trigger . The principal purpose of this trigger level is to ensure that pile-up does not occur in the TPC: it is the past–future protection. Similar protection intervals can be set as appropriate for the other sub-detectors. The level-2 reject decision can be issued at any time up to the full past–future protection interval. A level-2 accept cannot be issued until the full interval has elapsed.
- Event number: the trigger system distributes a unique identifier for each event. This event number consists of the orbit-number and the bunch-crossing number within a given orbit. The bunch-crossing number is counted by the TTCrx receiver chip and the orbit-number will be sent as a broadcast message by the trigger system to the detectors. All the event fragments generated by the sub-detector's electronics will be tagged by the event number for the subsequent sub-event and event building.
- Busy: each sub-detector communicates a single busy signal to the central trigger. The busy signal indicates that the sub-detector should not receive further triggers. The busy is set by each detector after the level 0 and is released as soon as the sub-detector is ready to receive the next trigger. The busy signal can be generated in the front-end electronics or at the receiving end of the corresponding DDLs.

A detailed description of the trigger system and the protocol with the sub-detectors can be found in Refs [12,13].

The architecture of the ALICE DAQ and trigger can be seen in Fig. 2.

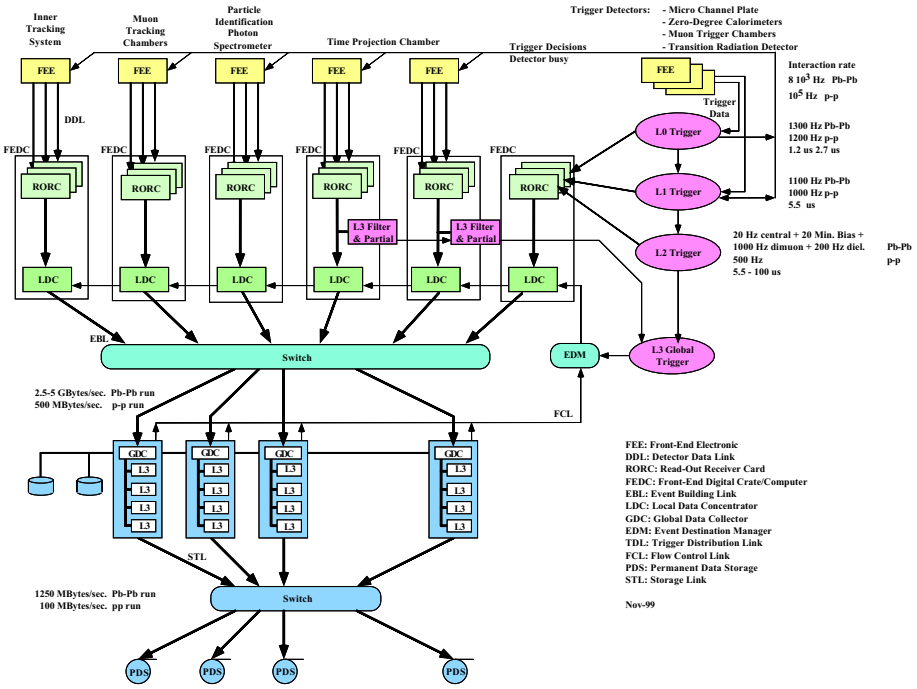

Fig. 2. The ALICE Trigger and DAQ architectures.

5 The data transfer

5.1 The Detector Data Link

The Detector Data Link (DDL) is the common hardware and protocol interface between the front-end electronics and the DAQ system. The DDL is used to transfer the raw physics data from the detectors to the DAQ and to control the detector front-end electronics or download data blocks to this electronics. The DDL consists of the Source Interface Unit (SIU), the optical fibre used as physical medium and the Destination Interface Unit (DIU). The DDL hardware and software is specified in several ALICE notes [14–16].

The data are transferred over the DDL and stored in the input buffer of the Read-Out Receiver Card (RORC). One Front-End Digital Crate or Computer (FEDC) contains one or several RORCs and one LDC. The LDC reads out the event fragments from the RORCs and assembles them into one sub-event.

The first DDL prototype was built using electronics chips used for the 1 Gbit/s Fibre Channel physical layer. The first RORC prototype was built as a VME board [17]. This set of prototypes was used for tests of the ALICE TPC and

was integrated with the DATE system. The configuration used was transferred data between two VME crates. The measured bandwidth over the DDL is 100 MBytes/s. The latency of transferring one command from one processor to the other was 6 μs and the sustained memory to memory data transfer including the VME readout is 34 MByte/s using 64-bit block transfer mode.

The DDL is independent of the form-factor of the FEDC because it is a daughter board of the RORC. The current prototype was based on VME and the next generation will be based on PCI. This prototype will also explore the possibility of using direct-memory access to the LDC memory. The final form factor of the FEDC has not yet been decided on account of the evolution in the field of I/O busses (see Section 11).

Several tools were developed to ease the integration of the DDL with the detector's electronics. In particular a simple hardware module was designed to simulate the sender side of the DDL SIU, the SIU simulator [18]. This module makes the initial debugging of the front-end electronics much easier than with a complete link connected to a computer. It simulates the simplest DDL transactions. It acts as a receiving null device or as a fast data source. A set of test libraries and programs was also developed [19].

6 Data compression

The increasing requirements will need better ways to reduce the data throughput and in particular better data compression algorithms. One possible way is to add dedicated hardware in the DAQ chain. The option of adding this hardware before the DDL is being investigated. It has the advantage of using the DDL bandwidth to its maximum while requiring the uncompress of the data for any access such as online monitoring.

Statistical analysis of the distribution of the TPC simulated data has been started. This allows the evaluation of the gain that could be obtained by using data compression techniques such as Huffman coding or vector quantisation [7]. The compression factors expected are of the order of 2 to 3.

7 The event building

7.1 The event building software layers

The sub-events prepared by the LDCs must be transferred to one GDC where the full event can be assembled. In ALICE, the event building is managed by the Event Building and Distribution System (EBDS) [20]. This distributed protocol runs on all the machines, LDCs and GDCs, participating as data sources or destinations. The goals of the EBDS are to synchronise all the LDCs on the choice of the GDC used as destination and to balance the loads on the different GDCs. The issue is to keep up with the data-flow while keeping the EBDS protocol overhead as low as possible.

We have used up to now, the de-facto standard TCP/IP as transport mechanism for the EBDS protocol and for the data transfer. This protocol is the standard to which the networking industry has converged and it is running on a wide range of hardware layers.

7.2 The event building hardware layer

The strategy is to build on what we have and to use commodity networking as far as possible. Ethernet is the dominant, ubiquitous LAN technology. We therefore used this technology to build our prototypes. However, as long as we can rely on a standard protocol such as TCP/IP, the final choice of the hardware layer used in the event-building network can be delayed.

8 The permanent data storage

8.1 The Mass Storage System

As mentioned in Section 2 on requirements, the ALICE experiment will rely on a huge Mass Storage System (MSS) with unprecedented bandwidth and storage capacities. The mass storage system is the software layer responsible for insulating user applications (DAQ and off-line systems) from the underlying hardware used to store the data: intermediate secondary disks, final tertiary storage, servers and robotics. The ALICE DAQ and offline systems rely on a logical interface of the MSS being a hierarchical file-system or a staging system. The MSS also has the key function to insulate the user applications from the evolution of the tertiary storage hardware.

In 1998, a common Mass Storage Project [21] was started by the ALICE DAQ and the IT/PDP group. With the available hardware, a sustained bandwidth of up to 35 MBytes/s was reached with eight parallel streams. The scope of the project was then enlarged to become the main test bed for all DAQ and offline tests (see Section 10).

Up to now, the ALICE DAQ prototypes have used as MSS the system developed by the HPSS consortium [22]. The HPSS system provided more than the functionalities needed. However, the server part of HPSS is not yet supported on a wide range of hardware platforms even though the situation is evolving. The server part is not available on the PC/Linux platform. HPSS cannot be the only solution and we want to investigate alternatives.

We plan therefore to test two other MSS systems: Castor [23,24], being developed at CERN, and Eurostore [25] being developed by a European consortium.

8.2 Secondary and tertiary storage systems

The Permanent Data Storage (PDS) comprises an aggregate of secondary and tertiary storage. The ALICE DAQ will use parallel streams to record the data after the event building. The current baseline option is that the PDS will be part of a Central Recording Facility (CDR) located in the CERN computing centre.

The data will be transferred between the ALICE DAQ system and the CDR facility over a dedicated network backbone as proposed by the LHC communication infrastructure working group.

In the computing centre, several tens of parallel data streams will be recorded on several magnetic tape units. It is anticipated to see products with a bandwidth of 20 to 50 MBytes/s. The nominal ALICE bandwidth would require between 25 and 60 of these devices during the heavy ion-run. It is also expected to see tape cartridges with capacities between 100 and 200 GBytes. The ALICE DAQ would therefore require between 40 and 80 volume movements per hour which is feasible with the current robotics.

9 Region-of-interest readout, online filtering and advanced data compression

9.1 Motivations

As described earlier, the updated physics requirements of ALICE have increased so much that the available storage bandwidth will not be able to store the data throughput needed by the ALICE physics programme. The addition of online processing could also improve this programme. Therefore, some online reduction of the overall data throughput has to be added to the original DAQ architecture. Three different approaches will be evaluated:

- For some types of triggers, it may be sufficient to read out a fraction of the sub-detectors. This is the case for dielectron triggers for which the data from the Region-Of-Interest (ROI) would suffice for dielectron studies.
- For some types of triggers, the statistics could be improved by inspecting more events online and discarding a large fraction of them.
- More advanced data compression schemes could be applied to the TPC data.

The ROI readout and the online filtering are possible thanks to the indications that will be provided by the TRD detector on the area where electron tracks have been detected.

9.2 Region-Of-Interest readout

The TRD detector will deliver after 5.5 μs (for the L1 trigger) an indication of the presence of electrons tracks and their location. Out of the 36 TPC sectors, on average, only 3 or 4 will contain such tracks and will constitute the ROI. The ROI information will be broadcast to all the LDCs. The TPC LDCs containing no ROI will generate no data. The TPC LDCs containing an ROI will read out all or only a fraction of the data. This reduction could reduce the TPC data by a factor comprised between 10 and 40.

The ROI readout is a simple data processing local to each TPC LDC and it can be performed with the CPU power available in the LDC. It requires ROI information to be received from the TRD.

The ROI readout could proceed in two steps. The LDCs that contain no ROI can discard the sub-event. The LDCs that contain an ROI can extracted the ROI from the sub-event and send it to the event building.

The same information delivered by the TRD detector could also be used to initiate a fast tracking of few electron tracks in the TPC data. This fast tracking could be used to verify the presence of the electron track and subsequently decide to archive the event or not. This online filtering could allow up to 200 events/s to be inspected and a fraction of them to be kept.

The online filtering must receive ROI information in the same way as the ROI readout. The tracking can be performed with the addition of limited CPU power made available in a common L3 processing farm. The current estimation is of the order of 40 kCU [6] that could be implemented as one additional CPU for each TPC sector.

The LDCs containing ROIs could send the sub-event to the L3 farm where the tracking could be performed. The global decision to keep or reject the event would then be used for the event building.

9.4 Advanced data compression

More advanced ways of compressing the data are being investigated for the TPC data. They consist of coding the data as a set of model parameters and deviations from this model. The gain is mainly obtained because the model is simple and requires few parameters and small deviations. The model used here is a local tracking model. The raw data are then encoded as the deviations of the clusters with this tracking model. The data stored consist of the parameters of the local tracks and the cluster deviations. It should be noted that the final tracking will be repeated for the offline reconstruction. This online tracking is expected to compress the data by a factor of up to 15.

The data processing power required for this data compression has been estimated to be of the order of 400 kCU. This could be avalaible in a L3 farm present between the GDCs and the data storage.

[6] The unit of CPU power used in this paper is the CERN Unit (CU). The approximate conversion with other CPU power units is the following: 10 CU = 40 MIPS = 40 SpecInt92 = 1 SpecInt95.

10 Prototyping and the ALICE Data Challenge

The ALICE DAQ project fellows a development process including early prototypes and regular releases of production quality prototypes.

Prototypes of the different components of the DAQ system are being developed by different groups. Some of them are used to support present activities of the sub-detector groups. This allows the functional features of the prototypes to be verified and fast user feed back which is essential during the development phase.

However, the use of prototypes in present sub-detectors activities requires limited performances compared to the final system. The ALICE DAQ and Offline project, together with the CERN IT division, have therefore initiated a series of 'Data Challenges'. At regular intervals, once or twice a year, a set of prototypes are assembled and a test run of a few days is executed on the test set-up.

The goals of these Data Challenges are to verify the progress of the different prototypes, to start the integration of these prototypes into a common system, and to asses the performances of the whole.

The first ALICE Data Challenge was carried out using 10 LDCs in an experimental area and 10 GDCs in the computing centre. The test ran for 6 days at a sustained bandwidth of 14 MBytes/s. A total of 7 TBytes of data in 15 000 files was collected. The target of the next ALICE Data Challenge is to reach a sustained bandwidth of 100 MBytes/s with 18 LDCs and 20 GDCs connected with Fast Ethernet.

11 Technology evolution

The ALICE DAQ system will be based as far as possible on commodity computing and communication products. This section is therefore devoted to some of the basic technologies used in these commodity products. The basic technologies required for most of the components are already available with adequate performance or are expected to come on the market before the experiment start-up [26]. In this section we review also the evolution of these technologies or the emergence of new ones that could influence the implementation of the final system.

It is reasonable to believe that by 2005 we will have access to computers based on processors running at between 1.2 and 2 GHz (Section A of Ref. [26]). These computers will be equipped with a memory of a few GBytes.

The question of memory bandwidth is critical for data acquisition applications. The dearth of main memory bandwidth is well known. It is due to the delay of the low-cost DRAMs access time over the rapid increase in processor clock frequency. The increase of cache size and performance has been sufficient to hide this effect. However, two new factors are strong incentives pushing for a higher memory bandwidth:

- The most common PC graphics card interface (Accelerated Graphic Port) will have an increased bandwidth from the present AGP (266 MBytes/s) to the AGP 4x (1 GBytes/s).
- The Personal Computer Interface (PCI) bandwidth has been increased from 133 to 532 MBytes/s by a doubling of its width from 32 to 64 bits and its frequency from 33 to 66 MHz.

As a consequence, several innovations in the design of DRAM memory subsystems reducing the access latency and increasing the bandwidth [27] will become available in the future PCs.

Future high-end PC systems will also include faster system busses. The default PC system bus running previously at 66 MHz was up to 100 MHz in 1998 and recently to 133 MHz resulting in a bandwidth of respectively 800 and 1100 MBytes/s [28]. Several options for the next generations of system busses are being proposed by industry: RAMBUS [29–31], Double Data Rate (DDR133) [32] and DDR-2 (DDR200) providing bandwidth of respectively 1.6, 2.1 and 3.2 Gbytes/s.

This issue will also be addressed by the designers of the next generation of CPU chips. The three consortia developing the three most widely used lines of processors are going in this direction. The Intel/HP IA-64 line will use the Intel 460GX chip which interfaces the CPU to four PCI buses each at 64 bits and 66 MHz and one AGP 4x port [33]. The next Intel processor generation will run above 1 GHz, creating a far greater demand on memory that the current RDRAM could address [29]. The Alpha 21364 will include four RDRAM channels to achieve 6 GBytes/s of total bandwidth to main memory [31]. The IBM Power4 CPU is even foreseeing 55 GBytes/s bandwidth between the CPU and the memory and between CPUs [34].

The PCI bus has imposed itself as de facto I/O bus standard in commodity and mid-range computers with several variations (32/64 bits, 33/66 MHz). Although recent strides have been made to improve it at up to 1 GBytes/s, there seems to be an agreement in the industry that the next generation of I/O busses will be of a different nature: it will become serial and based on a switch instead of a shared media. Two contending proposals have been made: Next Generation I/O (NGIO) supported by Intel, Dell, and Sun; and Future I/O (FIO) supported by Compaq, IBM, and HP. The two camps have reached an agreement to develop in common the System I/O bus (SIO) sponsored by the InfiniBandSM Trade Association (IBTA). SIO is a switched-fabric interconnect that departs from shared-bus architecture such as PCI with bandwidth. It is based on a serial wire running at 2.5 Gbits/s. It will be available for 1, 4, or 12 wires and a bandwidth of respectively 500 MBytes/s, 2 GBytes/s, and 6 GBytes/s

The first SIO-based products should appear on the market by the end of 2001. Even with the usual delays, it will most probably be available by the time of LHC start-up. The first products will be server-class machine with high I/O needs. This class of machine is today more expensive than PCs. It might however become affordable due to the emergence of large markets for this class of machine such as electronic commerce. In the longer run, the SIO could also be used on the commodity hardware that we are using today. The ALICE DAQ architecture will remain open to this new evolution.

11.3 Local Area Networks

The field of Local Area Networks (LANs) has exploded since 1996 with the introduction of Fast Ethernet and Gigabit Ethernet. According to the industry analyst International Data Corporation (IDC), more than 85 percent of all installed network connections were Ethernet at the end of 1997, representing more than 118 million interconnected PCs, workstations, and servers. The Internet and the all the new applications being developed are changing the network. The LAN traffic is growing at almost 100% per year, mostly from the Internet [35]. This technology is therefore attracting a large fraction of the development efforts by the LAN industry. The current generation of gigabit Ethernet switches allows aggregate bandwidth of up to hundreds of Gbits/s. In the future, Dense Wavelength Division Multiplexing (DWMD) will enable the transport and switching of Tbits/s on optical fibres [36]. This ensures a migration or growth path to the technologies, services, and performance that we will need in the future.

Our intention is to rely on a standard transport protocol if the performance is acceptable in order to be independent from the hardware layer used in the event-building network. We currently use the TCP/IP protocol. This protocol requires a relatively high CPU overhead which is acceptable for the current prototyping that is being performed in ALICE. This overhead is being reduced with the current trend of the Network Interface Card (NIC) to perform a significant fraction of the protocol in the interface card itself [37,38].

11.4 Mass Storage

The reference standard for MSS is the IEEE Reference Model for Open Storage Systems Interconnection. Its development has been very long and it is evolving at a slow pace. Several products conform to the model or part of it but none has implemented it. Moreover, the standard does not fully specify the interface between the various components. Therefore, the interoperability between different systems is not possible. Also, contrary to almost all the other technologies used in the ALICE DAQ system, the MSS is not a lively market. The question of portability of applications to another MSS is therefore critical and, to an even bigger extent, the question of portability of data (Section D of Ref. [26]). We want therefore to remain as independent as possible from the MSS. We will test several systems and avoid strong coupling between the DAQ and the MSS.

11.5 Secondary and tertiary storage - Storage Area Networks

Traditionally, the tertiary storage was two orders of magnitude cheaper than the secondary storage. The current evolution of secondary and tertiary storage costs are such that by 2005 the cost of disk storage could become competitive compared to tape storage if one includes the infrastructure cost (drives and robotics) (see Sections B and C of Ref. [26]). One could therefore envisage an alternative scenario where the active data of a year would be recorded onto disk during the data-taking period and archived to tape offline. This scenario would allow a faster access to all the active data of the year and would reduce the investment cost of the tertiary storage bandwidth. This makes the importance of the MSS even stronger.

There has recently been intense activity in the computing industry concerning Storage Area Networks (SANs). The SAN is a network where several storage systems and clients share data without an intermediate server. The SAN hardware layer is often fibre channel. There are still important issues to be solved: device sharing, coexistence of different operating systems on the same SAN, etc. However, this technology could be beneficial in a distributed environment

like ours and in particular at the interface between the DAQ and the central data recording.

12 Conclusions

The ALICE DAQ and trigger systems architecture are now being revised to take into account new requirements. On one hand, these new requirements indicate that the aggregate throughput of the ALICE DAQ will be higher than the maximum acceptable mass storage bandwidth. On the other hand, the addition of a TRD detector will allow new types of online processing. An online processing farm will therefore be added to the ALICE DAQ system to perform region-of-interest readout, online filtering, and advanced data compression. The DAQ prototyping will from now on include this online processing.

The current prototypes of the ALICE DAQ system are based on commodity technologies. It is anticipated that these technologies will provide adequate performance by 2005. Two points of concern remain though on the existence of an appropriate mass storage system and the on cost of tertiary storage hardware and media.

Acknowledgements

We wish to acknowledge the good support that we have received from the CERN EP and IT Divisions for the ALICE Data Challenges. We want in particular to thank the members of the EP/ESS, IT/CS and IT/PDP groups for their active participation.

References

[1] ALICE Collaboration, *Technical Proposal*, CERN-LHCC 95-71.

[2] ALICE Collaboration, *The Forward Muon Spectrometer - Addendum 1 to the Technical Proposal*, CERN-LHCC 96-32.

[3] D. Swoboda, P. Vande Vyvre and O. Villalobos, *The ALICE DAQ, Trigger and Detector Control System*, ALICE Internal Note 98-33.

[4] ALICE Collaboration, *A Transition Radiation Detector for Electron Identification - Addendum 2 to the Technical Proposal*, CERN-LHCC 99-13.

[5] The ALICE Collaboration, *Inner Tracking System - Technical Design Report*, CERN-LHCC 99-12.

[6] ALICE Collaboration, *Dimuon Forward Spectrometer - Technical Design Report*, CERN-LHCC 99-22.

[7] ALICE Collaboration, *ALICE Time Projection Chamber - Technical Design Report*, CERN-LHCC 2000-001.

[8] ALICE DAQ group, *The ALICE DATE V3.5*, ALICE Internal Note 99-46.

[9] R. Brun et al., *ROOT - An Interactive Object Oriented Framework and its Application to NA49 Data Analysis*, CHEP'97, Berlin, 1997.

[10] R. Brun and F. Rademakers, *ROOT: An Object-Oriented Data Analysis Framework*, Linux Journal, Issue 51, July 1998.

[11] RD-12 project, *RD-12 status report*, CERN-LHCC 2000-002.

[12] A. Bhasin et al., *New developments for the ALICE Trigger*, LEB'99, Snowmass, September 1999.

[13] P. Jovanovic, O. Villalobos Baillie and P. Vande Vyvre, *A description of the Protocol between the ALICE Central Trigger and ALICE sub-Detectors*, ALICE Internal Note 99-39.

[14] G. Rubin and P. Vande Vyvre, *ALICE Detector Data Link (DDL) - Interface Control Document*, ALICE Internal Note 96-43.

[15] G. Rubin, *DDL Physical and Signalling Layer Specification (PhI)*, ALICE Internal Note 97-04.

[16] G. Rubin, *ALICE DDL - Hardware Guide for the front-end Designers*, ALICE Internal Note 98-21.

[17] G. Rubin and J. Sulyan, *The Read-out Receiver Card (RORC) hardware user's guide*, ALICE Internal Note 97-14.

[18] G. Rubin and C. Soos, *ALICE Detector Data Link - Users Guide for the SIU Simulator*, ALICE Internal Note 99-03.

[19] G. Rubin, *ALICE Detector Data Link - DDL Test Programs and Libraries, Users' Manual*, ALICE Internal Note 99-24.

[20] W. Carena et al., *The use of ROOM in the design of data-acquisition software components.*, Real-Time '99, Santa Fe, 1999.

[21] http://wwwinfo.cern.ch/pdp/ps/msp/msp.html

[22] HPSS Consortium, http://www.sdsc.edu/hpss/.

[23] CASTOR Project, http://wwwinfo.cern.ch/pdp/castor/.

[24] J.P. Baud, *The Castor project status*, CHEP 2000, Padova, February 2000.

[25] The Eurostore Consortium, http://www.quadrics.com/eurostore/Main.html.

[26] The LHC Computing Technology Tracking Team, *The second PASTA report - The LHC Technology Tracking Team for Processors, Memory, Architectures, Storage and Tapes*, http://nicewww.cern.ch/ les/pasta/ttt_welcome.html, CERN, 1999.

[27] T. Mitra, *Dynamic Random Access Memory: a survey*, Department of Computer Science, State University of New York, February 1999.

[28] M. Kellog, *PC133: SDRAM Main Memory Performances Reaches New Heights*, IBM Microelectronics, Second Quarter 1999.

[29] M. Slater, *Direct RDRAM (almost) arrives in PCs*, Micro Design Resources, October 1999.

[30] IBM Application Note, *Direct Rambus Memory System Overview*, March 1999.

[31] L.Gwennap, *Alpha 21364 to Ease Memory Bottleneck*, Microprocessor Forum, October 1998.

[32] B. Ji et al., *A 1.6 GBps 1Gb Double Data Rate Synchronous RAM*, IBM Micro News, Third Quarter 1999.

[33] L. Gwennap, *Intel Discloses New IA-64 Features*, Micro Design Resources, March 1999.

[34] K. Diefendorff, *Power4 Focuses on Memory Bandwidth*, Microprocessor report, October 1999.

[35] J. Becknell, *High Speed Networking for Tomorrow*, http://www.aberdeen.com/middleeast/, AberdeenGroup, GITEX CAIRO, April 1999.

[36] N. Mokhoff, *Lucent router sets the stage for all-optical nets*, EDTN Network EETimes, November 1999.

[37] M. McGowen, *Gigabit to Gigabyte and beyond (Building NICs that scale)*, http://www.ods.com/essential/info/nics.shtml, Essential/ODS Networks, 1999.

[38] NN, *Gigabit Ethernet comes of age - Scaling network performance to 1000 Mbps*, 3Com Technical paper, June 1999.