ARCHITECTURE AND IMPLEMENTATION OF THE ALICE DATA-ACQUISITION SYSTEM

F. Carena, W. Carena, S. Chapeland, R. Divià, U. Fuchs, I. Makhlyueva, J.C. Marin, K. Schossmaier, C. Soós, P. Vande Vyvre, A. Vascotto for the ALICE collaboration, CERN, Geneva, Switzerland
T. Anticic, Ruder Bošković Institute, Zagreb, Croatia
O. Cobanoglu, INFN Turin, on leave from Istanbul University, Istanbul, Turkey
E. Dénes, KFKI-RMKI, Budapest, Hungary
F. Ozok, Istanbul University, Istanbul, Turkey
S. Vergara, Benemerita Universidad Autonoma de Puebla, Mexico

Abstract

ALICE (A Large Ion Collider Experiment) is the heavy-ion detector designed to study the physics of strongly interacting matter and the quark-gluon plasma at the CERN Large Hadron Collider (LHC). A large bandwidth and flexible Data-Acquisition System (DAQ) is required to collect sufficient statistics in the short running time available per year for heavy ion and to accommodate very different requirements originated from the large set of detectors and the different beams used.

The DAQ system has been designed, implemented, and intensively tested. It has reached maturity and is being installed at the experimental area for tests and commissioning of detectors. It is heavily based on commodity hardware and open-source software but it also includes specific devices for custom needs. The interaction of thousands of DAQ entities turns out to be the core of this challenging project.

We will present the overall ALICE data-acquisition architecture, showing how the data flow is handled from the front-end electronics to the permanent data storage. Then some implementation choices (PCs, networks, databases) will be discussed, in particular the usage of tools for controlling and synchronizing the elements of this diversified environment. Practical aspects of deployment and infrastructure running will be covered as well, including performance tests achieved so far.

INTRODUCTION

The ALICE experiment [1],[2] consists of 17 detectors aiming at studying the interaction of particles colliding at the LHC. It primarily targets heavy-ion reactions, but it will also be able to cope with proton-proton and proton-ion collisions.

The Pb-Pb collisions will occur during a few weeks per year, characterized by big events (86.5 MB), large bandwidth to mass storage (1.25 GB/s), low interaction rate (10 KHz), and complex triggers. On the other hand, the pp and pA collisions will occur during several months per year and produce relatively small events (2.5 MB) at a high interaction rate (200 KHz), needing less bandwidth and simpler triggers with increased selectivity.

ALICE will collect physics data in a large set of detector configurations: detectors will be able to work altogether or separately, in stand-alone operation or synchronized data taking.

The ALICE DAQ system must cope with these heterogeneous requirements to transform the 25 GB/s aggregated throughput from the detectors into a set of recorded physics data files.

ARCHITECTURE

The data flow is controlled by the following online systems:

- The Trigger (TRG) is in charge of selecting the events, initiating the detectors read-out, and synchronizing the experiment with the LHC machine clock.
- The Data Acquisition (DAQ) is responsible for the data-flow from the detector electronics to the permanent storage, and for the control of this data-flow.
- The High-Level Trigger (HLT) provides filtering information to optimize the amount of interesting data in the available bandwidth.

Each system, described in details in a Technical Design Report [3], is made of several pieces of hardware equipments and software components interacting as described in Figure 1.

The Trigger system

The Central Trigger Processor (CTP) receives the input from the trigger detectors and the LHC clock. For every bunch crossing, and according to the busy status of all the detectors, the CTP produces trigger decisions which are transmitted to each detector via its Local Trigger Unit (LTU). The LTU converts these decisions into messages which are distributed to the detector electronics via the Timing, Trigger and Control (TTC) broadcast system. In ALICE, different types of triggers are generated, involving different sets of Front-End Readout Electronics (FERO) to be read-out.

The Data Acquisition

The readout electronics of all the detectors is interfaced to the ALICE standard Detector Data Links (DDL). The DAQ will use more than 500 DDLs, each link being able to transport about 200 MB/s over the ~100 meters long fibres, from the detector in the experimental area to a counting room close to the surface.

Each DDL is connected on one side to the FERO by a Source Interface Unit (SIU). At the receiving side of the DDLs there are PCI boards, called DAQ Read-Out Receiver Cards (D-RORC), hosted by PCs, the Local Data Concentrators (LDCs). Each D-RORC hosts two DDL interfaces and each LDC can handle one or more D-RORCs.

The data read out by the LDCs, are then shipped to another pool of computers, the Global Data Collectors (GDC), which build the full events. The load of GDCs is balanced by an Event Distribution Manager (EDM) process. The events are then temporarily stored by the GDCs on a local Transient Data Storage (TDS) system, residing on an independent storage network to avoid traffic congestions. The files are later on migrated to Permanent Data Storage (PDS) in the remote CERN computing centre.

The DATE software [4] is a distributed system consisting of several processes executed on every node of the DAQ system where it manages the dataflow and the control.

Some DAQ Services (DS) machines provide various central facilities to the other computers, including configuration database, publish and subscribe information servers, monitoring and logging repositories. These important nodes also run the DATE runControl, responsible for synchronizing all the DATE processes.

The High-Level Trigger

On most of the D-RORCs, one of the two DDL links is used to transfer the received detector data to the HLT system. This information is processed in a farm of computing elements, the Front-End Processors (FEP). They read data from the DDL via an HLT Readout Receiver Card (H-RORC). The analysis results and filtering decisions are injected back in the DAQ by another set of DDLs attached to dedicated LDCs.

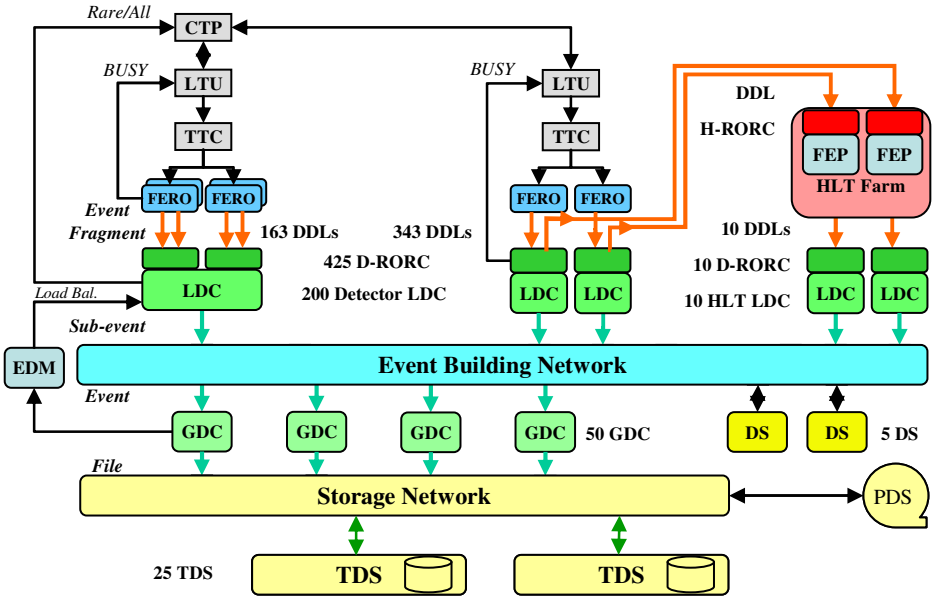

Figure 1: ALICE Data-Acquisition architecture

IMPLEMENTATION

The DAQ architecture is implemented by a combination of commodity items interacting with devices designed for specific purposes.

Transfer of data to the DAQ

Demanding requirements for the DDL imposed the development of custom hardware to meet the goals in terms of bandwidth and versatility.

The SIU is located inside the detectors; it works in a radiation environment and should therefore support a level of 13 Gy after 10 years of operation. Extensive tests have been performed to select radiation tolerant optical transceivers and FPGAs. It is necessary to minimize transmission faults, single event upsets and configuration losses. Successful results have been obtained [5], and a device of the Actel ProAsic 3 family will be used for the production series.

The D-RORC transfers data autonomously in the PC memory using DMA. This optimization avoids polling the PCI bus, leaving the 64-bit 66MHz bandwidth for data transfer. The D-RORC firmware includes facilities to generate data for testing purpose. The DDL Data Generator (DDG) is a hardware data source interfaced to the trigger via the TTC. This realistic emulation of the

detector electronics will be used during the commissioning phase.

Networking

The ALICE DAQ will perform the event building on a private network. The sub-events are transferred from the LDCs to the GDCs using the TCP/IP protocol through a standard Gigabit Ethernet network.

It was critical to qualify the selected router (Force10 E-1200) in real conditions. Figure 2 shows the performances achieved in October 2005, which largely meet the requirements. This hardware has demonstrated linearity in throughput as function of the number of data producers. The upper end of the plot shows a saturation on the consumers side (GDCs) when reaching the maximum wire speed.

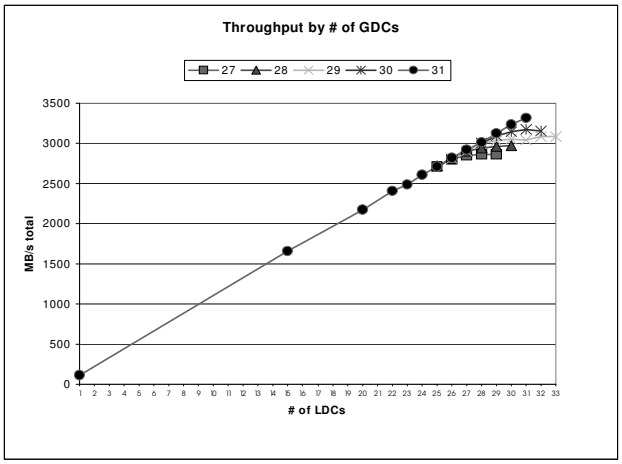

Figure 2: Event-building throughput

The GDCs are connected via a second interface card to another network to access the TDS. Several distributed file systems are being evaluated. The TDS will be implemented using disk arrays over Fibre Channel.

The uplink to the computer centre consists of a redundant pair of 10 Gb Ethernet, to provide a safe 20 Gb/s transmission link.

Access to the DAQ private network will be done through a few securely configured gateway machines. Some specific services (database, control) can interact transparently with hosts on external networks using IP routing on the gateways. The Linux IPtables software provides extended configuration options to select which traffic is allowed and to enforce access policies.

Computing infrastructure

All the DAQ computing equipment is installed in the same area, up in the pit near the surface. The space in the counting room being limited, a tight rack organisation has been chosen to fit the 300 computers needed for operation. Each of the 33 standard racks available can host a total of 56U devices in height. These devices consist of rack control units in charge of the temperature and smoke detection, Power Distribution Units (PDU) to control the electrical power in the rack, Keyboard Video Mouse (KVM) switches to remotely access the computers, network switches and routers, and finally the computers.

The total power consumption reaches 100kW, out of which 37 are on Uninterruptible Power Supply (UPS) having an autonomy of 10 minutes. The racks are equipped with special cooling doors (forcing a horizontal air flow through a radiator with cold water circulation) to evacuate the energy dissipated in this 70m^2 area. They are backed up by a legacy air conditioning system.

Control room

The experiment is operated from a separate surface control room equipped with workstations and monitors used by the shift crew. These computers run Linux and are equipped with several PCI dual-head display adapters (NVidia NVS-280 chips, selected for their good driver support). They allow a global view of all the applications controlling the experiment.

The X windows Xinerama mode has been selected for its transparent usage of windows across multiple screens. It is necessary to install enough RAM (1.5 GB), use the appropriate resolution (1280x1024), and disable fancy windows manager effects (shadings, animations, etc) in order to achieve good display performance on a 4 screens system. The display and geometry configuration is defined in the *xf86.config* file. Access to Windows machines from Linux has been tested with good performance in desktop operations using the Remote Desktop Protocol (RDP) client in 16-bit colour mode.

Computers

The base platform for the ALICE DAQ is Intel 32 bit PC running the Scientific Linux CERN (SLC) distribution. Different criteria have been taken into account to select the machines according to the DAQ tasks executed on them.

For the LDCs, the number of slots and architecture of the PCI busses is critical to fit as many DDL connections as possible in one machine. Figure 3 shows the data transfer speed, reaching a maximum of 1 GB/s on a PC

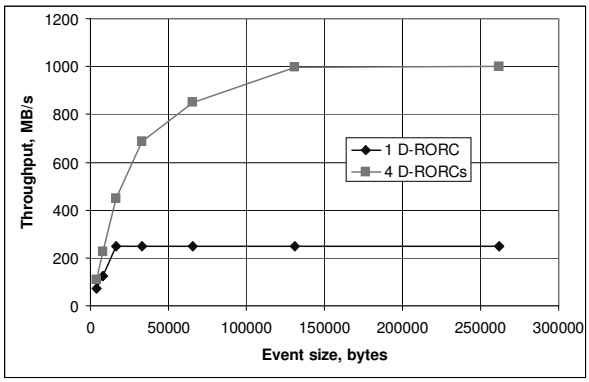

Figure 3: D-RORC data throughput

equipped with 4 D-RORCs saturating 4 PCI-64 slots.

For the GDCs, the overall CPU and memory performance is fundamental to build the events. As shown

in Fig. 2, the present generation of PCs is able to perform event-building at the maximum rate permitted by Gigabit Ethernet.

The best results for the DS servers running database applications have been obtained with computers using AMD Opteron Dual-Core 64-bit processors. Their outstanding performance, especially in multi-threaded applications, is due to the innovating memory bus architecture.

Figure 4 shows the results achieved by a test program with several threads concurrently inserting data in a MySQL table. The results achieved range from 10000 inserts/s on 2 x Xeon 2.8 GHz DDR 333 machines to 40000 inserts/s on 2 x Opteron Dual-Core 2.2 GHz DDR 400 systems. The type and organisation of the memory are indicated because critical in this mode of operation.

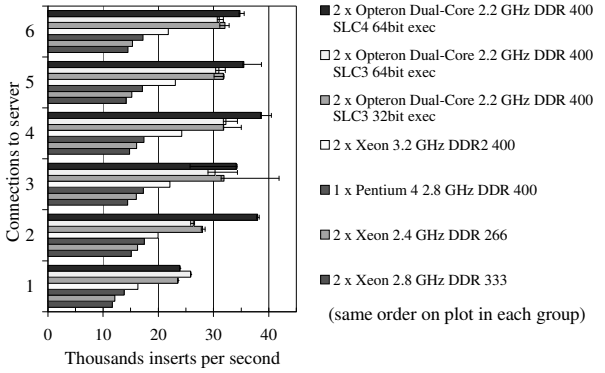

Figure 4: MySQL performance tests

The AMD Opteron Dual-Core 64-bit processors also provide an excellent compatibility with 32-bit code, which can run without recompilation, at the price of slightly slower speed.

Control and monitoring software

The core of the DAQ software is DATE described in a separate paper of this conference [6].

The Experiment Control System (ECS) [7] provides a unified view of the experiment and a central point from where all operations are initiated and controlled. It permits independent, concurrent activities on part of the experiment by different operators and coordinates the operations of the online systems (TRG, DAQ, HLT, and Detector Control System) within groups of detectors called “partitions”. Partitions can work independently.

The DATE runControl and ECS are based on Finite State Machines (FSM), describing the states and transitions of the software processes or experiment components, implemented using SMI [8]. The main FSM elements are running on the DS servers and communicate with remote items using the DIM protocol [9]. DIM is a publish-and-subscribe communication system based on TCP/IP.

DATE and ECS require a few open source software packages. A central CVS and RPM repository is used for the software code and configuration management. The Linux system and the DAQ application are installed over the network and automatically configured by dedicated scripts. AFFAIR (A Flexible Fabric and Application Information Recorder) is an integrated performance monitoring tool graphically reporting the status of system and DAQ metrics. The DATE infoLogger package provides facilities to generate, transport, collect, store and consult log and error messages. Data quality monitoring is insured by MOOD (Monitor of Online Data and Detector Debugger), with specific displays for each detector. These three flexible tools allow real time monitoring and tuning of the whole system.

MySQL has been the selected database to implement all the following features: store and distribute the DATE and ECS configurations, record the log messages, the experiment logbook and share data with other systems (e.g. the logbook run statistics with the Offline system, the list of active DDLs with the HLT). The reasons for using MySQL are good performance, minimal administration overhead, and open-source access.

CONCLUSION

The ALICE DAQ architecture is a distributed system of hundreds of hardware and software components, coordinated by a state-machine based control system. This design has been successfully used in a number of test-beams and Data Challenges. It has proved to be flexible enough to cope with the large range of configurations required for the experiment. The appropriate use of general purpose elements and implementation of ad-hoc solutions has permitted to achieve the required performance inside the experimental environment. The selection of the first batch of computers for the fabric is being finalized, and the installation for the final commissioning is under way. The full performance will be reached gradually by deploying the system in stages in order to benefit from the newest computing technologies.

REFERENCES

[1] ALICE Collaboration, Technical Proposal, CERN-LHCC-1995-71.

[2] http://aliceinfo.cern.ch/

[3] ALICE Collaboration, The Technical Design Report of the Trigger, Data-Acquisition, High Level Trigger, and Control System, CERN-LHCC-2003-062, January 2004.

[4] http://ph-dep-aid.web.cern.ch/

[5] E. Dénes et al., Radiation Tolerant Source Interface Unit for the ALICE Experiment, *Proc. of the Workshop on Electronics for LHC Experiments (LECC 2005)*, Heidelberg, Germany, 12-16 September 2005, 291-293.

[6] K. Schossmaier et al., The ALICE Data-Acquisition System DATE V5, CHEP06.

[7] http://alice-ecs.web.cern.ch

[8] http://smi.web.cern.ch

[9] http://dim.web.cern.ch