EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH
European Laboratory for Particle Physics

Publication
ALICE reference number
ALICE-PUB-2001-18 version 1.0
Institute reference number

Date of last change
2001-07-31

The ALICE Data Challenges

Authors:

J. P. Baud, W. Carena, F. Carminati, M. Collignon, F. Collin, R. Divia, J. D. Durand, S. Jarp, J. M. Jouanigot, B. Panzer, F. Rademakers, P. Saiz, K. Schossmaier, P. Vande Vyvre, A. Vascotto
for the ALICE Collaboration

The ALICE Data Challenges

J.P. Baud[1], W.Carena[2], F.Carminati[2], M.Collignon[1], F.Collin[1], R.Divia[2], J.D.Durand[1], S.Jarp[1], J.M. Jouanigot[1], B. Panzer[1], F. Rademakers[2], P. Saiz[2], K. Schossmaier[2], P. Vande Vyvre[2], A. Vascotto[2] for the ALICE collaboration
[1](CERN IT division, Geneva, Switzerland)
[2](CERN EP division, Geneva, Switzerland)

Abstract

Since 1998, the ALICE experiment and the CERN/IT division have jointly executed several large-scale high throughput distributed computing exercises: the ALICE data challenges. The goals of these regular exercises are to test hardware and software components of the data acquisition and computing systems in realistic conditions and to execute an early integration of the overall ALICE computing infrastructure. This paper reports on the third ALICE Data Challenge (ADC III) that has been performed at CERN from January to March 2001.

The data used during the ADC III are simulated physics raw data of the ALICE TPC, produced with the ALICE simulation program AliRoot. The data acquisition was based on the ALICE online framework called the ALICE Data Acquisition Test Environment (DATE) system. The data after event building, were then formatted with the ROOT I/O package and a data catalogue based on MySQL was established. The Mass Storage System used during ADC III is CASTOR. Different software tools have been used to monitor the performances. DATE has demonstrated performances of more than 500 MByte/s. An aggregate data throughput of 85 MByte/s was sustained in CASTOR over several days. The total collected data amounts to 100 TBytes in 100.000 files.

Keywords: AliROOT, CASTOR, CPU farm, DATE, disk server, Gigabit Ethernet, Mass Storage System, MySQL, ROOT

1 Introduction

The LHC Computing Review [1] recommended having data challenges of increasing size and complexity. The ALICE experiment and the IT division have jointly executed several data challenges. The goals of these regular exercises are to prototype the data acquisition and computing systems, to test hardware and software components of these systems in realistic conditions and to realise an early integration of the overall ALICE computing infrastructure. The third ALICE Data Challenge (ADC III) comprised prototypes of the ALICE data acquisition system, the ALICE offline computing, the CERN Tier 0-Tier 1 fabric, and the Mass Storage System.

The two most important functional goals of the ADC III were to reach a good stability of the system and of its performances over a period of a week and develop a set of online monitoring tools. The performance goals were to reach a stable bandwidth of 100 MByte/s over the complete chain for a week, an aggregate bandwidth of 300 MByte/s in the DAQ and a total of 80 TBytes of data stored in the Mass Storage System. The technologies that were planned to be used were: commodity PC's combined with a test of SMP servers and commodity networking for event building.

2 Testing of large computing fabrics

The LHC computing constitutes a big technological challenge for HEP. This challenge has to be addressed in a period of restricted funding situations. The expectation is therefore to realise the LHC computing fabric making extensive use of cheap commodity standard components. Several de-facto standard cheap components will contribute to this goal: PCs, IDE disks, the Linux operating system and switched Ethernet networking. A large part of the challenge

consists of building a high-performance, production quality and still manageable computing fabric with cheap computing and networking elements. For mainframes or supercomputers, the computer manufacturers mainly performed this integration of elements that he develops or buys from third parties. It is now largely the responsibility of the end-user to identify cheap components, to test them and to assemble them to realise a high performance computing fabric. The emergence of Linux as a de-facto standard in our environment constitutes an important ingredient of cheap computing fabric. Linux is a fantastic opportunity because it is an accepted standard and due to the creative and enthusiastic momentum that has been created around its development. Linux also introduces a complication compared to the mainframe situation because its road-map is uncertain. These factors render indispensable an early prototype of a fabric realised by integration of the above components. The prototyping work is an incremental process influenced both by the fast evolution of the technologies of its components and by the progress of the infrastructure software and the end-user applications. The ADC III is a typical example of test of a large computing fabric with end-users applications.

The objective of the ADC III regarding the network was to interconnect a processing farm of several dozens of machines, comprising CPU and Disk servers. The aggregate bandwidth needed inside the processing farm had to be of the order of 300 MByte/sec and the link to the Tape servers had to be in excess of 100 MByte/sec.

The network for the ALICE Data Challenge has been designed as an important step in the direction of the final set-up. In order to better understand the management and performance of a distributed switching system likely to be used in the coming LHC computing fabrics, a series of small switches has been used rather than one unique bigger central switch. It was also desired to mix various manufacturers and technologies and use multiple parallel links for interconnections. The two main Summit switches are connected to each other with four parallel Gigabit links to permit load sharing and provide slightly more than the required 300 MByte/s bandwidth. Different types of switches and one type of router from three different companies (Cabletron, Extreme Networks and 3COM) have been used. The gigabit Ethernet Network Interface Card (NIC) used is the NetGear GA620T.

3 Data flow

The event data used during the ADC III are simulated physics raw data of the ALICE TPC, produced with the ALICE simulation program AliRoot [2] before the beginning of the Data Challenge. These data were split into several sub-events and injected by several computers in the data acquisition fabric. The data acquisition software was based on the ALICE online framework called the ALICE Data Acquisition Test Environment (DATE) system [3].

The data flow architecture of DATE is organised along parallel data streams working independently and concurrently followed by an event builder stage where data are merged and eventually recorded as a complete event.

The computers running the DATE software are of two types:

- The LDC (Local Data Concentrator) is the front-end processor whose main purpose is to receive the data from the front-end electronics of a given detector (or section of a detector). The LDC data are injected on the event building network, through which they reach the GDC. During the ADC III, the LDCs were generating the data flowing through the system.
- The GDC (Global Data Collector) is a processor that performs the event-building function. It collects the various sub-events from the LDCs, puts them together and encapsulates them with the proper event structure. It also performs the recording function.

The LDCs and GDCs were implemented as CPU servers and the event building was performed using TCP/IP over the Fast/Gigabit Ethernet network.

The DATE run control was used for the overall control of the ADC III. All the activities attached to the runs such as the data formatting or the performance monitoring were started synchronously with the DATE run.

The data were then formatted with the ROOT I/O package [4] and a data catalogue based on MySQL was established. The Mass Storage System CASTOR [5] was managing the disk servers, tape servers and tape drives to provide a huge migrating filesystem. All the data archived during the ADC III were recorded as CASTOR data files.

4 Performances monitoring

The performance monitoring was one of the functional improvements of the ADC III. It was initially proposed to use the PEM software package for this purpose but it could not be made ready for the ADC III. An existing tool has therefore been used instead and a new one has been developed.

A farm-monitoring tool developed in the IT/PDP group is presenting at every moment the status of the farm with the main parameters (CPU load, network bandwidth). It also presents a history of these parameters.

DATE is instrumented with a monitoring system measuring on every node and at regular intervals a few critical variables of DATE itself (run number, number of events, data volume, rates) and of the computer system (system and user CPU load, network bandwidth). These measurements can be plotted and are also stored in a database (MySQL) to allow later display by web clients.

5 Performances

The maximum aggregate throughput reached in DATE was 550 MByte/s. This was obtained with a system of 13 LDCs and 13 GDCs and with data traffic of subevents between 50 and 60 kBytes. This test demonstrates the high level of performances of switched Ethernet networks and the stability of this network during long periods under heavy data traffic. It also shows the excellent level of interoperability achieved by the manufacturers of network components (NIC, switches and routers). A similar test performed with an ALICE-like data traffic has achieved 350 MByte/s of throughput. During all these tests, the CPU load on a LDC or a GDC running DATE has been measured to be of the order of 1.4 %/MByte/s on a dual-CPU PC.

The maximum throughput in DATE and ROOT was 240 MByte/s with data recording on the local disks of the GDCs. With the complete chain (DATE, ROOT and CASTOR), the maximum throughput was 120 MByte/s and 85 MByte/s on average during a week ($>$ 50 TByte/week). The 12 tape devices used could deliver a nominal installed bandwidth to tape of 120 MByte/s and 85 MByte/s are available to the end-user application over long periods. The IDE-based disk servers have shown a saturation with a simultaneous input and output traffic of 11 MByte/s. More work is needed to identify the sources (operating system, disk controller and driver, user application) of this bottleneck and to fix them. The total amount of data through DATE was 500 TBytes in DATE and 110 TBytes (100.000 files of 1 GBytes) were stored on mass storage. The metadata database was filled in with 10^5 entries.

One of the key design principles of the ALICE DAQ system is its capability to address needs over a large range of performances. The number of LDCs is fixed by the data sources (detectors) participating to the system. Likewise, DATE is currently also able to support multiple GDCs depending on the total aggregate bandwidth that the system has to support. This concept, known under the buzzword of scalability, has often be used and claimed. An important finding of a previous test was that this concept must be handled with extreme care. Each factor of two in the overall complexity of the system often put in evidence a new limitation

or saturation effect. The performance of different systems composed of a given number of LDCs in function of the number of GDCs is shown in Fig. 1.

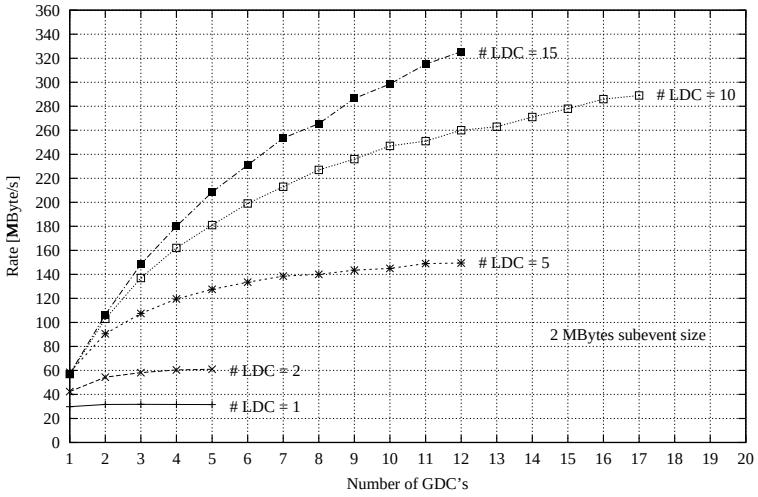

Figure 1: Aggregate bandwidth in function of the number of nodes

Two different types of saturation are visible. First, the current throughput limitation of an LDCs is of the order of 30 MByte/s. This limit is visible on each curve when the number of GDCs is large enough to be close to the maximum throughput generated by the LDCs. The second limitation is that the current maximum throughput of a GDC is 60 MByte/s. The measurement shows that this value is achieved for the first few GDCs added in the system. A complete presentation of all the results is available in [6].

6 Conclusion

During the third ALICE Data Challenge, the ALICE experiment and the IT division have achieved major performance milestones. The fabric has proven to be stable, and to deliver stable performance for periods of up to a week. The Mass Storage System has demonstrated to deliver 65 % of its installed nominal capacity over long periods. Future work is targeted at increasing this ratio. These results demonstrate the progress of ALICE towards the final DAQ and Computing systems using the three frameworks in use: DATE, AliRoot and ROOT. It also shows the progress of the CERN IT division in the direction of the large Tier 0/Tier 1 fabric.

References

[1] "Report of the Steering Group of the LHC Computing Review", CERN/LHCC/2001-004, February 2001.

[2] R. Brun, F. Carminati, F. Rademakers, "The ALICE Off-Line Strategy : A Successful Migration to OO", CHEP 2000, Padova, February 2000. Proceedings of .

[3] ALICE DAQ group, "The ALICE DATE User's Guide V3.7", ALICE Internal Note 2000-31.

[4] The ROOT team, "ROOT User's Guide 3.1", June 2001.

[5] J.P. Baud, "The CASTOR project status", CHEP 2000, Padova, February 2000.

[6] J.P. Baud et al, "ALICE Data Challenge III", ALICE Internal Note in preparation.