EUDET-Memo-2007-57-1

Calice Data Processing

R. Pöschl[*]

December 29, 2007

Abstract

The CALICE testbeam effort is largely applying standard software available for iLC Detector studies. The whole data analysis is made in a data format compatible with the general ILC Software. Moreover, existing frameworks such as marlin are employed for the CALICE software needs. The data management is entirely based on grid tools.

[*]LAL, Orsay, France

1 Introduction

Generally, the software concept for the processing of the CALICE data is driven by three main guidelines:

- Application of common ILC Software tools where possible and therefore benefiting from general developments of the ILC software. At the same time the application of these tools allow for the identification of the needs of the ILC Software for real data already at an early stage of the R&D phase.
- Since test beam data are taken at various locations they have to be available independent of the experimental site. Here, the grid was identified as the ideal tool to meet this goal. At the same time the grid offers significant computing resources needed for simulation and various re-processings of the data. In order to realize the data management and processing within the grid environment the virtual organization (vo) calice has been established. As of today this vo counts 60 members.
- As many users as possible are to get involved in the analysis effort. Therefore, entry points for an easy start-up for unexperienced users have to be provided.

2 Data Management

Data as acquired by the CALICE data acquisition system are stored on a local buffer at the experimental site. Locally, these so called native files are used for debugging purposes and online monitoring of the data quality. A server of type Dell Poweredge 1950 is used for the online monitoring. At the same time, the server is configured as a grid user interface and transfers the data to a mass storage, see Fig. 1, after the end of a run. The files are registered with their logical file names to the grid. By this, the data are available to the whole CALICE collaboration approximately 30 minutes after run end.
The primary mass storage is the DESY dCache system [1]. Here, roughly 35 TBytes are available for CALICE. This is the amount of data collected by the end of 2007 which includes the native data files, simulated files as well as files produced during the data processing steps described in the next section. In addition to the mass storage at DESY similar storage space has been allocated at other sites. Among these sites, there are major computer centers which will hold complete replicas of the data. In total, 13 sites support CALICE with storage and computing resources.

3 Data Processing Scheme

Note, that the CALICE DAQ system is optimized for a maximal acquisition speed. For technical reasons, information belonging to an event appears at different places in the data stream.

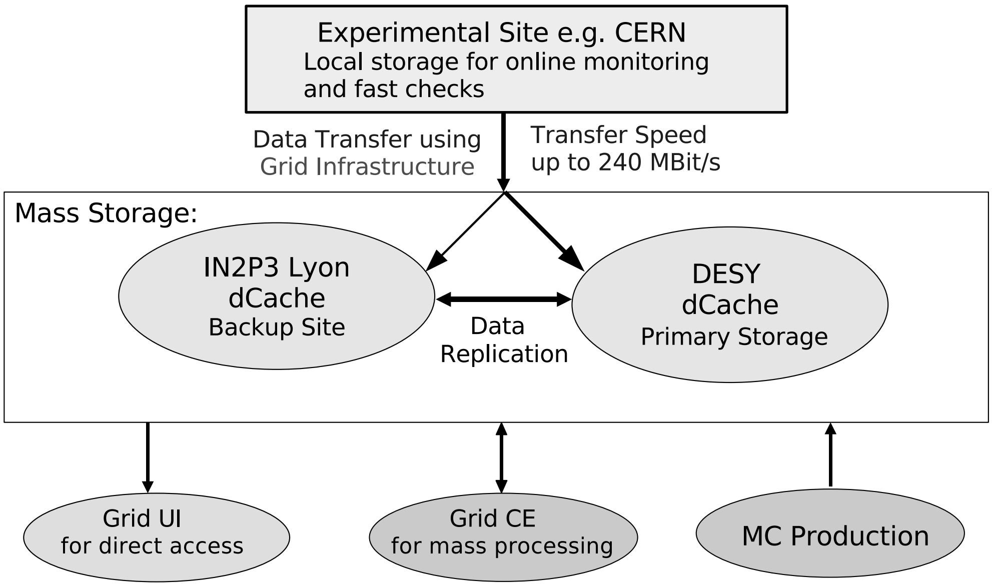

Figure 1: Data Management scheme as realized for CALICE Data. The DESY dCache is used as primary storage. Other main sites such as the Computing Center of the IN2P3 in Lyon/France hold replicas of the data or receive the primary copy in case of problems with the DESY mass storage during the data taking.

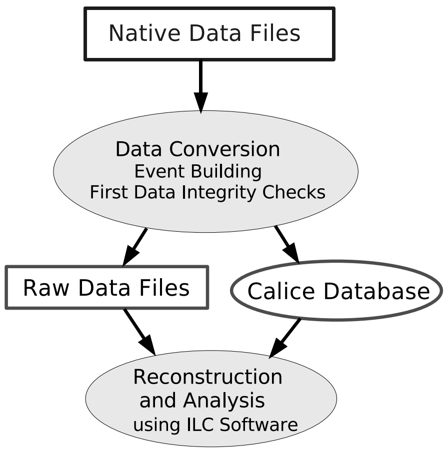

Figure 2: Overview on the processing chain for CALICE data.

To accommodate for that, the *native* data files are subject to a first processing step which acts as an event builder and provides first checks on the integrity of the data (see Figure 2). This step is also called the *conversion* step since here the native data are converted into a format compatible with the general ILC Software. As the data are read out without zero suppression the size per event is mainly driven by the number of cells in the setup. For each cell the ADC value is stored as a 16 bit value. During the data taking in 2007 the complete setup including the analogue hadronic calorimeter, a Silicon-Tungsten electromagnetic calorimeter and a Tail Catcher and Muon Tracker comprised approximately 17000 cells leading to an event size of about 34 kByte. By data compression, this size can be reduced by one third. During the conversion, a *database* is filled with detector configuration data and important conditions data such as e.g.

temperatures and voltage settings. This database holds also other conditions data such as calibration constants and details of the experimental setup. This conversion step is entirely made on the grid which allows for a parallel conversion of several runs and therefore for a fast availability of the data. The converted *raw* data are available two hours after the end of a given run. These files can be processed using regular ILC Software tools. The raw data are then subject to a *reconstruction* step of which the main output is calibrated calorimeter hits and tracks as measured by tracking devices which are part of the test beam setup. The latter files can be used to examine algorithms developed for a full ILC detector simulation on the measured test beam data. Vice versa, digitization algorithms developed for the test beam data and thus under realistic conditions, can be ported easily back into a full detector simulation.

4 Conclusion

CALICE has established a full chain of data processing based in ILC Software and grid tools. This way of organizing the software has lead to a wide spread analysis effort not restricted to the main detector experts. The strategy of using ILC Software tools will allow for a easy exchange between test beam results and results or algorithms obtained in general detector studies. This is in particular true for clustering algorithms which will be for the first time tested on data obtained with prototypes of ILC detectors. The CALICE collaboration will pursue this strategy throughout its running in 2007 and 2008. The experience obtained during the CALICE data taking has clearly revealed the need for a dedicated treatment of data types more related to hardware issues. For the next generation test beams the interface between the DAQ systems and the offline processing has to be better defined. First efforts into this direction are already undergoing.
As indicated above Conditions Data handling is a vital part of every experiment and the way how it is done in CALICE is clearly not optimal. The ILC community is herewith asked to allocate human and financial resources for the development of a well suited system for Conditions Data management.
In total the processing and analysis of the CALICE test beam data delivers important input to the development of the ILC software and computing environment with respect to technical aspects as well as for the development of Particle Flow Algorithms.

Acknowledgement

This work is supported by the Commission of the European Communities under the 6^{th} Framework Programme "Structuring the European Research Area", contract number RII3-026126.

References

[1] P. Fuhrmann V. Gülzow, "dCache Storage System for the Future", Conference Proceedings Europar, Dresden, 2006.
See also http://www.dcache.org/manuals/index.shtml