# TIER-1 AND TIER-2 REAL-TIME ANALYSIS EXPERIENCE IN CMS DATA CHALLENGE 2004

N. DE FILIPPIS*
Dipartimento Interateneo di Fisica dell'Universitá e del Politecnico di Bari e INFN, Bari, Italy
F. FANZAGO†
INFN, Padova, Italy
G. DONVITO, A. PIERRO, L. SILVESTRIS
Dipartimento interateneo di Fisica dell'Universitá e del Politecnico di Bari e INFN, Bari, Italy
A. FANFANI, C. GRANDI
Dipartimento di Fisica dell'Universitá di Bologna e INFN, Bologna, Italy
J.M. HERNÁNDEZ
CIEMAT, Madrid, Spain
D. BONACORSI
CNAF regional Centre of INFN, Bologna, Italy
M. CORVO
INFN, Padova, Italy

*Abstract*

During the CMS Data Challenge 2004 a real-time analysis was attempted at INFN and PIC Tier-1 and Tier-2s in order to test the ability of the instrumented methods to quickly process the data. Several agents and automatic procedures were implemented to perform the analysis at the Tier-1/2 synchronously with the data transfer from Tier-0 at CERN. The system was implemented in the LCG-2 Grid environment and allowed on-the-fly job preparation and subsequent submission to the Resource Broker as new data came along. Running job accessed data from the Storage Elements via remote file protocol, whenever possible, or copying them locally with replica manager commands. Details of the procedures adopted to run the analysis jobs and the expected results are described.

An evaluation of the ability of the system to maintain an analysis rate at Tier-1 and Tier-2 comparable with the data transfer rate is also presented. The results on the analysis timeline, the statistics of submitted jobs, the overall efficiency of the GRID services and the overhead introduced by the agents/procedures are reported. Performances and possible bottlenecks of the whole procedure are discussed.

## THE CMS DATA CHALLENGE 2004

During March-April 2004 the CMS collaboration performed a data challenge (DC04) which was a full-chain demonstration of data handling. The goals were to run CMS reconstruction for a sustained period at 25 Hz output rate from the online reconstruction farm, to distribute the data to the CMS Tier-1 centers and to analyze them in real-time at remote sites. The data challenge ran in a distributed LCG-2 Grid environment [1] in order to evaluate the robustness of the infrastructure in conditions as close as possible to that expected at CMS start-up [2].

The real-time analysis was designed to demonstrate that data could be analyzed as soon as they were transferred to a Tier-1 and to measure the time delay between the reconstruction at Tier-0 and the analysis at Tier-1/Tier-2s. Automatic data replication to Tier-2s was also achieved for offline analyses.

The CERN Replica Location Service (RLS) [3] provided the replica catalogue functionality for all the data distribution chains in DC04.

## THE REAL-TIME ANALYSIS STRATEGY

The real-time analysis was performed at Italian (INFN) and Spanish (PIC) Tier-1/Tier-2s. Software agents and automatic procedures were implemented to run the analysis in an LCG-2 Grid environment. Data were transferred from Tier-0 to Tier-1's, as described in Ref. [4] and were replicated by a dedicated agent to disk storage elements (SEs), as detailed in Ref. [5]. Whenever new files were available on disk the replication agent was also responsible to notify it with a drop box mechanism. The real-time analysis agent checked from the drop box for new files and triggered job preparation when all files of a given file-set (run) were available.

At INFN the data were made available for analysis on disk at CNAF Tier-1 and Legnaro Tier-2 depending on the event sample while at PIC all the data were analysed irrespective of their content.

* Nicola.Defilippis@ba.infn.it
† Federica.Fanzago@pd.infn.it

User interfaces (UIs) dedicated to analyses were setup at the previous sites. The CMS software was installed by a software manager at LCG-2 sites using a CMS distribution tool based on rpms. The CMS reconstruction and analysis program, ORCA [6] used COBRA program [7] as framework and POOL [8] as underlying persistency layer.

Physics Group oriented analyses of the PRS b/$\tau$ and muon samples [9] were performed. The ORCA executables and libraries for specific analyses of reconstructed (DST) data were provided by the PRS groups.

GridICE [10] was used as grid monitoring service at INFN and PIC. The CMS job monitoring was performed using BOSS tool [11].

### *The real-time analysis agent and the analysis job*

The real-time analysis agent was a cron job running on the UI performing the following tasks:

- getting input parameters (like analysis executable and dataset name) provided by the user via few configuration files;
- checking if a run (file-set) was ready to be analyzed. The availability of new data files was determined looking into the drop box. The information about the minimal set of files to be analysed was obtained using the COBRA *findcolls* command on files with collection container. This procedure extracted also the object identifiers of the collection, that had to be specified as input in the card of the ORCA analysis job. With the *findColls* command it was possible to use the virgin COBRA metadata files (containing no information about all the runs of a data sample), files necessary to read correctly data to analyze. The availability on a predefined SE of all the files of a run was checked querying the RLS. Also the relevant COBRA metadata distributed in a zipped format had to be available on the SE.
- preparing the job that is creating the scripts to run on the worker node and files needed by the analysis job (job script, card files) starting from a template and extracting the POOL catalogue of input data files querying the RLS. At the end a JDL file [12] defining the sandboxes, the requirements and the ranks in terms of input data was arranged to be submitted on the grid;
- submitting jobs through BOSS to the Resource Broker (RB). The input data into the JDL file drove the Broker to select computing resources with data stored or close to them.

The operations performed by the analysis job running on the worker node were:

- to setup the CMS environment to run ORCA in the LCG sites;
- to read input data from an SE via rfio whenever possible otherwise via replica manager commands;
- to download from an SE a zipped archive file with the COBRA metadata and their POOL xml catalog;
- to run the ORCA analysis executable;
- to stage the output file with histograms into an SE and register it into RLS.

### *Job monitoring and bookkeeping*

The GridICE grid job monitoring service stored general information about job submitted at LCG-2 sites. A CMS specific job monitoring was performed using the information of the BOSS database. Several set of information (called job-type) were stored in that database, some of them specific for the analysis (like the number of analyzed events, the kind of analysis executable, etc.) and the others LCG related (like the Replica Manager copy and the registration status of the output file).

A graphic interface (which used Qt libraries [13] and the MySQL API [14]) was also created to match the information about jobs over the Grid, using the Logging and bookkeeping component of the workload management system, with the BOSS data.

## STATISTICS AND RESULTS

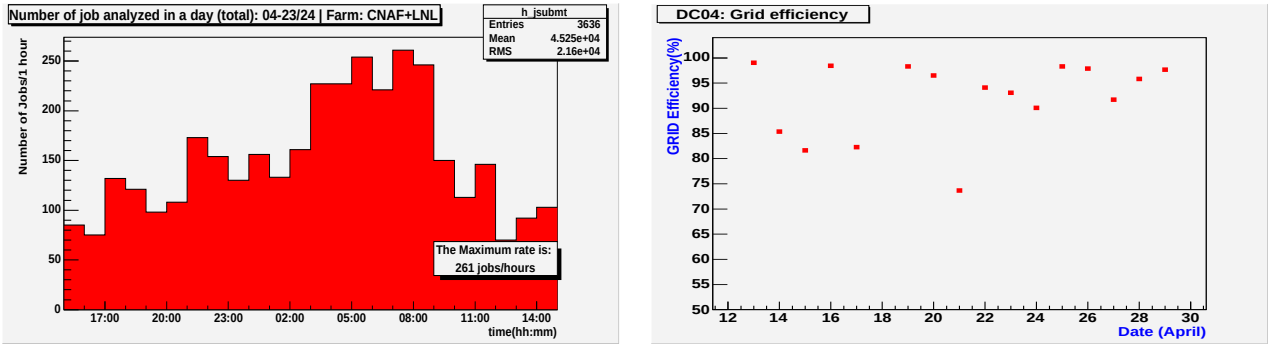
The real-time analysis at INFN started the 13th and ended the 29th of April.

The statistics and the results of the analyses and the behavior of jobs in LCG were derived using the job information stored in the BOSS database. The number of analysis jobs submitted at INFN sites was 15500: 8500 jobs at CNAF and 7000 at Legnaro.

At PIC the machinery was setup only at the end of DC04 over the last 78 hours uninterruptedly, analyzing all data coming from CERN. The number of analysis jobs submitted at PIC was 2000.

In Figure 1 a) the distribution of the job rate for the 23rd of April is shown. The rate was determined from BOSS counting the jobs having an end execution time falling in a given hour interval, taking into account all the datasets being analyzed at CNAF and Legnaro. The maximum rate of analysis jobs was about 260 jobs/hour. Taking into account that the number of events per job varied from 250 to 1000, depending on the dataset, the rate of jobs is translated into a rate of analyzed events which was of about 40 Hz at its maximum.

The Grid efficiency is defined as the percentage of finished jobs among all the jobs submitted to the Resource Broker. The Grid efficiency obtained in each day of running is reported in Figure 1 b); it was around 90-95% during the whole two weeks period except for few days due to the following problems: intermittent network problems at CNAF (April 14th and 15th), the RB disk being full causing the RB unavailability (nights of April 20th and




a) b)

Figure 1: a) Distribution of the number of finished jobs in an hour time interval, the job rate, computed for the 23rd of April. b) Efficiency of analysis jobs submitted to the Grid at INFN all over the two weeks period of running.

21st), the disk space of the UI being full (night of April 22nd), the Legnaro site disappeared from the Information System once (April 22nd-23rd).

The time spent by an analysis job varies depending on the kind of data and specific analysis performed, in any case the DST analysis are not very CPU intensive ranging from few to 30 minutes per job. The Grid initial overhead, defined as the difference between the job submission time and the time of start execution, was on average around 2 minutes, as shown in Figure 2 a).

A cross check of the job-level monitoring data in the BOSS database with GridICE was performed using the procedures described in Ref. [15]. The comparison was based on the number of submitted and running jobs according to both GridICE and BOSS. The information about jobs running in a farm with a start execution time in a given period was extracted from BOSS. GridICE is sensitive to transitions in the number of running and queued jobs on a farm. The time profiles of the number of running jobs as derived by GridICE and BOSS generally agree quite well, as shown in Figure 2 b) on the 23rd of April at CNAF.

The time delay between the appearance of the files at Tier-0 and their arrival on disk SE at Tier-1 (or Tier-2) was around 15 hours on average at CNAF; this large value was related to the tuning of the replication agent copying data to disks at Tier-1/2 and the replica agent operation affected by the problems on the CNAF tape stager. The time delay between the data availability on disk at Tier-1/2 and the start of analysis job was about 10 minutes at minimum. Instability in the RB, UI and analysis agents contributed to the time spread observed.

Better results about the timeline were obtained at PIC in the last 78 hours of DC04. The time delay between the availability at Tier-0 of a file and the analysis at PIC was 20 minutes on average, as shown in Figure 3. The minimum time was around 5 minutes. The main contributions to that time were:

- the time of transfer of the file from the Tier-0 to the Tier-1 that was 13 minutes on average;
- the time for replication from the CASTOR SE to the disk SEs that was less than 1 minute;
- the time for job preparation that was about 1.5 minutes;
- the time for job submission, including the local copy of the metadata files, that was about 3 minutes;
- the overhead for submitting to the grid the simplest job that was about 2 minutes.

## CONCLUSIONS

Real-time analysis at LCG Tier-1/2 consists of two weeks of quasi-continuous running of about 17500 of analysis jobs and a maximum rate of analyzed events reached of 40 Hz. The grid efficiency was larger than 90%. The average delay from data at Tier-0 to their analysis at Tier-1 was 20 minutes.

The data chain successfully met the DC04 goals of large scale file distribution to a set of destinations and subsequent analysis.

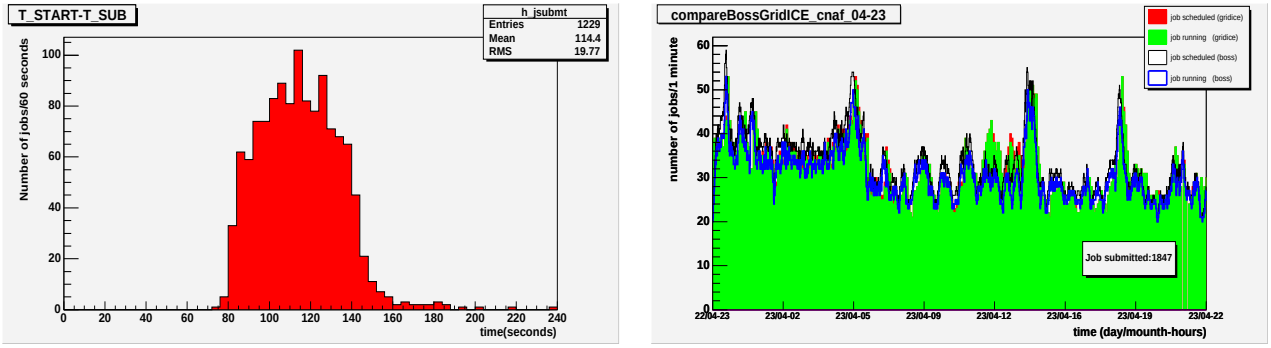
## ACKNOWLEDGEMENTS

We would like to acknowledge help and support of the LCG-EIS, LCG Deployment and CERN IT-DB teams. We would like to thank N. Magini, S. Cucciarelli, F. Ambroglini (from PRS/btau group) and M. Zanetti (from PRS/muon group) for the effort in providing the DST analysis code used in the real-time analysis at INFN. We thank S. Andreozzi, S. Fantinel, and G. Tortone for the grid monitoring with GridICE. At CERN we thank V. Innocente for his useful suggestions. We thank all the site managers at CNAF, Legnaro, PIC and CIEMAT, too.

## REFERENCES

[1] LCG Project: http://lcg.web.cern.ch/LCG

[2] A. Fanfani et al. ”Distributed computing Grid experiences in CMS DC04”, presented at CHEP04, Interlaken 2004.




a) b)

Figure 2: a) Time delay introduced by the grid measured as the difference between the job submission time and the time of start execution. b) BOSS and GridICE distributions of the number of scheduled and running jobs on the 23rd of April at CNAF.

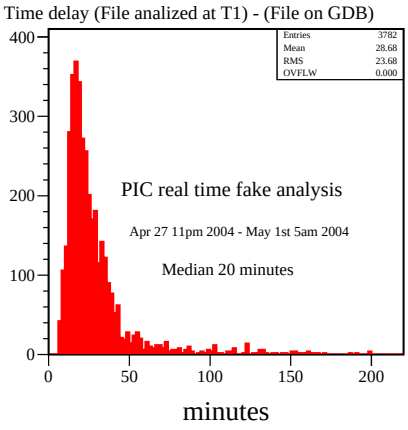



Figure 3: Real-time analysis turnaround time: time elapsed from the availability of a file for distribution at Tier-0 and the data analysis at PIC.


[3] User Guide for the EDG Local Replica Catalog 2.1.x http://edg-wp2.web.cern.ch/edg-wp2/replication/docu/r2.1/edg-lrc-userguide.pdf

[4] T. Barrass et al. ”Software agents in data and workfbw management” presented at CHEP04, Interlaken 2004.

[5] D. Bonacorsi et al. ”Role of Tier-0, Tier-1 and Tier-2 regional centers in CMS DC04”, presneted at CHEP04, Interlaken 2004.

[6] ORCA project: http://cmsdoc.cern.ch/orca

[7] V. Innocente ”COBRA - Coherent Object-oriented Base for Reconstruction, Analysis and simulation”, http://cobra.web.cern.ch/cobra/

[8] POOL (Pool Of Persistent Objects for LHC) Project: http://lcgapp.cern.ch/project/persist/

[9] http://cmsdoc.cern.ch/prsall.html

[10] S. Andreozzi et al. ”GridICE: a monitoring service for the Grid”, Presented at 3rd Cracow Grid Workshop, October 2003, Cracow, Poland. Available at http://server11.infn.it/gridice/.

[11] C. Grandi, A. Renzi ”Object Based System for Batch Job Submission and Monitoring (BOSS)”, CMS NOTE-2003/005 BOSS Home Page, http://www.bo.infn.it/cms/computing/BOSS/

[12] Job Description language HowTo. December 17th, 2001 http://server11.infn.it/workload-grid/docs/DataGrid-01-TEN-0102-0_2-Document.pdf

[13] http://www.trolltech.com/

[14] http://www.mysql.com/

[15] T. Coviello et al. ”Combined analysis of GRIDICE and BOSS information recorded during CMS-LCG0 production”, CMS Note 2004/XXX waiting for final approval.