View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Scientific Open-access Literature Archive and Repository

IL NUOVO CIMENTO Vol. 33 C, N. 6 Novembre-Dicembre 2010
DOI 10.1393/ncc/i2011-10805-2

Colloquia: IFAE 2010

Distributed computing in the LHC era

M. Paganoni
Università di Milano-Bicocca and INFN, Sezione di Milano - Milano, Italy

(ricevuto l'8 Ottobre 2010; pubblicato online il 15 Febbraio 2011)

Summary. — A large, worldwide distributed, scientific community is running intensively physics analyses on the first data collected at LHC. In order to prepare for this unprecedented computing challenge, the four LHC experiments have developed distributed computing models capable of serving, processing and archiving the large number of events produced by data taking, amounting to about 15 petabytes per year. The experiments workflows for event reconstruction from raw data, production of simulated events and physics analysis on skimmed data generate hundreds of thousands of jobs per day, running on a complex distributed computing fabric. All this is possible thanks to reliable Grid services, which have been developed, deployed at the needed scale and thouroughly tested by the WLCG Collaboration during the last ten years. In order to provide a concrete example, this paper concentrates on CMS computing model and CMS experience with the first data at LHC.

PACS 07.05.-t – Computers in experimental physics.
PACS 29.50.+v – Computer interfaces.

1. – The computing challenge at LHC

The high level-1 trigger rate and the large event size make the data treatment for LHC experiments a real, unprecedented computing challenge, as shown in fig. 1. The choice for facing it relies on a worldwide infrastructure of Grid sites that are part of WLCG (Worldwide LHC Computing Grid) [1, 2]. More than 150000 CPU cores and 50 petabytes of disk space are presently available in about 140 sites, classified in a multi-tiered hierarchy as proposed by the MONARC working group [3]. Typically 20% of the resources are placed at CERN (Tier-0 and CAF), 40% in large-size national computing centers (Tier-1 sites) and 40% in medium-size regional computing centers (Tier-2 sites). Figure 2 shows on a world map the WLCG sites, operated both from EGEE [4] and OGS [5]. In the last years a continuous effort was devoted, through Data Challenges, to demonstrate that Grid services and resources scale well beyond what is needed at the LHC startup. These activities culminated in the WLCG Common Computing Readiness Challenge in 2008 (CCRC'08) [6]. Monitoring and reporting processes improved substantially the reliability of the Grid infrastructure, by extensively testing all functionalities

© Società Italiana di Fisica

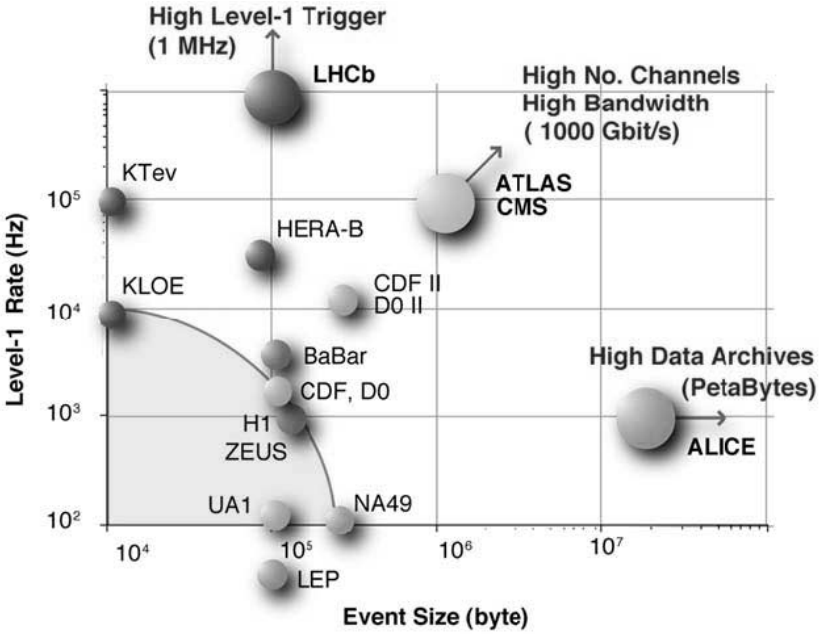

Fig. 1. – Level-1 trigger rate *vs.* event size for particle physics experiments.

of each Grid site before integrating it into the daily operations and by tracking its performance during the whole data-taking period. In this way LHC experiments provide their physicists communities easy and reliable data access and efficient data processing.

2. – The CMS distributed computing model

The CMS experiment has developed a distributed computing system capable of serving, processing and archiving the large number of events, amounting to several petabytes

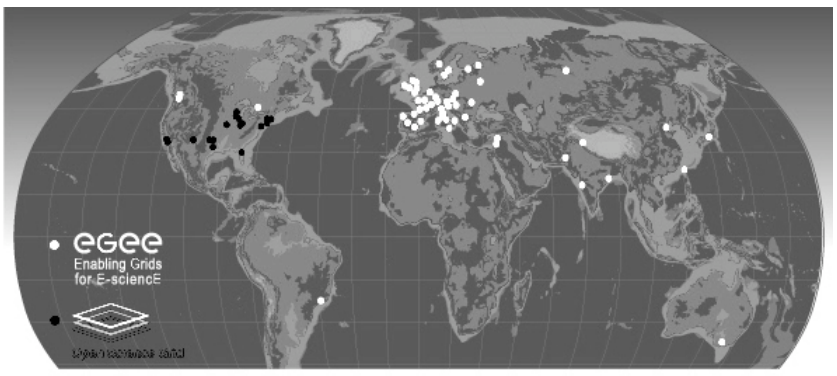

Fig. 2. – WLCG sites indicated on a world map.

TABLE I. – *Computing resources presently available for CMS.*

	CPU [kHS06]	Disk [PB]	Tape [PB]
Tier-0 and CMS CAF	55	3	9
Tier-1	100	11	20
Tier-2	192	12	-

per year, which is recorded during data taking. The CMS computing model [7, 8] foresees to place at CERN the activities with significant latency constraints, like express data processing, prompt feedback, alignment and calibration activities, storage of raw data backup. Seven Tier-1 sites (ASGC in Taiwan, FNAL in US, GridKA in Germany, IN2P3-Lyon in France, INFN-CNAF in Italy, PIC in Spain, RAL in UK) are responsible for serving a portion of the data and simulated events, for reprocessing and skimming activities and for the long-term custody of the data. For the first period of data taking 2-3 large reconstruction passes per year are foreseen. About 50 Tier-2 sites share their activities in equal parts between physics analysis and generation of simulated events, in a number matching the collision data. In the CMS computing model the location of the data drives the activities on the sites as the jobs are steered where the data have been pre-located. Table I summarizes the computing resources presently available for CMS worldwide.

3. – The main CMS workflow components

The CMS Workflow Management system (WM) manages the large-scale data processing that is optimized to let the final users access efficiently the data streams and perform their physics analysis. It supports all CMS necessary workflows (data re-reconstruction, calibration activities, Monte Carlo production, AOD production, skimming and physics analysis) and shields users from the full complexity of the underlying architecture. The CMS WM makes use in a coherent way of

- the CMS Grid Workload Management System (Grid WMS) which schedules jobs onto the distributed resources according to the CMS policy and priorities and monitors the jobs status;
- the CMS Data Management system (DM) which allows to discover, transfer and access data sets of different kinds across the Grid.

In order to manage and track the datasets, CMS has developed its own data management services:

- DBS, the dataset bookkeeping service, knows how a group of files forms a dataset and maps the files into logical quantities called data blocks, the smallest quantity we expect to track in the data transfer system and the global data catalog.
- PhEDEx [9], a reliable and scalable dataset replication tool, guarantees managed and structured data flow, by monitoring the data transfer and data integrity at the level of the file blocks.

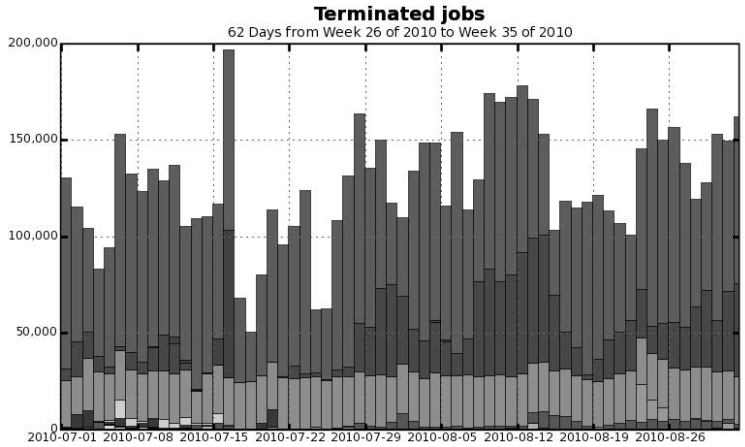

Fig. 3. – Number of terminated jobs per day, in all CMS Tier-2 sites during July-August 2010: testing jobs (light grey), simulated events production (black) and analysis jobs (dark grey).

Automatic functionality and reliability tests (SAM tests) are used in order to constantly monitor the quality of the sites. At the same time the Debugging Data Transfer (DDT) task force uses the PhEDEx LoadTest tool to generate data traffic transfer among sites and commission links between all CMS Tier centers. The full mesh, including cross-links between Tier-1 sites and upload/download links between Tier-1 and Tier-2 sites, has been commissioned.

During spring 2010 CMS moved more than 0.8 PB of custodial data from Tier-0 to Tier-1 sites, with daily-averaged aggregate CERN-outbound rates exceeding 650 MB/s. In the same period more than 3 PB were moved from Tier-1 to Tier-2 sites. Routes between Tier-2 sites, with daily-averaged rates as high as 0.8 GB/s, are proving to be very useful to optimize the overall transfer system and minimize the dataset transfer latency to a given destination.

4. – The CMS physics analysis at the Tier-2 sites

One of the most complex issues of the computing model is to ensure that data can be analyzed in an efficient way at the Tier-2 sites by a community of about 3000 physicists from all over the world. INFN has designed and developed CRAB (CMS Remote Analysis Builder) [10], a specific tool that allows end-users an easy and transparent access to the distributed data. CRAB has adopted a client-server architecture in order to automate the analysis workflow, leaving to the user just minor actions. In this way CRAB fully integrates into WLCG Grid infrastructure and into the CMS data and workload management system.

CRAB is in production and has been extensively used since Spring 2004. Peaks of daily submissions exceeding 100000 jobs/day for physics analysis at Tier-2 sites have been reached, as shown in fig. 3. The number of distinct CRAB users has exceeded 1500, with more than 500 individuals submitting jobs each day. CMS measured an average of about 80% success rate for jobs analysis, improving steadily towards the goal of more than 90% by optimizing the stage-out process of the produced output.

5. – Conclusions

CMS computing is working smoothly for LHC data taking at 7 TeV. All WLCG sites and the overall system are operating in a very reliable way and all the CMS workflows have proved to cope well with the data load.

REFERENCES

[1] Jacobs D. *et al.*, *WLCG Memorandum of understanding*, CERN-C-RRB-2005-01/Rev.
[2] *The Worldwide Computing Grid web portal*: `http://lcg.web.cern.ch`.
[3] Aderholz M. *et al.*, *Models of Networked Analysis at Regional Centres for LHC Experiments (MONARC) - Phase 2 Report*, CERN/LCB 2000-001 (2000).
[4] *Enabling Grids for E-sciencE web portal*: `http://www.eu-egee.org`.
[5] *Open Grid Science web portal*: `http://www.opensciencegrid.org`.
[6] Bauerdick L. and Bonacorsi D., *CMS results in the Combined Computing Readiness Challenge (CCRC08)*, *Nucl. Phys. B Proc. Suppl.*, **197** (2009) 99.
[7] The CMS Collaboration, *The CMS Computing Model*, CMS NOTE 2004-031.
[8] The CMS Collaboration, *The CMS Computing Project Technical Design Report*, CERN/LHCC 2005-023.
[9] Tuura L. *et al.*, *Scaling CMS data transfer system for LHC start-up*, *J. Phys. Conf. Ser.*, **119** (2008) 072030.
[10] Spiga D. *et al.*, *The CMS Remote Analysis Builder (CRAB)*, *Lect. Notes Comput. Sci.*, **4873** (2007) 580.