Use of Grid Tools to Support CMS Distributed Analysis

J. Andreeva, A. Anjum, T. Barrass, D. Bonacorsi, J. Bunn, M. Corvo , N. Darmenov, N. De Filippis, F. Donno, G. Donvito, G. Eulisse, A. Fanfani, F. Fanzago, A. Filine, C. Grandi, J.M. Hernández, V. Innocente, A. Jan, S. Lacaprara, I. Legrand, S. Metson, H. Newman, A. Pierro, L. Silvestris, C. Steenberg, H. Stockinger, L. Taylor, M. Thomas, L. Tuura, T. Wildish, F. Van Lingen

Abstract-- **In order to prepare the Physics Technical Design Report, due by end of 2005, the CMS experiment needs to simulate, reconstruct and analyse about 100 million events, corresponding to more than 200 TB of data. The data will be distributed to several Computing Centres. In order to provide access to the whole data sample to all the world-wide dispersed physicists, CMS is developing a layer of software that uses the Grid tools provided by the LCG project to gain access to data and resources and that aims to provide a user friendly interface to the physicists submitting the analysis jobs. To achieve these aims CMS will use Grid tools from both the LCG-2 release and those being developed in the framework of the ARDA project. This work describes the current status and the future developments of the CMS analysis system.**

I. Introduction

The Compact Muon Solenoid (CMS) [1] is one of the four particle physics experiments that will collect data at the Large Hadron Collider (LHC) being build at CERN. While the detector will not take data until 2007, hundreds of physicists around the world, members of the CMS collaboration, are currently taking part in computing intensive Monte Carlo simulation studies of the detector and its potential for uncovering new physics. The CMS collaboration has a long term need to perform large-scale simulation, in which physics events are generated and their manifestations in the CMS detector are simulated. These simulation efforts support detector design and the design of the real-time event filtering algorithms that will be used when CMS is running. Furthermore they provide a way for designing the reconstruction and analysis frameworks needed to process large amounts of events that will be available when the detector starts collecting data. The challenge for the CMS computing infrastructure is therefore to cope with the very large computational and data access requirements. The size of the resources required, the complexity of the software and the physical distribution of the CMS collaboration naturally imply a distributed computing and data access solution. The Grid paradigm is one of the most promising solutions to be investigated, and CMS is collaborating with many Grid projects around the world in order to explore the maturity and availability of middleware implementations and architectures. CMS decided to actively participate in the Grid projects since their outset, with the aim of understanding how the Grid can be useful for CMS and how CMS software needs to be adapted in order to maximize the benefit of using Grid functionalities and tools.

The preparation and building of the computing system, capable of managing the data being collected, pass through sequentially planned steps of increasing complexity, named data and physics challenges. The Data Challenge for CMS during the year 2004 (CMS DC04) was planned to reach a complexity scale equal to about 25% of that foreseen for LHC initial running. Its goal was to run CMS reconstruction at CERN for sustained period at 25Hz input rate, distribute the data to the CMS regional centres and analyse them at remote sites.

To meet this challenge a large simulated event production of about 50 million events was undertaken during the preceding months: the so-called Pre-Challenge Production (PCP). Within the PCP, prototypes of CMS distributed productions based on Grid middleware were deployed. The prototypes were based on early deployed systems of LCG [2] (LCG-0 and LCG-1), where most of the features used come from the EU implemented middleware, and on grid environment like Grid3[3] as used by the USMOP system [4] in the USA. Large scale productions were performed using these prototypes, demonstrating that it is possible to use them for real data production tasks [5].

During the CMS DC04, data distribution and data analysis at the remote sites ran in a prototype Grid environment using several LCG-2 tools. Automatic procedures were implemented to submit analysis jobs as new data came along and they could be integrated with the Grid services, with good performances,

A. Fanfani, C. Grandi are with the Physics Department of University of Bologna and INFN-Bologna viale Berti Pichat 6/2 , I-40127 Bologna, Italy (e-mail:fanfani@bo.infn.it).

T. Barrass, S. Metson are with Bristol University, United Kingdom.

J. Bunn, I. Legrand, H. Newman, C. Steenberg, M. Thomas, F. Van Lingen are with California Institute of Technology, USA.

J. Andreeva, F.Donno, A. Filine, V. Innocente, A. Jan, H. Stockinger (also INFN-Padova) are with CERN, Geneva, Switzerland.

J.M. Hernandez is with CIEMAT, Madrid, Spain.

N. De Filippis, G. Donvito, A. Pierro, L. Silvestris are with University and Politecnico of Bari and INFN-Bari, I-70126 Bari , Italy.

D. Bonacorsi is with INFN-CNAF, Bologna, Italy.

M. Corvo, F. Fanzago, S.Lacaprara are with Physics Department of University of Padova and INFN-Padova, Padova, Italy.

G. Eulisse, L. Taylor, L. Tuura are with Northeastern University, USA

A. Anjum is with National University of Science and Technology, Pakistan.

T. Wildish is with Princeton University, USA.

enabling CMS to gain experience on Grid-enabled data analysis and to identify potential bottlenecks.

The next challenge, due by the end of 2005, will be the preparation of the CMS Physics Technical Design Report that will require analysis of hundreds of TB of data. In order to provide access to the whole data sample and analysis services to all the worldwide dispersed physicists, CMS is developing a layer of software that uses the Grid tools to gain access to data and resources and that aims to provide physicists with a user friendly interface for submitting analysis job. A combination of generic Grid tools and specialized CMS tools will be required to manage the data and optimize the use of computing resources.

II. Analysis tasks

The main aspects of CMS distributed analysis can be decomposed into the following high-level tasks: data access, analysis strategy and tools, job monitoring.

A. Data Access

The task of providing to the end users the ability to access the data requires a Data Location service, to allow the users to discover what data exists and where, and a Data Transfer service to accommodate the needs of data distribution to multiple sites.

B. Analysis Strategy and Tools

An analysis strategy to efficiently access both data and computing resources has to be defined. CMS software distribution and installation at remote sites requires developing tools to cover the possible scenarios for software installation in a Grid environment, e.g. the need to execute private user code. Tools that act as an interface to the physicist to allow creation and submission of jobs to local and remote resources have to be provided.

C. Job Monitoring

Monitoring services are needed to monitor and archive information about site resources (collecting e.g. host-related metrics, network activity, disk-partition space, etc…). Services to monitor the data management and workload management system are also required. Application job monitoring should allow the user to extract job specific information and thus to monitor in real time the status of the application.

The analysis tasks covered during the CMS DC04 "scheduled" analysis (Physics Group's organized analysis) in the LCG environment and those currently being addressed to provide an end-to-end analysis system are described in the following sections.

Both the scheduled and the end-user analysis are done by the ORCA (Object-oriented Reconstruction for CMS Analysis) application [6] that uses the CMS COBRA framework [7]. The persistency layer used by the CMS framework is POOL [8].

III. Analysis in DC04

The main aspects of the CMS Data Challenge in 2004 were:

- Reconstruction of data in the CERN Tier-0 farm for sustained period at 25Hz
- Data distribution to Tier-1,Tier-2 sites
- Data analysis at remote sites as data arrive
- Monitor and archive resource and process information

The aim of the challenge was to demonstrate the feasibility of the full chain.

The reconstruction jobs were submitted to a computer farm at CERN and the produced data were stored on a Castor [9] stage area, so files were automatically archived to tape.

A. Data Access

A data distribution system was developed by CMS for DC04, built on top of available Grid point to-point file transfer tools, to form a directed and scheduled large-scale replica management system [10]. The distribution system was based on a structure of semi-autonomous software agents collaborating through the Transfer Management DataBase (TMDB). Several data transfer tools were supported: the LCG Replica Manager tools, native SRM (Storage Resource Manager) [11] and SRB (Storage Resource Broker) [12]. A series of "export buffers" at CERN were used as staging posts to inject data into the domain of each transfer tool. Agents at Tier-1 sites replicated files, migrated them to tape and made them available to associated Tier-2s. The final number of replicas at the end of DC04 was ~3.5 million. The data transfer (~6TB of data) to Tier-1s was able to keep up with the rate of data coming from the reconstruction at Tier-0 with good performances. The total network throughput was limited by the small size of the files being pushed through the system [13].

The LCG-Replica Location Service (RLS) [14] catalogue at CERN was used to locate the physical instances of the data files CMS-wide and to classify data as function of their POOL metadata. The transfer tools relied on the Local Replica Catalog (LRC) component of the RLS as a global file catalog to store the physical file locations. The Resource Broker queried the LRC to submit analysis jobs close to the data. The Replica Metadata Catalog (RMC) component of RLS was used as global metadata catalogue, registering the files attributes of the reconstructed data and querying it (by users or agents) to find logical collection of files. Roughly 570k logical filenames were registered in the RLS during DC04, each with typically from 5 to 10 Physical File Names and 9 metadata attributes per file (up to ~1 KB metadata per file). Some performance issues inserting and querying information, in particular concerning the metadata catalogue component of RLS, were identified. The time for the queries on metadata was too slow, requiring for example several hours to find all the files belonging to a given "dataset" collection.

B. Analysis Strategy and Tools

The analysis of the reconstructed data in real-time with their arrival was performed at several Tier-1 and Tier-2 sites, in

Italy and Spain, using the LCG infrastructure. A set of software agents and automatic procedures were developed to allow analysis job preparation and submission as new data came along [15]. The CMS software required for analysis (ORCA) was pre-installed across LCG-2 sites by the CMS software manager via Grid jobs. The ORCA analysis executable and libraries for specific Physics Group's analysis were sent together with the job. The LCG-2 Resource Broker was used to submit analysis jobs selecting the LCG-2 CMS resources hosting the data. The job output was stored on a data server and registered to the RLS and thus made available to the whole collaboration.

The analysis was run quasi continuously for two weeks submitting a total of more than 15000 jobs, with a job efficiency of 90-95%. During the last days of DC04 running an average latency of 20 minutes was measured between the appearance of the file at CERN and the start of the analysis job at the remote sites. The LCG submission system could cope with the rate of data coming from CERN.

C. Job Monitoring

MonaLISA [16]and GridICE [17] monitoring services were used to monitor the system, collecting detailed information about nodes and information of service machine (Resource Broker, Computing and Storage Element) with the possibility of notification in case of problems. The CMS specific job monitoring was done using BOSS [18]. An analysis specific job-type was defined to collect information like the number of analyzed events, the dataset being analyzed and other metrics.

IV. End-User Analysis

The current CMS activities concerning distributed analysis are in an R&D phase focused on providing an end-to-end analysis system. In general end-user analysis is a chaotic, non-organized task, carried on concurrently by many independent users that do not have a deep knowledge of distributed environment problematic. CMS is testing several prototypes of tools that act as an interface to the physicist.

PhySh (Physics Shell) is an end user shell with the aim of reducing the number of different tools and environments that the CMS physicist must learn to interact with to use the data and computing services. The interface to the user is modelled as a virtual filesystem interface, since filesystem interfaces are what most people are familiar with when dealing with their data. PhySh is an extensible "glue" interface among different services already present or to be coded, like locating physics data of interest, copying/moving of event data to new locations, accessing software releases/repositories, and so on. PhySh is based on the Clarens [19] Grid–enabled web service infrastructure. Clarens was developed as part of the Grid-Enabled Analysis environment (GAE)[20]. Clarens servers leverage the Apache web server to provide a scalable framework for clients to communicate with services using the SOAP and XML-RPC protocols.

A. Data Access

File-based data access is not satisfactory from the perspective of the end-user, since the user wants to access collection of files (dataset) rather than single files. CMS specific dataset catalogues to describe the dataset characteristics and allow for locating of replicas of datasets are under development: RefDB and PubDB.

The Reference Database (RefDB)[21] is a Dataset Metadata Catalogue containing information like which (logical) files is a dataset made of, the dataset status and so on. The Publication Database (PubDB) manages information about dataset catalogues, allowing the users to locate the data of a dataset and to know how to access them. The information in PubDB can be used by users or Workload management system to decide where to submit the job analyzing the data. The natural design is to have distributed PubDBs, one per site serving the data, allowing the site data-manager to manage their local catalogues coherently and in a uniform manner with the other CMS sites. The user can discover the available datasets querying RefDB. A global map of all dataset catalogues is held in RefDB through the links to the various PubDBs.

The CMS bulk data transfer management system is PhEDEx (PHysics Experiment Data EXport) [22] [10]. PhEDEx is a project born of CMS' experience during DC04. It retains the same architecture, relying on a central blackboard to enable the exchange of information between a series of distributed agents. The principal current aim of PhEDEx is to incorporate the CMS distribution use cases of subscription pull of Monte Carlo data, where a site subscribes to all data in a given set and data is transferred as it is produced, and of random pull, where a site or individual physicist just wishes to replicate an existent dataset in a one-off transfer.

B. Analysis Strategy and Tools

The current approach in CMS is to concentrate on simple analysis scenarios and learn from the implementation of simple use cases. The adopted analysis strategy using the Grid is to submit analysis jobs to the data, as described in Fig.1. The end-user typically wants to access a dataset in order to analyze it with his private code. The user on the User Interface provides the dataset and the private code, and a combination of specialized CMS tools and Grid components should take care of resource matching and submission. The Workload Management system [23] finds suitable computing resources (i.e. Computing Elements) to execute the job. Data discovery is one of the most important aspects in the match-making process. A Data Location Interface is being developed to allow a uniform "query interface" to locate the data. This interface will provide to the Workload Management System the functionality to query several catalogues: the LCG Replica Location Service to perform file-based data location; the experiment specific dataset catalogues to perform dataset-based data location; or any upcoming data catalogue.

Several user-friendly tools dealing with job preparation, job splitting and job submission are under development. These

tools are being integrated with middleware and tools already available and on new one being developed in several Grid projects: LCG, EGEE [24], Grid3, OSG[25].

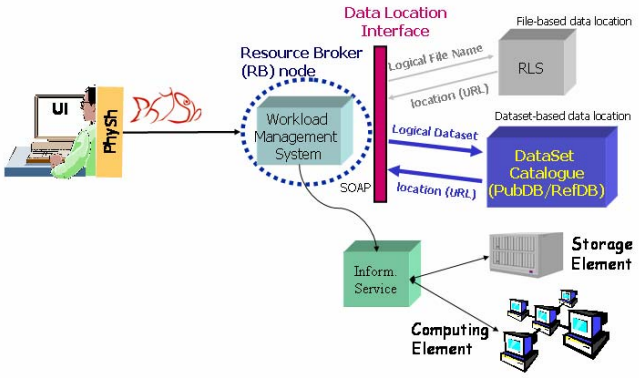

Fig. 1. Description of the analysis job submission in Grid. The user on the UI provides dataset and private code. The Workload Management System discovers the data location, via the Data Location interface to several catalogues, and submits the jobs on resources hosting the data.

C. *Job Monitoring*

MonaLISA, GridICE and BOSS can be used to monitor the system facilities and services, and the application job monitoring that is particularly relevant for the end user.

V. CONCLUSIONS

CMS is exploring the maturity and availability of middleware implementations and architectures of many Grid projects to provide access to the whole data sample and to required services to all the worldwide dispersed CMS physicists.

In CMS' Data Challenge 2004 the LCG environment provided the functionalities for distributed computing: global file and metadata catalogues, Grid point-to-point file transfer tools and infrastructure for data analysis. The major issues were related to the performances of data catalogues. Despite these problems, an average latency of 20 minutes was measured between the appearance of the file at CERN and the start of the analysis job at the remote sites.

Current work is focused on developing a layer of software that uses the Grid tools to gain access to data and resources and that aims to provide physicists with a user friendly interface for submitting the analysis jobs. Feasibility studies and development of prototypes are underway to provide a consistent interface to the physicist, a flexible application framework and a set of back-end services. This activity includes components from several Grid projects like LCG, EGEE, GRID-3, Clarens.

VI. REFERENCES

[1] CMS Experiment: http://cmsdoc.cern.ch/cms
[2] LHC Computing Grid Project: http://lcg.web.cern.ch/LCG
[3] I. Foster et al. "The Grid2003 Production Grid: Principles and Practice", 13th IEEE Intl. Symposium on High Performance Distributed Computing 2004 and references therein.
[4] The MOP Project, http://www.uscms.org/s&c/MOP
[5] A. Fanfani et al. "Distributed Computing Grid Experiences in CMS DC04", proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, Interlaken Switzerland 2004.
[6] ORCA (Object-oriented Reconstruction for Cms Analysis): 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] Castor (CERN Advanced STORage manager): http://cern.ch/it-div-ds/HSM/CASTOR/
[10] T. Barrass "Software agents in data and workload management" proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, Interlaken, Switzerland 2004
[11] SRM Project: http://sdm.lbl.gov/srm-wg
[12] A. Rajasekar et al., "SRB, managing distributed data in a Grid", Computer Society of India Journal, special issue on SAN, 33(4):42-54, 2003
[13] D. Bonacorsi "Role of Tier-0, Tier-1 and Tier-2 Regional Centers during CMS DC04",proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, Interlaken, Switzerland 2004
[14] D. Cameron et al "Replica Management in the European Data Grid Project" Journal of Grid computing 2004 In Print.
[15] N. De Filippis et al "Real-time analysis at Tier-1 and Tier-2 in CMS DC04" proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, Interlaken, Switzerland 2004
[16] I.C. Legrand et al, "MonALISA: a distribute monitoring service architecture", proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, La Jolla, California 2003
[17] S.Andreozzi et al "GridICE: a monitoring service for the Grid", 3rd Cracow Grid Workshop, Oct 2003
[18] C.Grandi, A.Renzi "Object Based System for Batch Job Submission and Monitoring (BOSS)", CMS NOTE-2003/005
[19] C. Steenberg and E. Aslakson, J. Bunn, H. Newman, M. Thomas, F. Van Lingen "The Clarens Web service architecture" proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, La Jolla 2003, California
[20] Grid-Enabled Analysis : http://ultralight.caltech.edu/gaeweb
[21] V. Lefebure, J. Andreeva "RefDB: The Reference Database for CMS Monte Carlo Production", proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, La Jolla, California 2003
[22] CMS PhEDEx: http://www.cern.ch/cms-project-phedex
[23] Workload Management System (WMS) in EDG and LCG-2: http://server11.infn.it/workload-grid/
[24] Enabling Grid for E-sience in Europe (EGEE)Project: http://public.eu-egee.org/
[25] R. Pordes et al. "The Open Science Grid", proceedings at Computing in High Energy and Nuclear Physics (CHEP) conference, Interlaken, Switzerland 2004