iAvailable on CMS information server

CMS CR 2006/080

CMS Conference Report

15 October 2006

The first year of LHC Physics Analysis Using the Grid: Prospects from CMS

I. Fisk

Fermi National Accelerator Laboratory, Batavia, IL. USA

Abstract

The CMS computing model has been distributed since early in the experiment preparation. In order for the experiment to succeed, CMS needs to develop efficient distributed analysis techniques using grid services. CMS has an active program of development and deployment to ensure the experiment can perform analysis using a worldwide infrastructure of computing clusters already at the beginning of LHC operation. In this presentation the status, plans, and prospects for CMS analysis using the grid are outlined.

1 Introduction

1.1 The CMS Computing Model

CMS has had a distributed computing model from early in the experiment planning. The decision of the experiment was motivated by a variety of factors. The large quantity of data the experiment will collect requires a corresponding investment in computing, storage, and data serving infrastructure. From an infrastructure point of view, it is difficult to construct a single centralized facility that can support all four LHC experiment while wide distribution can share the load. In addition, the distribution of computing to university and national laboratories allows hardware, infrastructure, and expertise to be leveraged.

The current proposal in the CMS computing model implies that 20% of the resources will be located at CERN. The Tier-0 resources include primary reconstruction facilities and a small number of analysis resources for calibration and other computing activities with significant latency constraints. About 40% of the computing resources are located at 7 Tier-1 computing centers. These facilities are primarily located at national labs and are responsible for archiving and serving a portion of the data and simulated events. They also perform re-reconstruction on data events and provide computing resources for skimming and selection. The final 40% of the computing resources are located at Tier-2 computing centers. The Tier-2 centers are the primary location for analysis activities and are predominantly located at universities. There are not sufficient resources at any one center to complete the analysis tasks of the experiment. Grid analysis will have to succeed in CMS from the opening of the experiment. [1]

In the CMS model the location of the data drives the activities on the sites. This leads to a distributed computing system in which the activities and required functionalities for each site are largely predictable. The experiment data is divided into streams based on trigger information and hosted at experiment specified sites. In this model opportunistic computing is largely restricted to activities like event simulation that are CPU bound and make output storage requirements, but input is restricted to parameters and conditions information. The proposed model is not as transparent or as flexible as some earlier grid models, but it is a system CMS believes can be built. It does not prevent more dynamic computing models from evolving, if functionality and capacity are available.

1.2 Simplifying Constraints

In order to make realistic requirements to the existing grid services and to ensure that data can be analyzed using distributed computing centers from the start of the experiment, CMS is applying constraints to the grid enabled sites to reduce the complexity. The first is simply that analysis jobs are only executed on sites that support the experiment. The sites that host analysis jobs also host some lightweight experiment services and configurations. The experiment data management services have local agents at the sites that control the resident data.

The second simplification is driving activities with data location. Processing requests are sent to sites with data and all data objects that will be accessed by an analysis job are fully specified at the time of job submission. In the baseline computing model, the replica of data for load balancing between sites is performed by policy and not automatically. For CMS all data access is made over local access protocols and does not assume the ability to automatically stage in data. The exception to this policy is the calibration information, which is accessible through read-only database caches that use web services to cache and refresh database queries.

The final simplification is an assumption about the consistency of the environment a processing request can expect when arriving at the remote site. The basic CMS software is installed on the site allowing only the user modifications to the standard distributions to be sent with the jobs.

2 Data Management for Analysis

In the CMS model analyses are performed on datasets. The dataset name is the contact used by analysis tools to specify the individual files that an analysis application will access, and subsequently the site or sites that would be able to satisfy the processing request. Datasets may be created by the experiment as event reconstruction, from analysis groups from complete data streams at Tier-1 centers, or user created skimming and selection jobs at Tier-1 or Tier-2 centers.

In order to manage and track the datasets, CMS has developed three data management services. A dataset is registered in the CMS dataset bookkeeping service (DBS). The global instance of this is a database with a defined service interface and authorized access for writing. In addition to the global instance local scope instances for individual users, groups, simulation, and reconstruction tasks exist. The DBS knows how a group of files forms a dataset. CMS files are anticipated to be 5-10GB on average, so a large dataset may run into the thousands of files. The DBS also maps the files into logical quantities called data blocks. The data block is the smallest quantity we expect to track in the data transfer system and simplifies the requirements to the global data catalog. Instead of tracking the location of every file, the experiment can track the location of the blocks, which are reduced in number by a factor of 100-1000. The blocks are registered into the Dataset Location Service (DLS) [2]. The DLS is currently based on the grid catalog technology Local File Catalog (LFC). The final data management service is the replication service called PhEDEx (Physics Experiment Data Exporter) [3]. PhEDEx moves data between sites and handles the subscription of datasets between sites. It relies on Storage Resource Manager(SRM) [4] to provide a consistent interface to the sites and the File Transfer Service (FTS) to perform the file transfer.

3 Specifying and Submitting Analysis Applications

Once a user has identified a dataset from the DBS and the corresponding data blocks have been located at a site from information in the DLS the analysis jobs must be specified and submitted. In July of 2005 CMS introduced the CMS Remote Analysis Builder (CRAB) [5]. CRAB was originally developed exclusively by INFN, though more recently has grown into a global effort with contributions from the US and the UK. CRAB provides a simple interface to the user to specify the dataset, the application, the input parameters and the number of events per process.

A user can query the DBS to determine available datasets. The current query capabilities are fairly primitive, but are improving. The identified dataset is defined by a number of data blocks. CRAB handles the interaction with the data management components, identifying the sites that serve the blocks and providing the user with the ability to preferentially select or veto sites. CRAB handles the job preparation by comparing the local user software environment to the reference software environment. The differences are assembled into an archive that is distributed with the job. Finally CRAB performs the job submission through the appropriate grid infrastructure. Specified jobs are sent either to the European DataGrid (EGD) developed resource broker for the European Grid for Enabling E-science (EGEE) [6] resources or Condor-G for the Open Science Grid (OSG) [7] resources. The resource broker has more functionality for matching resources, while Condor-G is faster for direct submission.

3.1 Calibration

In order to have successful analysis submission in a distributed environment CMS requires access to up-to-date calibration and alignment information for all sites processing analysis jobs. To meet this need CMS is deploying an infrastructure of read-only caches to the central Oracle conditions and calibration database. In this system entire DB queries are cached, which can be very efficient if queries are frequently the same. The system is called Frontier [8] and was developed for CDF distributed computing. It uses the same infrastructure of SQUID serves used for web site caching. While Frontier is still in testing, the performance numbers are promising and the experience with wide-scale deployment and operations continues to be good.

4 Status

Currently CRAB submission for user submitted analysis applications has reached more than 100 thousand jobs per month. The submissions tends to peak most strongly before Physics Technical Design Report submissions. While the majority of the access has so far has been to Tier-1 centers, in the computing model the bulk of analysis resources for individual users is located at Tier-2 centers. An increasing number of Tier-2 centers have successfully hosted data samples and have accepted analysis jobs. The demonstrated scale by users is a small fraction of the total number of jobs we expect to process at the beginning of physics running.

4.1 Future Directions

CMS expects to process approximately 100k jobs per day when the experiment is actively running in 2008. While this estimate has many uncertainties and is based on the number of active physicists and the number of times data is accessed during analysis development, it does provide a target goal for the submission infrastructure. CMS has seen a ramping up the number of jobs submitted through CRAB in the context of the WLCG Service Challenges. The currently sustainable rate is around 15k jobs per day through an infrastructure of 4 Resource brokers. We believe to meet the 2008 job submission goals we need to switch to the next generation of brokering infrastructure. Tests are on-going and there is significant validation work left to do.

In addition to achieving the higher scale there are a number of open areas of work to enhance the current infrastructure functionality. A frequent user comment is the lack of robust problem diagnosis and debugging. The job monitoring and tracking has improved in CMS, but users still have less information than in a traditional batch queue. This makes it difficult to determine where the failure occurred and hard to distinguish infrastructure failures from application or user problems. In general the experiment and grid infrastructures must continue to improve in reliability. The CMS goal for grid submission success this year is 90%, but much higher efficiency is desired in the final system.

Currently CMS only has very rudimentary capabilities to distinguish activities for prioritization. In order to manage experiment activities, CMS would like to be able to specify the utilization of resources based on the experiment scientific priorities. Some of the underlying grid technology exists do to at least coarse divisions, but the deployment in an operationally sustainable way is difficult.

5 Outlook

CMS is designing a grid analysis system where resource utilization is based on data location, which simplifies the decision on where applications should be run. Hopefully, it also makes realistic expectations of the current functionality of grid services. The first years of analysis are likely to be a strenuous test of the grid and experiment service infrastructure. In order for the experiment to succeed the distributed computing infrastructure must work from the beginning. We believe we have chosen a computing model that is deployable. Significant progress is being made, but there is a large ramp in scale, performance and reliability that must be achieved before the experiment begins real data analysis.

References

[1] .G.L. Bayatian*et al.* "CMS Computing Technical Design Report," CERN-LHCC-2005-0232005

[2] "Dataset Location Home Page" http://cmsdoc.cern.ch/cms/LCG/DLS/ 2006

[3] L. Tuura *et al.*, " PhEDEx high-throughput data transfer management system" *Prepared for Computing in High-Energy Physics (CHEP '06), Mumbia, India, 13 Feb. - 17 Feb 2006*

[4] T. Perelmutov, J. Bakken and D. Petravick, FERMILAB-CONF-04-473-CD *Prepared for Computing in High-Energy Physics (CHEP '04), Interlaken, Switzerland, 27 Sep - 1 Oct 2004*

[5] "CMS Remote Analysis Builder" http://cmsdoc.cern.ch/cms/ccs/ws/www/Crab/ 2006

[6] "EGEE Project Home Page " http://public.eu-egee.org/ 2006

[7] "Open Science Grid Project Home Page " http://www.opensciencegrid.org/ 2006

[8] S. Kosyakov *et al.*, "Frontier: High performance database access using standard Web components FERMILAB-CONF-04-367-CD *Presented at Computing in High-Energy Physics (CHEP '04), Interlaken, Switzerland, 27 Sep - 1 Oct 2004*