

LPC CAF Description

Ken Bloom, Oliver Gutsche,
David Mason, Liz Sexton-Kennedy

June 10, 2016

1 Introduction

The U.S. CMS community includes about 1000 scientists (roughly 28% of the collaboration), most of whom are involved in analyzing CMS data. The seven U.S. CMS Tier-2 sites that were selected in 2005 to provide U.S. CMS analysis facilities and other computing functions were sized to provide for the analysis needs of about 40 people each. As the collaboration grew and the size and complexity of the data grew, the Tier-2 centers also grew. However these seven Tier-2 centers, while important for the U.S. CMS physics analysis capabilities, have never been sufficient by themselves for the community. The U.S. CMS Software and Computing Operations Program baseline plan has always called for supporting substantial analysis work at the Fermilab facility. Original plans called for a facility to service the needs of about 100 people; in reality, ten years later, we have about 150 active users, among about 750 people with user accounts. This facility is called the LPC CAF. The resources are funded through the U.S. CMS Operations Program, and are operated by the Fermilab Scientific Computing Division.

2 Services Required

2.1 Login services

The LPC CAF requires login nodes for users to access the system; there are currently about two dozen such nodes. Any user approved by the LPC leaders or their designees will have login access to these nodes. Features of the login service include:

- A single login point with distribution of interactive users across the cluster
- Protection against long-running and high-I/O interactive jobs of single users
- Interactive submission to the LPC CAF batch system, including through the CMS CRAB3 job submission tool

- Access to CMS software via CVMFS and grid UI
- POSIX access to home directories from the central FNAL home directory infrastructure
- POSIX access to interactive storage for ntuple analysis and other analysis activities
- Non-POSIX access and sufficient network bandwidth to LPC mass disk storage, Tier-1 facility mass disk storage, and the CMS AAA (xrootd) data federation

2.2 Batch services

The batch system has access to the LPC CAF worker nodes, which currently provide about 5000 cores for processing. Features of the batch nodes include:

- Access to CMS software via CVMFS and grid UI
- No POSIX mounts
- High bandwidth network access to LPC mass disk storage, Tier-1 mass disk storage, and the CMS AAA data federation
- Job submission both through HTCondor, allowing for flexible requests in the number of cores and memory, and through the CMS CRAB3 tool
- Monitoring and accounting of user priorities
- Tools to modify user priorities

LPC leaders and U.S. CMS Operations Program leaders jointly hold the right to request that batch resources be migrated between the Tier-1 facility and the LPC CAF. The optimal sizes of the two facilities could change over time given the demands of the CMS experiment. SCD will migrate resources between the facilities within one week of a request.

2.3 Storage services

Users will have access to

- A home area that is served from the central FNAL home directory infrastructure. This area is backed up and can be recovered in case of hardware failure. The default user quota is 10 GB.
- Data and scratch areas also available via NFS mount on the interactive nodes. The default user quota on data area is also 10 GB. The scratch area has no quota, however files are auto deleted after a two weeks or 3 days, depending on location. Currently the data and scratch areas have 130 TB and 75 TB allocated to them respectively.

- A multi-petabyte, distributed storage system. POSIX access is not required. The current implementation is a 4.2 PB EOS store. The system has the ability to monitor and account for disk usage and file replication, minimally at the user and group level, and preferably at the directory level. The default user quota is 2 TB. Users must have the ability to grant access of specified portions of their storage to other users. Groups of users that have access to group storage areas can be formed. This includes quota, permission and replica management. The storage system is minimally accessible through native protocols, gridftp and xrootd.

2.4 Additional services

The multi-petabyte distributed storage system is an endpoint for the CMS PhEDEx data transfer system, under the site name T3_US_FNALLPC. The PhEDEx services need to be kept up to date and operational. The same system is included in the global data federation via xrootd, and accepts incoming data transfers via gridftp from the CRAB3 ASO system.

Currently a specialized CRAB3 service for job submission into the LPC condor cluster must be maintained.

Currently the DataView and Vulcan user mapping systems must be maintained for user monitoring of storage quotas and mapping CMS DN's to user accounts.

3 Technical Description of Usage

Typical batch jobs make use of CMS software available on worker nodes via CVMFS plus relevant user code that is sent to the node by HTCondor. Sample workflows include:

- Ntuple production from the CMS miniAOD format. A standard workflow consists of about 25K jobs that each run several hours. Each job runs on one input file with output copied to the to the distributed storage system.
- Skimming of ntuples to smaller ntuples, which are reduced both in the number of events and the information kept for each event. Each job reads multiple files and runs several hours, with output copied to the to the distributed storage system.
- Adding of histogram files within ROOT for easier interactive use.
- Monte Carlo production, which takes no input and produces miniAOD output. Typical samples require 20-200 jobs.
- Accessing user generated histograms and ntuples in data/theory model fitting and limit setting

More resource-intensive uses of the login service include:

- Generation of plots, which in some cases can be I/O intensive with the same file opened multiple times.
- CMS code development, which in some cases requires compilation of hundreds of packages, creating libraries of order 10 GB in the data area, and running unit tests and profiling tools.
- Providing non grid certificate authenticated access to ntuples and derived data to outside computing resources (i.e. personal or institutional computers/laptops)

4 Support for User-Facing Services

One or more specific named persons will serve as liaisons¹ between LPC CAF users and those who maintain and operate the services. In general, 0.5 FTE from LPC support staff will be available to assist in this task. Responsibilities of the liaisons include the following:

- Creation of new user accounts upon request of LPC leaders or their designees, and management of existing user accounts, group affiliations, shell preferences, etc.
- Management of requests for changes in disk quota on all storage systems, both for individual users and defined groups of users.
- Management of disk space use, *e.g.* notifying users who have gone over quota, cleanup of shared disk areas, *etc.*
- Data manager role within PhEDEx for the T3_US_FNALLPC site.
- Documentation of the CAF and its services for users.
- The first level of user support is through the lpc-howto mailing list, where users report problems and ask for help. Problems and issues related to the services of the LPC (*e.g.* malfunctioning hardware or poor performance) are identified by the named liaisons, who then open tickets and follow up. LPC CAF users will in general not interact with the ticketing system; instead the liaisons will do so on their behalf.
- Management of open service tickets.
- Coordination of downtime scheduling between LPC leaders and system administrators, and announcements of the downtimes.
- Organization of regular LPC CAF user meetings.

¹These “liaisons” are distinct from the “CS liaisons” who are contacts between the experiments and SCD.

In general, all of these support functions will be limited to business hours. However, LPC leaders can request a higher level support at critical times in the lifecycle of the CMS experiment, *e.g.* preparations for major conference periods (winter conferences, summer conferences, end-of-year jamboree). These will occur no more than three times per year, and requests can be made for a two-week interval of increased support for each period.

5 Acknowledgements

We thank Kevin Pedro for his feedback this document.