LPC CAF Description

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

August 15, 2016

1 Introduction

The LHC Physics Center (LPC) is a regional center designed to engage more members of the U.S. CMS collaborating institutions in analysis of LHC data and in upgrade-related work. The LPC lowers the barrier to remote participation, creates a vibrant intellectual environment by providing: proximity to a broad range of detector and object expertise under one roof; extensive and expert software support; access to outstanding computing resources; basic training and advanced classes; office space; and an economical location for researchers. The LPC is lead by the LPC coordinators.

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. Many of these users are not resident at Fermilab, but instead at their home universities or CERN. This facility is called the LHC Physics Center (LPC) CMS Analysis Facility (CAF). The resources are funded through the U.S. CMS Operations Program and managed by the U.S. CMS Tier-1 manager. The U.S. CMS Computing Resource Allocation Advisory Board reviews semi-annually the LPC resource requests by the LPC coordinators and provides resource allocation guidelines to the U.S. CMS Tier-1 manager. The U.S. CMS Tier-1 manager consults frequently with the LPC coordinators about short-term changes to the resource needs, for example due to the conference schedule. The LPC CAF resources are operated by the Fermilab Scientific Computing Division (SCD). Support of the computing resources and the user community is shared between SCD and LPC support staff (see Section 4).

2 Services Required

A functional diagram of the LPC CAF is shown in Figure 1.

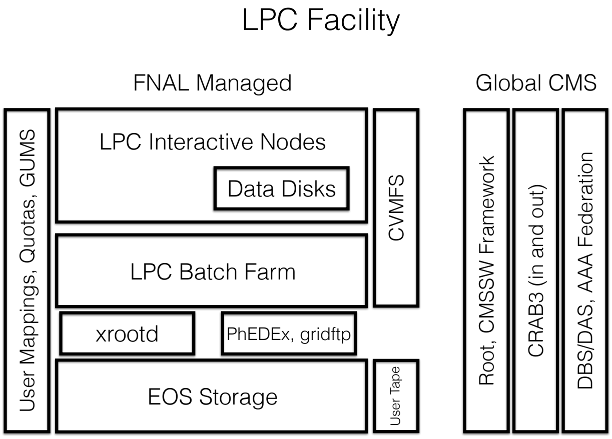

Figure 1: Functional diagram of LPC CAF.

2.1 Login services

The LPC CAF requires login nodes (see "LPC Interactive nodes" in Figure 1) for users to access the system; there are currently about two dozen such nodes. Any user approved by the LPC coordinators 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 access to a 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 access to a 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

The U.S. CMS Tier-1 manager holds the right to request that batch resources be migrated between the Tier-1 facility and the LPC CAF. 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 and POSIX-mounted on the LPC interactive nodes. This area is under backup and can be recovered in case of hardware failure. The default user quota is 10 GB.
- Data and scratch areas also available via POSIX-mounts 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 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, but users can request larger quotas. Currently, about 200 users have files in the system with total sizes on the order of terabytes, with the top twenty of those users using 0.5 PB. (Many other users use much less space in the system.)

  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, and is encouraged as the default mode of operations.

This includes quota, permission and replica management. Group and individual quotas are assigned and managed by the LPC coordinators. The storage system is minimally accessible through native protocols, gridftp and xrootd.

- Providing non grid certificate authenticated access to ntuples and derived data to outside computing resources (i.e. personal or institutional computers/laptops) is a new request as of July 2016.

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.

A specialized CRAB3 service for job submission into the LPC condor cluster must be maintained, but that might change in the future.

The DataViewer and Vulcan user mapping systems must be maintained for user monitoring of storage quotas and mapping CMS DN's to user accounts, but that might change in the future.

A facility for users to copy files to tape and retrieve them needs to be supported. Currently this is handled through tickets to storage support. The tape space needed for user files is requested by the LPC coordinators, reviewed through the U.S. CMS Computing Resource Allocation Advisory Board process and allocated by the U.S. CMS Tier-1 manager.

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. Example 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 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 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
- Emulation of the Level-1 trigger for HL-LHC upgrade studies, a special case which uses specialized high-memory cores and can be memory-intensive in spurts.

It is also possible for workflows to be run outside the CAF batch system (and outside of Fermilab) and to access the distributed storage system through the xrootd protocol and the CMS global data federation.

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.

4 Support for User-Facing Services

The support for all user-facing services of the LPC and user computing activities at the LPC is provided by SCD, where LPC support staff is contributing to support user activities. Services are provided by individual service providers in SCD as documented in Table 1.

The support of the user activity is provided by the "Distributed Computing" service area in SCD. In general user support is located between the users and the individual service providers. Its task is to triage reported user problems and requests, and either work with the user to resolve problems with using the LPC services and facilities, or work with the service providers to resolve problems with the services or facilities themselves. 0.5 FTE from the LPC support staff and 1.0 FTE from the SCD support staff will be available for these activities[1]. The SCD support effort will be distributed among several members of the SCD support team, to allow for enlarged coverage during vacation time or similar situations. One member of the SCD team will be named explicitly to be a direct contact for the LPC users and the LPC coordinators to discuss smaller and big-picture issues. It is not expected that the named contact solves all problems, but is dedicated to facilitate a close and personal support for LPC users. If the person fulfilling the role of named contact changes, the LPC coordinators and the U.S. CMS Tier-1 manager need to be informed.

Support responsibilities include:

- Creation of new user accounts upon request of LPC coordinators or their designees, and management of existing user accounts, group affiliations, shell preferences, etc.

[1] As of July 2016, the SCD support staff is 1.5 FTE. This is composed out of the 1.0 FTE to support the LPC plus 0.5 FTE to support EOS. The 0.5 FTE EOS support is planned to be transitioned to the Scientific Data Storage and Access service offering.

- Management of requests for changes in disk quota on all storage systems, both for individual users and defined groups of users, upon authorization by the LPC coordinators.
- Management of disk space use, *e.g.* notifying users who have gone over quota, cleanup of shared disk areas, *etc.*
- Documentation of the CAF and its services for users.
- Management of open service tickets or conversion of support email list discussions into tickets.
- Coordination of downtime scheduling between LPC coordinators and system administrators, and announcements of the downtimes.
- Availability for office hours on a semi-weekly basis.
- Organization of regular LPC CAF user meetings.

The respective responsibilities of the LPC and SCD support staff is guided by the following:

- Established user workflows such as those described in Section 3 are supported by the SCD support staff
- Establishing new workflows to help users accomplish new tasks which have not yet been executed on LPC facilities are supported by the LPC support staff

User support requests are submitted in two ways. The preferred way to submit user support request is through a ticket in the Fermilab Service Desk (SNOW). Tickets are submitted either through a web interface or by sending e-mail to a gateway[2]. LPC users also have access to the lpc-howto mailing list, where they report problems and ask for help. Problems and issues related to established workflows discussed on the list are converted into tickets by LPC support staff and then resolved by SCD support staff. New workflows are discussed on the list or individually with the users by the LPC support staff. The LPC support staff then teaches the SCD support staff about newly implemented workflows and the support responsibility changes over. Both LPC and SCD support staff are involved in educating the user community how to best use the ticket system to get the fastest response to their queries. In addition, a status board will be provided where general problems are listed and their solution announced [3].

As an example, a user reports a slowdown of the EOS file system in the lpc-howto email list. The LPC support staff catches the problem and opens a

[2] As of July 2016, the e-mail gateway has not yet been delivered.

[3] The status board is a new request but is needed to guarantee effective communication with a large and diversely located user community. A good example is the CERN-IT SSB `https://cern.service-now.com/service-portal/ssb.do?area=IT`

ticket on behalf of the user. LPC support staff debugs the problem, if necessary contacting the user to ask for more information. If the problem is with the infrastructure itself, SCD support staff passes the problem on to the service provider, posts a short description of the problem on the status board, and follows the solution of the problem and in the end closes the issue including post on the status board. If the problem is related to an established workflow, SCD support staff works with the user to resolve the problem and advises how to use services and facilities in a more efficient way. If the problem is related to a new workflow, LPC support works with the user to establish a new procedure how to use services and facilities for this new task.

The data manager role within PhEDEx for the `T3_US_FNALLPC` site is counted within the area of user support but is carried out by Fermilab scientific staff on the CMS experiment.

In general, all of these support functions will be limited to business hours with best effort support during off-hours. However, LPC coordinators can request a higher level support at critical times in the life cycle 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 SCD Service Offerings

SCD is supporting infrastructure and services for its experiments and Virtual Organizations (VOs). SCD organized and documented these as service offerings in the service desk: `https://fermi.service-now.com/list_service_areas.do?division=scientific_computing`. Table 1 provides an overview of services needed to operate the LPC CAF (both infrastructure and support) mapped to SCD service offerings. It also includes non-SCD functions like the LPC support, and the support of the PhEDEx data management, and services that are provided by global CMS.

6 Acknowledgements

We thank Kevin Pedro for his feedback this document.

Service	Service Area
Login Service and Hardware	High Throughput Computing
Login Node config (Puppet)	Scientific Computing Systems
Home area management: backups, usage and quota handling	Network Storage
NFS Storage Service	Scientific Server Infrastructure
NFS Storage Hardware	Scientific Server Infrastructure
NFS Storage User/Group Management	Distributed Computing
EOS Service	Scientific Data Storage and Access (add EOS)
EOS Hardware	Scientific Data Storage and Access (add EOS)
EOS User/Group Management	Distributed Computing
PhEDEx Service	Scientific Data Storage and Access
PhEDEX Hardware	Scientific Computing Systems
PhEDEX Data Management/Monitoring	Fermilab scientific staff on CMS experiment
HTCondor cluster Service	High Throughput Computing
HTCondor cluster Hardware	Scientific Computing Systems
CRAB3 Server (JobRouter)	Global CMS
CRAB3 User Access	High Throughput Computing
User Interaction	Distributed Computing + LPC support
User Account Generation Workflow	Distributed Computing
LPC User support through SNOW	Distributed Computing
LPC User support through community email list	LPC
LPC monitoring	Distributing Computing (add LPC monitoring)
LPC user meetings	LPC organizes
LPC Coordination of downtime	LPC

Table 1: Mapping of LPC services to SCD service offerings