PROCEEDINGS OF SCIENCE

Any data, any time, any where

Sudhir Malik[1]
University of Nebraska-Lincoln
Lincoln, NE 68588
E-mail: malik@fnal.gov

In this paper we will discuss progress on providing access to CMS experiment data through the use of wide area transfer protocols directly from remote storage to a running application. This program of work has involved the deployment of infrastructure on facilities in both the US and in Europe, the optimization of the application to make more efficient use of a higher latency connection, and the demonstration of a variety of use cases to showcase the value of this functionality. The project is at the level where several of the intended applications are used in production running, and several of the more ambitious techniques are in the prototype phase.

36th International Conference on High Energy Physics
July 4-11, 2012
Melbourne, Australia

1

Speaker

PoS(ICHEP2012)522

1. Introduction

A robust and distributed computing infrastructure has played a key role behind the discoveries by the CMS experiment [1] and over 200 publications in the past 3 years [2]. It has been successful dealing with petabytes of data per year and has allowed physicists to reconstruct and analyze data quickly and deliver physics results in a timely fashion. The physics measurements performed by a global collaboration of 3000 physicists rely on a distributed computing model that is coherently organized into four layers or Tiers [3] located in institutions around the world. However, there are several challenges with the current computing model. Despite a large network of grid computing, CMS resources to process and store its data are still limited. It requires storage systems that host datasets to be co-located with the processors that analyze them and restricts analysis to centers that have the resources and expertise to operate large storage infrastructures, thus excluding many university clusters (Tier-3) from being useful for physics analysis. Following the recent discovery of Higgs-like particle [4] at CMS and ATLAS experiments, the trigger rates are expected to increase in order to study its properties. With a limited budget, the best way to meet the challenge is not in expanding the resources but to adjust and optimize them. Using current technologies we must build tools for data access infrastructure that will remove the requirement of co-locating storage and processing resources. This will lead to the user experience being the same whether the data being analyzed is halfway around the world or right next-door. This could truly enable any user to analyze any data at any time, anywhere and allow the use of "opportunistic grid" as well as "commercial cloud" resources.

2. Current CMS computing model

The CMS computing model is highly distributed, both by design and by necessity. The rate of producing interesting physics events in CMS is one in a trillion. Therefore, realizing the full physics potential of the CMS requires sifting through petabytes of data quickly and efficiently. No single institution can afford the computing power needed for this. Handling the CMS data involves reconstruction of the raw data from the detectors, producing simulated data and the physics analysis of the reconstructed datasets from both the detector and the simulation. The sharing of the load of processing petabytes of generated and an equal amount of derived data is done by a complex topology of tiered computing centers across the globe (Fig. 1). The Tier-0 (T0) center located at CERN does the prompt calibration and reconstruction of the raw data, and archives a copy of raw and reconstructed data. The Tier-1s (T1) perform re-reconstruction of the data and simulated event production. They archive a copy of the reconstructed data. The typical bandwidth between CERN and Tier1s is ~ 2 gigabytes per second. Tier-2s (T2) have large enough disk pools and CPU space to be able to serve as the primary source of physics analysis by users. The Tier3 (T3) serves as source of user analysis, and store user defined event content and reduced datasets with no commitment of resources beyond the institute-owned and operated clusters.

Each year 2010-2012 has brought increasing challenges for computing. There has been an increase in luminosity to 6 x 10 ^{33}cm^{-1}sec^{-1} producing more collisions to analyze and a pileup

size of > 16. Pileup interactions occur when there are multiple collisions within the same event crossing increasing the size of the event. Disentangling the pileup contribution to look at the single collision events needs more computing power.

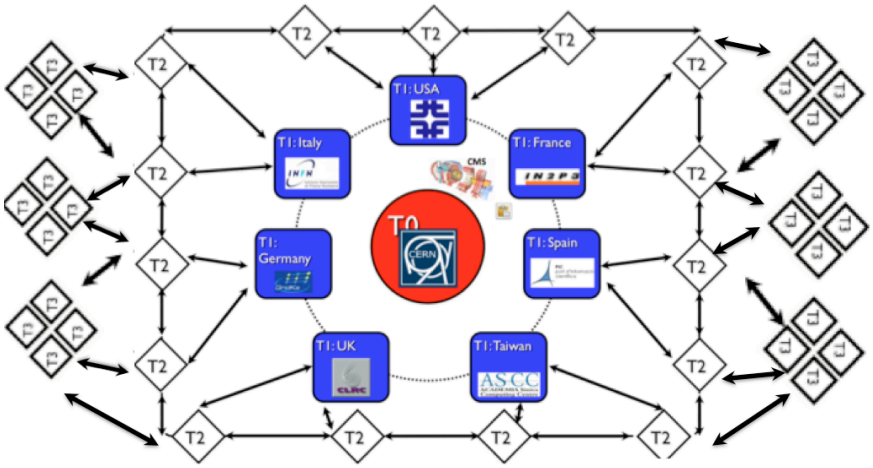

Figure 1: CMS tiered (T0, T1, T2, T3) computing architecture.

Computing has kept up with the size of reconstructed event in RECO [5] data format of 0.8 MB/event and in a more reduced format called AOD [5] of 0.2 MB/event and a high trigger rate of ~400 Hz. The pressure to produce a result in time for the biggest particle physics conference adds to the stress on computing. An average of 400 users submit analysis jobs per week (Fig. 2) over the computing grid to look and analyze the datasets that have grown to be over 30,000. The analysis jobs per week grew exponentially to 100K and leveled of reaching the era of resource limitation (Fig. 3)

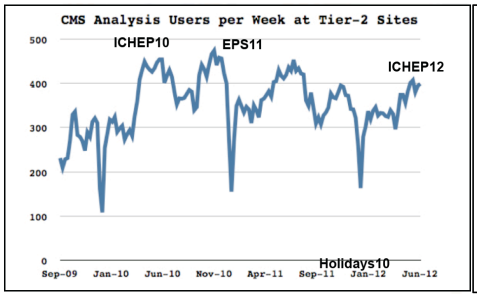

Figure 2: CMS analysis user/week.

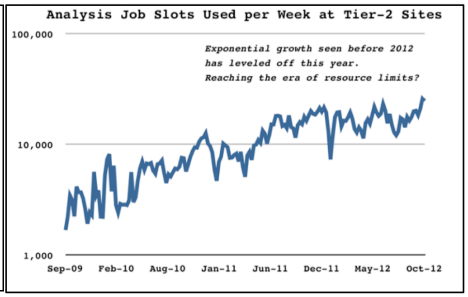

Figure 3: CMS analysis jobs/week.

3. Limitations with the computing model

The Tier-2 facilities are the primary resource for all physics analysis and need to host the data accessible to users. This is where the various physics groups, and individual physicists have disk space allocated, produce group-level skims, develop and test new reconstruction algorithms and perform physics analysis. However, it requires a human to make access request, and another one, the receiving site's data manager, to accept the request thus slowing distribution of data. It also restricts analysis to centers with resources and expertise to operate large storage infrastructure. Physicists tend to insist on having the data they think they will need on disk for faster access. Given the size and distributed nature of the collaboration, making the

distinction between what is expected to be needed and what is actually used is difficult. Some datasets not or rarely used occupy space that could be filled with more popular datasets and sites remain underutilized. But some sites hosting highly accessed datasets are routinely saturated. They provide more opportunistic resources than pledged or have larger queues of pending job. If data gets deleted at a site chosen by a job, that job fails.

The current model renders Tier-3s useful for physics data analysis only for the local institution. They are driven entirely by the scientific goals of the institutions that own and operate them. There is a great heterogeneity in size and functionality of Tier-3s ranging from being just a few computer nodes to clusters on the scale of a Tier-2 facility. The Tier-3s challenges include lack of operational expertise, limited storage space, and the challenge of moving data into the site. The requirement of data co-location under-utilizes both the computational and intellectual resources of a Tier-3 institution. Physicists need improvements that will reduce the operational costs of storage, strengthen training and support, and ease access to the data. We also want to expand the usage of resources beyond those controlled by physicists at Tier-3 to opportunistic resources that might be located on their campus, or anywhere else in the world. To do so will require technologies that do not yet exist in CMS or in grid computing in general, including the ability to transparently add computing resources in grids and clouds.

4. Strategy and solution

The CMS computing organization has made significant progress in optimizing reading of data files over the network and the additional cost compared to reading a file in the same room is very little. Therefore one is inclined to relax the requirement of co-location of data and CPU and to think of big. We need to build software tools with infrastructure with no requirement of co-locating storage and processing resources. They should be reliable with no I/O error or failure for end-user unless no site can service the request and able to catch failures early and re-directing I/O to a different site. The underlying system actions like catalog, lookups and reconnections should be transparent to the users and automatic. The workflow to access data "close by" or halfway around the world should be same. Any solution should seamlessly integrate with CMS application framework, must not degrade event processing significantly and complement the currently deployed data management system. This should be replace by a more dynamic data access and placement, including having the ability to access data in our Tier-2 caches from Tier-3, cloud and other resources the CMS project does not explicitly own. We should make use of "opportunistic grid" and "commercial cloud" to help reduce financial and operational burden. These steps would lead to analyze any data, any time, anywhere.

To achieve these goals CMS has begun using a distributed architecture based upon the Xrootd [6] protocol and software developed by SLAC. Without replacing current CMS data access methods for production it will greatly reduce the difficulty of data access for physicists on the small or medium scale. It would break the "data-locality" and provide fallback option for grid jobs in case of overflow (deal with resource limitation). If a Tier-2 loses a file or it gets corrupted for some unknown reason, the CMSSW application will automatically redirect (using network redirectors from direct local file system access to the less efficient national, and

eventually global scale Xrootd service instead of presenting a failure to the end-user. It would also provide disk-free data access system for Tier-3s. Fig. 4(left) shows the schematic view of system of network redirectors where first (1) a user application attempts to open the file in the regional redirector. If the regional redirector does not know the file's location, it then queries all of other regional sites (2). Site A responds that it has the file, so the redirector redirects (3) the client to Site A's Xrootd server. Finally, the client contacts Site A (4) and starts reading data (5). Fig. 4(right) shows the same except a user application is querying the regional redirector when the desired file is not within the region. The user is redirected to Site C (5) and successfully opens the file (6 and 7). Remote access gives us data for one site. We need a federation to access all sites across all CMS sites. Fig. 5 shows schematic of a federation.

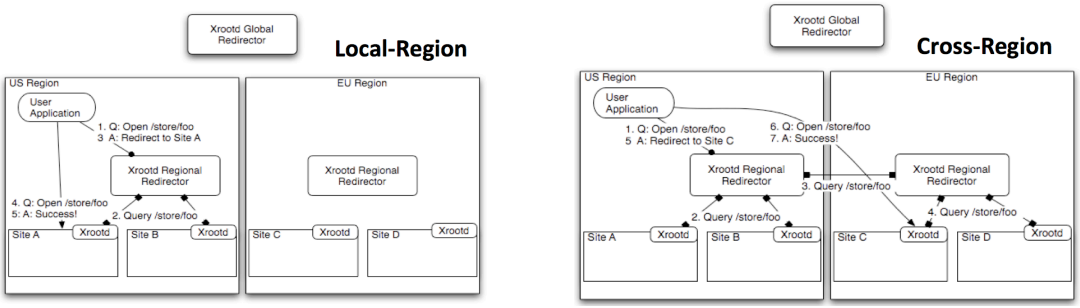

Figure 4: Schematic view of system of redirectors redirecting locally and across the region.

5. Operational experience

A prototype [7] of Xrootd architecture has been tested in the CMS sites in U.S.A. and has evolved to include more CMS sites worldwide and all the relevant storage technologies. The redirectors are located at UNL (Nebraska, U.S.A.), Bari (Italy) and CERN. During April 2012, it recorded over 300 unique users, 900K file transfers and 300TB of data movement. Any new site can integrate via installing a plugin specific to their storage system. To provide a uniform namespace, a site must export the global filename and not the local filename. This is achieved through Xrootd plugin that does the mapping through a list of mapping rules and regular expressions. To authorize data access request a Grid Security Infrastructure (GSI) based authentication is used that has a plugin for mapping the GSI credentials and the DN (Distinguished Name) and VOMS (Virtual Organization Membership Service) attributes are passed to the site mapping service. Two monitoring streams – one for summary (Fig. 6) and other for detailed federation monitoring are available at this website: http://xrootd.t2.ucsd.edu/dashboard/

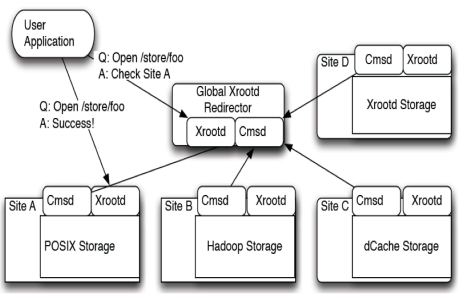

Figure 5: Federation schematic.

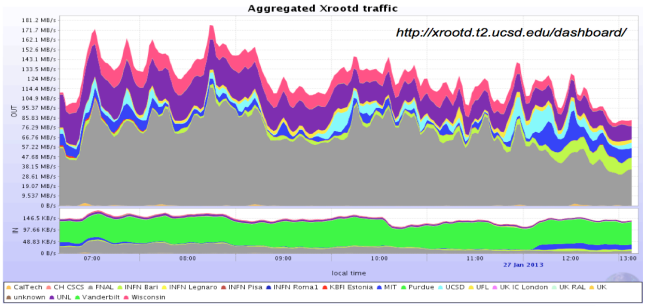

Figure 6: Aggregated CMS Xrootd traffic.

The federation of redirectors has been used to implement fallback option for file access for the grid jobs. If grid job fails to open a file, Xrootd tries to read again from the redirector. There is cost paid in terms of loss of efficiency but the job does not crash. This implementation uses federation to re-download a broken file from another site. This is called Storage Healing. If a remote files is read many times and given that remote I/O is expensive in terms of bandwidth, instead of asking the redirector to fetch the file every time, one can request a local Xrootd install. The local Xrootd will stage the file to a local disk and on the next access the file will be local. This is called File Caching. It has not been yet implemented in CMS.

The grid submission systems now used by CMS bind the user's job to a specific grid site, making it difficult to dynamically take advantage of opportunistic cores when available. However, we do have some experience with a dynamically, late binding system at Univ. of Wisconsin Tier-2 that uses Condor flocking mechanisms to export excess jobs to the large opportunistic resources available throughout the campus. CMS group at Univ. of California at San Diego also operates a Condor glidein-based [8] analysis system that dynamically gathers computing resources from Tier-2s to be used for data analysis jobs through the CMS CRAB [9] infrastructure. Both are not optimal but one can extract best of the two to find solution for on-demand CPU. This would need to provide access to CMS software releases, solve the data access problem, and make the data access solution scalable, or at least regulate its scale. Having CMS software releases pre-installed on all sites is fundamentally not an option for opportunistic resources. We need to efficiently use combination of WAN (Wide Area Network) bandwidth and access to network and storage resources.

6. Conclusions

The current CMS computing model is a great success and currently able to meet the demand of its users. The challenge in near future is perhaps not in expanding the total resources available, but using them optimally. Xrootd represents such a direction. After successfully performing prototype test of Xrootd at CMS sites, it is will be expanded to all the sites. It is also working to meet the challenge of optimally using the opportunistic CPU and storage resources. The success of CMS in adapting its computing model to improve resource use suggests that these efforts will be successful in the future too. This will truly lead to access of any data, any time and any where for CMS users. Thinking beyond CMS and HEP, this could serve as a prototype for the other "big data" applications beyond HEP.

References

[1] The CMS Collab., *The Compact Muon Solenoid Technical Proposal*, CERN/LHCC 94 -38, 1994.

[2] The CMS Collab., *Physics Papers Timeline*, http://cms.web.cern.ch/org/physics-papers-timeline

[3] The CMS Collab., *CMS Computing Model*, https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel

[4] The CMS Collab., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Physics Letters B 716 (2012) 30–61.

[5] S Malik (for CMS Collab.), *CMS Analysis Deconstructed*, 2012 J. Phys.: Conf. Ser. 396 032073.

[6] SLAC/CERN, *XRootD*, http://xrootd.slac.stanford.edu/

[7] The CMS Collab., *Using Xrootd Service for Remote Data Access*, https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookXrootdService

[8] The CMS Collab., *GlideinWMS*, http://www.uscms.org/SoftwareComputing/Grid/WMS/glideinWMS/doc.prd/index.html

[9] The CMS Collab., *Data Analysis with CRAB*, https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookRunningGrid