Comprehensive Monitoring for Heterogeneous Geographically Distributed Storage

N Ratnikova[1], E KaravakisE[2], S Lammel[1], T Wildish[3]

[1] Fermi National Accelerator Laboratory, US
[2] CERN, CH
[3] Princeton University, US

E-mail: natalia.ratnikova@cern.ch, tony.wildish@cern.ch

Abstract.

Storage capacity at CMS Tier-1 and Tier-2 sites reached over 100 Petabytes in 2014, and will be substantially increased during Run 2 data taking. The allocation of storage for the individual users analysis data, which is not accounted as a centrally managed storage space, will be increased to up to 40%. For comprehensive tracking and monitoring of the storage utilization across all participating sites, CMS developed a Space Monitoring system, which provides a central view of the geographically dispersed heterogeneous storage systems. The first prototype has been deployed at the pilot sites in summer 2014, and has been substantially reworked since then. In this presentation we discuss the functionality and our experience of system deployment and operation on the full CMS scale.

1. Introduction

SpaceMon system provides monitoring for the space occupied by various directories in the CMS namespace directory tree at the geographically distributed CMS sites.

Information is retrieved directly from the local site storage systems in the form of storage dumps [1] produced by various information providers tools, depending on the storage technology.

Aggregated space usage information is stored in the DMWMMON oracle database at CERN.

Space usage records are uploaded to and retrieved from the database via web based data service interface.

Access to information store is based on certificate authentication. Site admins are allowed to upload records for their respective site. Any CMS VO registered user can retrieve the information.

World map showing locations of CMS collaborating universities and institutes.

2. Site information providers

CMS sites use various storage technologies for maintaining CMS data: Castor, dCache,DPM, EOS, Hadoop, LStore, Lustre, StoRM.There is no unique solution for retrieving the local storage information for all sites. On the posix-compliant file systems, such as EOS, Lustre, or hadoop systems mounted with FUSE, the find command can do the work.For dCache, direct database query on the namespace database is often more efficient and causes less load on the system than traversing the whole namespace via NFS client tools. The impact on the production

storage namespace can be completely avoided if the storage dumps are produced on the hot standby server, onto which the namespace database is dynamically replicated for the backup purpose.Sites sharing the same storage instance between multiple virtual organizations may optimize by dumping only a sub-branch of the namespace under CMS-owned directory.Another possible way to optimize is to do aggregation on the database level instead of producing the whole storage dump. SpaceMon client provides an API to read in the aggregated record in a pre-defined format.CMS site admins exchange tools and solutions for creating storage dumps via an open source repository on github [2].

3. Site client software

3.1. The spacecount utility

- improved prototype.

3.2. The SpaceMon framework

- eliminate dependency on PhEDEx for independent packaging and release management
- replace Namespace framework that supports various technologies by Format framework, which implements similar idea for various storage dump formats.
- RecordIO module provides functions (interfaces?) for handlig the space information

4. Central infrastructure

4.1. DMWMMON Database as Information store

4.2. Data Service as interface to Information store

4.3. SiteDb as authentication authority

4.4. DashBoard as es end-user visualization tool

The goals of the visualization are to present the space monitoring information in a convenient form, enabling the users to check space usage across the sites, explore the historical views or drill down into a particular directory in the CMS storage namespace, and eventually help planning and optimize the storage resource usage.We have recently started work with WLCG developers to implemented the visualization of the sites space monitoring information in the CMS dashboard.

Design is similar to that of the ATLAS DDM storage accounting [5] based on rucio storage summaries [6]. The challenge in CMS case is to represent uniformly the monitoring data asynchronously pushed by the sites.

5. Deployment campaign

Systematic deployment of CMS Space Monitoring at CMS Tier-2 sites started late spring 2014.Site deployment path includes three steps: 1. Setup Information Provider for site storage2. Install Space Monitoring client to parse and aggregate storage dump and to upload the info3. Set up a scheduler to upload weeklyDetailed instructions [3] for the site adminswere tested and verified at a few pilot sites.Information Providers tools have been verified and improved with several site admins contributions, and now stay stable.The central Data service installation was enhanced to fix the problems discovered.Most frequently reported site specific problems were related to the authenticated upload. A list of troubleshooting steps has been provided for the sites. Still this area required most individual help from the developers. While working on improving the troubleshooting, we are also looking at other secure upload solutions. Status of Space Monitoring deployment at the sites and the dates of the most recent upload are reflected in the CMS Dashboard.

6. Summary

CMS space monitoring is designed to provide a central view of space occupied by all CMS data stored in heterogeneous storage systems deployed at CMS sites.Storage dumps produced locally by the sites include user areas and CMS production data not accounted in the CMS central data catalogues.Central infrastructure has been deployed as CMS web service at CERN and started collecting data uploaded by the sites.In the course of the deployment campaign all system components benefited from the improvements based on feedback from the sites.Remaining steps are: complete deployment at CMS sites and provide the visualization.

References

[1] Grandi C, Stickland D and Taylor L 2005 The CMS Computing Model *CERN-LHCC-2004-35/G-083, CMS note 2004-031*

[2] Egeland R, Wildish T and Metson S 2008 Data transfer infrastructure for CMS data taking, *XII Advanced Computing and Analysis Techniques in Physics Research (Erice, Italy: Proceedings of Science)*

[3] Knobloch J *et al.* 2005 "LHC Computing Grid Technical Design Report" CERN-LHCC-2005-024

[4] Garonne V, Molfetas A, Lassnig M, Barisits M, Stewart G A, Beermann T 2012 "The ATLAS Distributed Data Management project: Past and Future", these proceedings

[5] Shoshani A, Sim A and Gu J 2002 "Storage Resource Managers: Middleware Components for Grid Storage", Nineteenth IEEE Symposium on Mass Storage Systems'

[6] The CMS Collaboration 2008 "The CMS experiment at the CERN LHC" JINST **3** S08004

[7] Bonacorsi D 2007 "The CMS computing model" *Nucl. Phys.* B *(Proc. Suppl.)* **172** 53-56

[8] Giffels M and Guo Y Data Bookkeeping Service 3 - A new event data catalog for CMS, submitted to CHEP 2012

[9] Egeland R, Wildish T and Metson S 2008 "Data transfer infrastructure for CMS data taking" *XII Advanced Computing and Analysis Techniques in Physics Research* (Erice, Italy: Proceedings of Science)

[10] Ball G 2011 PhD thesis, chapter 8 "Computing Monitoring Pages", https://workspace.imperial.ac.uk/highenergyphysics/Public/theses/Ball.pdf

[11] The D3 javascript library http://mbostock.github.com/d3/

[12] Stagni F *et al.* 2012: "LHCbDIRAC: distributed computing in LHCb", these proceedings

[13] Lopienski S 2008 "Service level status - a new real-time status display for IT services" *J. Phys.: Conf. Ser.* **119** 052025

[14] Baud J-Ph et al 2005 Performance analysis of a file catalog for the LHC computing grid HPDC 14 91-99

[15] Charpentier P *et al.* 2012 "The LHCb Data Management System", these proceedings

[16] Serfon C 2010 "Data management tools and operational procedures in ATLAS: Example of the German cloud," J. Phys. Conf. Ser. **219**, 042053 (2010).

[17] N. Magini, N. Ratnikova, P. Rossman, A. Sanchez-Hernandez and T. Wildish, "Distributed data transfers in CMS," J. Phys. Conf. Ser. **331**, 042036 (2011).

[18] Lanciotti E 2011 "Storage elements dumps and consistency checks versus file catalogues" https://twiki.cern.ch/twiki/bin/view/LCG/ConsistencyChecksSEsDumps

[19] Bauerdick L and Sciaba A 2012 "Towards a global monitoring system for CMS computing", these proceedings

[20] Millar P *et al.* 2010 "Dealing with orphans: Catalogue synchronisation with SynCat" J. Phys.: Conf. Ser. 219 062060

[21] Sanchez-Hernandez A, Egeland R, Huang C-H, Ratnikova N, Magini N and Wildish T, 2012 "From toolkit to framework - the past and future evolution of PhEDEx", these proceedings

[22] Castor, http://castor.web.cern.ch/castor/

[23] dCache, http://www.dcache.org/

[24] DPM (Disk Pool Manager) https://twiki.cern.ch/twiki/bin/view/LCG/DataManagementDocumentation#DPM

[25] Peters A-J 2011 "The EOS disk storage system at CERN", ACAT conference proceedings

[26] BeStMan (Berkeley Storage Manager) https://sdm.lbl.gov/bestman/

[27] Lustre http://www.lustre.org/

[28] L-Store (Logistical Storage) http://www.accre.vanderbilt.edu/mission/services/lstore.php

[29] Cavalli A *et al* 2010 "StoRM-GPFS-TSM: A new approach to hierarchical storage management for the LHC experiments" *J. Phys.: Conf. Ser.* **219** 072030

[30] XRootD, http://xrootd.slac.stanford.edu/

[31] The Protovis javascript library http://mbostock.github.com/protovis/

[32] The Highcharts javascript library http://www.highcharts.com/

[33] Matplotlib python 2D plotting library http://matplotlib.sourceforge.net/