Table of Contents

CMS Xrootd Architecture

This is the homepage for the Xrootd-based federations in CMS.

Documentation

For Users

We have the following user documentation available also:

- Xrootd Client Usage - How to utilize the current infrastructure, on your desktop, in ROOT or in a CRAB job.

For Admins

The following documentation is aimed at the sysadmins of CMS sites:

- Planning your Xrootd hardware deploy.
- How to integrate Xrootd into your site. Select the appropriate one for your SE technology.
    - ♦ HDFS - Joining a HDFS site to the global Xrootd federation.
    - ♦ dCache - native Xrootd doors - Joining a dCache site to the global Xrootd federation.
        - ◊ dCache - Xrootd proxy - Joining dCache to the global federation using a site proxy. Also applies to dCache sites whose LFN->PFN mapping is more complex than adding a prefix.
    - ♦ POSIX - Joining a POSIX (Lustre, GPFS) site to the global Xrootd federation.
    - ♦ DPM - Instructions for a DPM site; link to the external DPM site. Main page still lack some of the generic info like contact etc, but should be a good start!
        - ◊ More tuning hints - in order to get better overall performance check out this page.
    - ♦ XRootD-HTTP - Configure/Enable your XRootD copy over HTTP protocol
- Checklist for production sites. Requirements for a site to reach (and maintain) production status.
    - ♦ Changes for the TFC. Required changes for TFC at sites to support Xrootd monitoring.
    - ♦ Changes to the site-local-config.xml. Changes necessary to support generic file monitoring for CMSSW.
- Xrootd Monitoring Tests. An overview of the monitoring tests performed.
- Throttling in Xrootd. An overview of what "throttles" are available to the Scalla Xrootd.
- Configuring Fallback Access. How to configure fallback access at your site.
- Configuring CMS Generic File Monitoring. How to configure the generic file monitoring at your site.
- Joining Federation (production or transitional). How to configure site redirector manager or server look up files in the hierarchy of redirectors (a.k.a. AAA).

For Operators

- Operations Guide
- Troubleshooting Guide
- Site Support Guide

Introduction

CMS is exploring a new architecture for data access, emphasizing the following three items:

- **Reliability**: The end-user should never see an I/O error or failure propagated up to their application unless no USCMS site can serve the file. Failures should be caught as early as possible and I/O retried

or rerouted to a different site (possibly degrading the service slightly).
- **Transparency**: All actions of the underlying system should be automatic for the user catalog lookups, redirections, reconnections. There should not be a different workflow for accessing the data ``close by" versus halfway around the world. This implies the system serves user requests almost instantly; opening files should be a ``lightweight" operation.
- **Usability**: All CMS application frameworks (CMSSW, FWLite, bare ROOT) must natively integrate with any proposed solution. The proposed solution must not degrade the event processing rate significantly.
- **Global**: A CMS user should be able to get at any CMS file through the Xrootd service.

To achieve these goals, we will be pursuing a distributed architecture based upon the Xrootd protocol and software developed by SLAC. The proposed architecture is also similar to the current data management architecture of the ALICE experiment. Note that we specifically did not put scalability here - we already have an existing infrastructure that scales just fine. We have no intents on replacing current CMS data access methods for production.

We believe that these goals will greatly reduce the difficulty of data access for physicists on the small or medium scale. This new architecture has four deliverables for CMS:

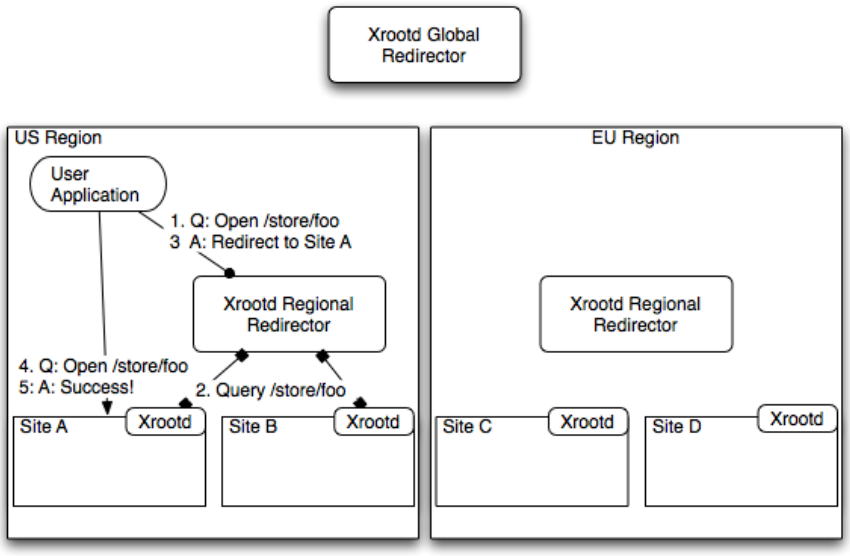
1. A production-quality, global xrootd infrastructure.
2. Fallback data access for jobs running at the T2.
3. Interactive access for CMS physicists.
4. A disk-free data access system for T3 sites.

Architecture

To explore the xrootd architecture, we put together a prototype for the WLCG, involving CMS sites worldwide and all the relevant storage technologies. This prototype wrapped up in January 2011, and we are moving to a regional redirector-based system. This injects another layer into the hierarchy which will make sure requests keep in a local network region if possible.

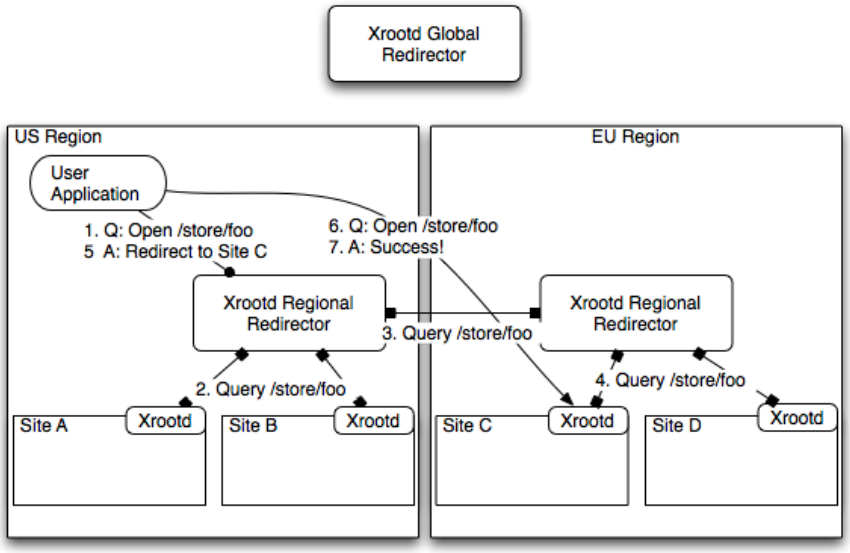
Local-region redirection

The image below shows the communication paths for a user application querying the regional redirector when the desired file is within the region. First (1), the user application attempts to open the file in the regional redirector. If the regional redirector does not know the file's location, it will then query all of the logged-in sites (2). In this diagram, 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). This is all implemented within the Xrootd client; no user interaction is necessary.



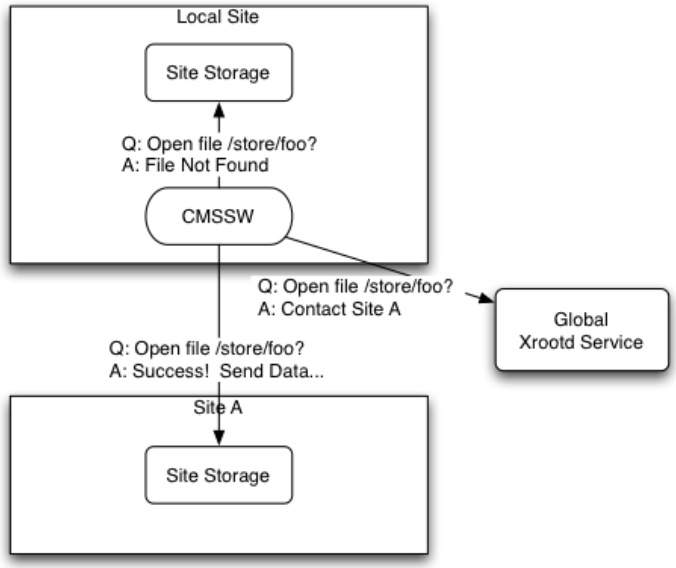
Cross-region redirection

The image below shows the communication paths for a user application querying the regional redirector when the desired file **is not** within the region. This proceeds as in the previous case, except all local sites respond they do not have the file. Then, the regional redirector will contact the other regions (3); if the file location is not in cache , the other regional redirector will query its sites (4). In this example, the user is redirected to Site C (5) and successfully opens the file (6 and 7).



Fallback Access

In the prototype, most sites won't use Xrootd as their primary method; instead, they will use it primarily as a fallback. The image below shows how the file access would work for such a site:



Notes for Project Staff

Participating Sites

US:

1. T1_US_FNAL
2. T2_US_Nebraska
3. T2_US_Caltech
4. T2_US_UCSD
5. T2_US_Purdue
6. T2_US_Wisconsin
7. T2_US_MIT
8. T2_US_Vanderbilt
9. T2_US_Florida
10. T3_US_FNALLPC

UK:

1. T2_UK_London_IC

Italy:

1. T2_IT_Legnaro
2. T2_IT_Bari
3. T2_IT_Pisa

Germany:

1. T2_DE_DESY

Switzerland:

1. CERN EOS

Improving CMSSW I/O

CMSSW has traditionally been very sensitive to latency. In order to make remote streaming feasible, we have been working closely with the CMSSW and ROOT team to provide guidance and code to remove this sensitivity.

The following is a list of changes:

- ROOT TTreecCache functioning (some items landed in 3.3; true functionality was in 3.6).
  - ♦ Squashing accompanying memory leak
- ROOT TTreeCache on by default; Delivered in 3.7
- Fix broken caching on RAW files. Delivered in 3.8 and 3.9
- Fallback protocols in CMSSW. Delivered 3.9
- Xrootd stagein calls. Delivered 3.9
- Removal of non-Event TTrees. Important for high-latency links. Delivered 3.9
- Fix broken caching for Lumi and Run trees. Upcoming (4.2)
- Addition of secondary cache for learning phase. Upcoming (4.2)
- Validation of ROOT 5.26+ auto-clustering. Upcoming (4.2)
- Validation of ROOT 5.32 TFile.Prefetching. Patches sent to ROOT - ROOT 5.34?
- Allow limited backward seeks. Upcoming (5_2)
- Combine read coalescing and vector reads. Upcoming (6_0)
- Switch from TXNetFile to XrdAdaptor. Upcoming (6_0)

Several of these improvements were implemented by others, but benefit us and are listed here.

Tests and Issues

XRootD related

- Tests for the Xrootd Demonstrator (back to 2010 initiative) we've performed are documented on this page.
- We are also trying to document all the issues we observe with the xrootd-based system here: CmsXrootdIssues.
- We record the CMSSW/ROOT I/O improvements needed here: CmsRootIoIssues.

XRootD-AAA related

- This page documents the open scalability tests we've performed.
- Tommaso's findings during CSA14 and AAA drilling

Presentations and Workshops

- Presentations:
  - ♦ CMS Offline+Computing week - scroll down for "AAA Session"
  - ♦ CMS Use of a Data Federation
  - ♦ CMS Lessons Learned & What We Would Have (Done) Differently
  - ♦ Data Federations: CMS Status and Plans
  - ♦ Transcending the AA data access patterns
  - ♦ Performance Tests of DPM Sites for CMS AAA
- XRootD Workshop in UCSD 2015:
  - ♦ https://indico.cern.ch/event/330212/other-view?view=standard
- OSG AHM 2014: Storage Federations (see Friday's agenda)
  - ♦ https://indico.fnal.gov/conferenceDisplay.py?confId=7207

Project Deliverables and Milestones

Project timeline for the US region.

- Progress report for 9 February 2011.
- Progress report for 23 February 2011.
- Progress report for 02 March 2011.
- Progress report for 16 March 2011.
- Progress report for 13 April 2011.
- Progress report for 27 April 2011.
- Progress report for 25 October 2011.
- Progress report for 01 November 2011.
- Progress report for 15 November 2011.
- Progress report for 29 November 2011.
- Progress report for 18 December 2011.
- Progress report for 10 January 2012.
- Progress report for 7 Feb 2012.
- Progress report for 3 April 2012.

This topic: Main > CmsXrootdArchitecture
Topic revision: r54 - 2017-07-31 - MarianZvada