Experience With Wide Area Lustre

OSG All Hands Meeting March 2011

Dave Dykstra, Fermilab
dwd@fnal.gov

Work supported by the U.S. Department of Energy under contract No. DE-AC02-07CH11359

Outline

- Brief intro to ExTENCI project
- Wide Area Lustre subproject goals
- Test setup
- Results so far with CMS application
- Plans

ExTENCI project

- ExTENCI – Extending science Through Enhanced National CyberInfrastructure
- 2 year NSF-funded project
- OSG & Teragrid cooperating
- Four subprojects, many small roles:
  - Wide Area Lustre
  - Virtual Machines & Clouds
  - Overlay Job Scheduling
  - Education & Outreach
- Testing with 7 major projects' applications

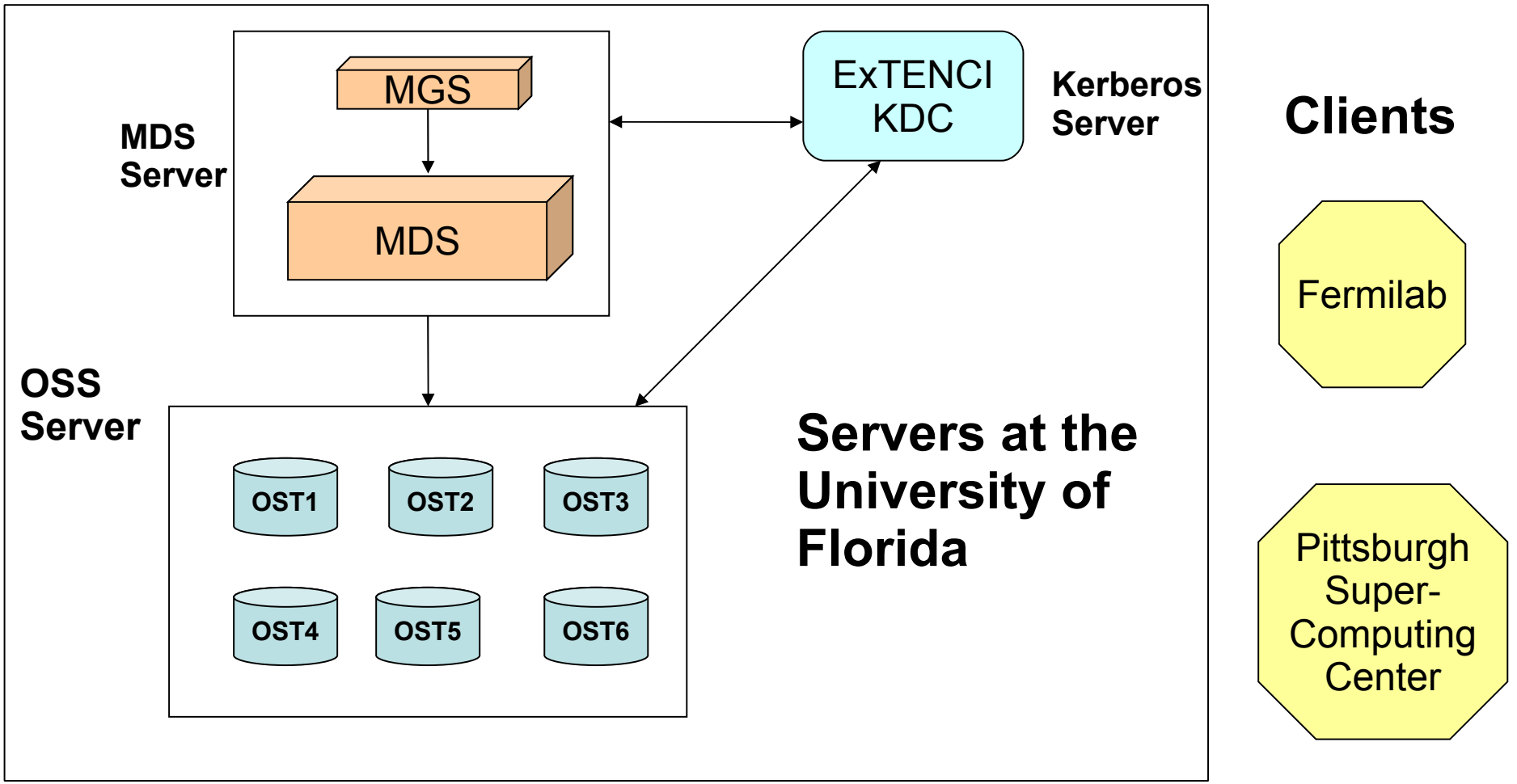
Wide Area Lustre subproject

- Based on Lustre 2.0
- Goals:
  - Verify that Kerberos can effectively be used for access, including (by the end of the project) tying in with existing Kerberos infrastructure
  - See how well it performs
  - CMS: find out if WA Lustre can help with Tier 3 sites
    1. Easy, effective access to large collision data files?
    2. Easy, effective access to software distribution?

Promising Lustre raw performance

- Using the ExTENCI server infrastructure and 2 local Lustre clients, measured 720-750MB/s for read and 430-450MB/s for write (with FDT package, FDT server local at University of Florida and FDT clients at SC10 New Orleans)
- Published Teragrid 2007 test demonstrated 1GB/s for 2 reads or writes over 10 gigabit link between Indiana University and Oak Ridge National Lab in Tennessee
  - Using Myrinet

Wide Area Lustre test setup



Accomplishments so far

- Able to mount remotely both at PSC & FNAL
- Kerberos access works within single extenci.org domain
- One release of CMS software installed on the server
- Performance tests done with remote CMS app

CMS application test

- Only got as far as initial standard test of generating two cmsRun configurations using cmsDriver.py
- Using CMSSW_3_9_7
- >80K file accesses, >70% don't exist
- 20K are open(), 4.6K succeed, 3.2K unique
- Total space used by unique opens: ~7MB

CMS application results

- From FNAL or PSC: test takes 100 minutes first try, 90 minutes second try
  - Similar times at FNAL for cmsRun first step (which is closer to a real job)
- Running on local Lustre client at UF: 53 seconds first try, 27 seconds second try
- Same test using instead CernVM-FS on same FNAL client with server at CERN: 7-3/4 minutes first try, 13 seconds second try
  - First try on second client: 16 seconds (shared squid)

Plans

- Will try to tune Lustre further, but unlikely will gain 2 orders of magnitude for software access without major new client-side caching
- Will proceed to test remote access to large event files
- Later integrate cross-realm Kerberos access
- Set up server also at FNAL

Summary

- OSG & Teragrid working together
- Wide Area Lustre appears to not be suited for software distribution, but still may be for event data distribution