OSG All Hands Meeting

Contribution ID: 51

Data Storage and Analysis for the CMS Tier1: Principles and Practice

Monday, 19-Mar-2012

Primary authors & Presenter:

Catalin L. Dumitrescu

Session classification : Site Technologies

Introduction

(Why is data management important?)

- LHC generates 10-15 PB of data per year
  - (20million CDs – a 14 miles stack)
- Experiments and collaborators are spread all over the world
- Fermi National Accelerator Laboratory provides a larger fraction of the CMS resource share

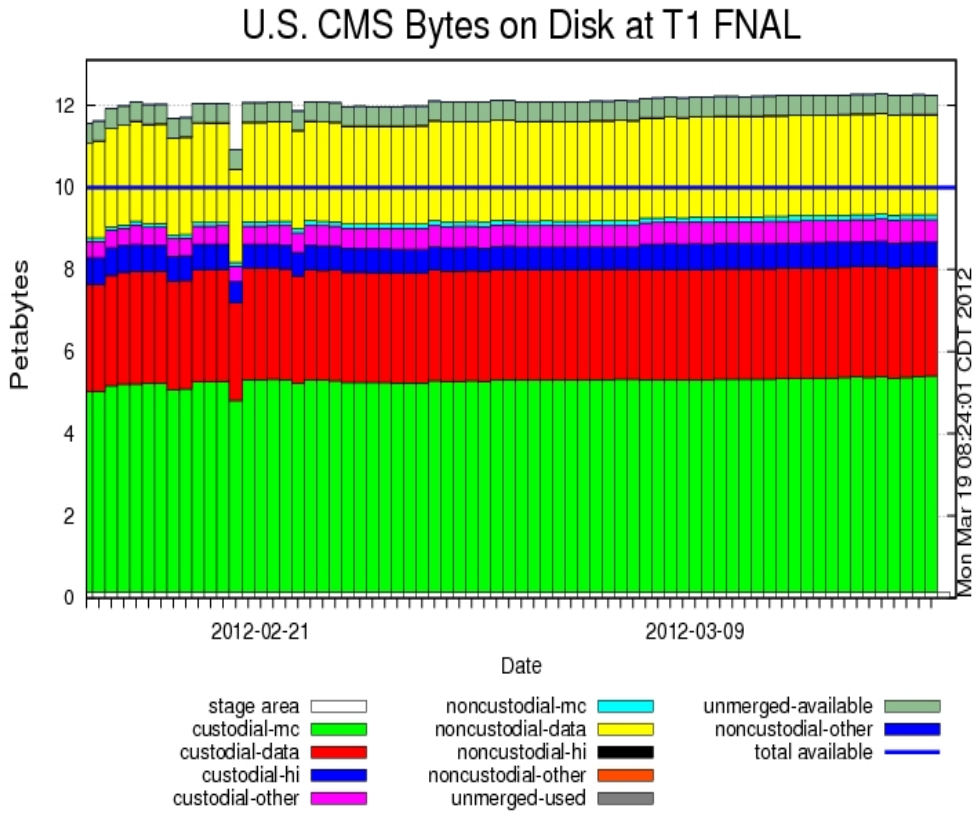
Presentation Overview

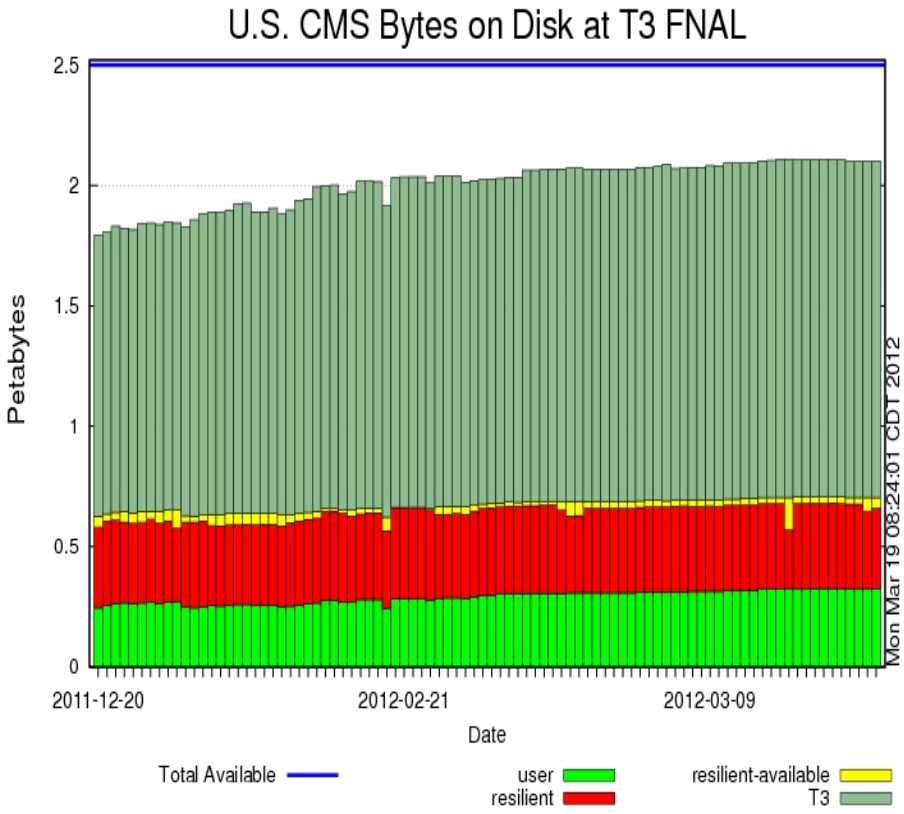
- Introduction
- Deployed Systems
- Principles
- Deployment Details
- Issues & Challenges
- Performance Tunings
- Results
- Next Steps
- Future Directions
- Conclusions

Data Technologies at the T1

- **dCache with PNFS and T10KC tape system**
  - 15 PB disk ; 5TB / tape (?)
- **Xrootd**
  - Provides read-only access to dCache and EOS data
  - No additional storage

Space Distribution





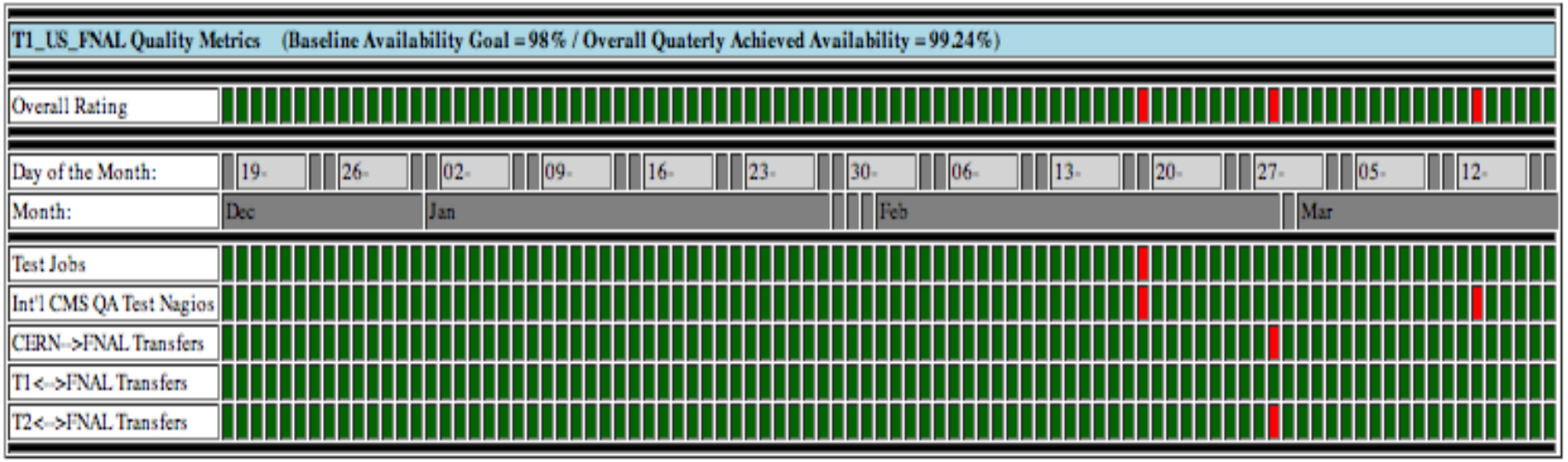
Data Technologies at the T1

- **BlueArc**
  - Used primarily for users' home and data areas (FNAL & CERN)
  - 200 TB disk split over two heads and 3 file systems
- **Lustre**
  - Used for temporary production files to avoid dCache overload
  - 7 OSSs, 42 OSTs = 230TB disk
- **EOS**
  - Under testing for replacing various others data areas
  - 420 TB disk currently

Principles

- **Achieve Availability Agreements**
  - 98% during collision taking
  - 97% during downtimes
- **Consistency and Uniformity for the Data Servers**
  - We deal with hundreds of data servers and 20 PT of data
  - Automation of operations with minimal admin intervention (i.e., automated reboots when nodes are down)
- **Increase QoS by reducing overloads or queuing times**
  - Sustain required data rates
  - Server all requests immediately

Availability



Legend:
- OK
- Scheduled Downtime
- Failure
- Weekend Day
- No data (n/a)

Corrections:

```
T1_US_FNAL,2012-02-18,SAMAvailabilityNagios,red,75%
T1_US_FNAL,2012-02-18,JobRobot,red,75%
T1_US_FNAL,2012-02-28,GoodT1linksfromT0,red,75%
T1_US_FNAL,2012-02-28,GoodT1linksfromT2s,red,75%
T1_US_FNAL,2012-03-12,SAMAvailabilityNagios,red,75%
```

Explanations:

```
2012-02-19: Network failure caused drop in QoS
2012-02-28: failed transfers caused by FTS migration from SL4 to SL5
2012-03-13: Nagios tests failure / New CERN monitoring system in place
```

Issues & Challenges

- Crashes, overloads, HW & SW upgrades
  - 2 days downtime in three months is over the agreement
  - 2-3 kernel releases per quarter, network maintenance, other needed reboots (repairs)
  - Local farms are large – over 12k batch slots total
    - Easily can overload any namespace
- Budget operations to for unexpected downtimes
  - Other parties can affect our availability (GUMS, BDii)
  - Software hidden bugs
  - Users can 'abuse' the system within the rules we set
    - QoS dropping and complains from the rest of the users

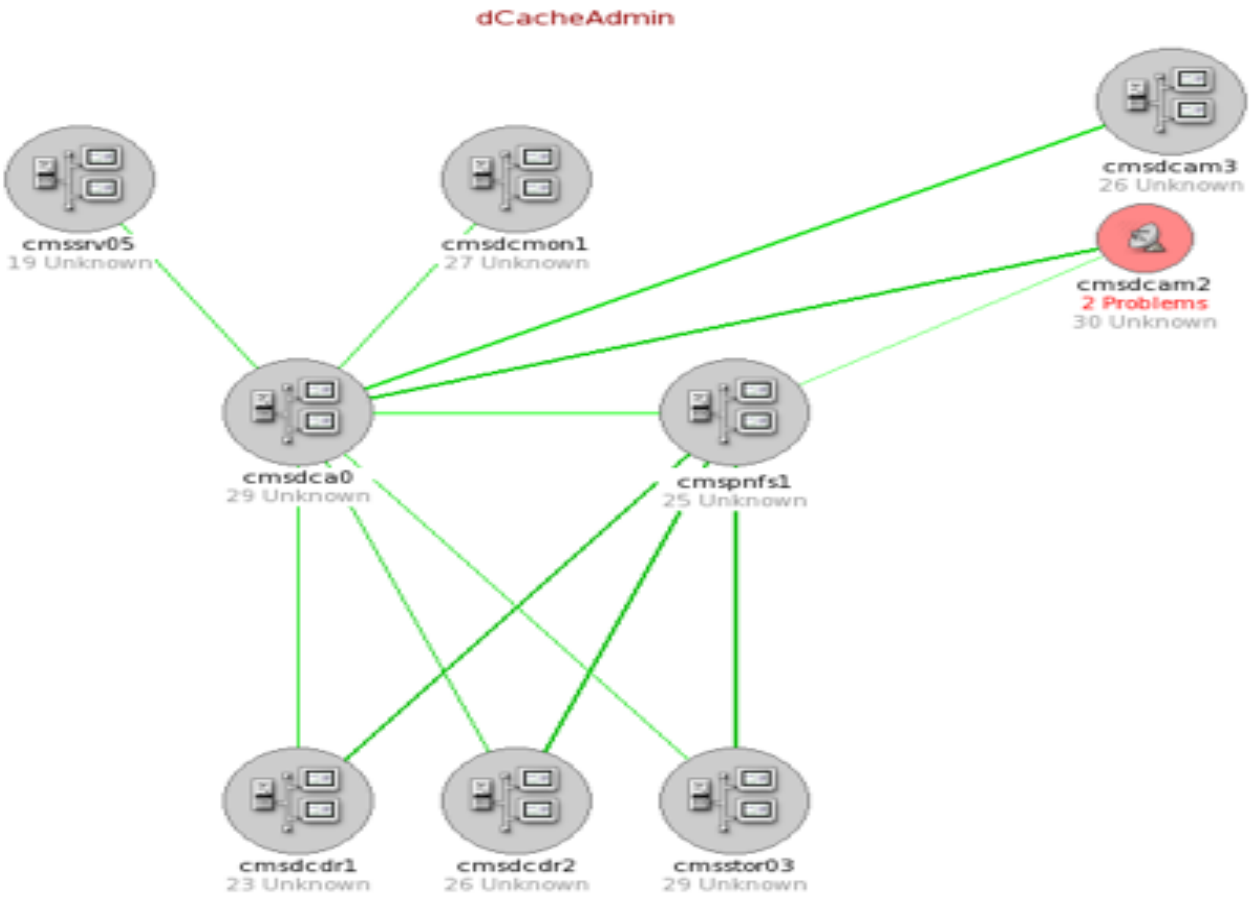
Pro-Active Steps

- Advanced Monitoring and Alert Categorization
  - Zabbix, home grown tools
- Performance Tunings and High-End Servers
- Separation of high demanding operations
  - Small temporary files are saved on Lustre
  - Resilient and scratch spaces are slowly moved to EOS
- User Education, Monitoring and Enforcement

Advanced Monitoring

- Hardware Monitoring
  - Node pinging and restarts
- Components Monitoring
  - Monitor all ftp doors up, all pools up, all services running
- Service QoS Monitoring
  - Monitor transfer quality
  - Alarm on high error rates

Integrated Monitoring



Performance Tunings

- dCache/SRM
  - Use additional caching and translation mechanisms
  - High-end node for the PNFS server
  - Distributed SRM with 0 queuing
  - Allow access through Xrootd directly to data
- Lustre
  - Already an improvement for small production files
  - Monitor load and act on overloaded OSSs/OSTs
  - Use larger memory buffer on the worker nodes
- BlueArc
  - Additional caching mechanisms have been investigated
  - Recommend sometimes using instead EOS

User Education / Usage Monitoring

- dCache/SRM
  - Inform about system limitations like < 1000 files / dir
  - Monitor and notify if happens
  - Explain why using dcap protocol locally has advantages
- BlueArc
  - Provide users with Condor config examples to avoid direct access
  - Encourage using CRAB that has all the knobs
  - Monitor open file descriptors and held jobs for users causing problems
- Lustre
  - Ask for low number of skim jobs that trash the system

Results

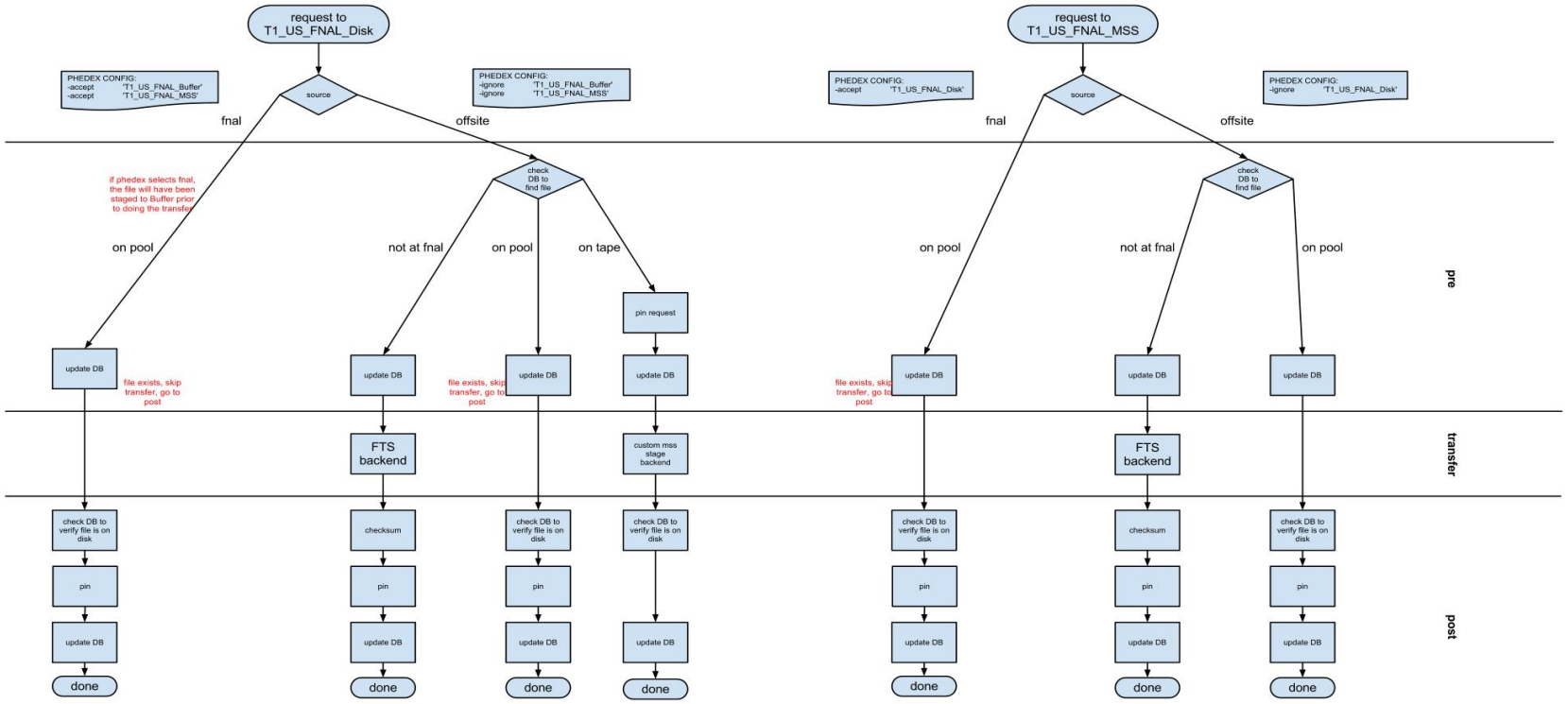
- dCache/SRM
  - Passing availability metrics without problems
  - Deployed 15PB of storage (dCache) and there were 0 downtimes for the last year
  - Sustained high data transfer rates and file counts from all sites
- BlueArc
  - Maintained acceptable transfer rates allowing interactive usage
- Lustre
  - Avoided dCache namespace overloads with skim jobs
  - Delays and overloads affect now only specific production jobs

Next Steps

- Move EOS to production or find alternatives
- Replace Lustre with EOS (lower overall maintenance)
- Grow EOS and grow total storage space at FNAL
- No migration to PNFS for a while (risk, time)
- Downsize BlueArc and focus on increased performance

Future Directions

- Research for the right storage system
- Downsize dCache supporting tape only and grow EOS



Conclusions

- We have a working solution
- We need better solutions for growing the space further (doubling, …)
- dCache has proven to work so far
  - Major upgrades and changes are still to be done

Other Deployed Technologies

- CVMFS (Stratum 1 servers) – software distribution
- PhEDEx, FTS - data transfer management
- Condor - computing resource management
- glideinWMS for job submission
- Various monitoring tools / JobMon, JobView, CondorView
- VMs/Xens