OSG All Hands Meeting

Future Storage Options for Fermilab/CMS Tier 1

Monday, 11-Mar-2013

Primary Author & Presenter:
Catalin L. Dumitrescu

Introduction

- Data Management is Important
  - LHC has generated useful data (10-15PB/year)
  - In 2015 higher energies are planned

- Fermilab Tier1 continues to provide a larger fraction of the CMS resource share (>40%)

- 2000 local and production users access data
- *Remote data access has gain importance through the AAA project*

Presentation Overview

- Introduction & Principles Review
- Deployed Systems & Ongoing Issues

- New CMS Requirements
- Ongoing Challenges

- System Growth & Simplification Plans
- Storage Evaluation Results
- Conclusions

Principles Review

- Availability Agreements
  - 98% during collision taking
  - 97% during downtimes

- Consistency and Uniformity for Data Servers
  - hundreds of data servers / 40 PB of data
  - automation in case of failure is a must

- QoS remains important
  - sustainable performance
  - rich feature-set for users and production

Deployed System

- dCache 1.9.5 with PNFS
  - bypassed weaknesses seen over years
  - PNFS performance is monitored carefully
- Lustre still used for small temp area

- xrootd 3.2.7 underneath / remote access
- EOS 0.2.29 / alternate user home areas
- BlueArc for home and data areas

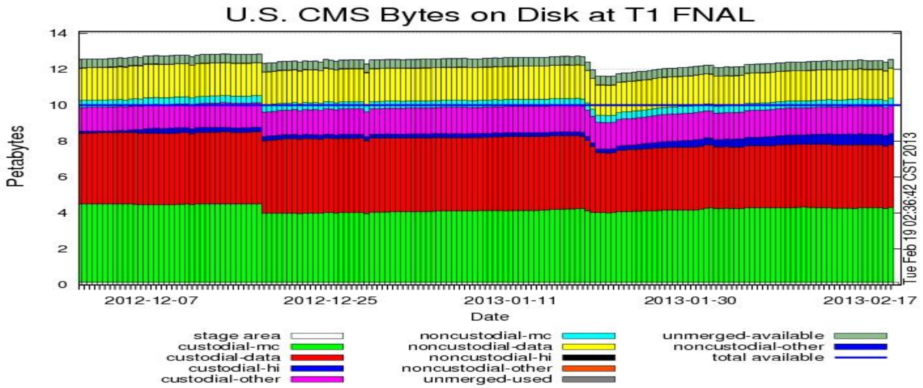
- Total: 5 technologies == difficult to manage

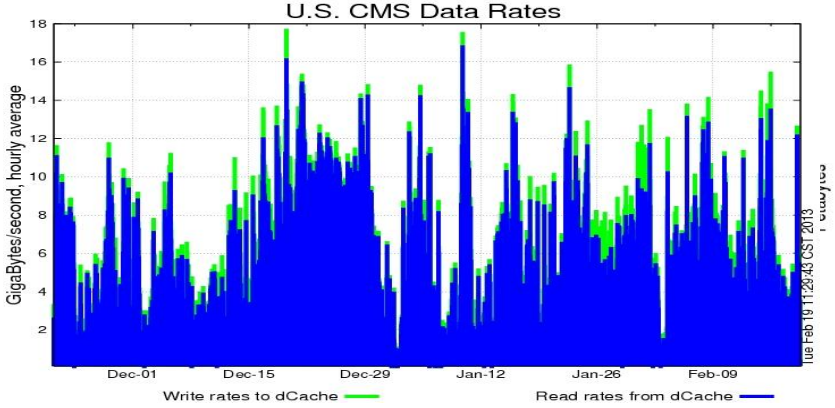
Achievements

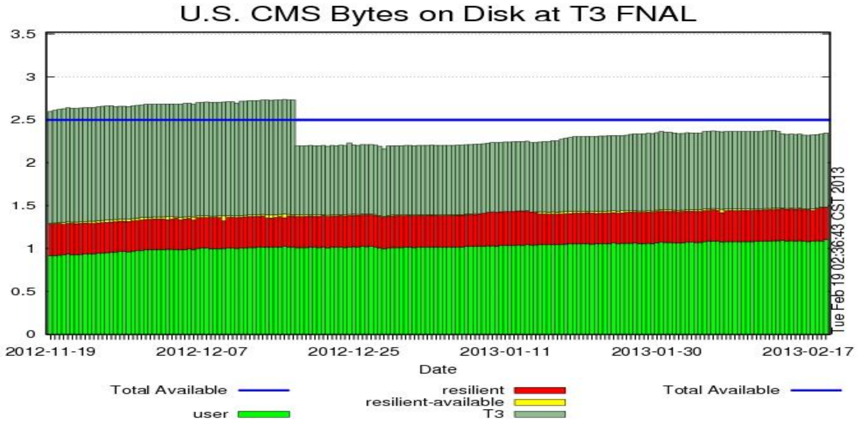
- Overall
  - deployed 17PB of storage and 40PB on tapes
  - pass the availability metrics all the time
  - top site for 2012 availability metrics
- dCache & Lustre
  - provide data above users / production expectations
  - access to 40PB of data with 0 downtimes
- EOS
  - highly performant compared to other systems
  - transparent upgrades (at any time)

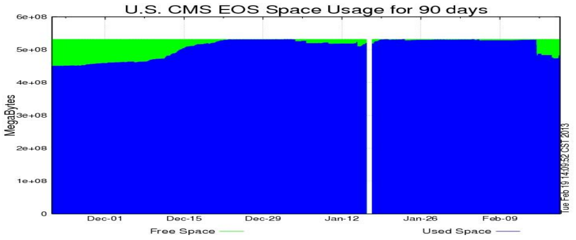
Space Distribution - 17PB / 40PB

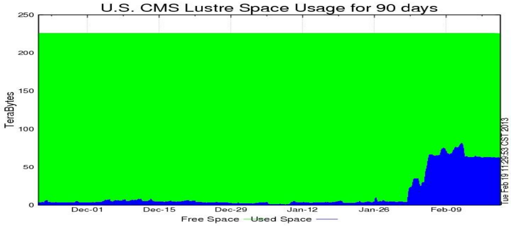
- dCache - 15 PB
- Lustre - 200 TB
- EOS - 520 TB
- BlueArc - 250 TB

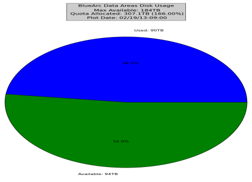












New CMS Requirements

- CMS Operations want control via PhEDEx
  - file staging to disk and saving to tape
  - common solutions for simplified data handling

- New protocols and algorithms require also storage reevaluations

- Storage space increases 20% every year (?)

Ongoing Issues

- **dCache**
  - fragile PNFS - better alternatives available
  - sync to the next golden release
- *Lustre*
  - cannot afford network saturation
  - configuration changes (bugs) bring system down
- **EOS**
  - CERN support only
  - production validation still pending
- Overall (including BlueArc)
  - too many systems to be maintained
  - HW space splitting over different technologies
  - ongoing performance tunings / user education

Challenges for 2013-2014

- On the fly system upgrade
  - 0 downtimes, easy upgrades
- Helpful monitoring and interfacing tools
- QoS provisioning

- Reduced homegrown tools, performance tunings and local monitoring
- Increased production farms and new remote access patterns (AAA project)

System Growth & Plans

- Target is 18-20PB on a single technology

- Support for new protocols (xrootd, POSIX)
- Higher performance and reliability from one single storage (instead of dCache + Lustre)

- Upgrades through migration:
  - build a new instance - 80% of the space
  - reduce the tape backend instance - 20%

Evaluation Criterias

- Minimal performance requirements
    - 100Hz for operations
    - 0.7GB/s for tape writing
- reliability
    - less unplanned & planned downtimes
    - data available when needed and with minimal effort
- POSIX interface (users)
    - EOS has proved its importance
- CMS needed protocols
    - xrootd is largely used for production / CMSSW
    - POSIX interface is useful

Considered Solutions

- dCache 2.2.7
  - handles large amounts of data, POSIX interface, performance, good support and long term development plans
- EOS 0.2.29
  - POSIX interface, xrootd, easy deployment on SLF5 or SLF6
- Hadoop 2.0
  - OSG support, additional tools available, POSIX interface
- Lustre 1.8.6
  - POSIX interface

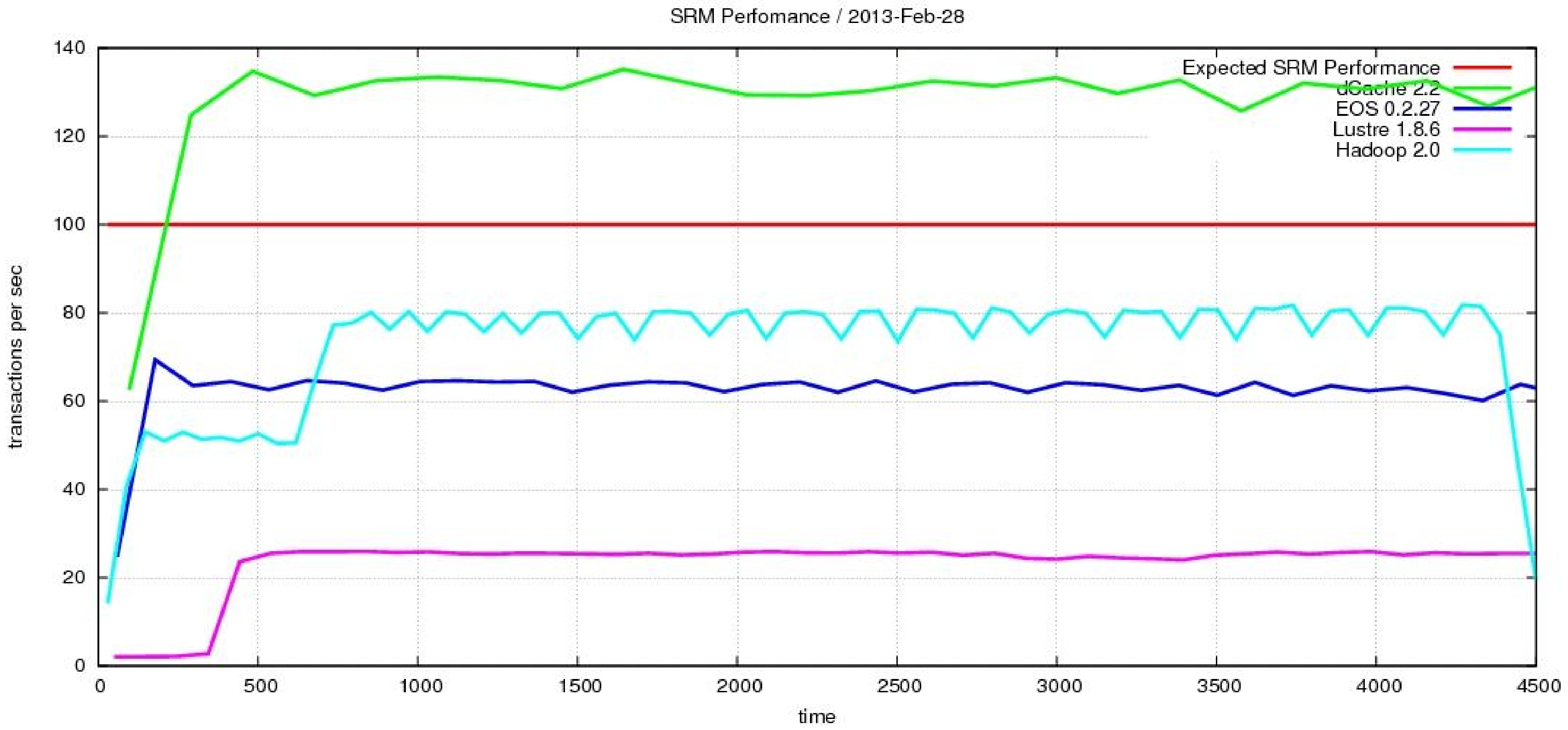
Testing Setup and Approach

- Environment
  - 270 test nodes connected over 1GB/s
  - 1 to 100 testing threads / node
  - pool of 100 files
  - load increase every 1 second

- Advantages
  - identification of service saturation
  - identification of breaking point
  - easy to find *performance vs. clients*

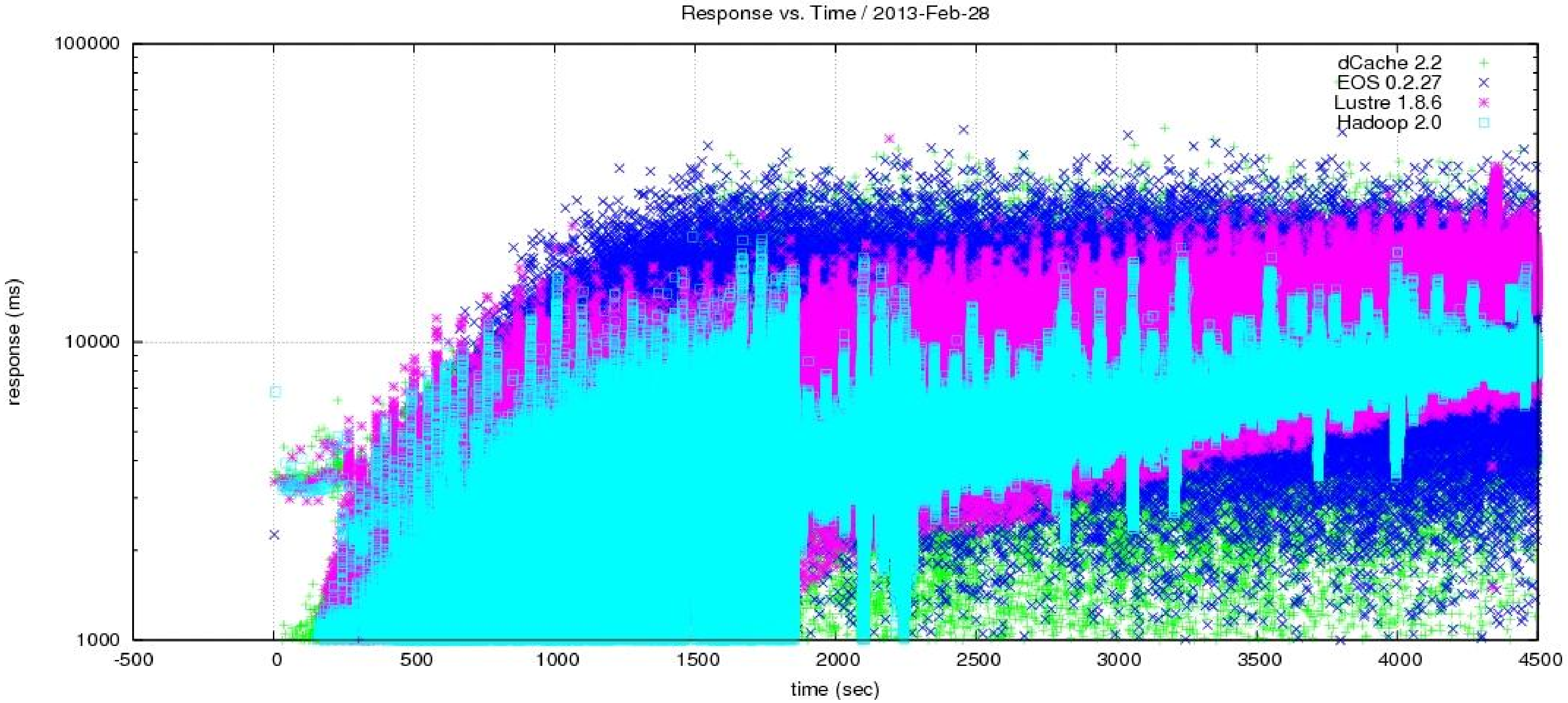
Evaluation Results - SRM

- OPs for distributed load from 300 nodes ; thousands of threads



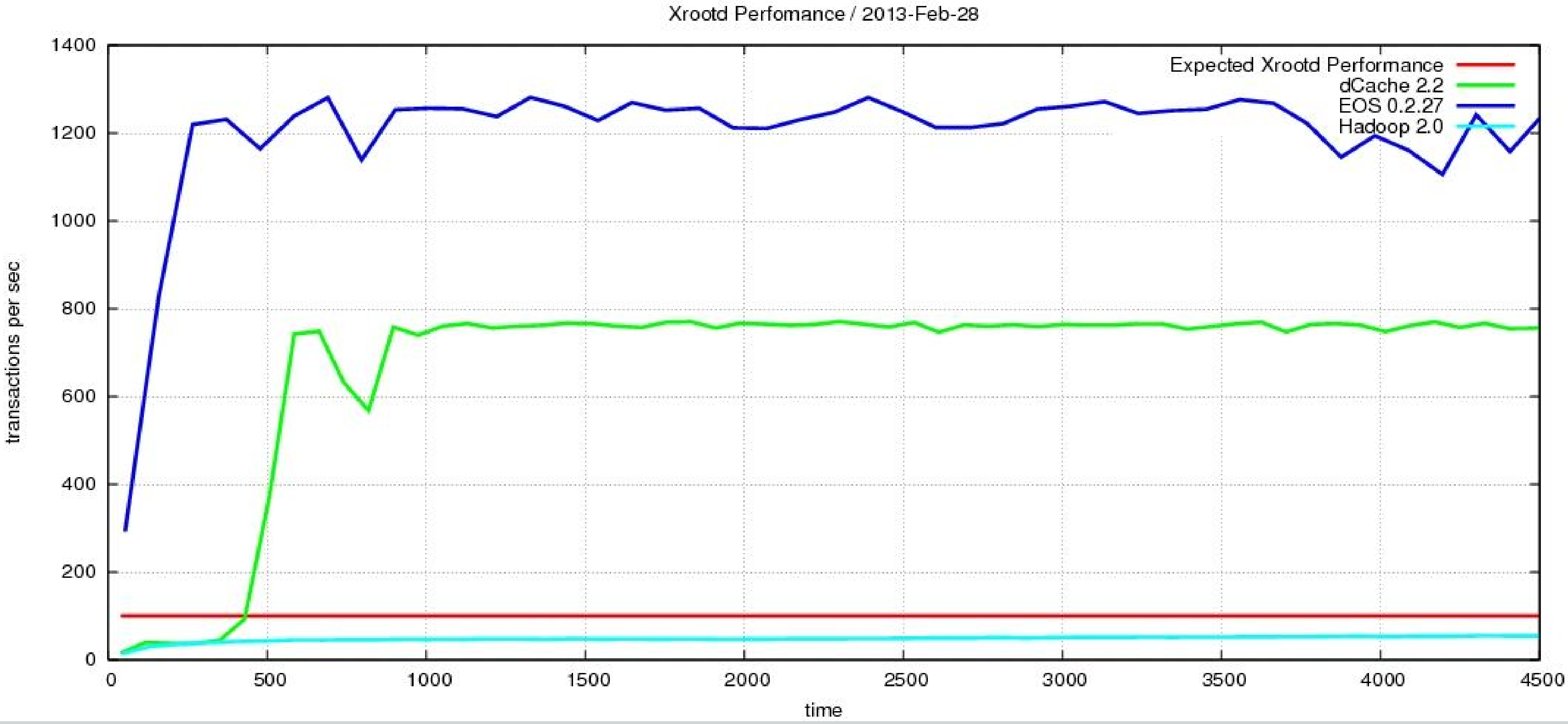
Evaluation Results - SRM

- Response time for the same load



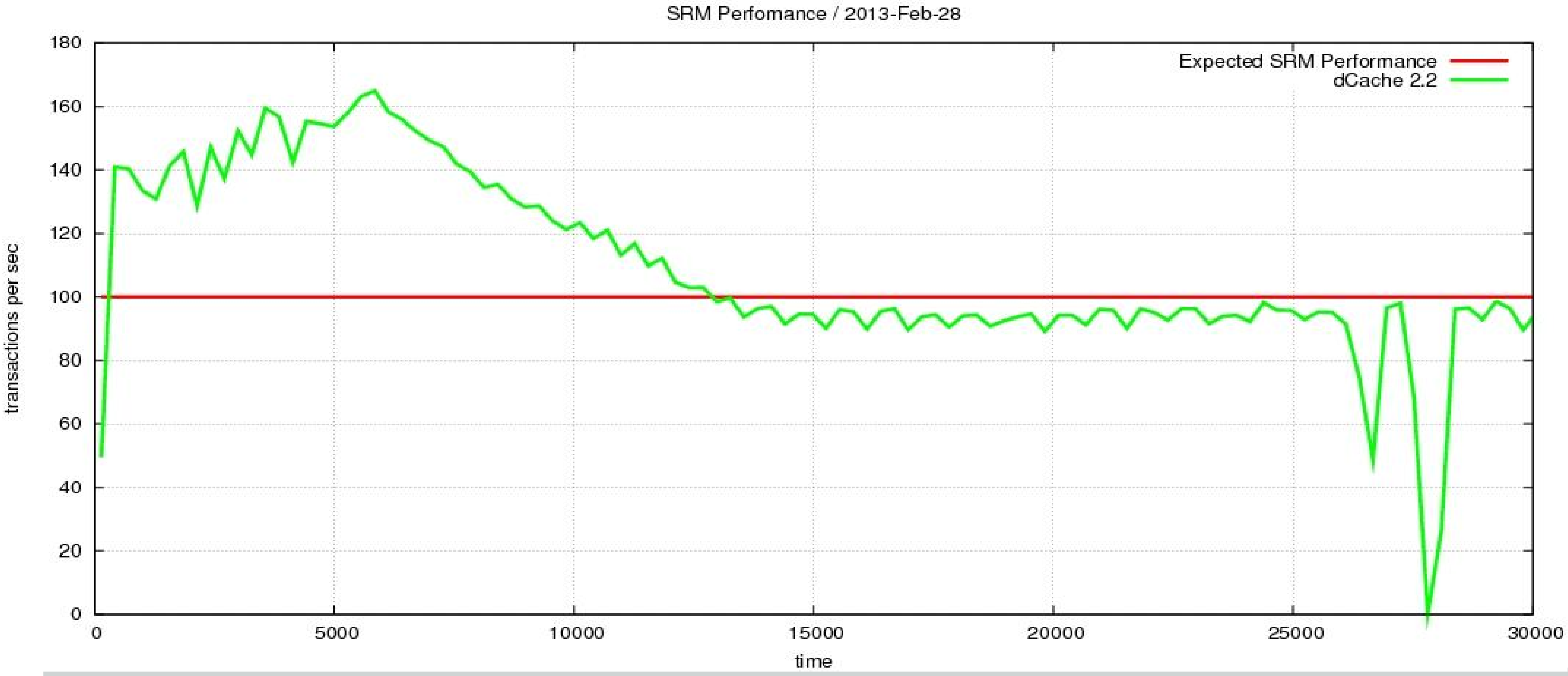
Evaluation Results - xrootd

- xrootd OPs for clients from 300 nodes and thousands of threads



Evaluation Results - dcap

- dCache / dcap evaluation for clients running on 300 nodes



Planning for the Future

- Authorization schemas
  - SSL implementation
  - GSI evolution support
  - GUMS evolution support
- Protocols
  - SRM scalability / development
  - xrootd
  - other protocols
- Easy of use
  - support for known protocols and interfaces
  - easy of deployment on various OSs

Deploying with the Future in Mind

- Why splitting?
  - plan with safety in mind
  - possibility for replacement
- Why one (or few) technologies?
  - learning curve reduction
  - keeping with updates and less effort
- Why dCache?
  - performance is acceptable
  - support and development plans are strong
  - new technologies incorporation is ongoing
  - Enstore integration is unique

Conclusions

- It is difficult to predict
  - next steps are expected to provide a stable system for at least 1 to 2 years
- Testing and results are important
  - help in ensuring that dCache scales if right protocols are used
  - improve requests for development directions
- Collected experience is important
  - dCache has worked
  - EOS is liked by users and very easy to manage

Questions?