



dCache

Dmitry Litvintsev, Fermilab

OSG Storage Forum, September 21, 2010

Happy Birthday dCache



On 09/16/2000 dCache project was accepted by
DESY computer review board

What is dCache

- Distributed Peta Byte Disk Storage with single rooted filesystem providing location independent file access
- A cache front-end to Tertiary Storage to optimize media I/O
- An implementation of Grid Storage Element with standard data access protocols, Authentication and Authorization, Information Provider and SRM

dCache Concepts

cell - basic component of dCache

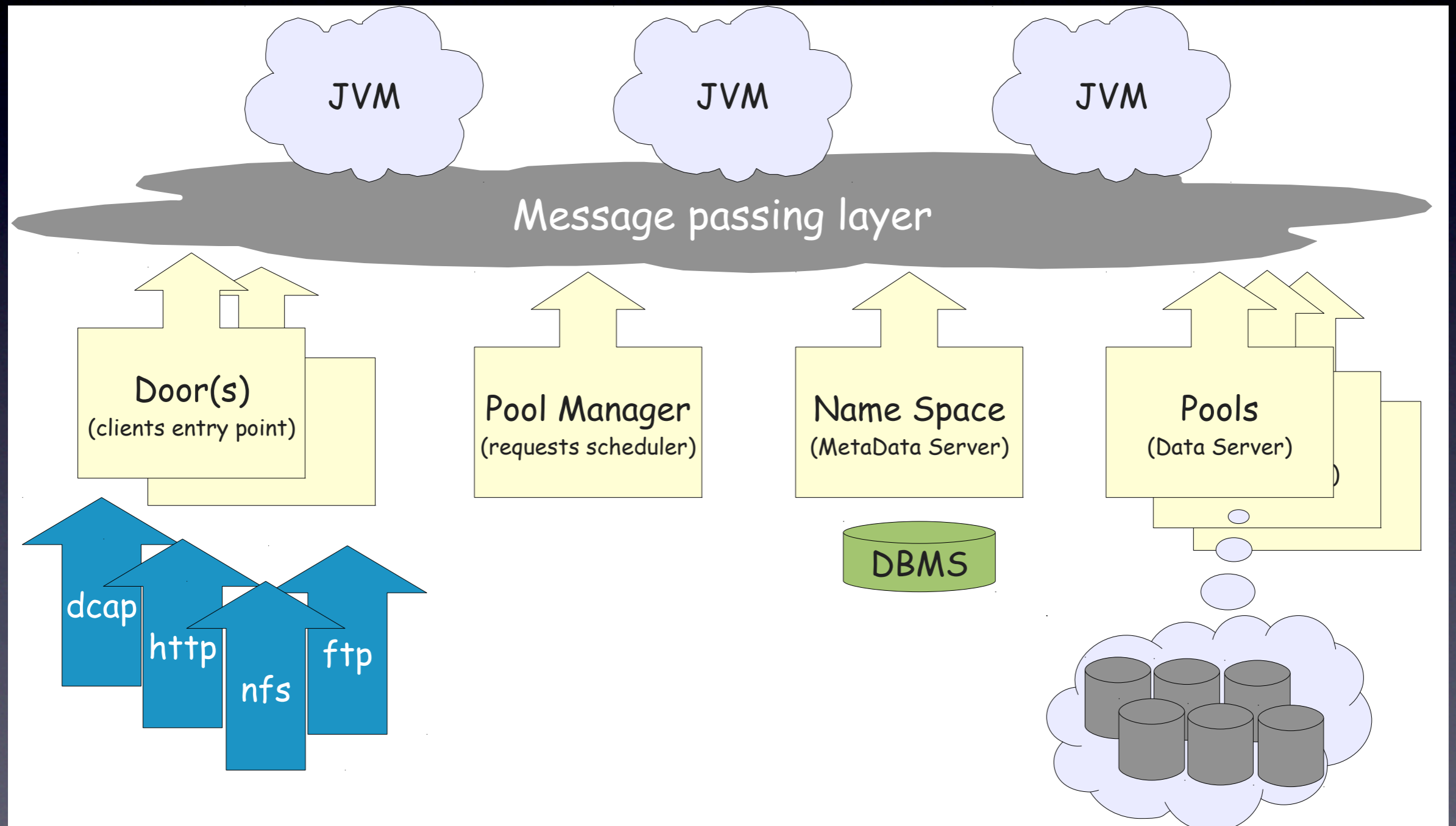
domain - container, hosting one or more cells. A domain runs within single JVM. Must have unique name in dCache instance

cells communicate via proprietary cell messaging on top of TCP/IP

concrete cells implement specific servers:

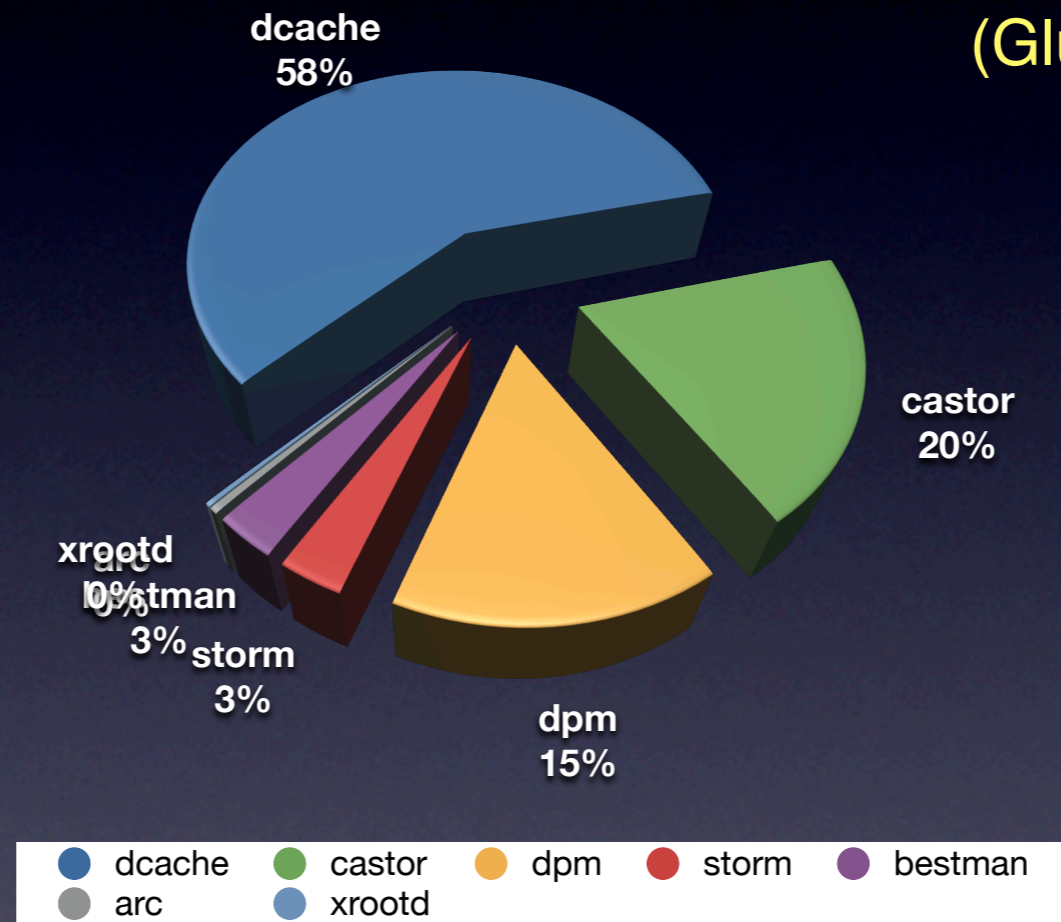
- **I/O Doors** - client access points, implement particular access protocol and perform authentication if required. **(gsi,kerberos)ftp**, **WebDAV**, **http(s)**, **nfs v4.1**, **(gsi,kerberos)cap**, **gridftp**, **xrootd**, **SRM** etc.
- **PnfsManager** - interface to namespace
- **Pool** - provides physical data services over contiguous disk area
- **PoolManager** manages the collection of pools
- Authorization and other resource management - **gPlazma**, **SpaceManager**, **PinManager**, **SrmSpaceManager**, **ResilientManager** etc.

Basic dCache Design



Total Online Space Share

52.5 PB in dCache
(GlueSETotalOnlineSize)

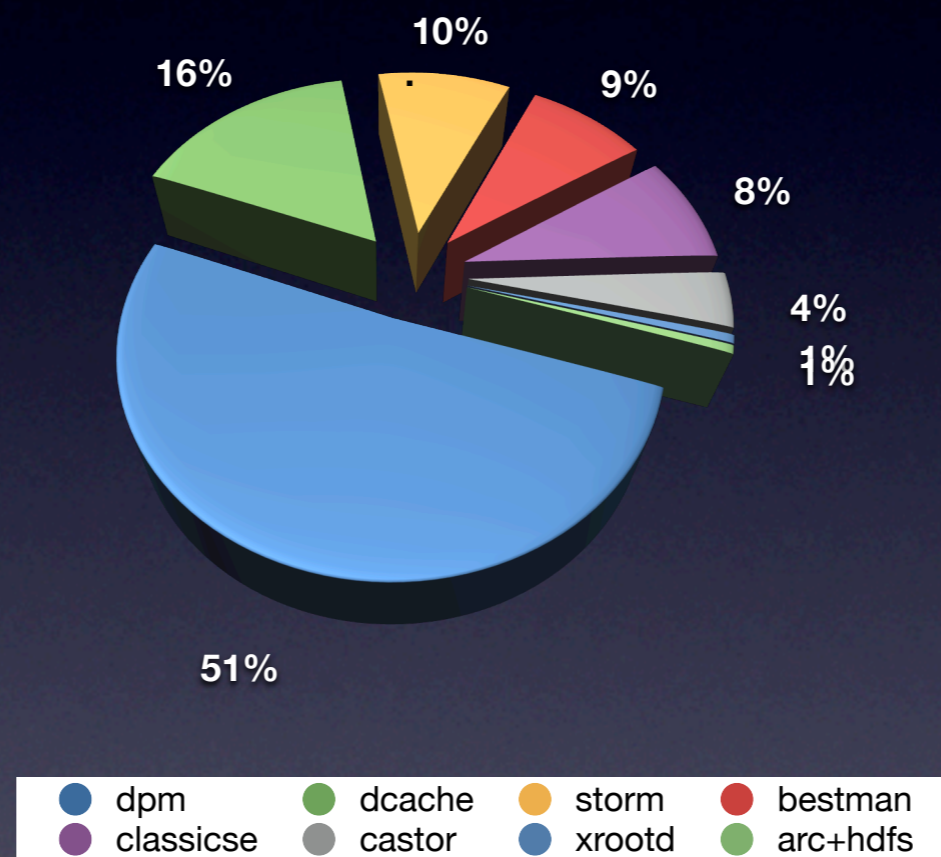


Extracted from

```
ldapsearch -LLL -x -H ldap://lcg-bdii.cern.ch:2170 -b -o grid '(&(objectClass=GlueSE))'
```

Popularity

77 dCache SEs



Extracted from

```
ldapsearch -LLL -x -H ldap://lcg-bdii.cern.ch:2170 -b -o grid '(&(objectClass=GlueSE))'
```

dCache @ Fermilab

	total online	precious	on tape
CDF	0.4 PB	0	6.3 PB
public	0.1 PB	6 TB	2.5 PB
CMS	8.1 PB	0.2 PB	6.9 PB

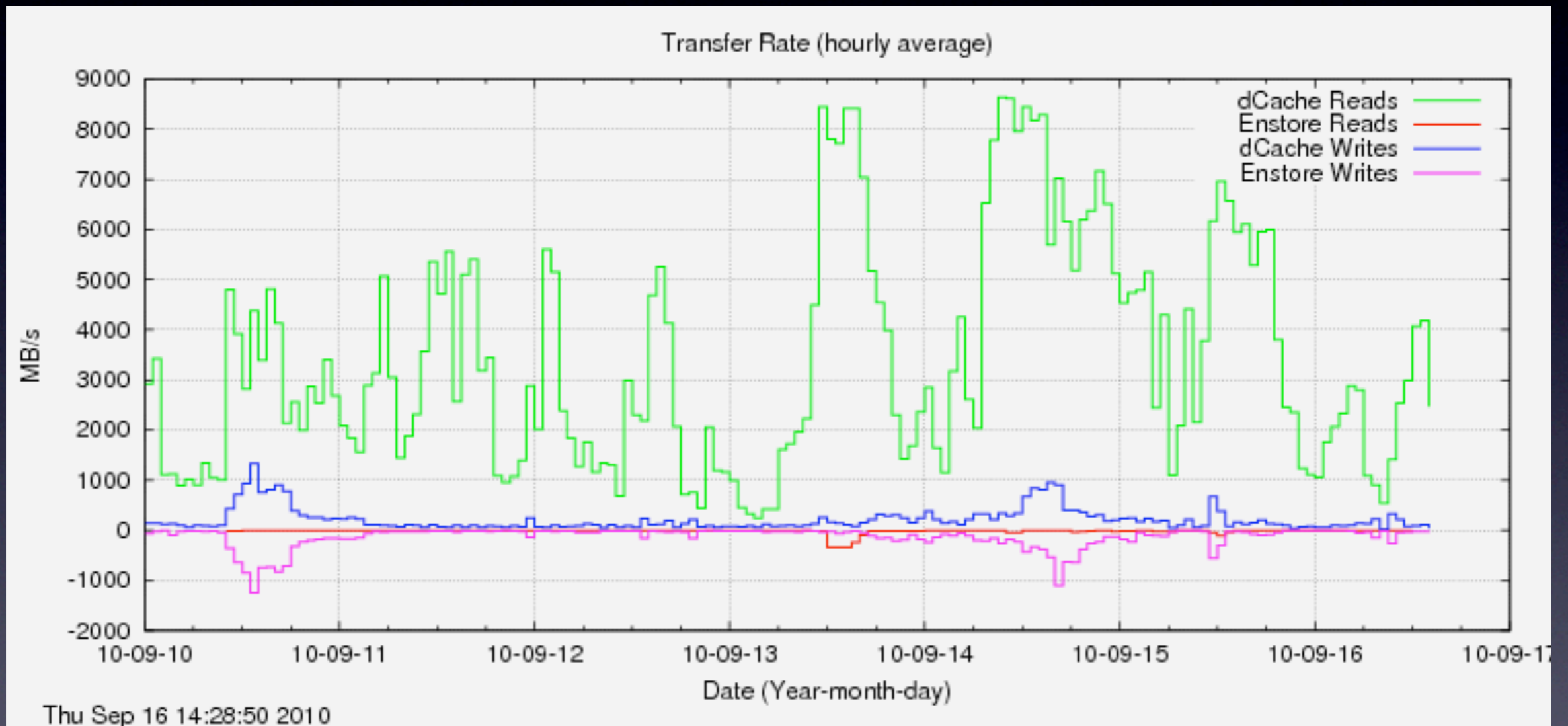
httpdDoors:

<http://cdfdca.fnal.gov:2288/>

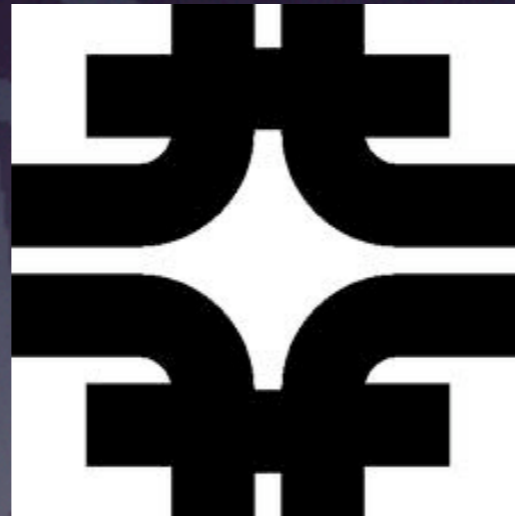
<http://fndca.fnal.gov:2288/>

<http://cmsdcam.fnal.gov:2288/>

Transfer Rates CMS T1, FNAL

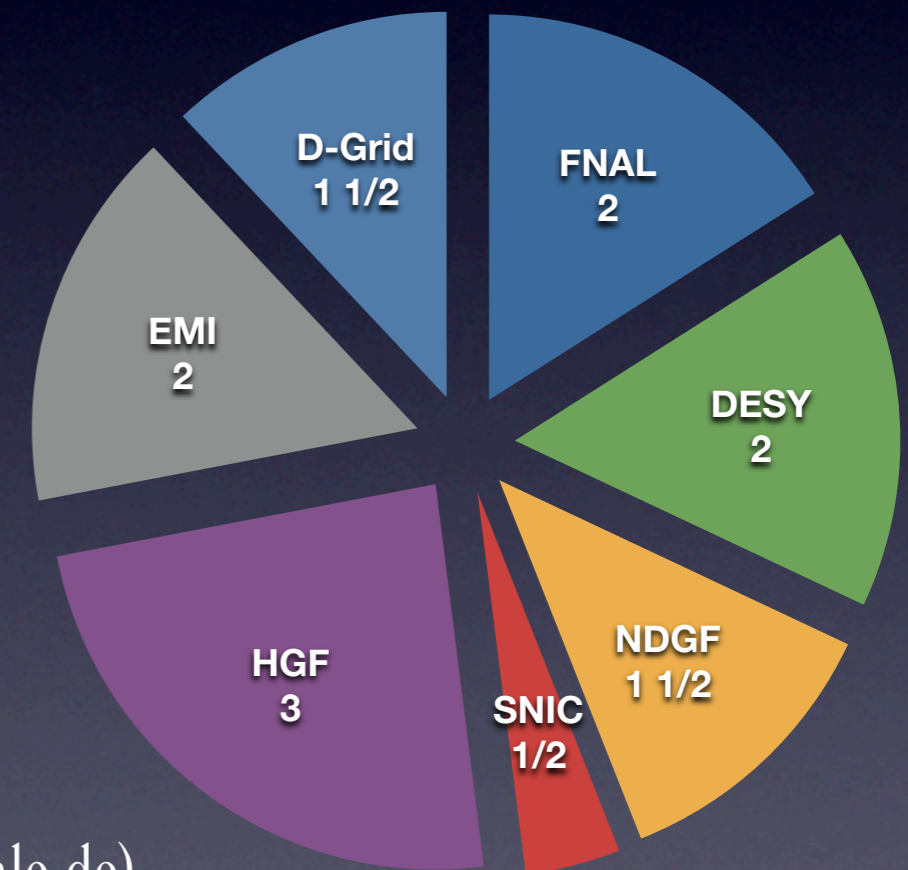


dCache Collaboration





dCache Funding

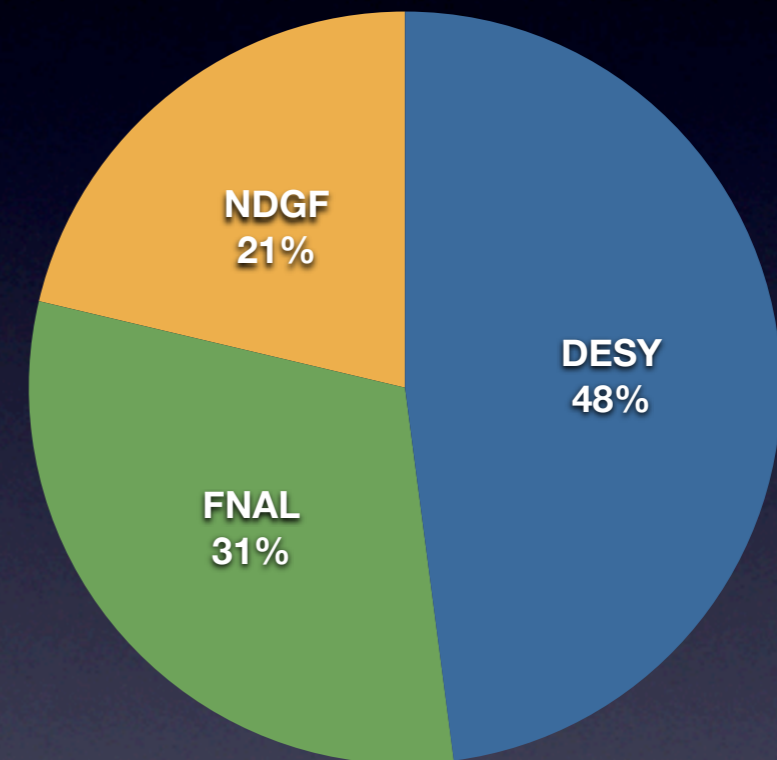
- Labs:
 - FNAL
 - DESY
- Organizations:
 - NDGF (www.ndgf.org)
 - European Middleware Initiative (EMI) (www.eu-emi.eu)
 - Swedish National Infrastructure for computing (SNIC) (www.snic.vr.se)
- German Government:
 - Helmholtz Alliance “Physics at the Terascale” (www.terascale.de)
 - German D-Grid “Integration Project II” (www.d-grid.de)



Collaborative Development

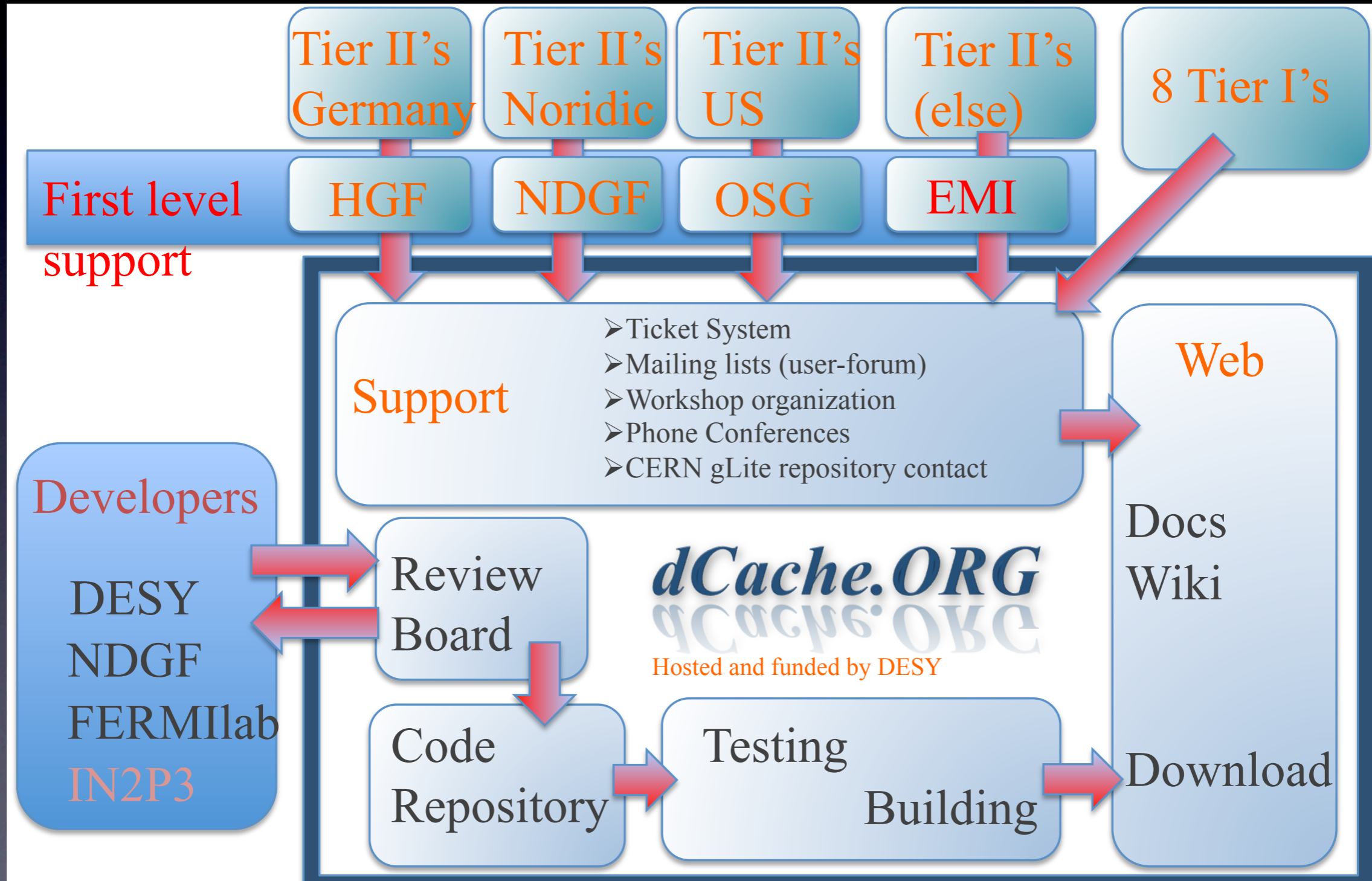
- Pre-commit peer code review
use [reviewboard](#) by Google
- Weekly developers tele confs.
- Minutes and docs on Wiki.
- SVN code repository + [Mercurial](#) for patch management.
moving to [Mercurial](#) 
- Automated building and testing with Hudson.
moving to [Maven](#) 
- Trunk is always releasable.
- Release manager controls patches to branches.

trunk activity

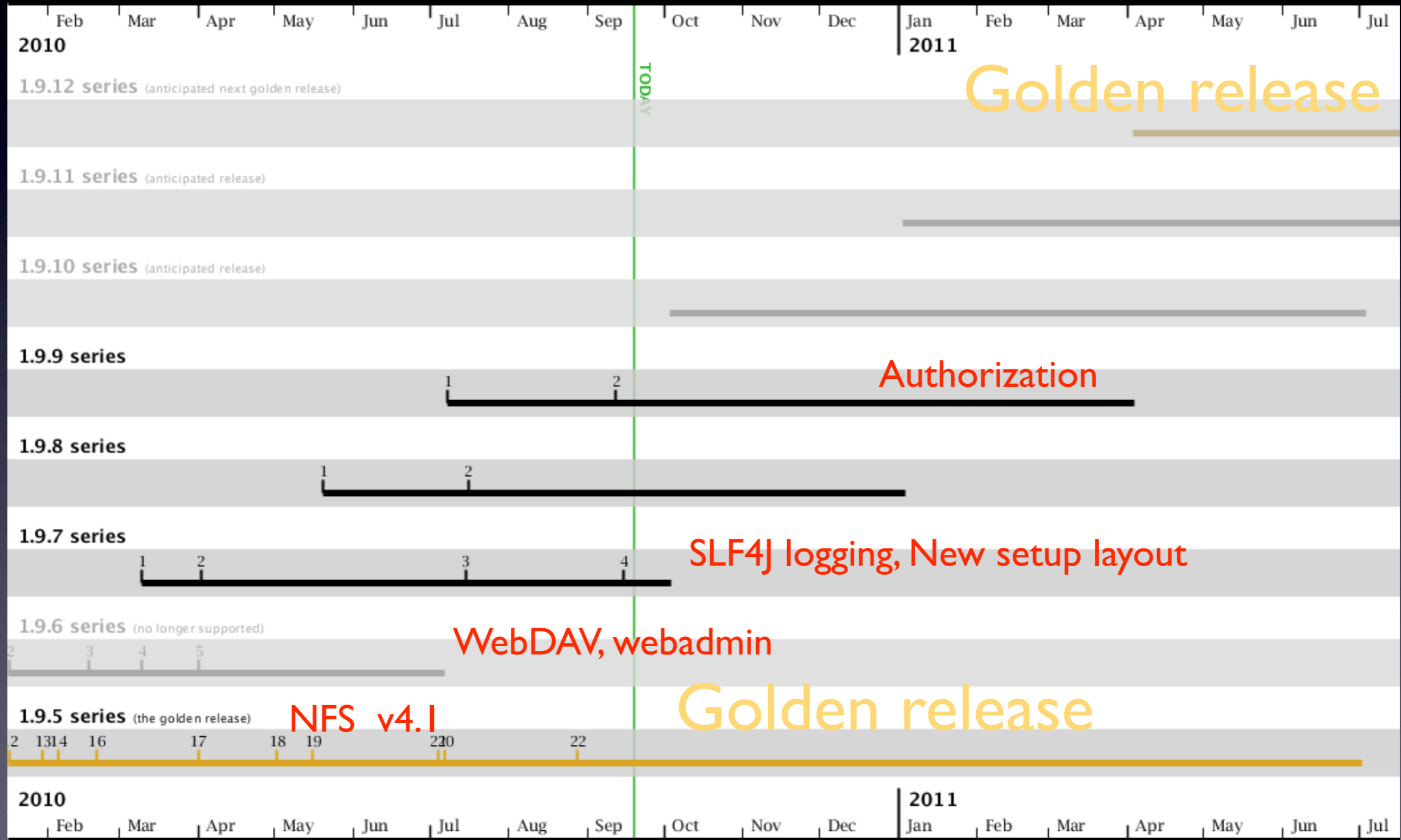


About 1.2K commits per year

dCache Process



dCache Release Schedule



Feature releases

Evolution from Within

- Gradual introduction of newer, widely accepted technologies
 - Spring Framework
 - JMS tunnel for inter-cell communications
 - SLF4J replaced of log4j
 - Jetty to host Web Services
 - Extensive use of Java generics
- A lot of code cleanup and re-factoring
- Deadlock resolution and bug fixes :)

Namespace speedup

- Limitations of PNFS:
 - metadata access only thru NFS server. A bottleneck.
 - metadata stored as BLOBs (no metadata query functionality)
- **Chimera** is a replacement for PNFS. Available since 1.8.0-15.
 - Chimera is Java API, access library and RDBMS providing direct access to metadata bypassing NFS interface
 - **No Need to mount NFS anywhere in dCache**

Fermilab in collaboration with PIC is preparing Enstore migration to Chimera

Going Standard

- New user communities require standard access protocols on multiple OSes
 - HTTP
 - FTP
 - NFS v4.1
 - WebDAV

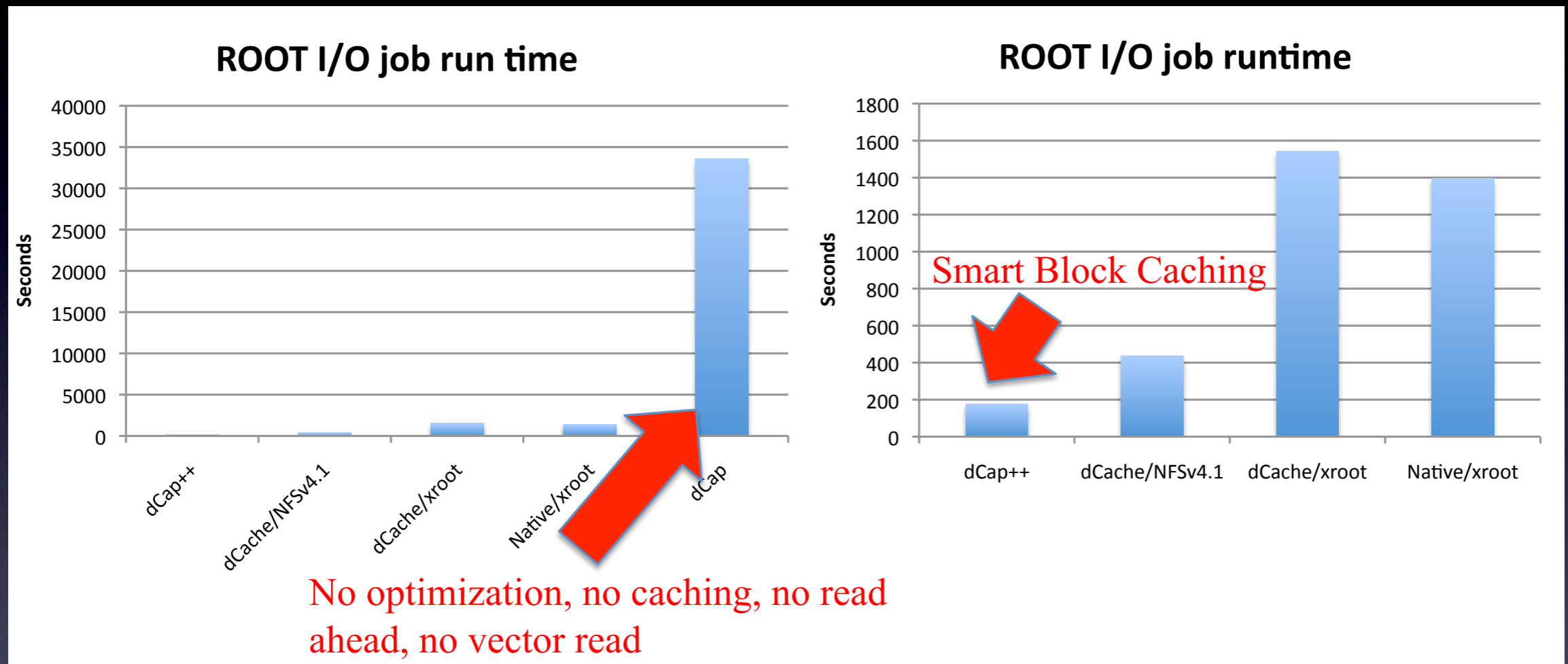
NFS 4.1 (pNFS)

- Compared to [dcap/rfio/xrood/gsiftp](#) NFS v4.1
 - industry standard
 - clients are provided and maintained by others
 - client caching (regular file system cache)
- Compared to NFS v2:
 - Compound RPC calls (multiple ops, one RPC call)
 - Security GSS API is part of the specs
- Support of client redirect to disk pools
- pNFS can be mounted on worker node as any other NFS and data can be directly accessed using real POSIX I/O (for in-kernel clients)

pNFS Server in dCache

- Nameserver and I/O available since 1.9.5
- Immutable files only (as always in dCache)
- No striping
- Security:
 - Kerberos since 1.9.9
 - X509 being evaluated
 - Full ACL support (via admin interface)
 - Automatic tape restores disabled (to protect tape system)

From Hepix



Access: reading every 100th event out of 53K events from a non optimized Atlas event file

NFS v4.1 Working Group Activities :

<https://twiki.cern.ch/twiki/bin/view/EMI/Emijra1DataDetailsNFS41>

More to be presented @ CHEP

WebDAV Door

- **Web-based Distributed Authoring and Versioning (WebDAV)** - extension of http allowing **file operations** (browsing, upload and download) and **namespace operations** (rm, mv, cp)
- Supported in Windows, Mac OS X, Linux.
- WebDAV door in dCache available since 1.9.6
 - x509 client certificate based authentication, or
 - username/password basic authentication (http or https)
 - RFC proxy certificate and VOMS attributes support soon

Browse, Drag & Drop on the Grid

WebDAV Door











Location: http://dmsdca03:2880/pnfs/fnal.gov/data/testers/NULL/litvinse/unavailable/

[S:/ 131.225.84.42]



/pnfs /fnal.gov /data /testers /NULL /litvinse /unavailable

File System

Name	Last Modified
 file.data	Thu Sep 09 15:00:41 CDT 2010
 file1.data	Thu Sep 09 15:23:03 CDT 2010
 file2.data	Thu Sep 09 15:26:03 CDT 2010
 file3.data	Fri Sep 10 09:54:27 CDT 2010
 file4.data	Fri Sep 10 11:19:48 CDT 2010
 file5.data	Fri Sep 10 16:31:30 CDT 2010
 file6.data	Wed Sep 15 14:50:06 CDT 2010
 subdir/	Wed Sep 15 15:15:41 CDT 2010

www.dCache.org

dCache.org - File System - Konqueror

Location Edit View Go Bookmarks Tools Settings Window Help

Location: <http://dmsdca03-2880/pnfs/fnal.gov/data/testers/NULL/lityinse/unavailable/>

dmsdca01.fnal.gov - Konqueror

Open 'http://dmsdca01.fnal.gov/data/testers/NULL/lityinse/unavailable/file6.data'?
 Type: application/octet-stream

Do not ask again

Save As... Open With... Cancel

Name	Last Modified
file.data	Thu Sep 09 15:00:41 CDT 2010
file1.data	Thu Sep 09 15:23:03 CDT 2010
file2.data	Thu Sep 09 15:26:03 CDT 2010
file3.data	Fri Sep 10 09:54:27 CDT 2010
file4.data	Fri Sep 10 11:19:48 CDT 2010
file5.data	Fri Sep 10 16:31:30 CDT 2010
file6.data	Wed Sep 15 14:50:06 CDT 2010
subdir/	Wed Sep 15 15:15:41 CDT 2010

www.dCache.org

dCache.org - File System - Konqueror

Location: <http://dmsdca03-2880/pnfs/fnal.gov/data/testers/NULL/litvinse/unavailable/>

dmsdca01.fnal.gov - Konqueror

Open 'http://dmsdca01.fnal.gov/data/testers/NULL/litvinse/unavailable/file6.data'?
 Type: application/octet-stream

Do not ask again

Save As...

Save As - Konqueror

/home/litvinse/

Desktop
 Home Folder
 Storage Media
 Network Folders

ne-fermi-config-for-cdf
 ne_info
 dcache-tools
 Desktop
 desy
 docs
 Downloads
 e
 eclipse
 elib
 enst
 enst

Location: Save

Filter: Cancel

Automatically select filename extension

	file3.data	Fri Sep 10 09:54:27 CDT 2010
	file4.data	Fri Sep 10 11:19:48 CDT 2010
	file5.data	Fri Sep 10 16:31:30 CDT 2010
	file6.data	Wed Sep 15 14:50:06 CDT 2010
	subdir/	Wed Sep 15 15:15:41 CDT 2010

www.dCache.org

Location Edit View Go Bookmarks Tools Settings Window Help

Location: <http://dmsdca03-2880/bnfs/fnal.gov/data/testers/NULL/litvinse/unavailable/>

dmsdca01.fnal.gov - Konqueror

Open 'http://dmsdca01.fnal.gov/...'

Type: application/octet-stream

Do not ask again

Save As...

Desktop

Home Folder

Storage Media

Network Folders

litvinse@uqbar:~

File Edit View Terminal Tabs Help

```

drwxr-xr-x  48 litvinse cdf      20480 Sep 17 16:29 Desktop
-rw-----   1 litvinse cdf       9383 Sep 17 17:38 .ICEauthority
lrwxrwxrwx   1 litvinse cdf         44 Sep 17 17:38 .DCOPserver_uqbar.fnal.gov_
:0 -> /home/litvinse/.DCOPserver_uqbar.fnal.gov__0
-rw-rw-r--   1 litvinse cdf         63 Sep 17 17:38 .DCOPserver_uqbar.fnal.gov_
_0
-rw-----   1 litvinse cdf         31 Sep 17 17:41 .mccoprc
-rw-----   1 litvinse cdf         480 Sep 18 11:19 .Xauthority
-rw-----   1 litvinse cdf      23088 Sep 18 11:22 .bash_history
drwx-----   6 litvinse cdf       4096 Sep 18 13:37 .purple
-rw-----   1 litvinse cdf         844 Sep 18 13:57 .lessht
-rw-----   1 litvinse cdf       2741 Sep 18 14:03 kcal00918.crt
-rw-----   1 litvinse cdf       1024 Sep 18 14:03 .rnd
-rw-rw-r--   1 litvinse cdf       2341 Sep 18 14:03 kcal00918.p12
drwxr-xr-x  21 litvinse cdf       4096 Sep 18 14:09 .gimp-2.2
drwx-----   2 litvinse cdf       4096 Sep 18 14:15 .gconfd
-rw-rw-r--   1 litvinse cdf      61821 Sep 18 14:17 webdav.png
-rw-rw-r--   1 litvinse cdf      15396 Sep 18 14:19 webdav1.png
-rw-rw-r--   1 litvinse cdf      31884 Sep 18 14:20 webdav2.png
-rw-----   1 litvinse cdf      36479 Sep 18 14:20 .recently-used
-rw-r--r--   1 litvinse cdf      12763 Sep 18 14:20 .xsession-errors
-rw-rw-r--   1 litvinse cdf    1024000 Sep 18 14:20 file6.data
drwx--x--x 166 litvinse cdf      49152 Sep 18 14:20 .
[litvinse@uqbar ~]$

```

www.dCache.org

WebAdmin

- Next generation admin web portal
- Run as `JettyCell` - a cell with embedded Jetty Server *jetty://*
- Web content created using Apache Wicket web application framework



WebAdmin

WebAdmin

The screenshot shows a web browser window with the following elements:

- Browser Menu:** File, Edit, View, History, Bookmarks, Tools, Help.
- Navigation Buttons:** Back, Forward, Reload, Stop, Home.
- Address Bar:** <https://dmsdca03:8442/webadmin/login/wicket:interface/:5:LoginForm::IFormSubmitListener::>
- Search:** Google search bar.
- Bookmarks:** Most Visited, Aol, Customize..., DmitryLitvintsev < ... CDF, Labs, XXX, Swim.
- Header:** dCache logo featuring a falcon.
- Navigation Menu:** Home (selected), Cell Services, Pool Usage, Pool Queues, Poolgroups, Pool Admin, Cell Admin, Info Xml.
- Content Area:** dmsdca03
- Footer:** Login :  Logout : 

WebAdmin

The screenshot shows a web browser window with the following elements:

- Browser Menu:** File, Edit, View, History, Bookmarks, Tools, Help.
- Navigation:** Back, Forward, Reload, Stop, Home buttons.
- Address Bar:** <https://dmsdca03:8442/webadmin/login>
- Search:** Google search bar.
- Bookmarks:** Most Visited, AOL Customiz..., DmitryLitvintsev <..., CDF, Labs, XXX, Swim.
- Header:** dCache logo featuring a falcon.
- Navigation Menu:** Home, Cell Services, Pool Usage, Pool Queues, Poolgroups, Pool Admin, Cell Admin, Info Xml.
- Section:** **dmsdca03**
- Form:** "Use Guest/guest as Guest-login" with fields for Username (testers), Password, and a checked "Remember Me" checkbox. Buttons for Log In, Reset, and Certificate Log In are present.
- Image:** A faint falcon logo in the background of the login area.

WebAdmin

Cell Admin - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Back Forward Reload Stop Home Google

Most Visited Customize... DmitryLitvintsev < ... CDF Labs XXX Swim

dCache

Home Cell Services Pool Usage Pool Queues Poolgroups Pool Admin **Cell Admin** Info Xml

Cell Admin

spacemanagerDomain SrmSpaceManager

Submit Last Command:ls

Response of SrmSpaceManager@spacemanagerDomain

Reservations:
1668501 voGroup:ha voRole:a retentionPolicy:REPLICA accessLatency:NEARLINE linkGroupId:53989 size:10 created:Tue Oct 14 12:25:28 CDT 2008 lifetime:8640000000ms expiration:Mon Jul 11 12:25:28 CDT 2011 description:null state:RESERVED used:0 allocated:0
1669946 voGroup:testers voRole: retentionPolicy:REPLICA accessLatency:ONLINE linkGroupId:53989 size:10 created:Wed Oct 15 20:16:19 CDT 2008 lifetime:8640000000ms expiration:Tue Jul 12 20:16:19 CDT 2011 description:null state:RESERVED used:0 allocated:0
2678521 voGroup:/alice voRole:production retentionPolicy:CUSTODIAL accessLatency:NEARLINE linkGroupId:53988 size:1048576 created:Tue Mar 10 15:42:42 CDT 2009 lifetime:-1ms expiration:NEVER description:DTEAM_RAW state:RESERVED used:0 allocated:0
4828508 voGroup:testers voRole:testers retentionPolicy:CUSTODIAL accessLatency:NEARLINE linkGroupId:4648504 size:286331148723 created:Mon Jun 14 15:28:08 CDT 2010 lifetime:-1ms expiration:NEVER description:null state:RESERVED used:0 allocated:0
total number of reservations: 4

Find: Previous Next Highlight all Match case

Done

WebAdmin

Pool Usage - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Back Forward Reload Stop Home Google

Most Visited Customize... DmitryLitvintsev < ... CDF Labs XXX Swim

dCache

Home Cell Services **Pool Usage** Pool Queues Poolgroups Pool Admin Cell Admin Info Xml

Disk Space Usage

Enabled

Selected	Name	Domain Name	Enabled	total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/ used/ free)
<input checked="" type="checkbox"/>	hal9000_1	dmsdca01Domain	true	819200	819197	0	
<input type="checkbox"/>	hal9000_2	dmsdca01Domain	true	819200	819197	1	
<input type="checkbox"/>	hal9000_3	dmsdca01Domain	true	819200	818222	0	
<input type="checkbox"/>	hal9000_4	dmsdca01Domain	true	819200	818222	0	

x Find: Previous Next Highlight all Match case

Done

WebAdmin

Pool Property Tables - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Back Forward Reload Stop Home Google

Most Visited Customize... DmitryLitvintsev < ... CDF Labs XXX Swim

dCache

Home Cell Services Pool Usage Pool Queues **Poolgroups** Pool Admin Cell Admin Info Xml

Pool Groups of PoolManager

PoolGroup	Enabled	total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/ used/ free)
read_pg	true	1638400	1636444	0	
write_pg	true	819200	819197	0	
write_pg_1	true	819200	819197	1	

Pool Group:

Cell View **Space Usage** Request Queues

Name	Domain Name	Enabled	total Space/MiB	Free Space/MiB	Precious Space/MiB	Layout (precious/ used/ free)
------	-------------	---------	-----------------	----------------	--------------------	---------------------------------

Find: Previous Next Highlight all Match case

Done

dCache/SRM

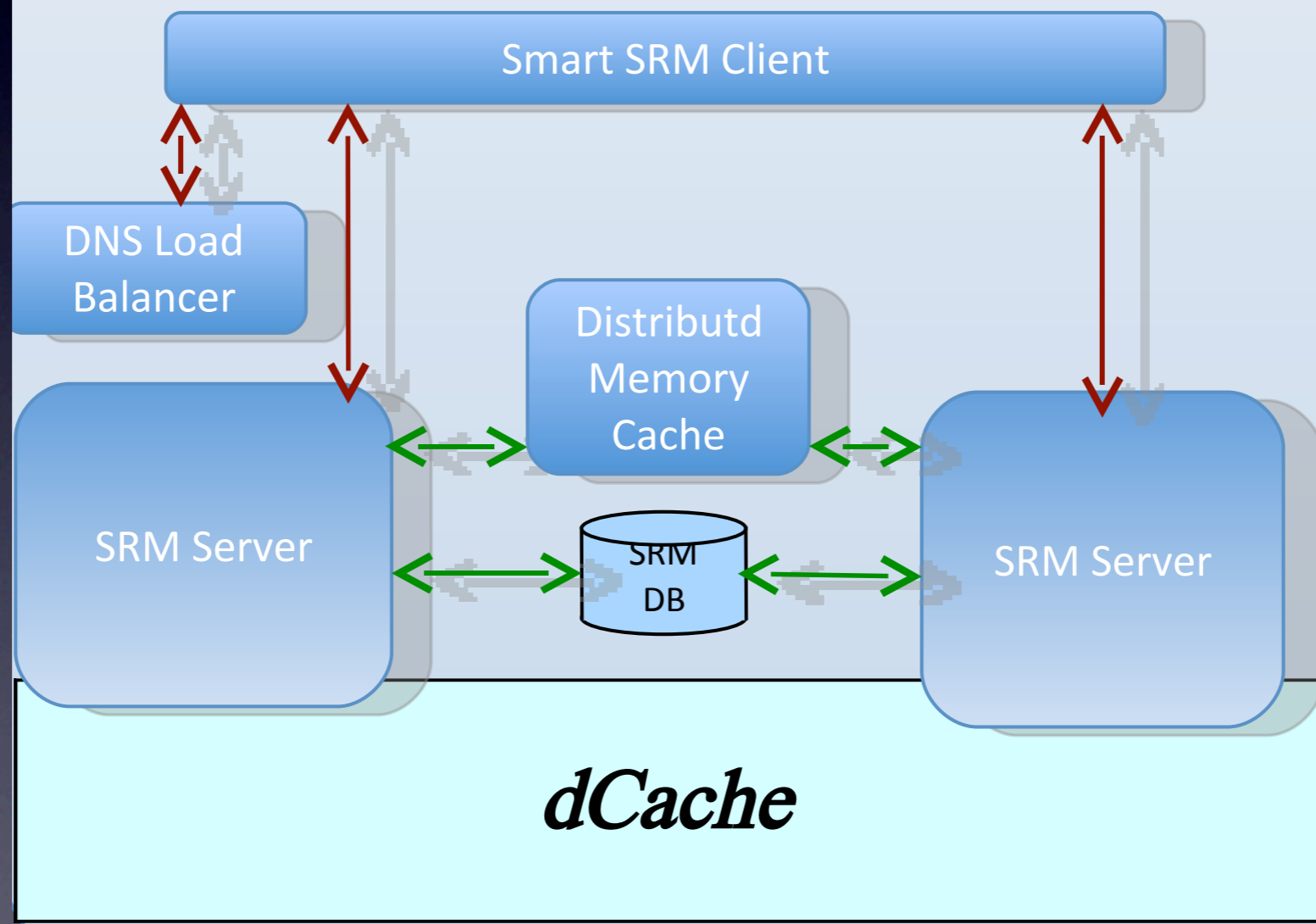
- SRM is a web service interface to an SE that provides a set of space management, permission, directory and transport functions that allow transparent interaction with heterogeneous storage on the Grid.
- SRM/dCache is one of the implementations of SRM v2.2, part of dCache distribution.
- SRM Client is also available from dcache.org

SRM/dCache issues

- High CPU load due to GSI Authentication and Credential delegation
 - mitigated by (public,private) key caching
 - considering https as a long term solution
- Blocking SRM functions (especially srmls) executed in large volumes saturate connections limits
 - implemented asynchronous srmls. Unfortunately cannot do the same for rm, mv, mkdir etc. (protocol limitation)
- SRM is a single point of entry to WLCG SE
 - a possible bottleneck
 - a single point of failure
 - distributed SRM may be a solution

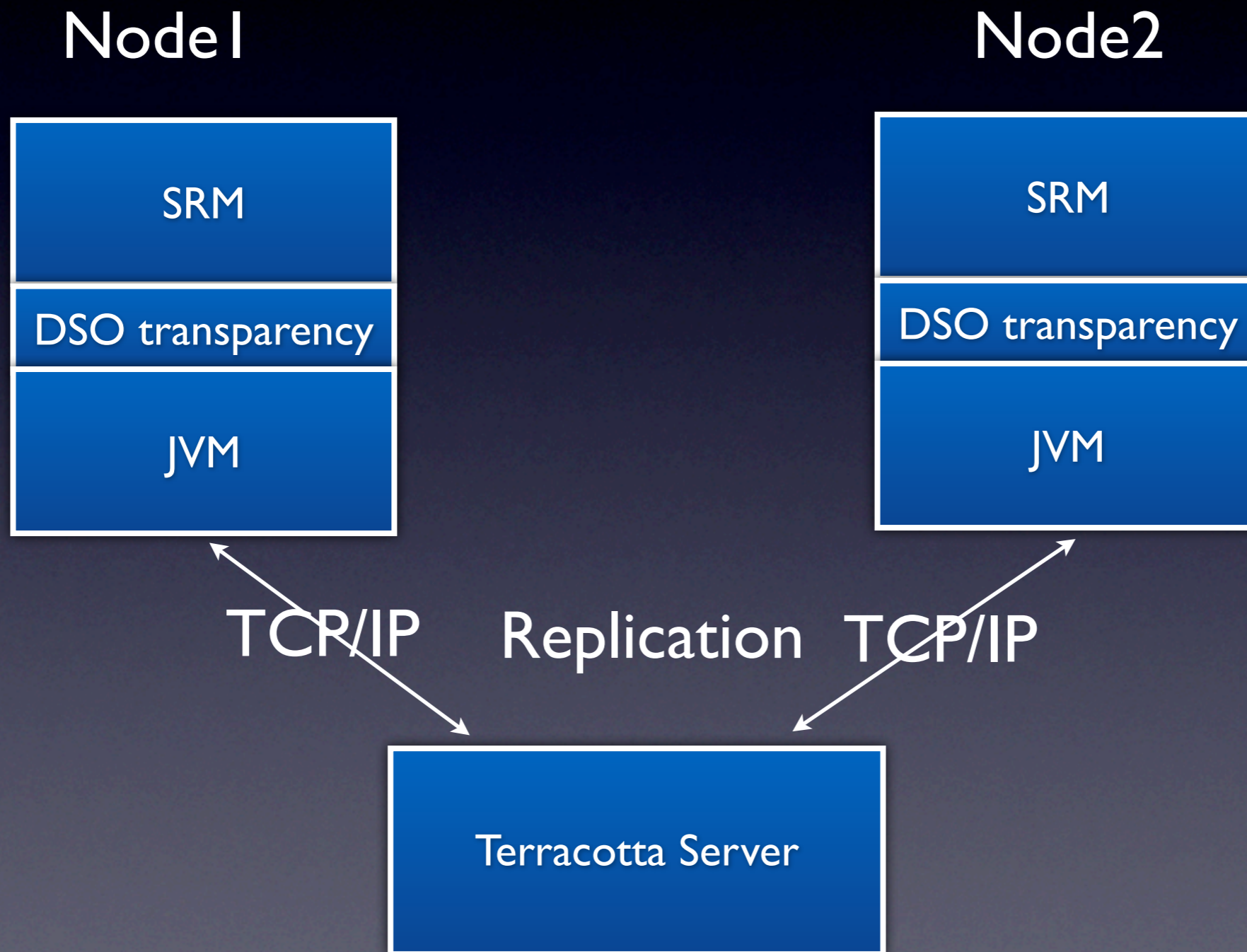
SRM/dCache

Scalable dCache SRM Server

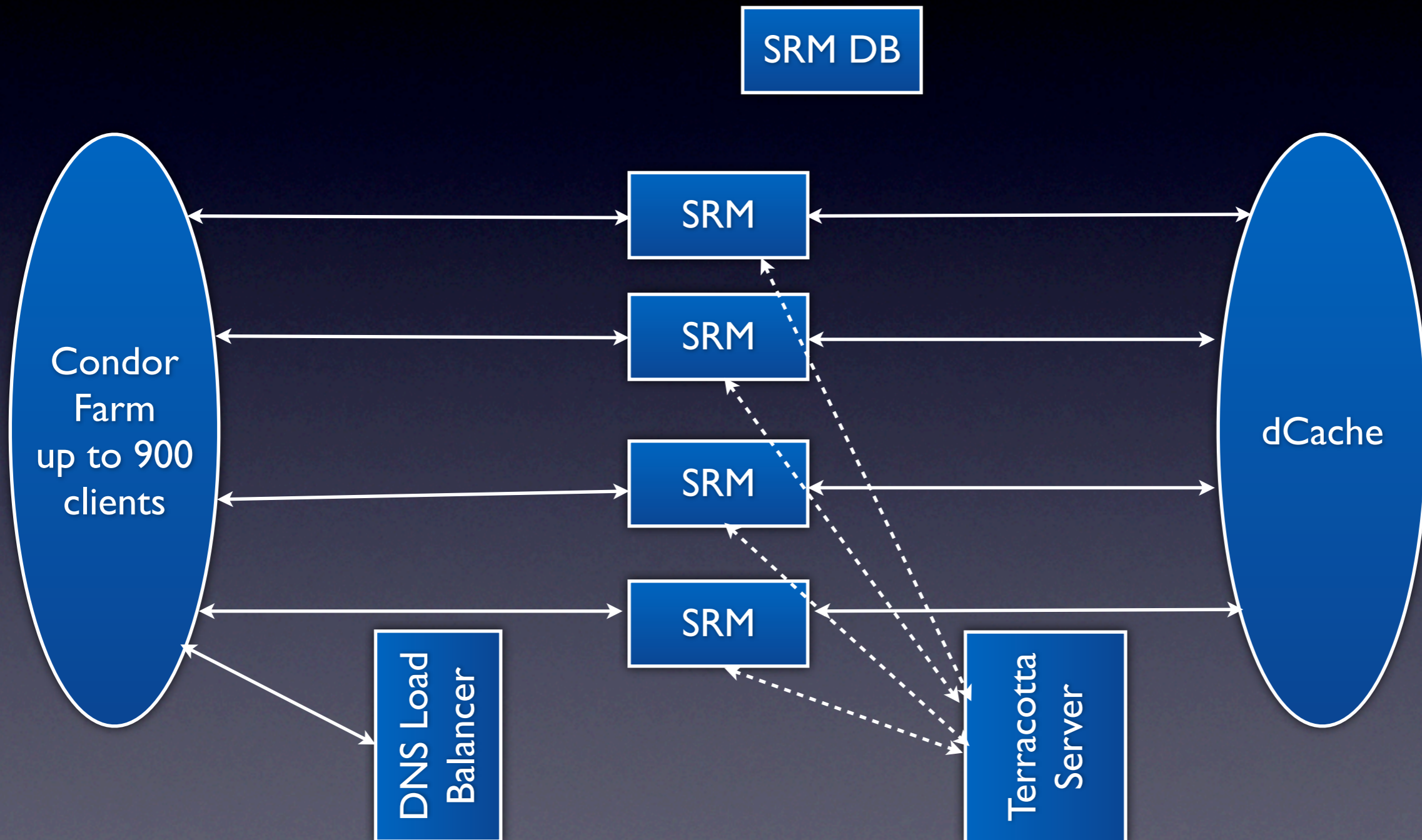


Terracotta DSO

Open Source JVM level clustering

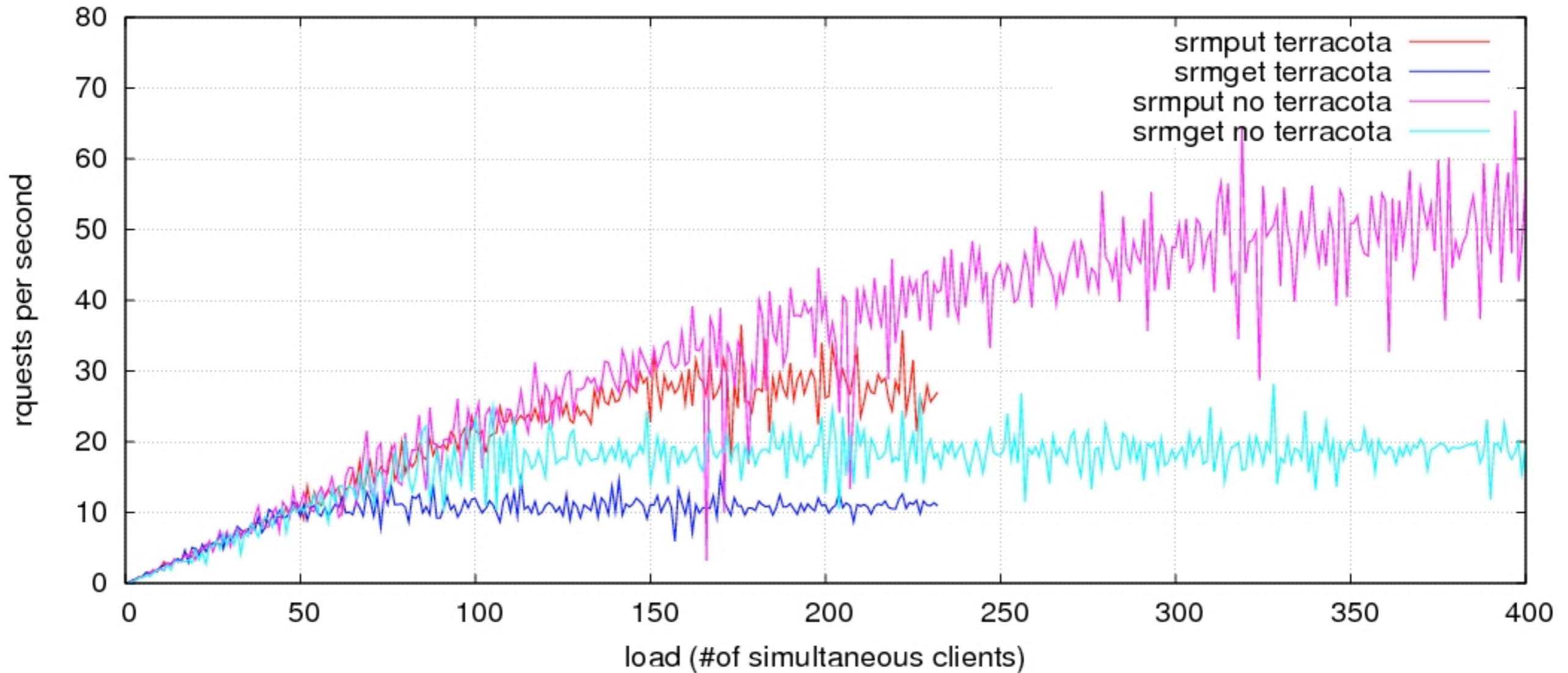


Distributed SRM test



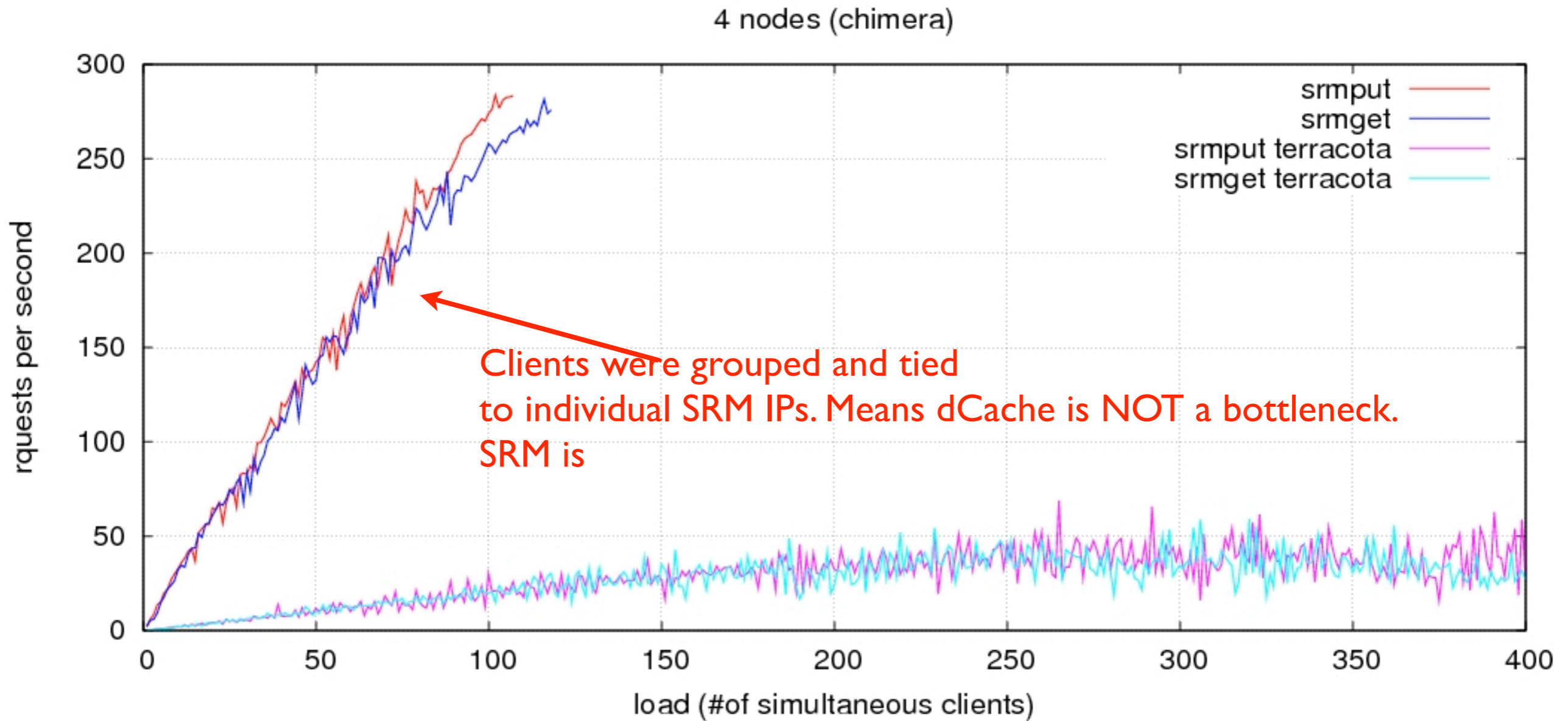
dCache/SRM test

Single node performance (PNFS)



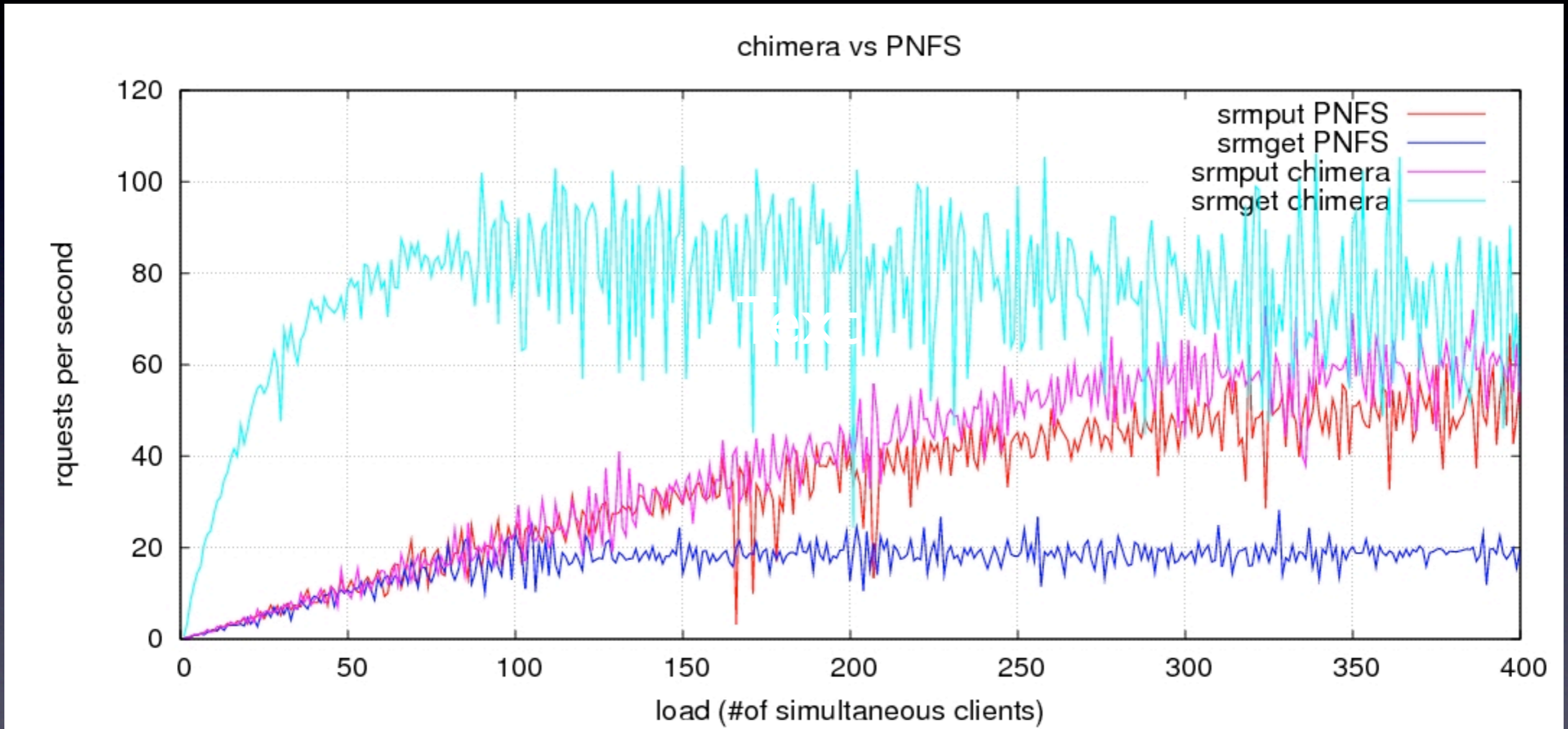
NB: this test was done on PNFS

dCache/SRM test



Not very encouraging :(

Bonus Slide



chimera vs PNFS

Preliminary Conclusions

- Terracotta out of the box and no changes on SRM side is not of any help. Issues may be in SRM implementation.
- Consider other extreme - thin SRM layer on top of database that keeps request state. Multiple SRM portals. Rely on underlying DB implementation scalability.
- We can achieve scaling by running multiple SRMs in parallel (e.g. SRM per VO). dCache is not a bottleneck.

Summary

- dCache is experiencing rapid evolution from within towards modern technologies
- Code development and release management processes have become more robust resulting in improved code quality and maintainability
- dCache provides managed storage at Petabyte scale
- dCache SE has met real data challenge from LHC
- dCache is embracing standard protocols to attract larger scientific community.
- We are actively working on areas that require improvement.