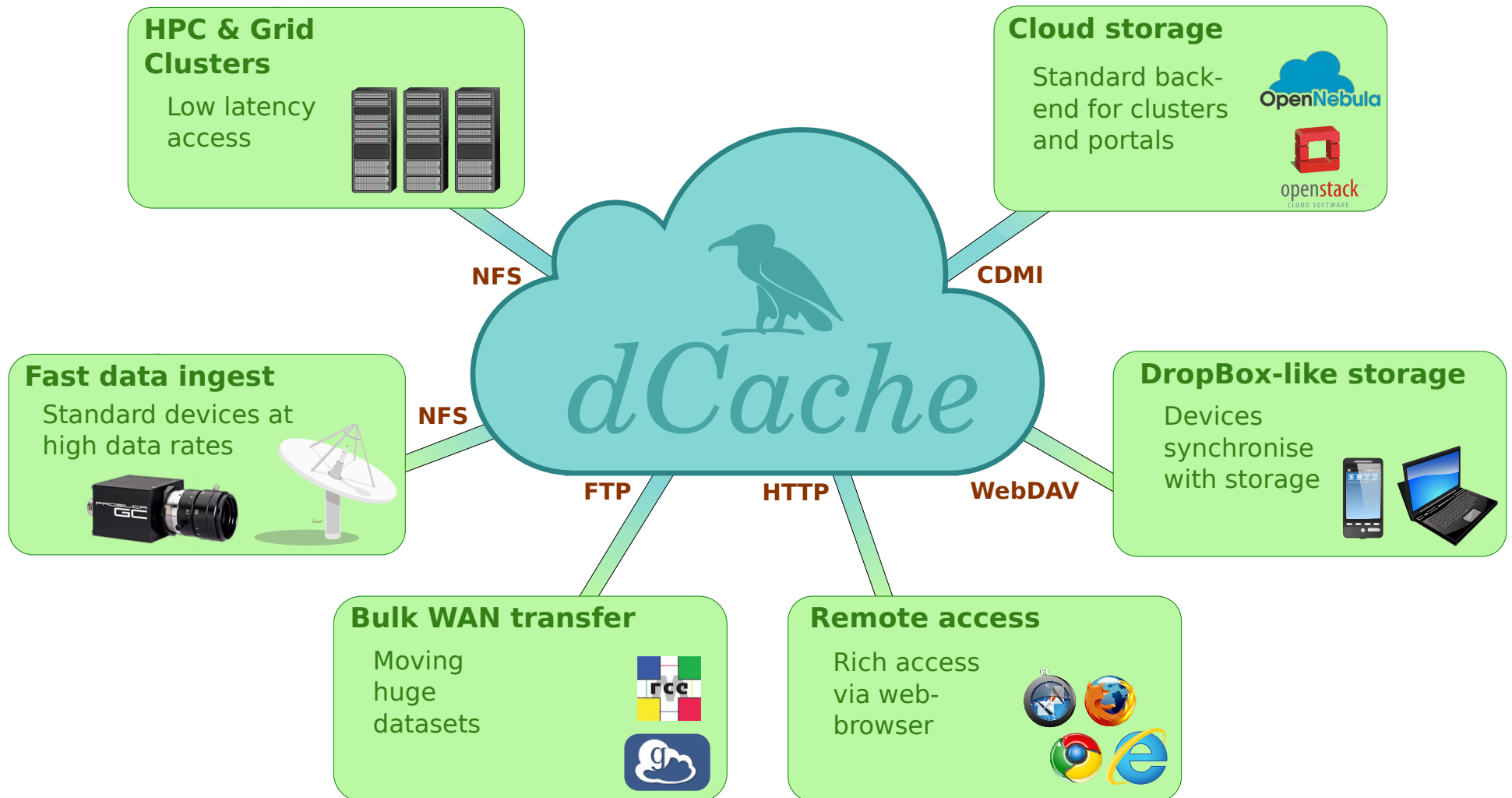


pNFS - two years in production

Tigran Mkrtchyan for dCache Team



Problem domains



the mission

“... to provide a system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods.”

16 Sep. 2000

Michael Ernst, Patrick Fuhrmann, Martin Gasthuber, Rainer Mankel

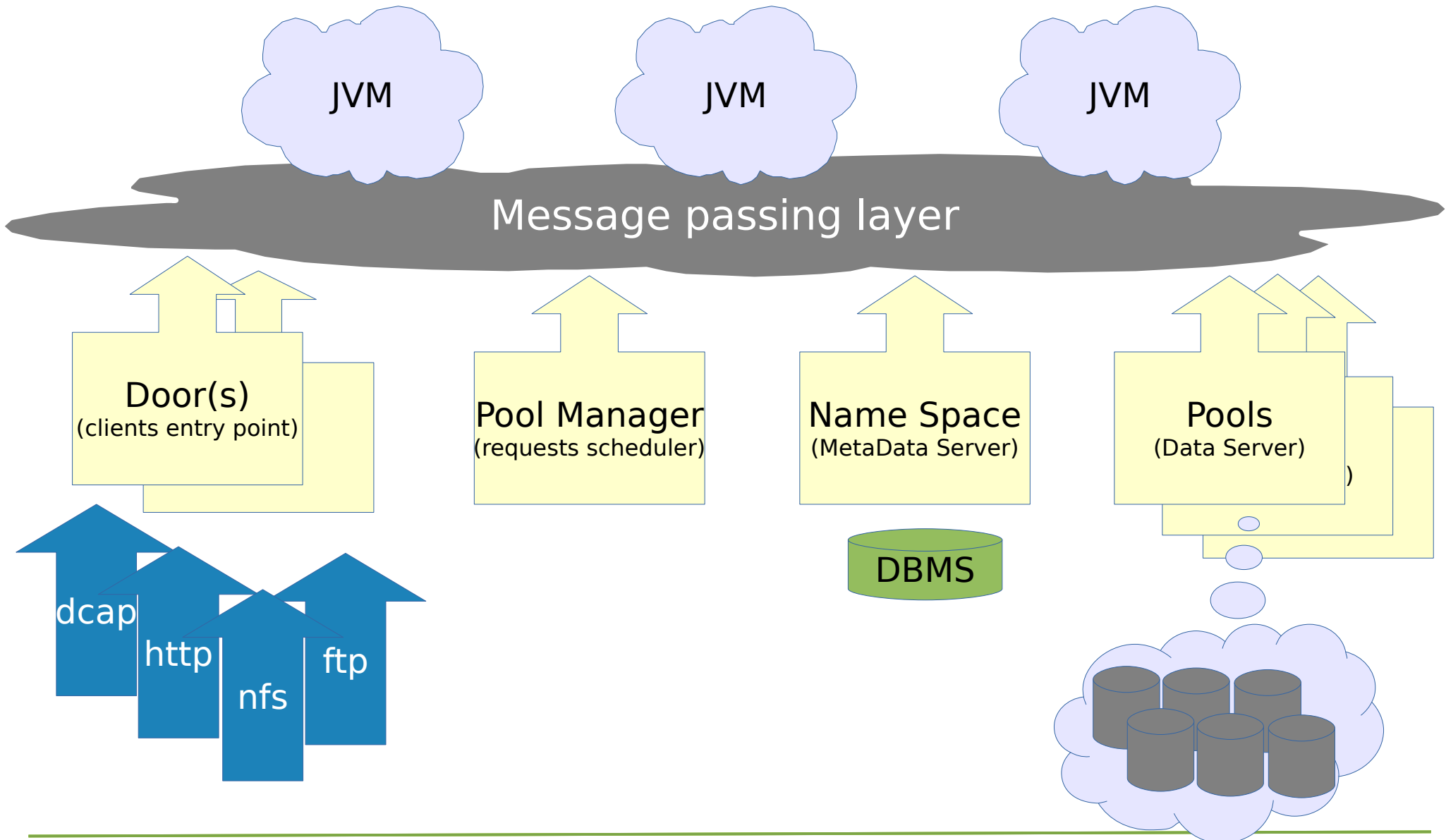
pnfs problem statement

“...Separating storage data flow from file system control flow effectively moves the bottleneck away from the single endpoint of an NFS server and distributes it across the bisectional bandwidth of the storage network between the cluster nodes and storage devices.”

pnfs problem statement

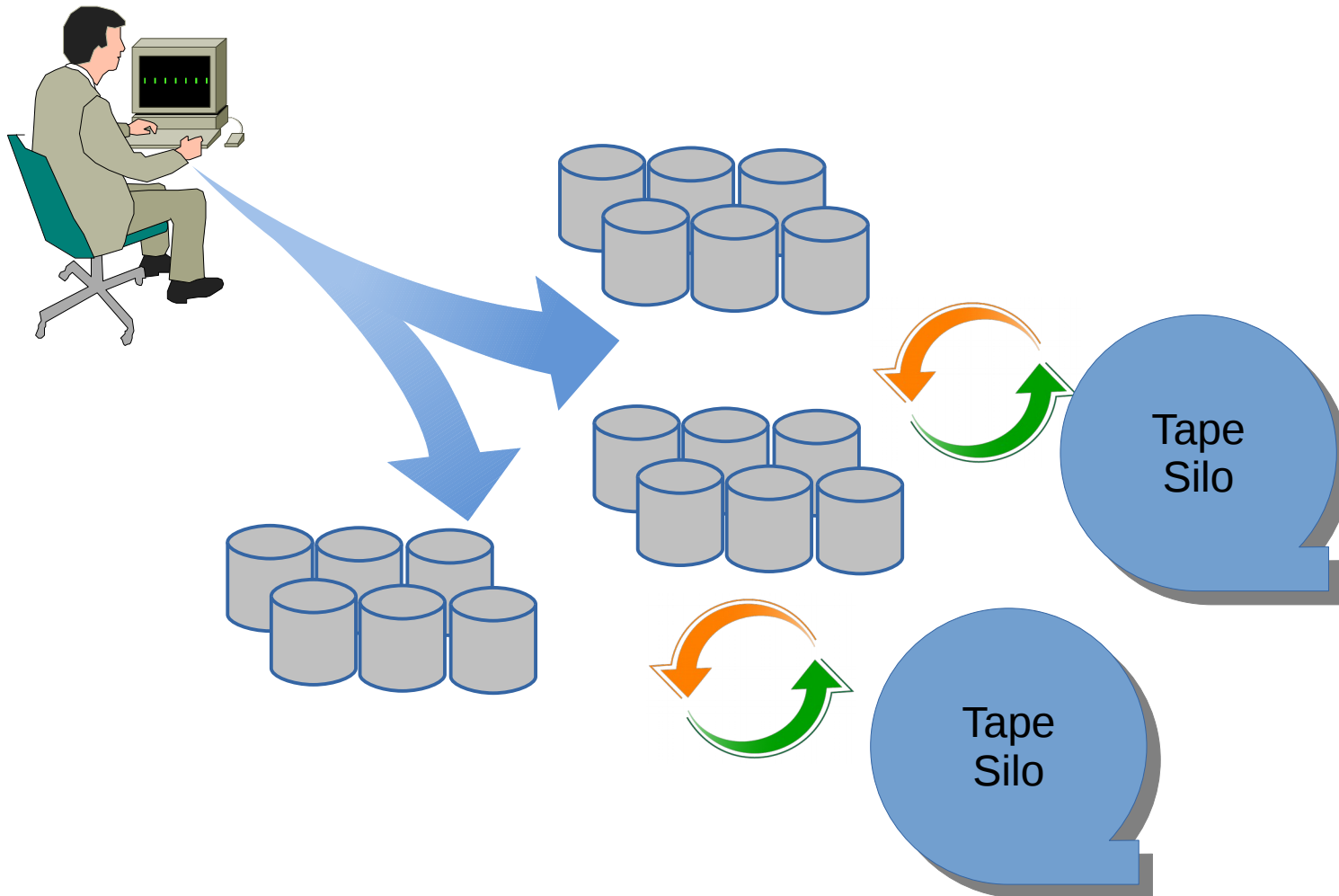
Garth Gibson et al., July 2004

dCache in one slide



Tape migration

Disk <-> Tape migration behaves similar to write-back cache:

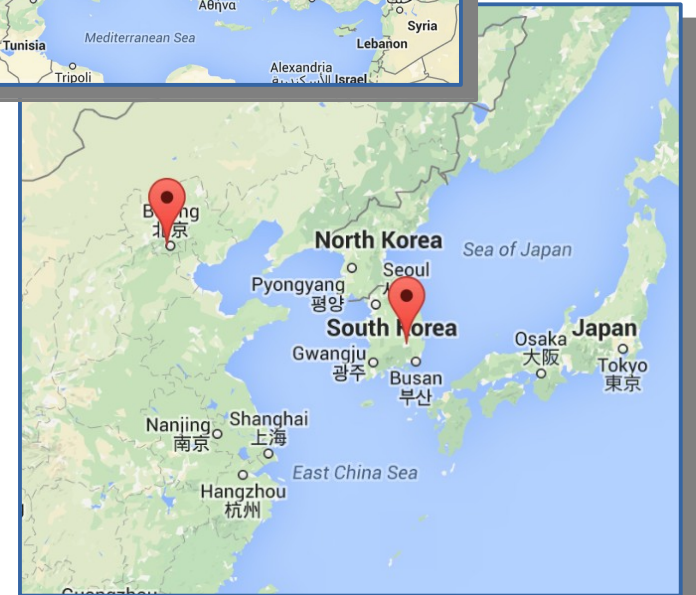
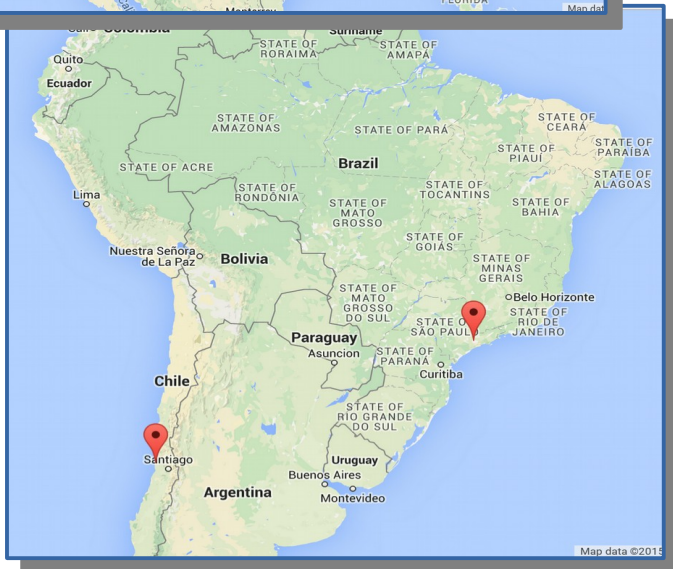
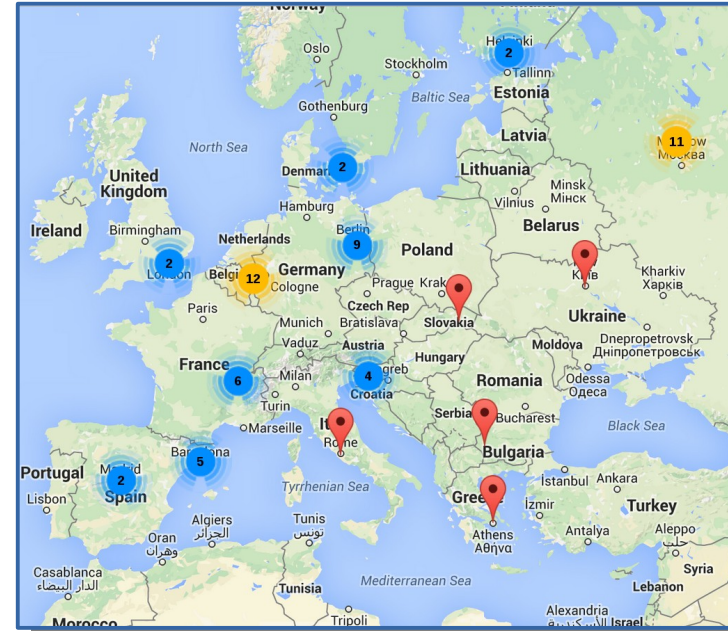
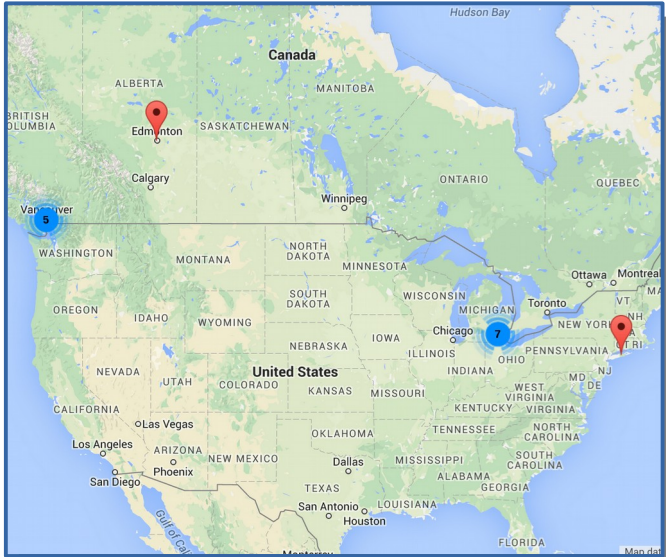




Some Numbers

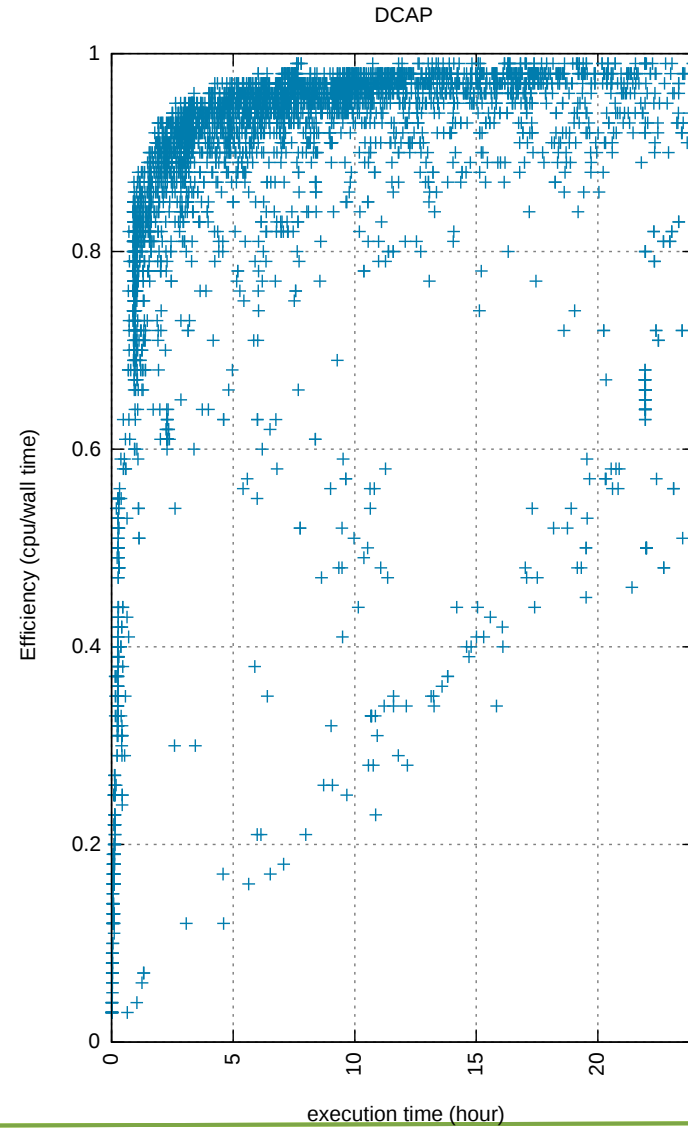
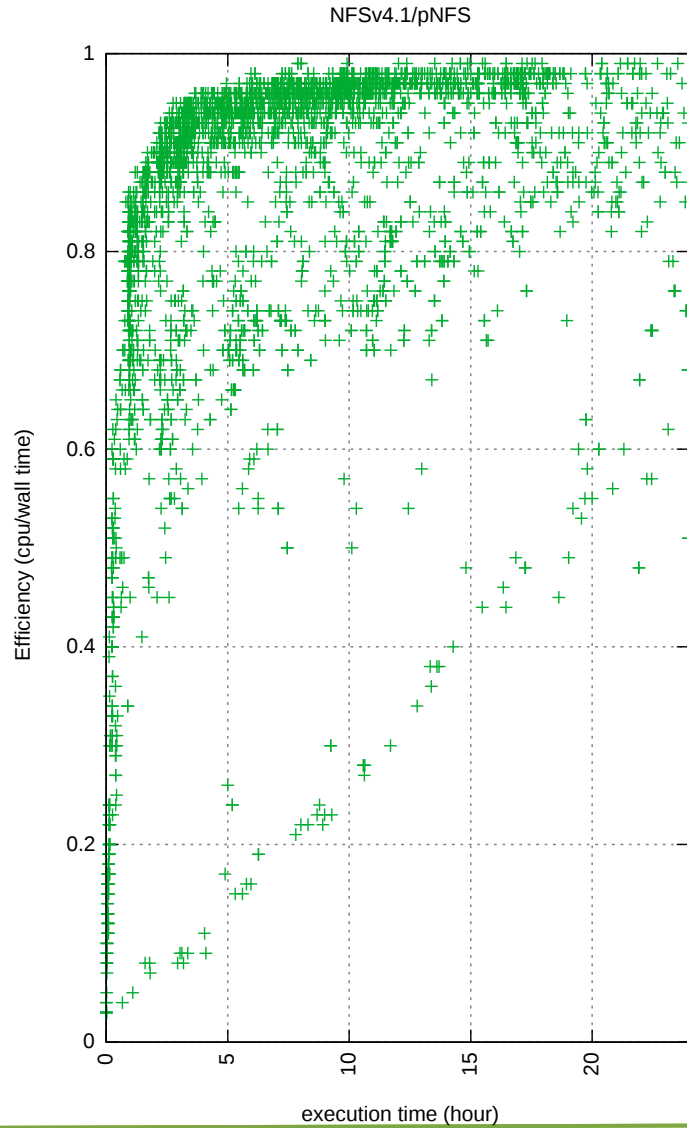
- ~10PB in total
- ~1200 DS (~300 hosts)
- ~10K Cores (~800 hosts)
 - SL6 (RHEL6)
 - handful RHEL7
- ~ 400 TB per day (~90% read)

~50% of LHC data around the world

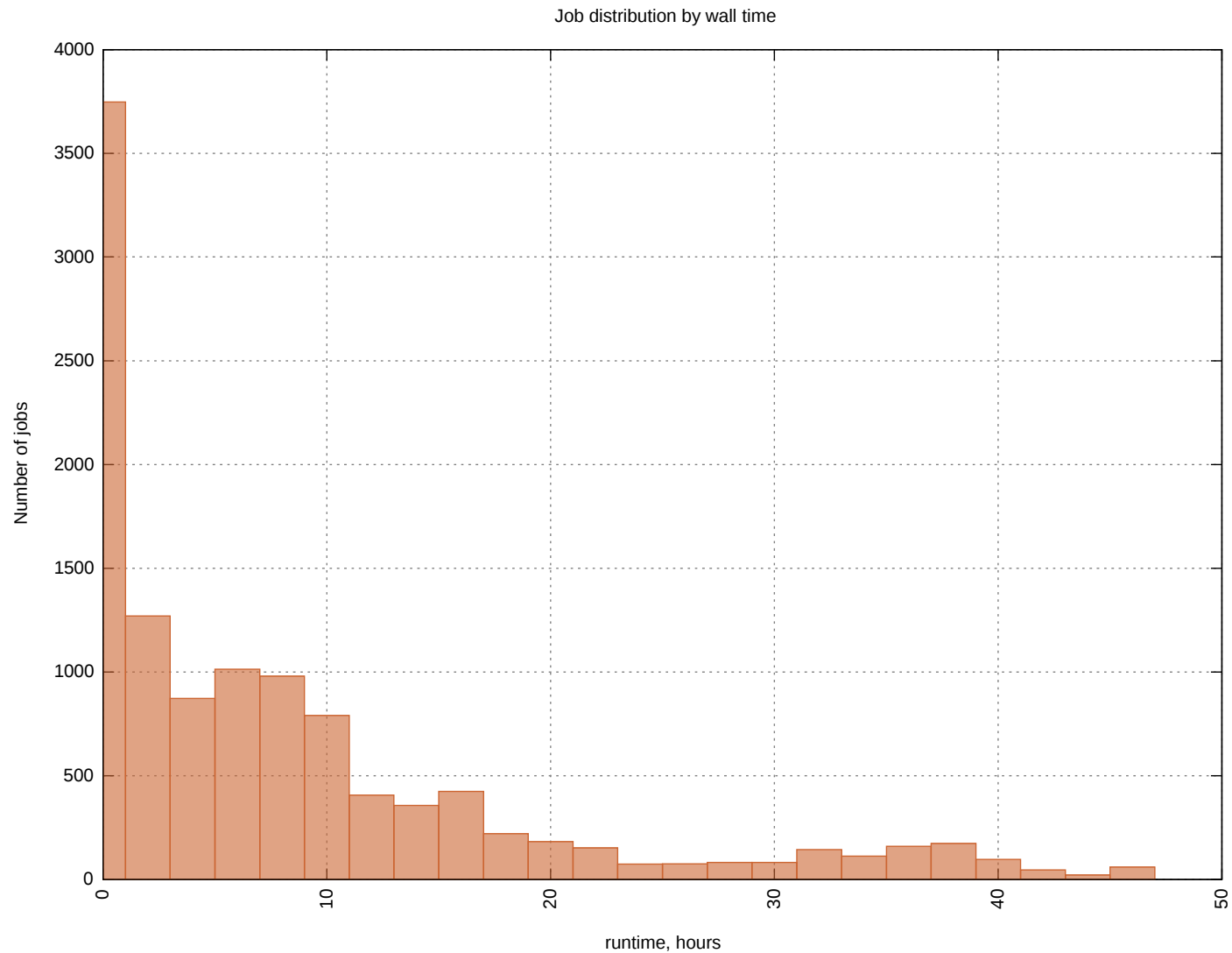


Job Efficiency

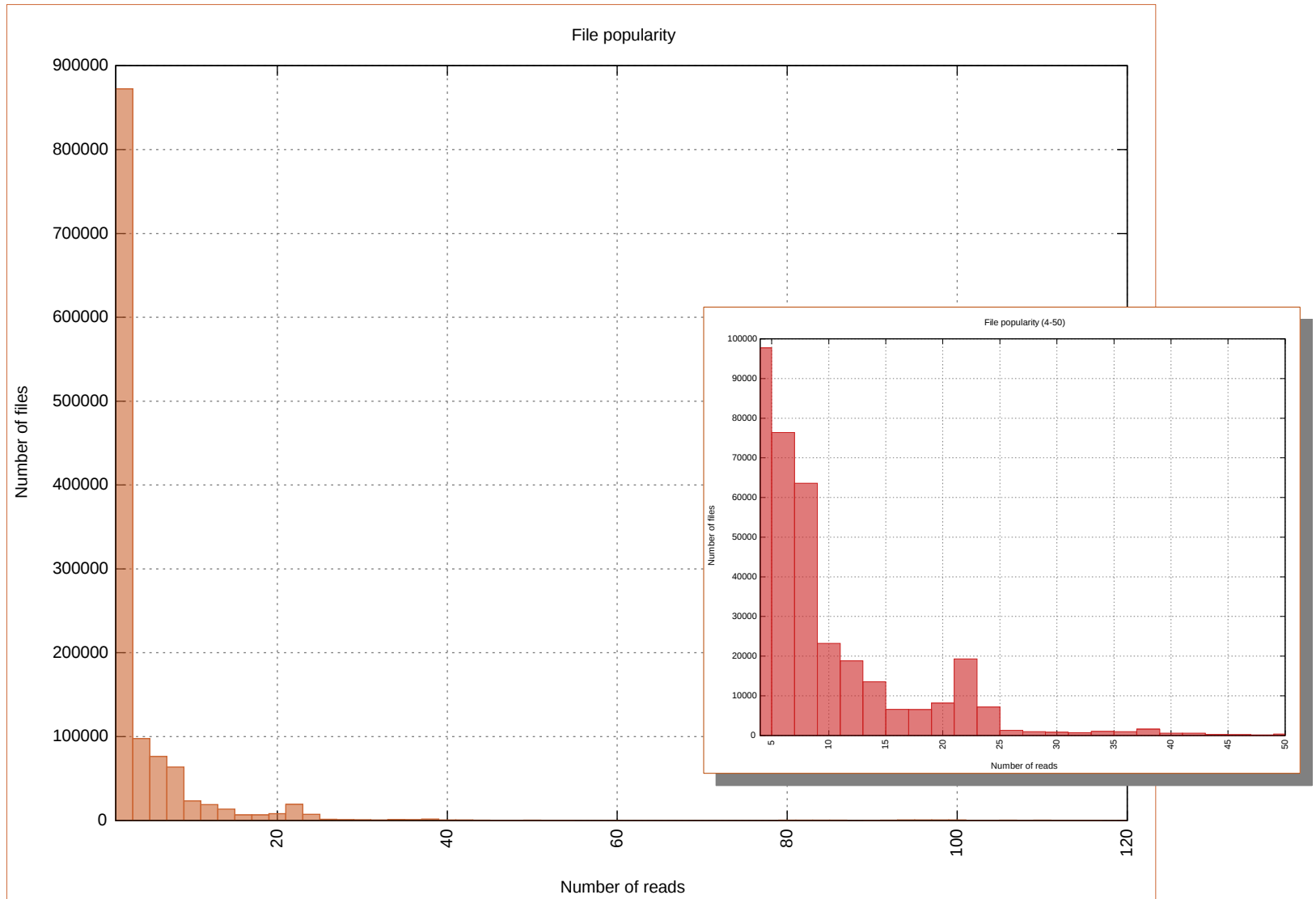
CMS job efficiency by access protocol



Job distribution by runtime



File Popularity



www.nobelprize.org/nobel_prizes/physics/laureates/2013/

Physics Prizes < 2013 >

▼ About the Nobel Prize in Physics 2013

- Summary
- Prize Announcement
- Press Release
- Advanced Information
- Popular Information
- Greetings
- Award Ceremony Video
- Award Ceremony Speech
- ▶ François Englert
- ▶ Peter Higgs

All Nobel Prizes in Physics
All Nobel Prizes in 2013

Share this:

The Nobel Prize in Physics 2013





Photo: A. Mahmoud
François Englert
Prize share: 1/2

Photo: A. Mahmoud
Peter W. Higgs
Prize share: 1/2

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"*

▶ Contact | Press | Sitemap | FAQ | Terms

Copyright © Nobel Media AB 2015

Nobel Laureates
© The Nobel Foundation

Explore the Nobel Prize Talks Podcast

NOBEL PRIZE QUIZ
Test your knowledge about the Nobel Prize

Discover features and trivia about the Nobel Prize

Sign up for Nobelprize.org Monthly

Operational experience

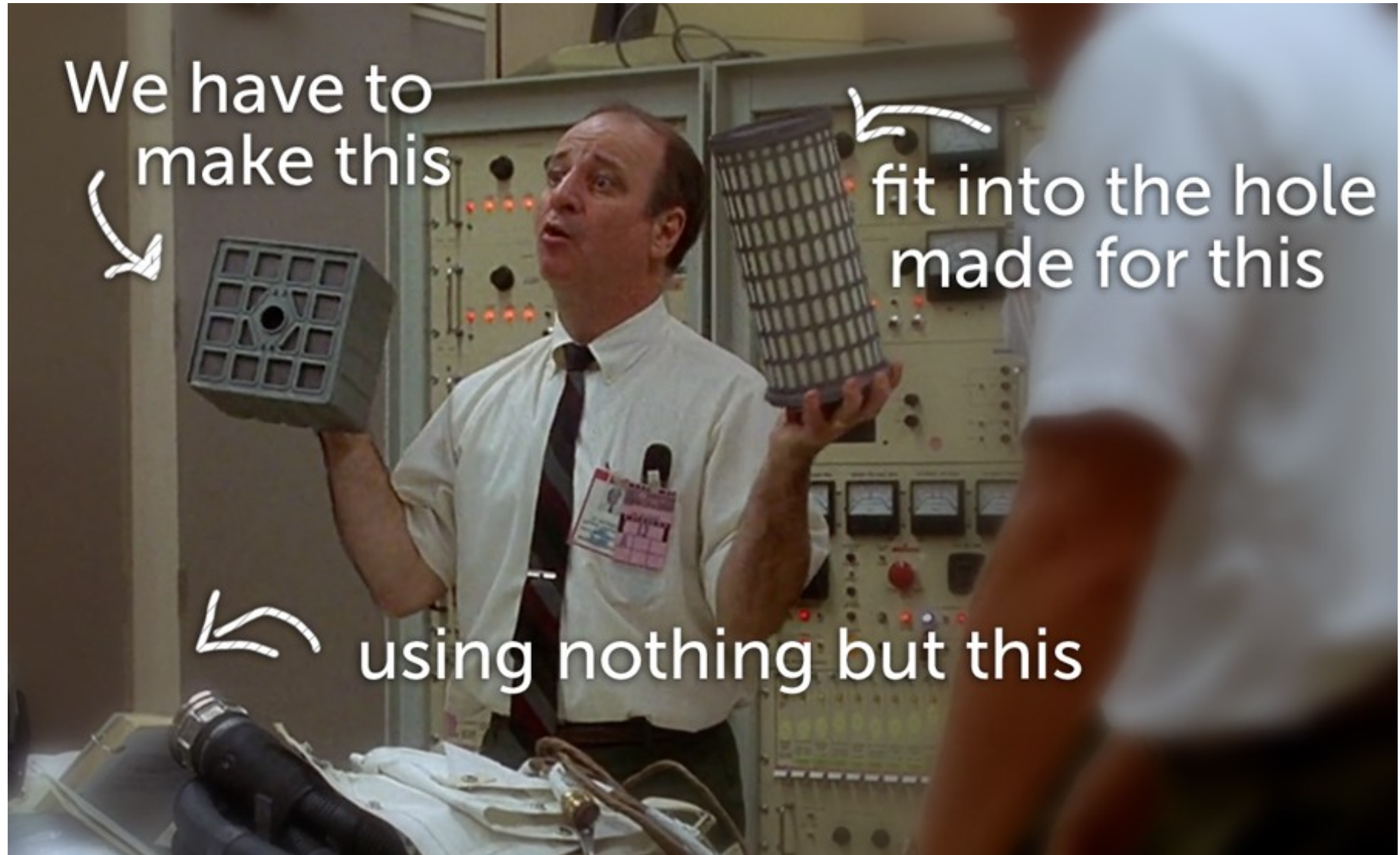
- performs well as long as it running
 - Hard to get under control when fails
 - mostly to understand '**WHY?**'
-

pNFS - two years in production
or

pNFS – 'p' for PAIN

Tigran Mkrtchyan for dCache Team





Observed problems

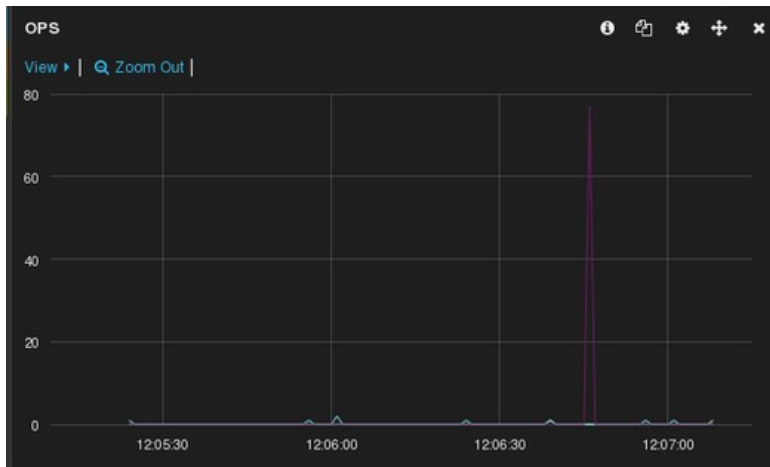
- Dual stack DS (IPv4 + IPv6)
 - Multi-homed DS (more than one IPv4 address)
 - DS behind firewall
 - Unstable DS (Host, Disk, Network errors)
 - Weak clients (VMs)
 - Distributed nature (rule #1)
 - Spec violation (infinite state recovery)
 - DS blacklisting
 - Stale kernel threads
-

Conclusion

- We see us as a clear winners of pNFS
 - enables us to expose our storage system to standard clients
 - No other commercial/opensource servers with comparable installations in production
 - Most of the problems we was able to solve
 - A new community get a posix-like access to shared storage
 - RHEL6 (SL6) our main platform for next 4 yeas
 - IPv6 slowly become a reality
-

Future plans

- Delegations
 - better experience for interactive users
 - FlexFiles
 - Mirroring (especially on write)
 - error reporting
 - Locks
-



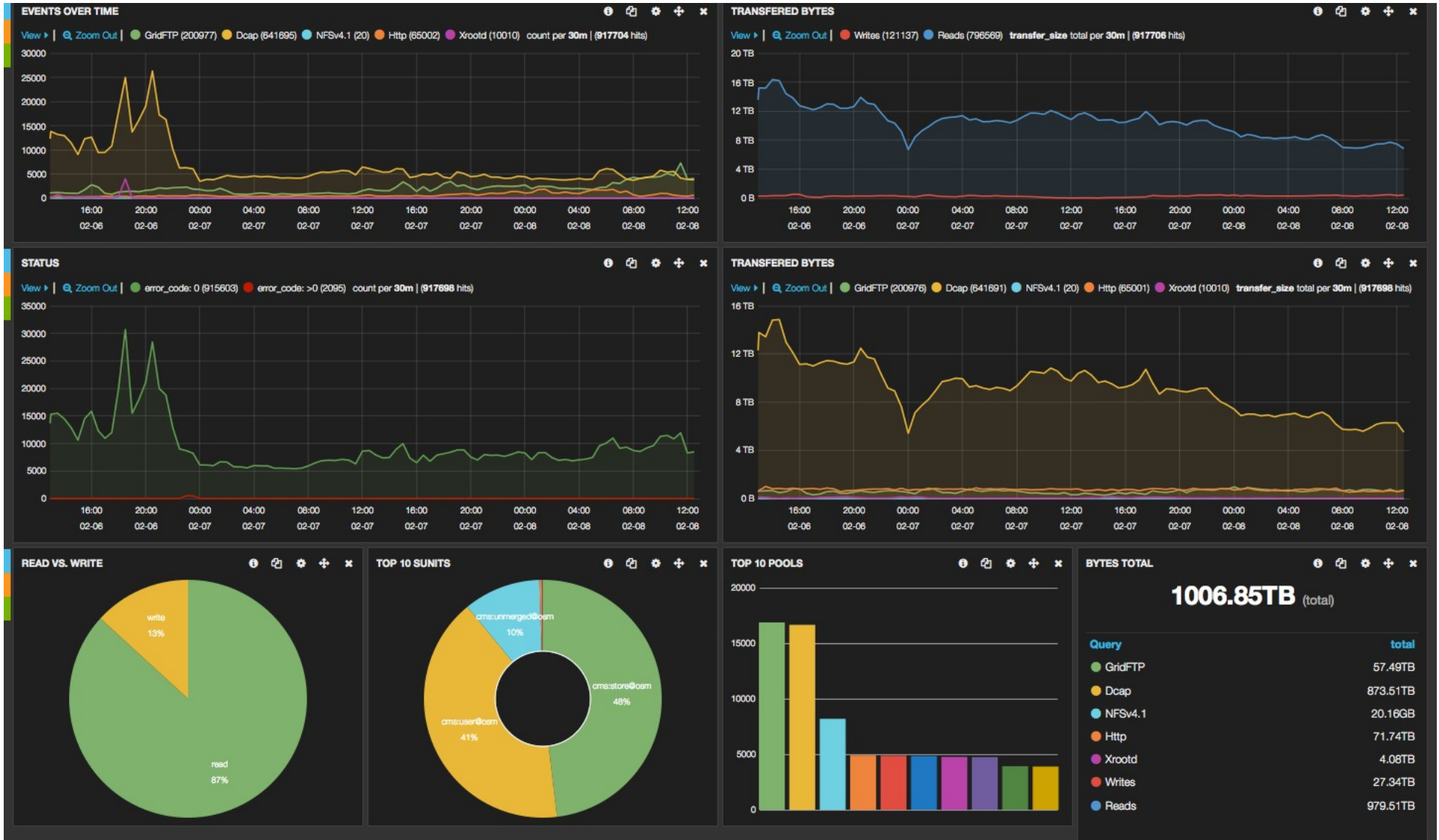
STATS

0.003 ms (mean)

Query	min	max	mean	std_deviation
*v4_LAYOUTRETURN	0.000 ms	0.001 ms	0.000 ms	0.000 ms
v4_ACCESS	0.000 ms	0.068 ms	0.001 ms	0.004 ms
v4_CLOSE	0.000 ms	0.024 ms	0.003 ms	0.008 ms
v4_CREATE_SESSION			0.000 ms	0.000 ms
v4_DESTROY_SESSION			0.000 ms	0.000 ms
v4_EXCHANGE_ID			0.000 ms	0.000 ms
v4_GETATTR	0.000 ms	0.067 ms	0.002 ms	0.003 ms
v4_LAYOUTGET	0.023 ms	0.024 ms	0.024 ms	0.001 ms
v4_LOOKUP	0.000 ms	0.002 ms	0.001 ms	0.001 ms
v4_OPEN	0.001 ms	0.062 ms	0.003 ms	0.001 ms
v4_READ			0.000 ms	0.000 ms
v4_REMOVE			0.000 ms	0.000 ms
v4_RENAME	0.003 ms	0.003 ms	0.003 ms	0.000 ms
v4_SETATTR	0.001 ms	0.002 ms	0.002 ms	0.000 ms
v4_WRITE			0.000 ms	0.000 ms

ADD A ROW

Live view



Slides with Problems

IPv6

- Dual stack DS
 - Client doesn't support IPv6
 - Client takes first entry in multipath list

 - Put IPv4 addresses before IPv6
 - RHEL6 client updated to pick first IPv4 address
-

IPv4

- Dual Home DS
 - Client takes first entry in multipath list
 - Discover interface which will be used and put it as a first in multipath list
-

pNFS!

- DSes behind firewall
 - Implemented extra export option 'nopnfs'
 - always returns NFSERR_LAYOUTUNAVAILABLE
-

DS errors

- Host crash
 - DISK errors
 - Network

 - Yes, yes.....proxy-IO
-

Clients in a VM

- Client with limited resources
 - Client can't digest requested data
 - Hmm...Session limits per client?
 - Physical host
-

Distributed Nature

- Network glitches
 - All client (ALL!!!) fall-back to IO through MDS
 - Interactive users unhappy
 - *And my phone rings*
 - dedicated low-latency MDS for interactive users
 - dedicated high-throughput MDS for cluster nodes
-

Infinite state recovery

- Client and server can't agree
 - Infinite state recovery loop
 - added special command on MDS to forget the client
 - client tries reboot recovery
-

DS blacklisting

- client blacklists DS
 - All access to that DS use proxy-io
 - added special command on MDS to generate new device id
-

Spec is complicated

- Every time I read find something new
 - Every time I re-read I understand differently
 - tanks for Connectathons and Bakeathons to get it right
-

Job Efficiency

CMS job efficiency by access protocol

