dCache: Powerful Storage Made Easy

Paul Millar
(on behalf of the
dCache.org team)

dCache

- dCache is powerful storage system, providing what people need, now.
- We're also pushing the envelope of what storage systems can do.
- Funding from the three partner institutes (DESY, Fermilab, NDGF) and from EMI, Physics at the Terascale and D-Grid

Powerful?

Powerful: lots of data

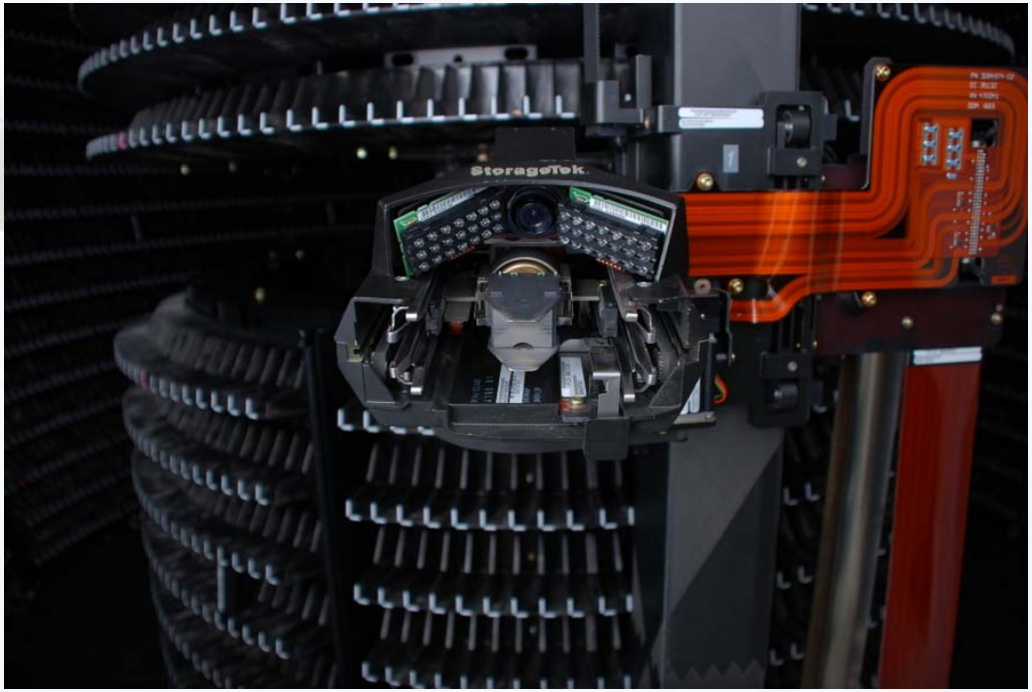
- Over 50% of storage for WLCG (>75 PB) is achieved with dCache; over 70 instances spanning the globe.
- 17 of these sites provide over a petabyte of data capacity, each.
- The largest two sites (BNL, Fermilab) have over 9 PB of disk capacity, each.
    - more being added all the time

Powerful: distributed

- A dCache instance has a flexible deployment and can be highly distributed. Exemplar are:
  - **Swegrid** a dCache instance that is spread over Sweden.
  - **NDGF** a dCache instance spread over **5 countries** (Norway, Denmark,
    Sweden, Finland,
    Slovenia), **8 sites**,
    6 of which have tape
    back-end.

Powerful: transparent access to tape

- dCache interfaces to (almost) any tape system
  - Enstore, HPSS, OSM, TSM, ···
- Users are unaware that files are stored on tape (other than some files take a while to open)
- Supports prestaging

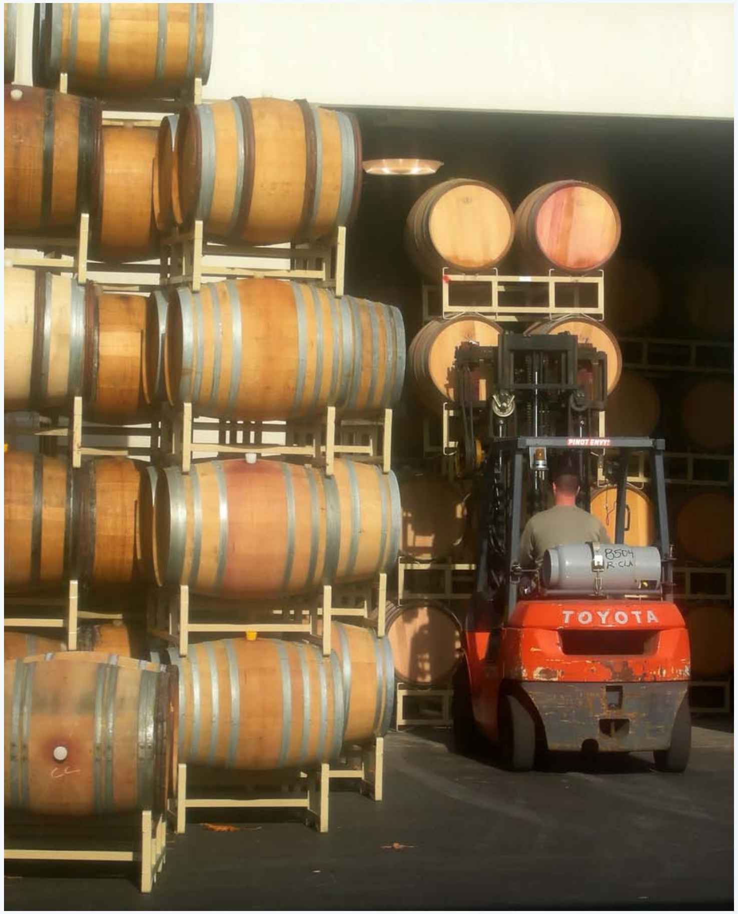


Powerful: elastic storage

- dCache can scale out to use a cloud back-end, tested with Amazon S3.
- Can host complete dCache instance within a cloud, tested with Amazon S3 and EC.
  - All access protocols available, including SRM, GridFTP and WebDAV

Powerful: managed data

- Pools can be assigned to any number of VOs
- Pools be used for writing (disk or tape), for reading, for staging back from tape .. or any mixture thereof
- Can bind behaviour to subtree of namespace
- Can enforce n-copies for redundancy.
- Supports load-balancing, overload protection and



Powerful: supporting *all* HEP needs

- dCache supports all HEP requirements: SRM, GridFTP.
- The HEP community have various proprietary/legacy protocols that they use for historical reasons
- dCache also supports the proprietary xrootd protocol with a production-hardened implementation.

New configuration

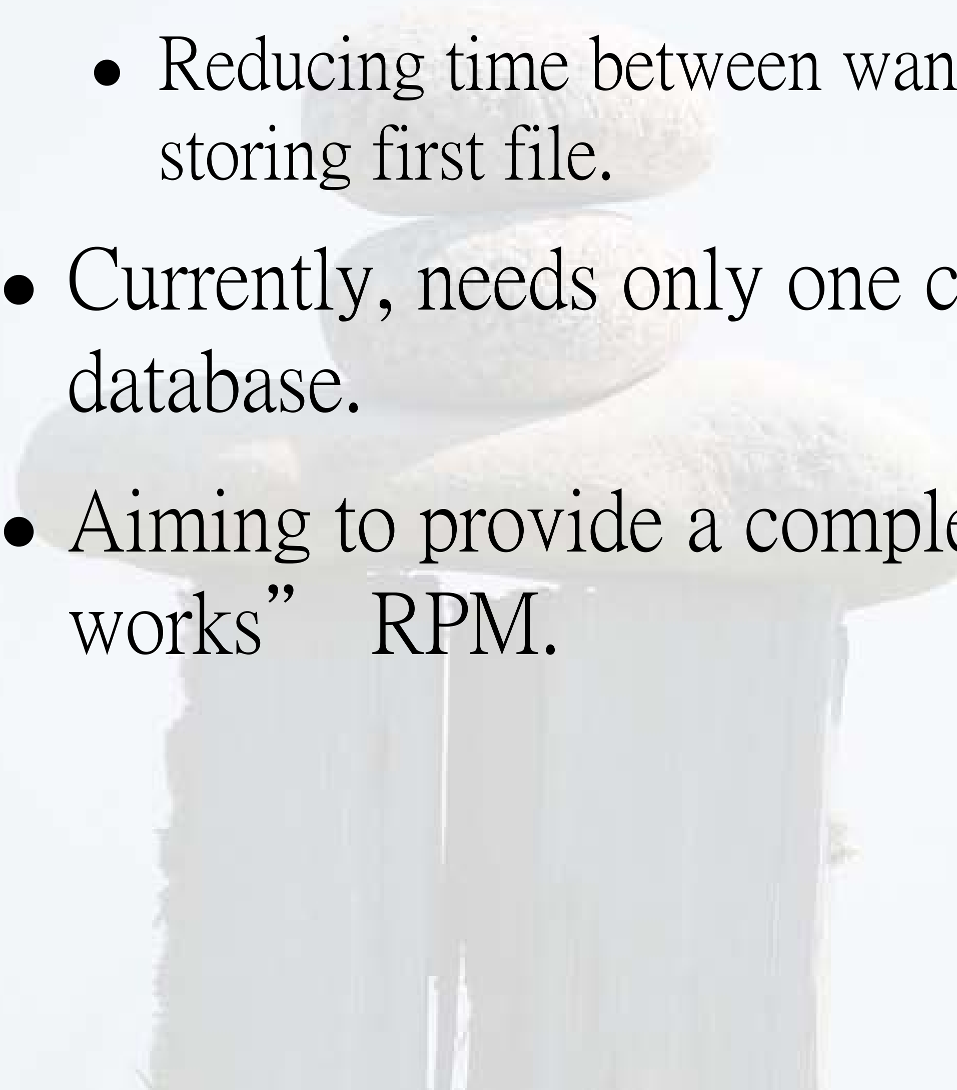
- We've invested significant effort into making dCache configure more flexible, supporting a broader range of configuration.
- … and, at the same time, we've made configuration even easier.
- Early adopters have been very enthusiastic.

Namespace

- Based on monitoring, testing and feedback from sites, we've delivered, in time, improvements to dCache's namespace performance.

- Part of this work is "Chimera"
- Sites are migrating or have migrated; majority having already switched.
- Sites adopting Chimera have found improved performance

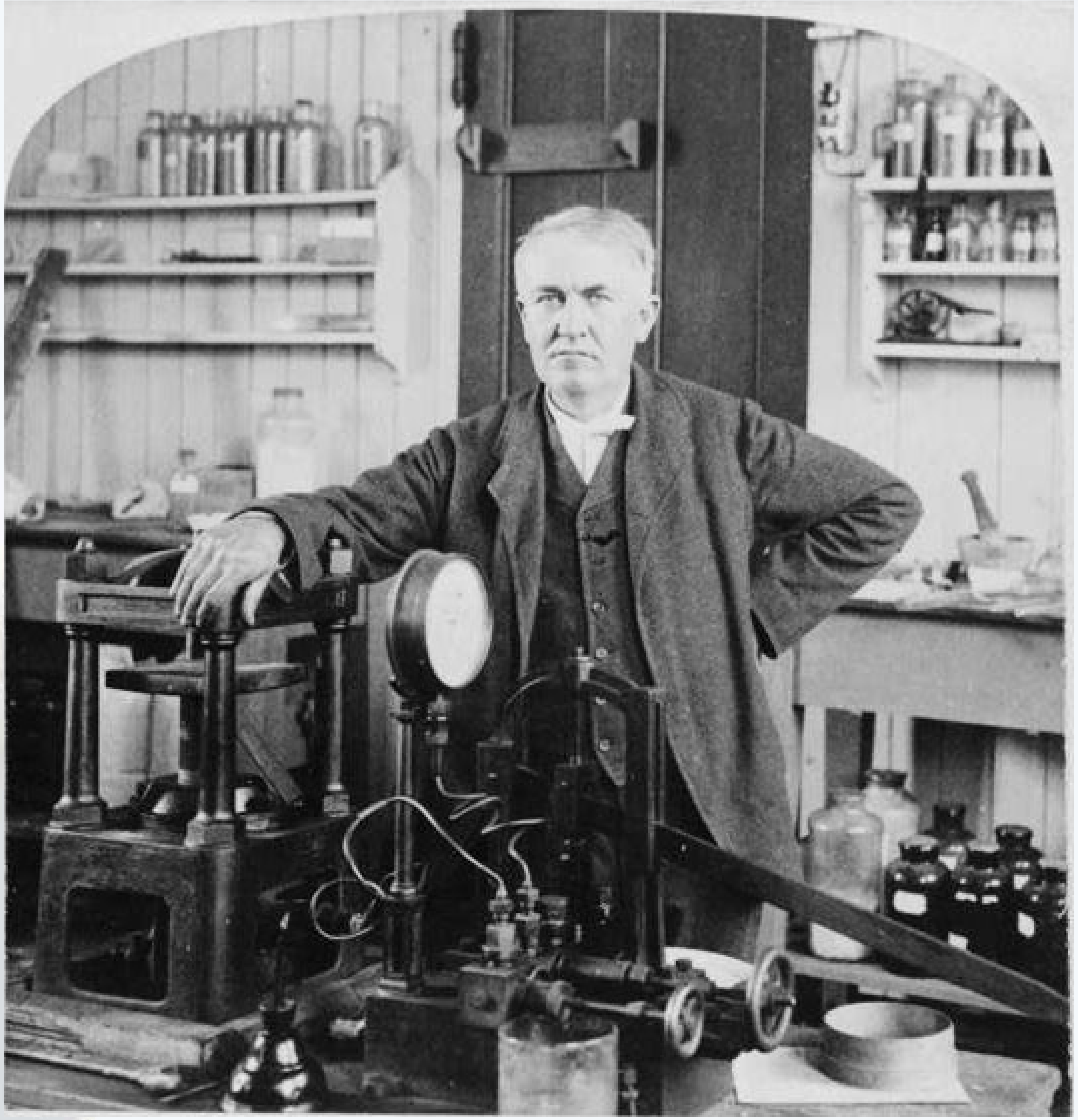
Deployment



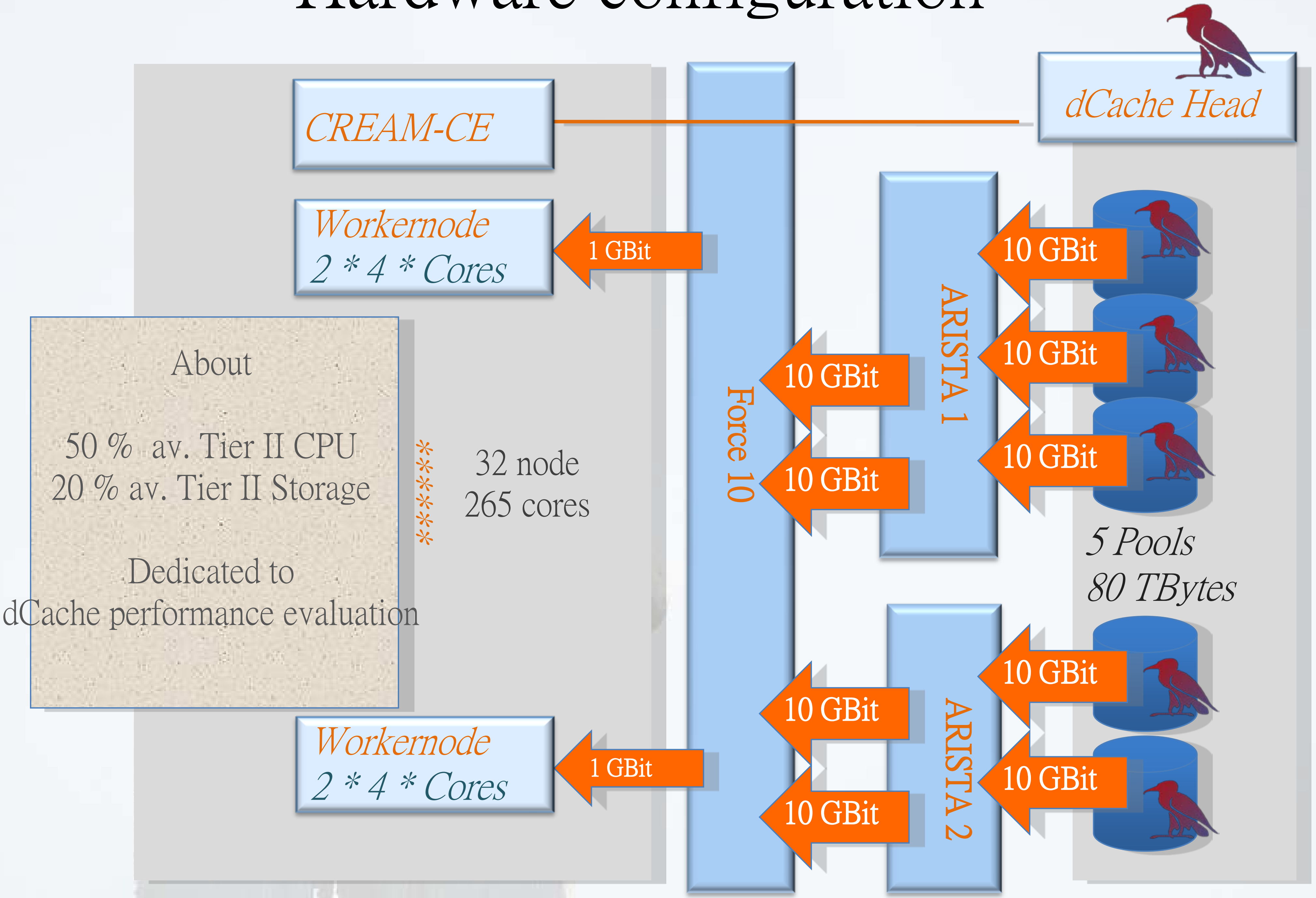
- Working on helping install smaller dCache instances.
  - Reducing time between wanting to have a dCache and storing first file.
- Currently, needs only one command and to deploy a database.
- Aiming to provide a complete out-of-box “just works” RPM.

The DESY Grid Lab

Operated by
Yves Kemp
Dmitri Ozerov



Hardware configuration



CREAM-CE

dCache Head

Workernode
*2 * 4 * Cores*

1 GBit

Force 10

ARISTA 1

ARISTA 2

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

5 Pools
80 TBytes

******

32 node
265 cores

Workernode
*2 * 4 * Cores*

1 GBit

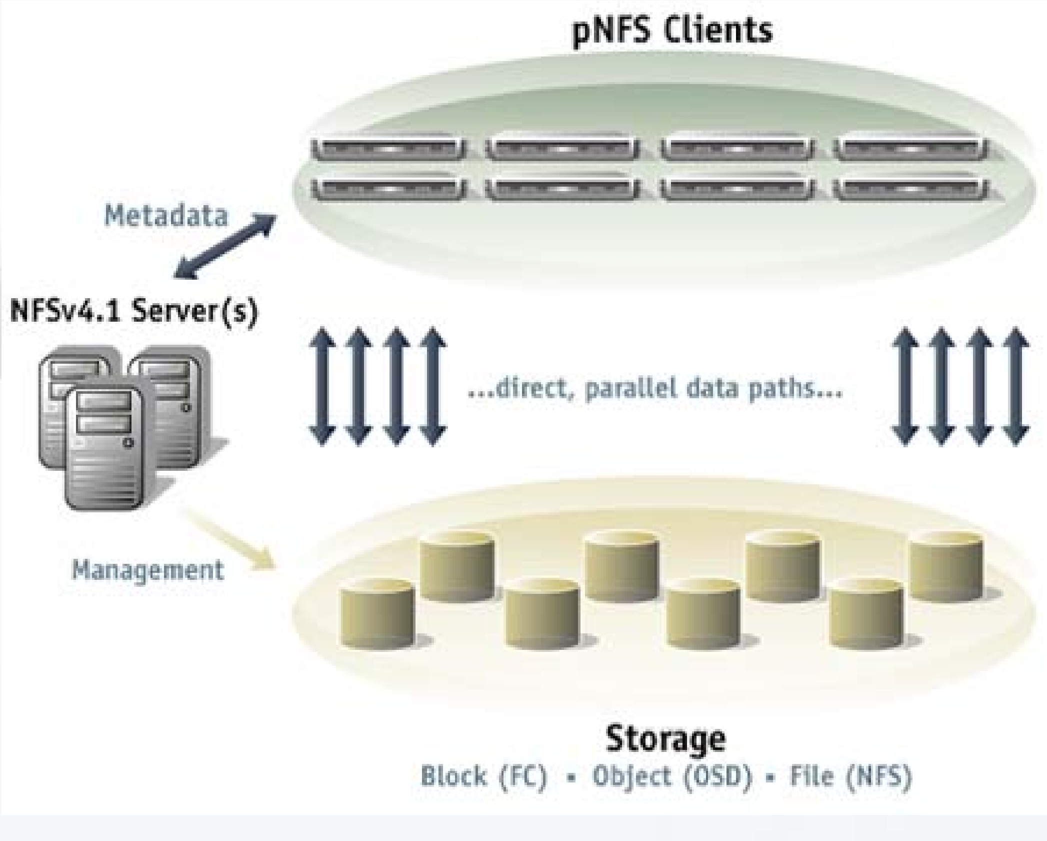
About

50 % av. Tier II CPU
20 % av. Tier II Storage

Dedicated to
dCache performance evaluation

Pushing the envelope: NFS v4.1

- dCache.org is member of CITI group of NFS v4.1 server and client implementers
  - We take part in NFS v4.1/pNFS connectathon, bakeathon events
- dCache is first available NFS v4.1 server,
- 2.6.38 has pNFS support; expect RHEL/ SL 6.2 to work out-of-box.

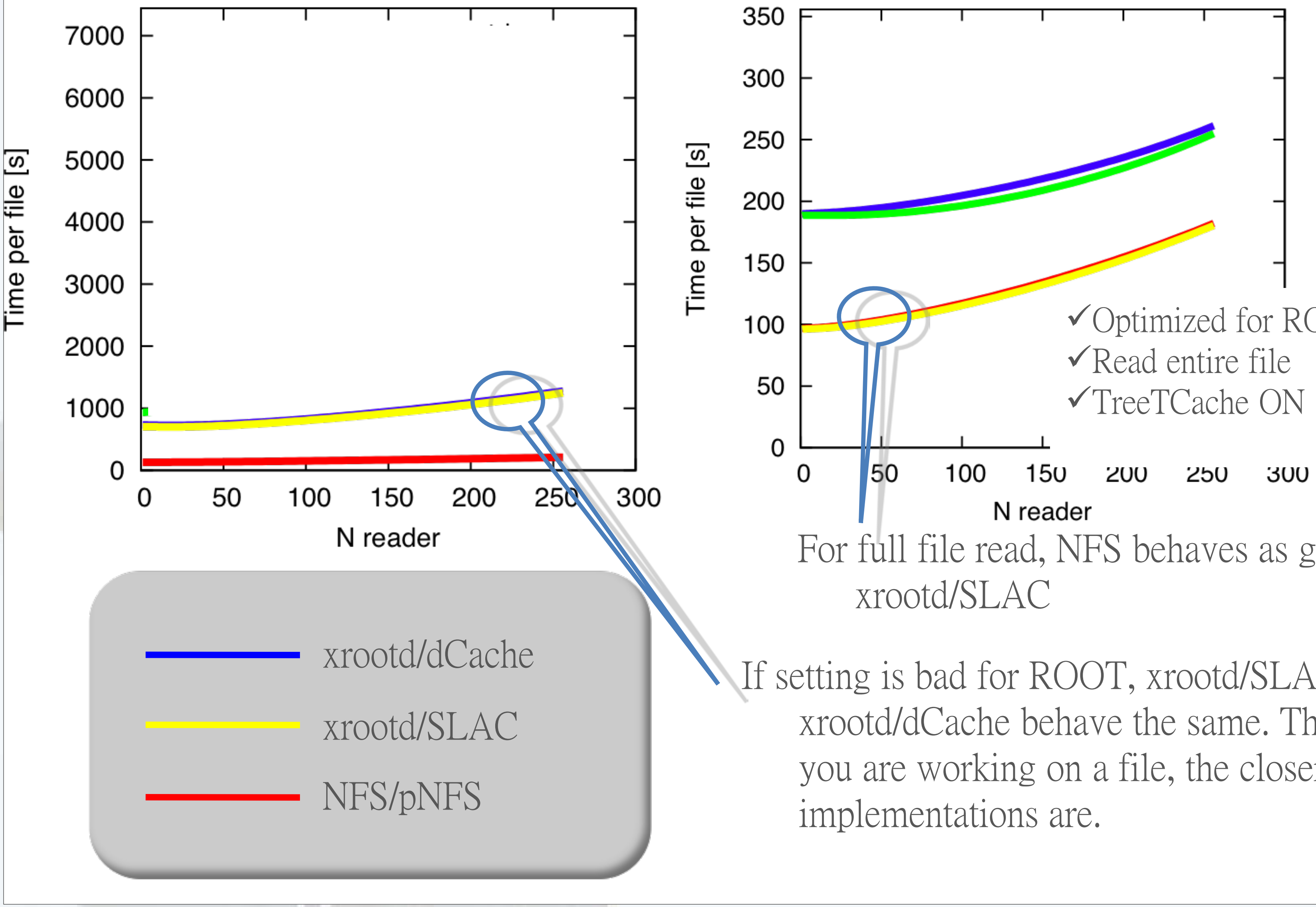

pNFS Clients
Metadata
NFSv4.1 Server(s)
...direct, parallel data paths...
Management
Storage
Block (FC) • Object (OSD) • File (NFS)

xrootd vs NFS 4.1

Reading entire file.

Worst Case for ROOT

Best Case for ROOT



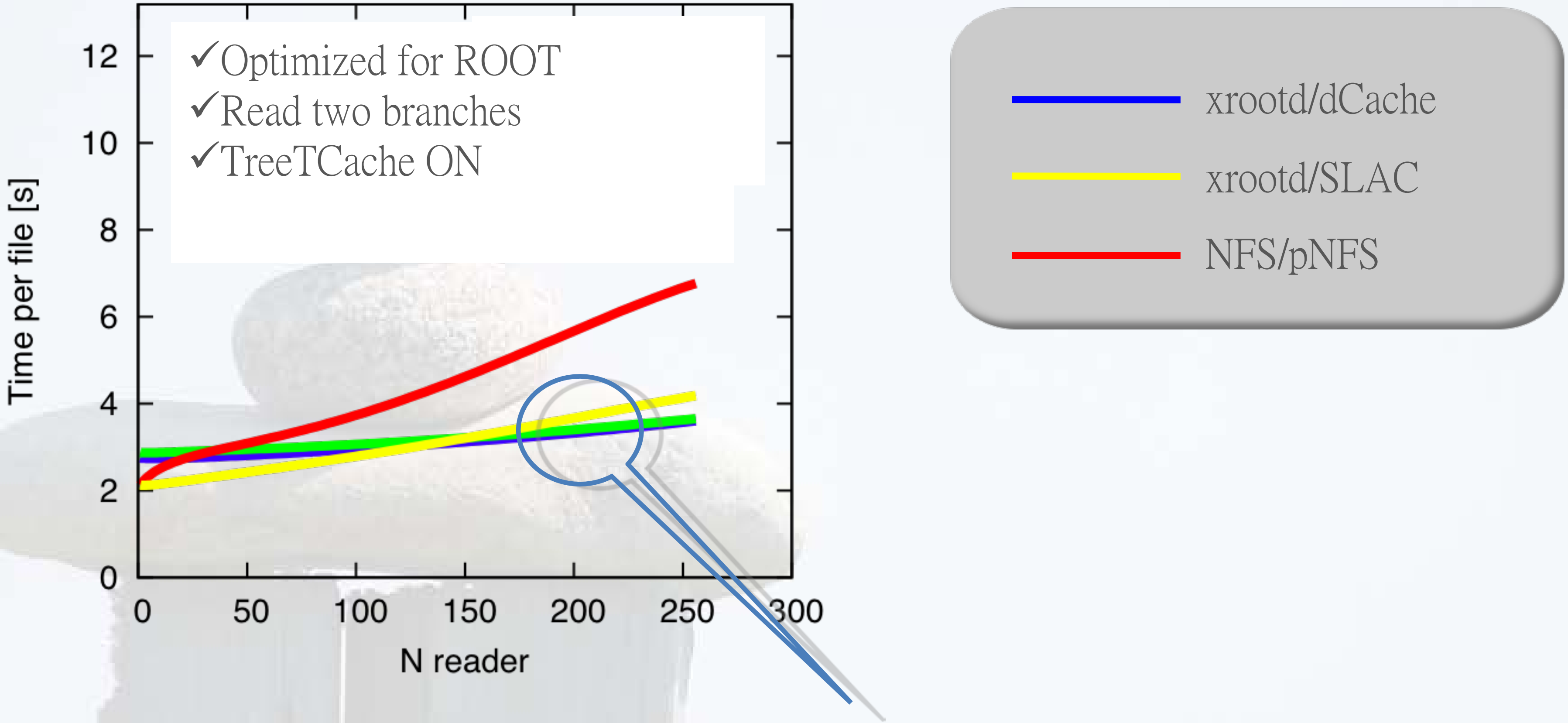
✓Optimized for ROOT
✓Read entire file
✓TreeTCache ON

For full file read, NFS behaves as good as xrootd/SLAC

If setting is bad for ROOT, xrootd/SLAC and xrootd/dCache behave the same. The longer you are working on a file, the closer both implementations are.

xrootd vs NFS 4.1

We trying to find a case where NFS 4.1 is really bad (and found one)



✓Optimized for ROOT
✓Read two branches
✓TreeTCache ON

The vector read effect. The ROOT driver doesn't do vector read for POSIX IO (i.e. NFS v4.1) but does so for xrootd.

Moral of the story...

Life is difficult

WebDAV

- Not everyone wants to use X.509 certificates and GridFTP to access data remotely.
    - Some people might want to read files using their web-browser.
- Support HTTP/HTTPS for reading and browsing.
- Support WebDAV for reading/writing/..
    - (all OSes have at least one WebDAV client)
- Supports different authentication schemes:
    - anonymous and X509 available now,
    - other authentication methods are being added.

Conclusions

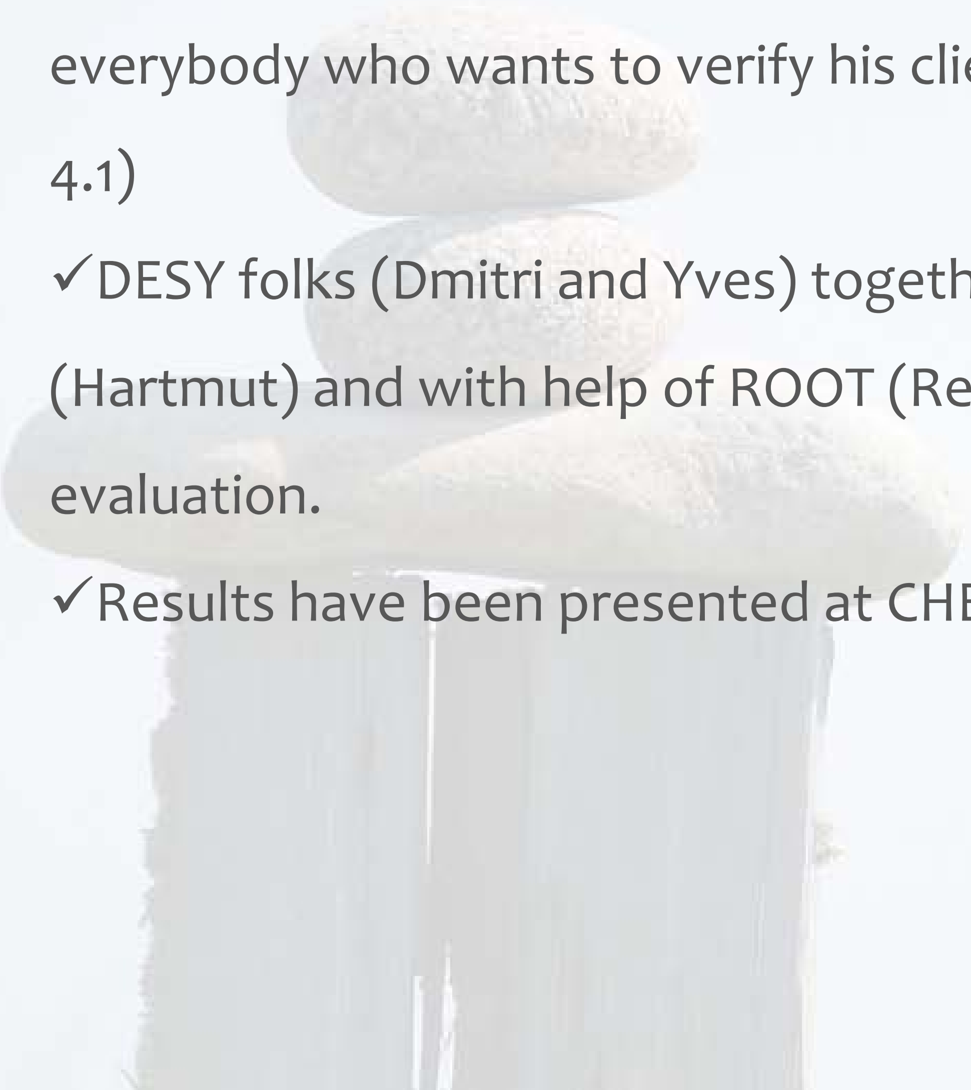
- We continue to make improvements for our existing users:
  - Making storage simpler for site-admins and existing users.
- We're also pushing the envelope of what you can expect from storage system:
  - Dedicated hardware to test real-world scenarios
  - Engaging and adopting standards to allow new communities to use large storage

Image credits

- “Zen stones” by grakki / Jon
- “Powderhorn silo” by naezmi
- “Clouds” by Karin Dalzial
- “Hard Disk” by Jeff Kubina
- “Meerkats” by Jamie Campbell
- “in-with-the-new” by Marie Richie
- “Swiss Army Knife” by AJ Cann
- “Switch it” by smlp.co.uk
- “Wooden boxes” by Asier Villafanca

Backup slides

DESY Grid Lab



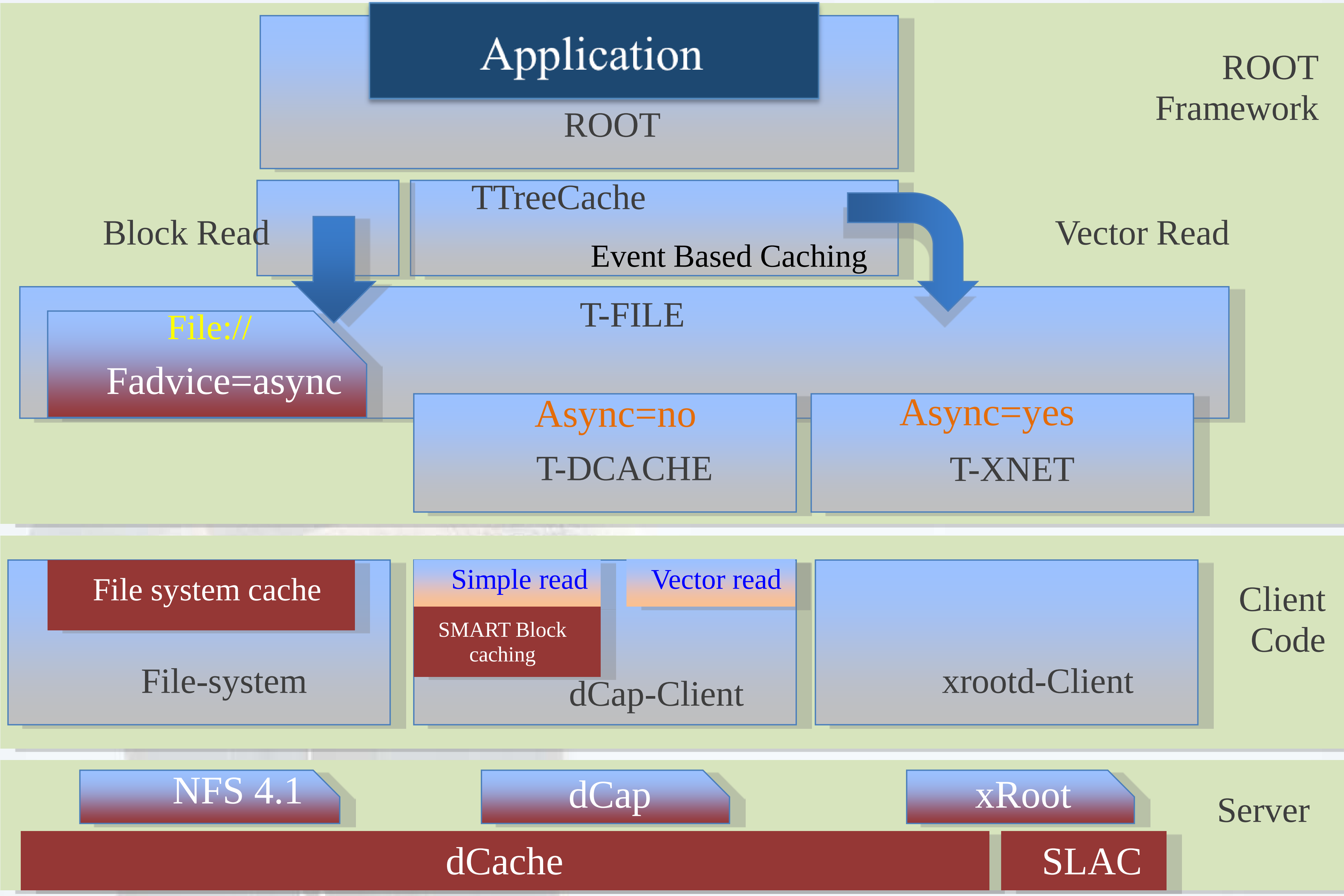
✓Since mid of last year, DESY provides a Tier II like test stand with dCache/pNFS server and pNFS enabled SL5 worker nodes.

✓This test stand is REAL and not paperwork and is available for everybody who wants to verify his client/framework against pNFS. (NFS 4.1)

✓DESY folks (Dmitri and Yves) together with ATLAS (Johannes), CMS (Hartmut) and with help of ROOT (Rene) have been running all kind of evaluation.

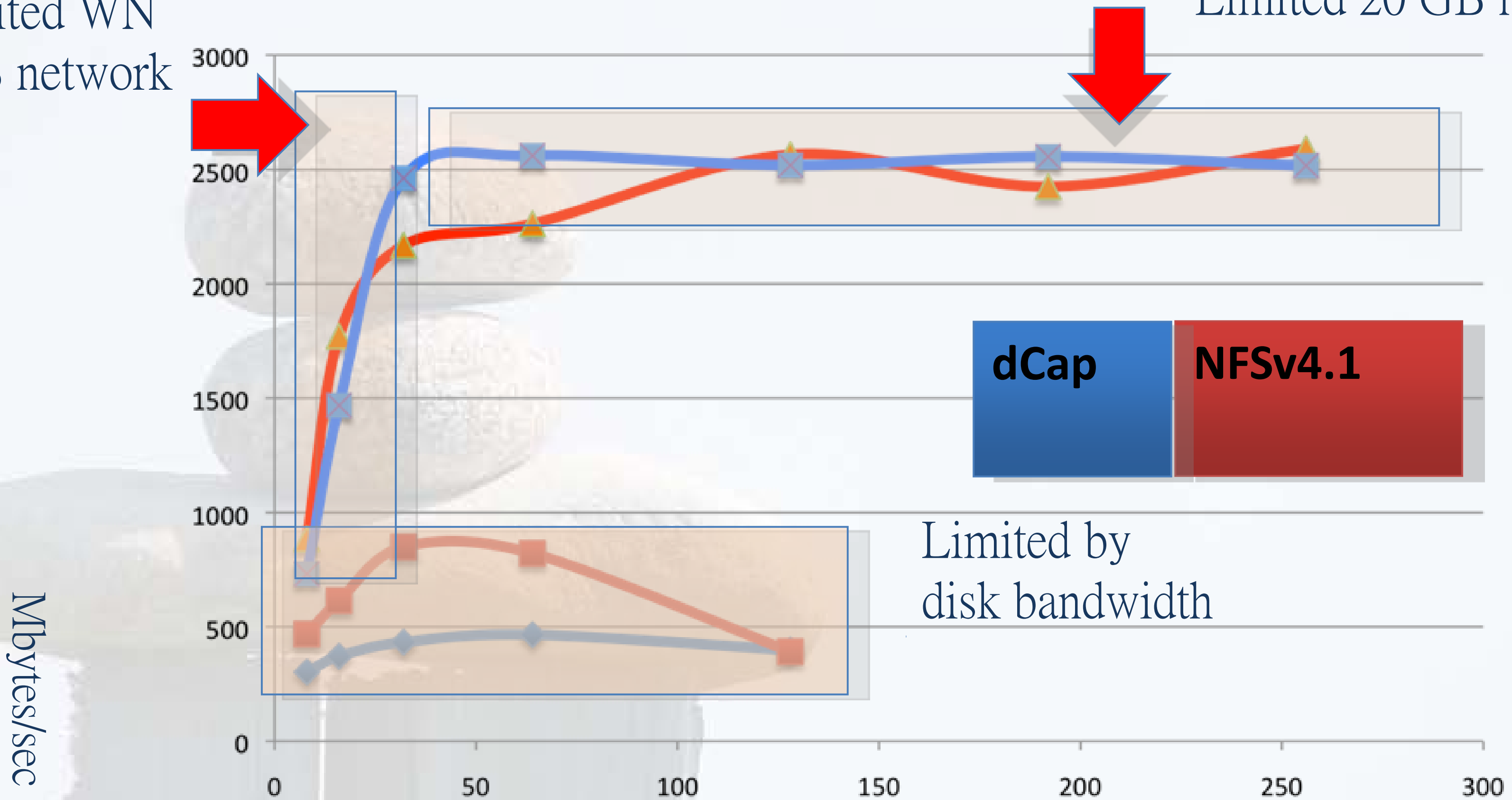
✓Results have been presented at CHEP’10 and at 2010 Spring HEPIX.

ROOT I/O Framework



Limited only by network and disk

Removing server disk congestion effect by keeping all data in file system cache of the pool.



Limited WN
1GB network

Limited 20 GB network

Limited by
disk bandwidth

Number of threads

Total throughput doesn’ t depend on the protocol.