EUROPEAN MIDDLEWARE INITIATIVE

dCache.org HELMHOLTZ ASSOCIATION DESY NDGF FERMIlab

DGI Extension

Tanja Baranova (dCache.org)
Jean-Philippe Baud (CERN)
Johannes Elmsheuser (LMU Munich)
Yves Kemp (DESY)
Maarten Litmaath (CERN)
Tigran Mkrtchyan (dCache.org)
Dmitri Ozerov (DESY)
Ricardo Rocha (CERN)
Andrea Sciaba (CERN)
Hartmut Stadie (DESY, CMS)

NFS 4.1 / pNFS activities in dCache

Patrick Fuhrmann

Content

- ✓ Why should you be interested in pNFS.
- ✓ What is the status and the timeline for 2011 ?
  - ✓ Availability of the different components !
  - ✓ Protocol verification and performance evaluation !
- ✓ Some results from the NFS 4.1 / pNFS evaluation at Grid-Lab
- ✓ What is the pNFS funding model for deployment ?

Why should you be interested in pNFS

Stolen from : http://www.pnfs.com/

Benefits of Parallel I/O

- ➢Delivers Very High Application Performance
- ➢Allows for Massive Scalability without diminished performance

Benefits of NFS (or most any standard)

- ➢Ensures Interoperability among vendor solutions
- ➢Allows Choice of best-of-breed products
- ➢Eliminates Risks of deploying proprietary technology

Two aspect from our perspective

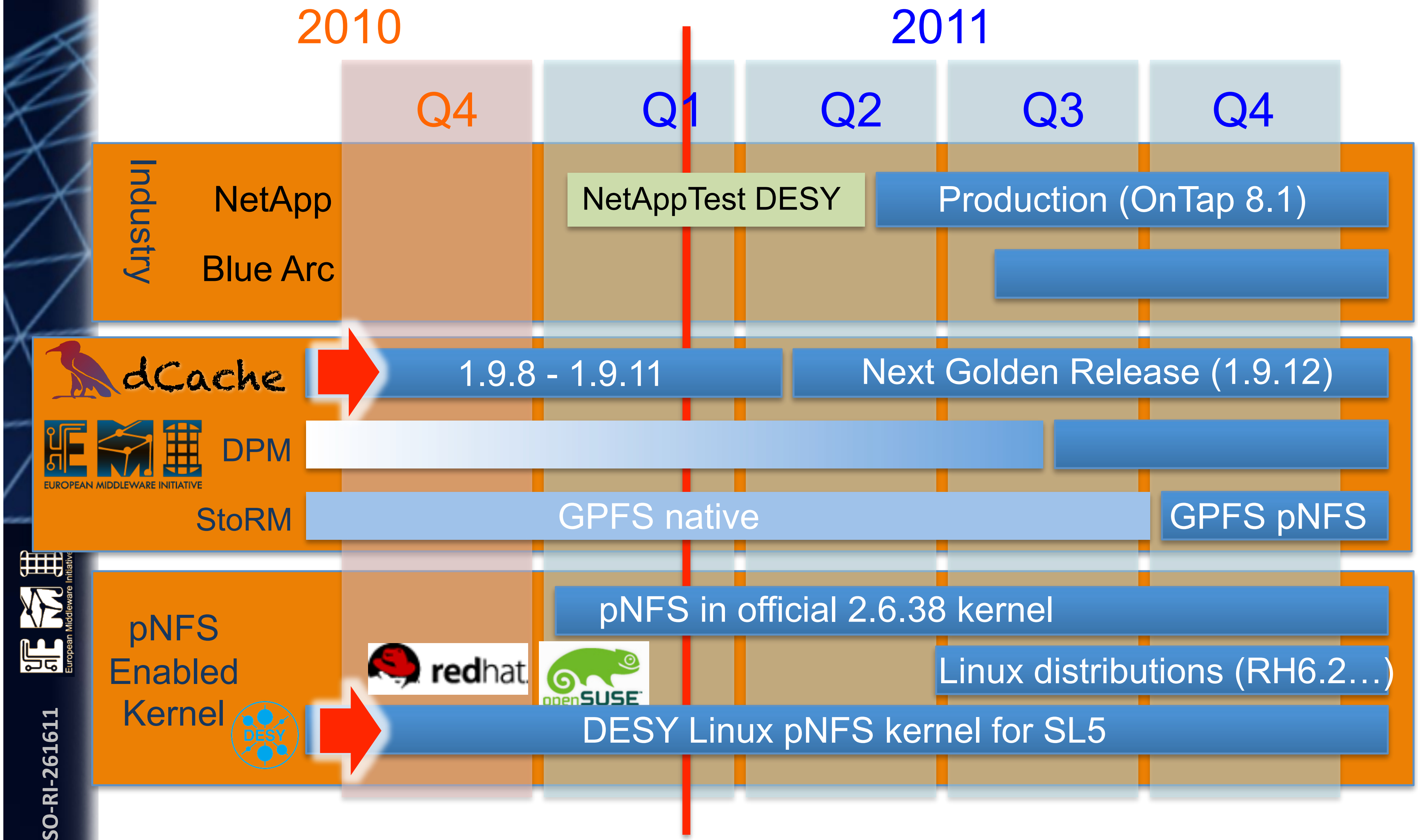
Simplicity

- ✓ Regular mount-point and real POSIX I/O
- ✓ Can be used by unmodified applications (e.g. Mathematica..)
- ✓ Data client provided by the OS vendor
- ✓ Smart caching (block caching) development done by OS vendors

Performance

- ✓ pNFS : parallel NFS (first version of NFS which support multiple data servers)
- ✓ Clever protocols , e.g. Component Requests

Availability for production use



2010

2011

Q4

Q1

Q2

Q3

Q4

Industry

NetApp

NetAppTest DESY

Production (OnTap 8.1)

Blue Arc

dCache

1.9.8 - 1.9.11

Next Golden Release (1.9.12)

DPM

StoRM

GPFS native

GPFS pNFS

pNFS Enabled Kernel

pNFS in official 2.6.38 kernel

Linux distributions (RH6.2…)

DESY Linux pNFS kernel for SL5

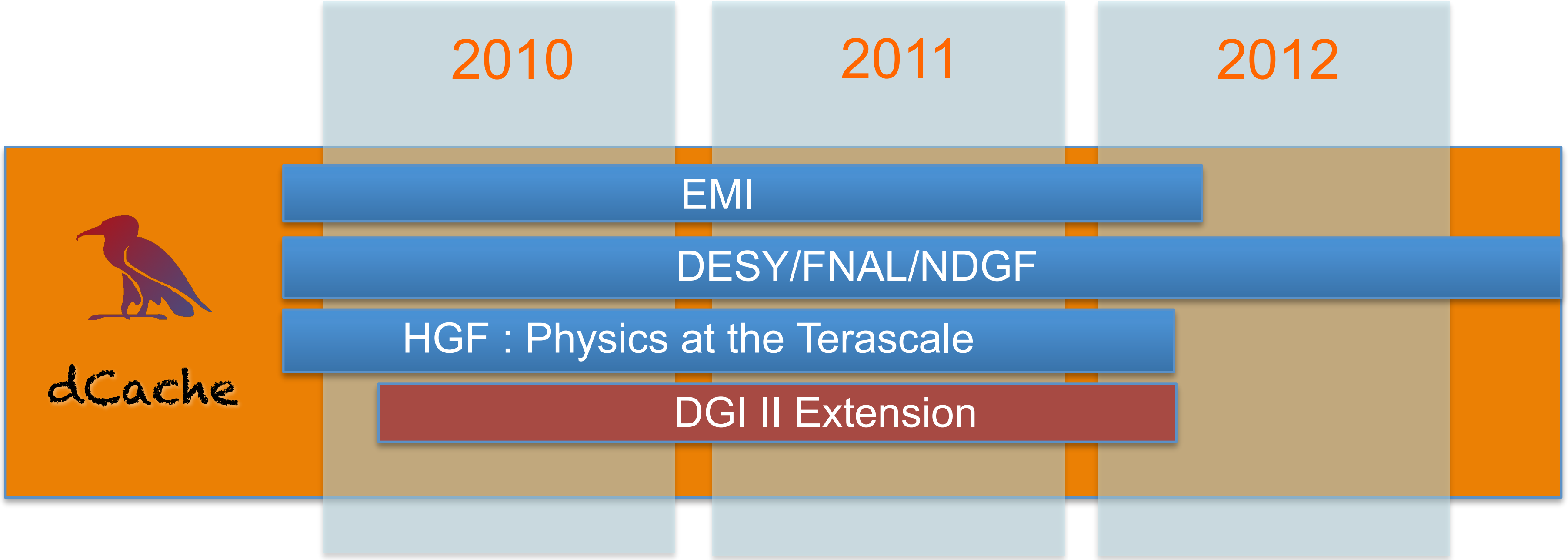
pNFS support in SL5/6

- Full NFS 4.1/pNFS client available in 2.6.38
- Back port into RH6 expected with RH6.2, shortly after it will be in SL6.2.

Funding models

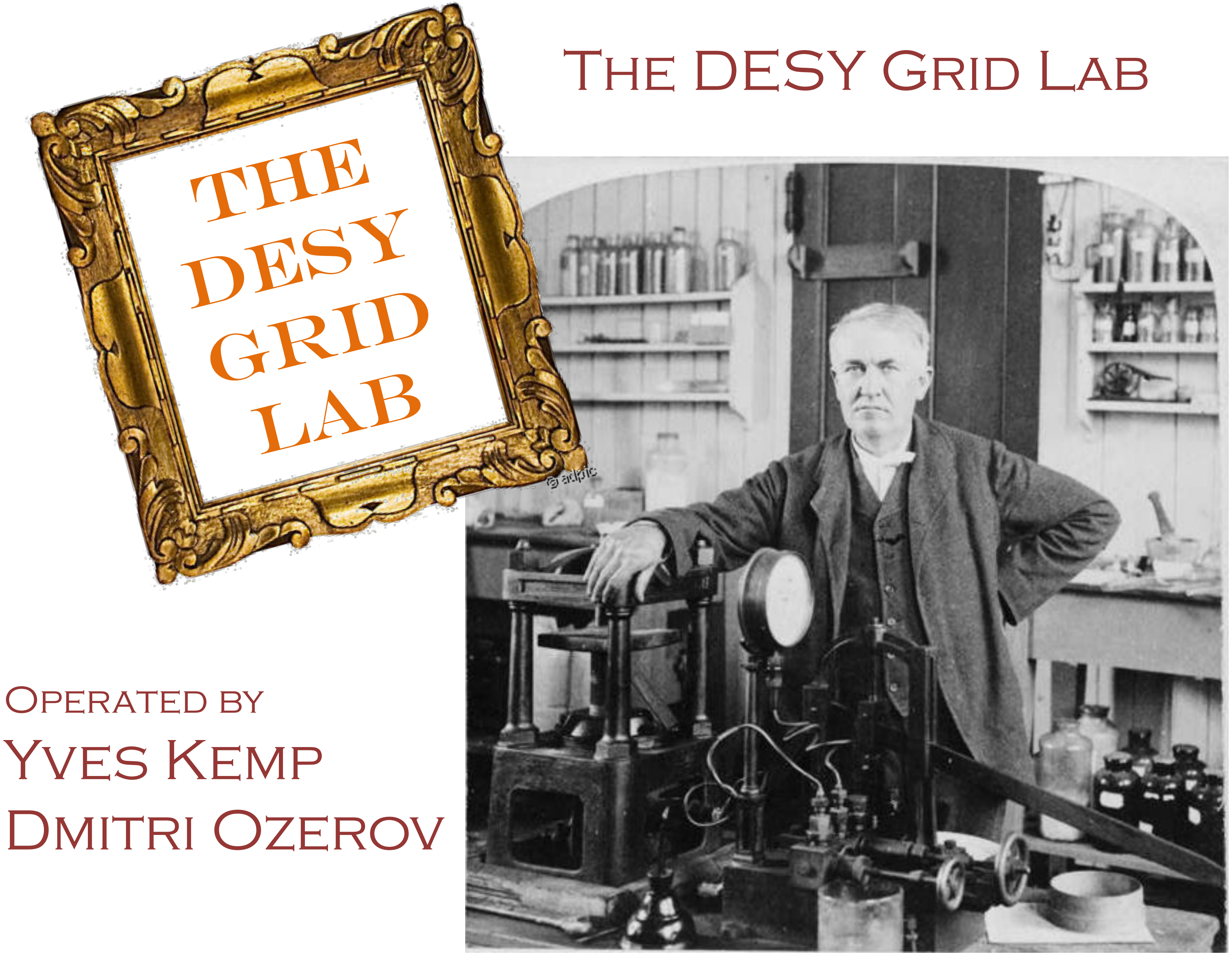
Funding model for deployment and support !

Funding for development and deployment



Funding gives plenty of headroom for pNFS development and deployment.

The DESY Grid Lab



Operated by

Yves Kemp

Dmitri Ozerov

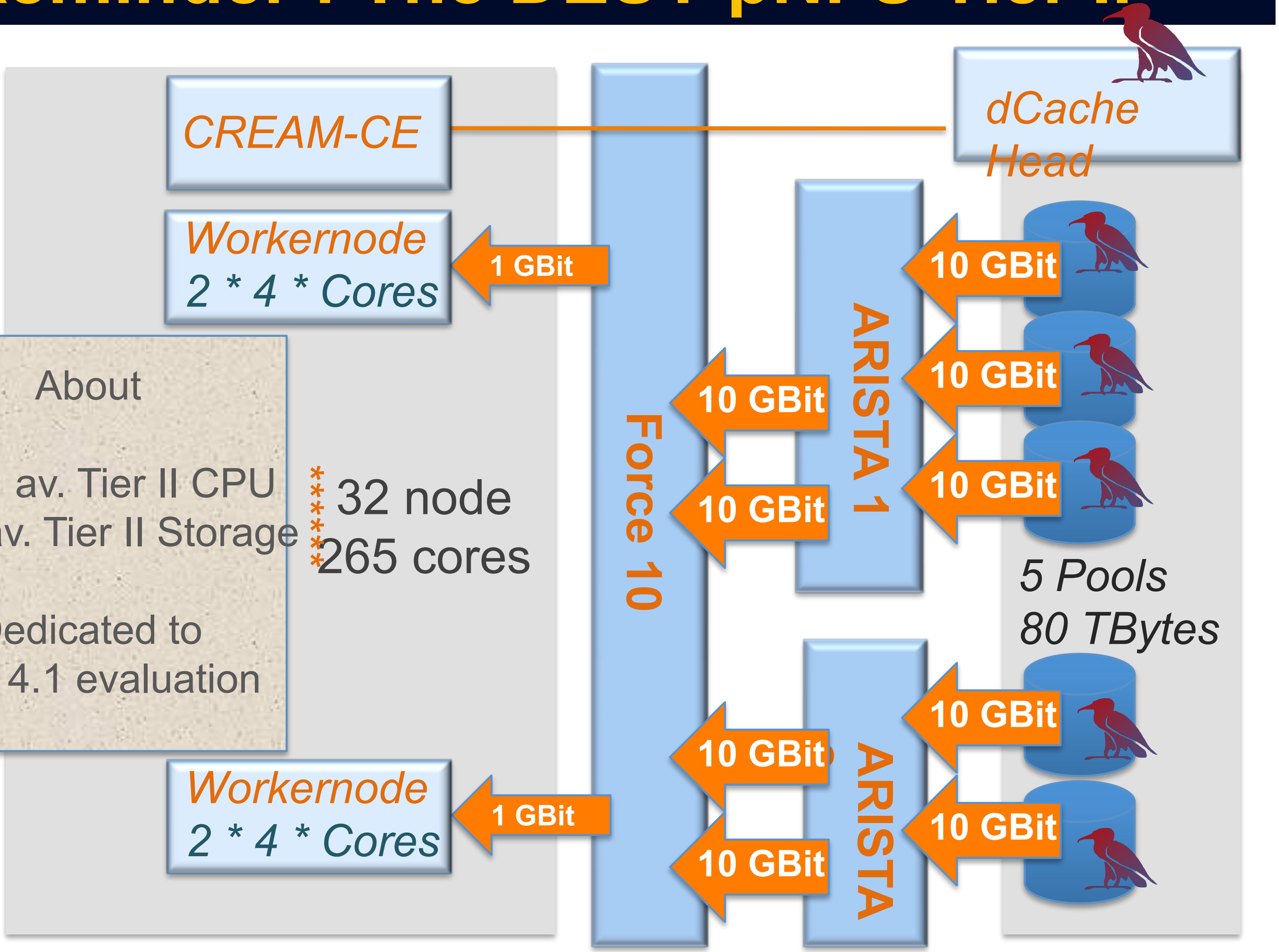
DESY Grid Lab

- ✓ 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.

Disclaimer

This presentation is about comparing NFS 4.1 with xrootd (SLAC and dCache). The results for the dCap protocol should be just ignored as we used an old version of the dcap client which doesn't yet support smart block caching, introduced by Günter Duckeck.

Reminder : The DESY pNFS Tier II



About

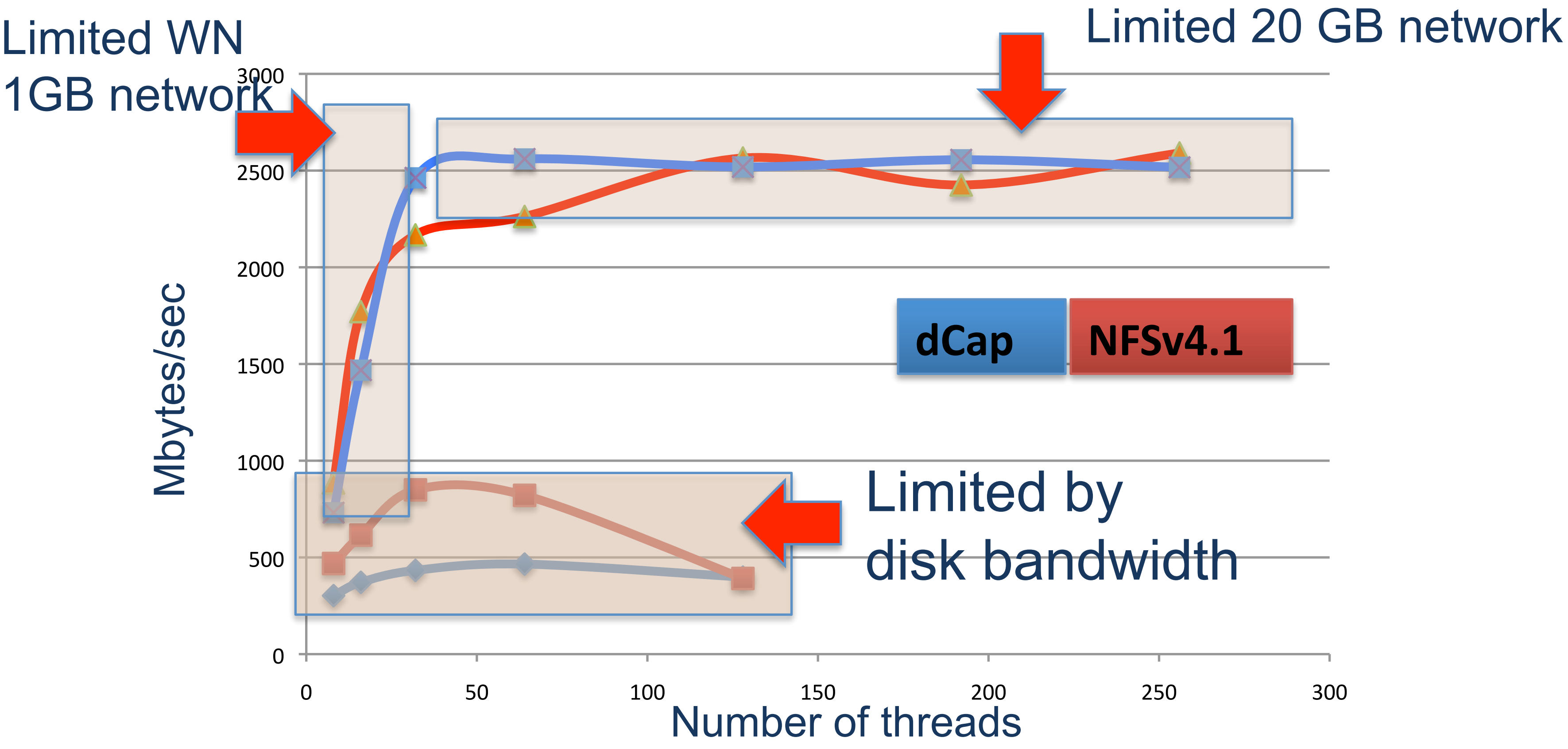
50 % av. Tier II CPU
20 % av. Tier II Storage

Dedicated to
NFS 4.1 evaluation

***** 32 node
265 cores

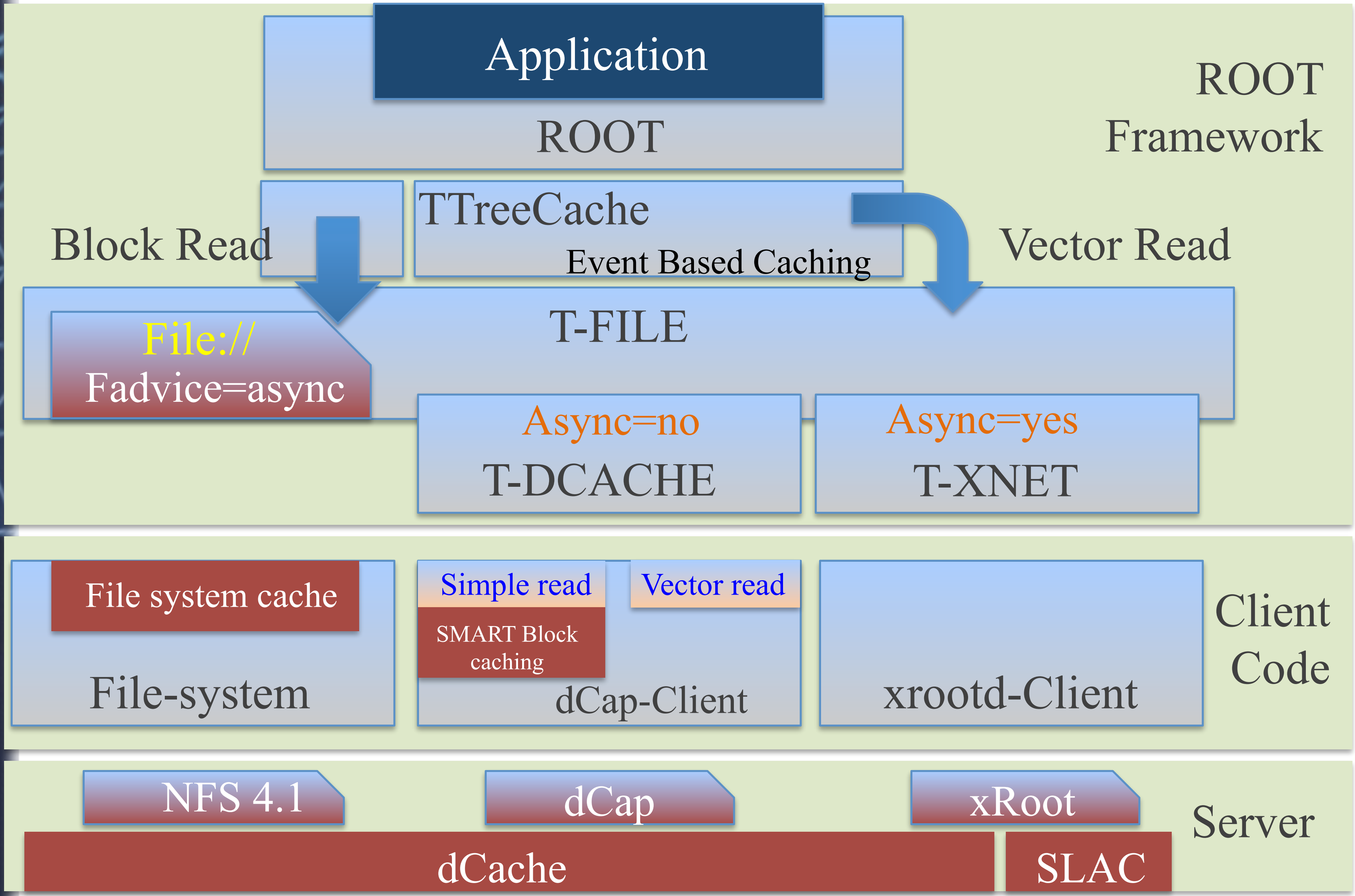
Limited only by network and disk

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



Total throughput doesn’t depend on the protocol.

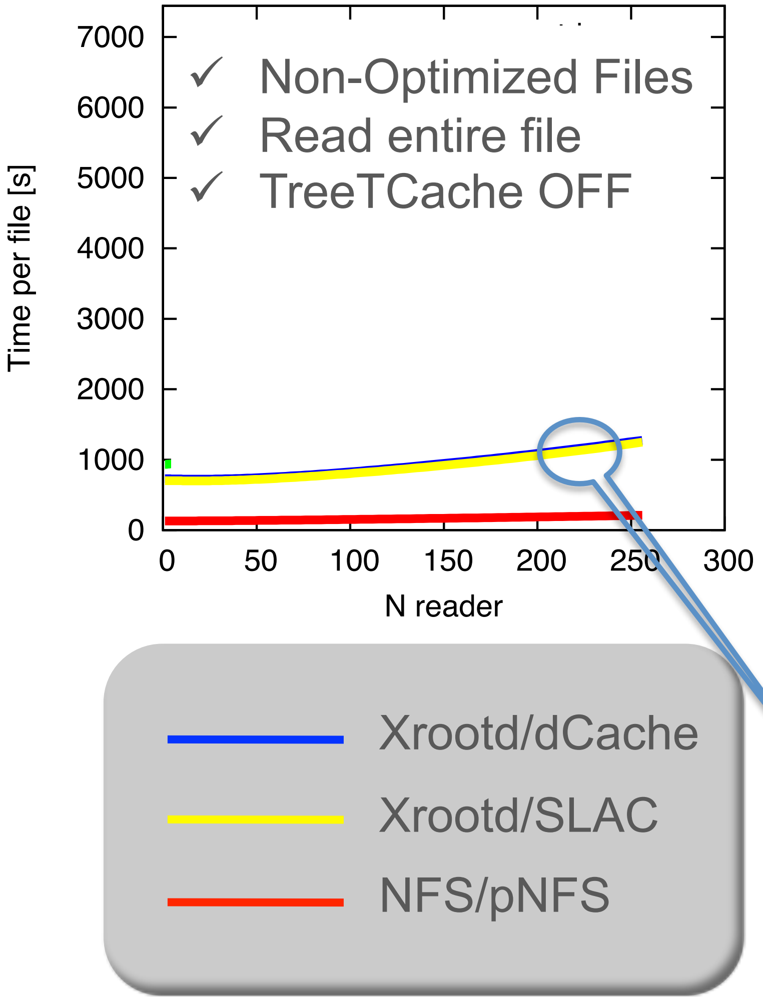
ROOT I/O Framework



xRoot / NFS 4.1

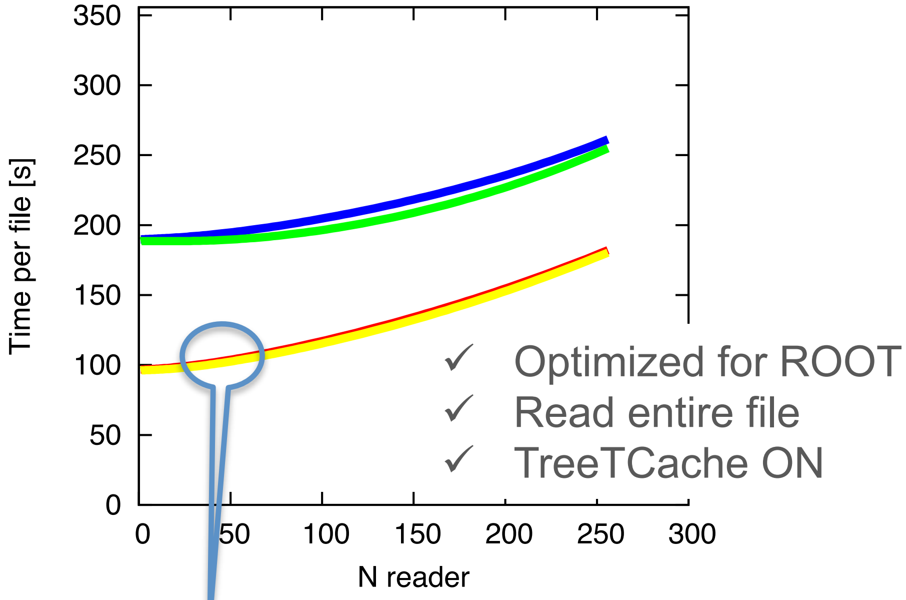
Reading entire file.

Worst Case for ROOT



- ✓ Non-Optimized Files
- ✓ Read entire file
- ✓ TreeTCache OFF

Best Case for ROOT



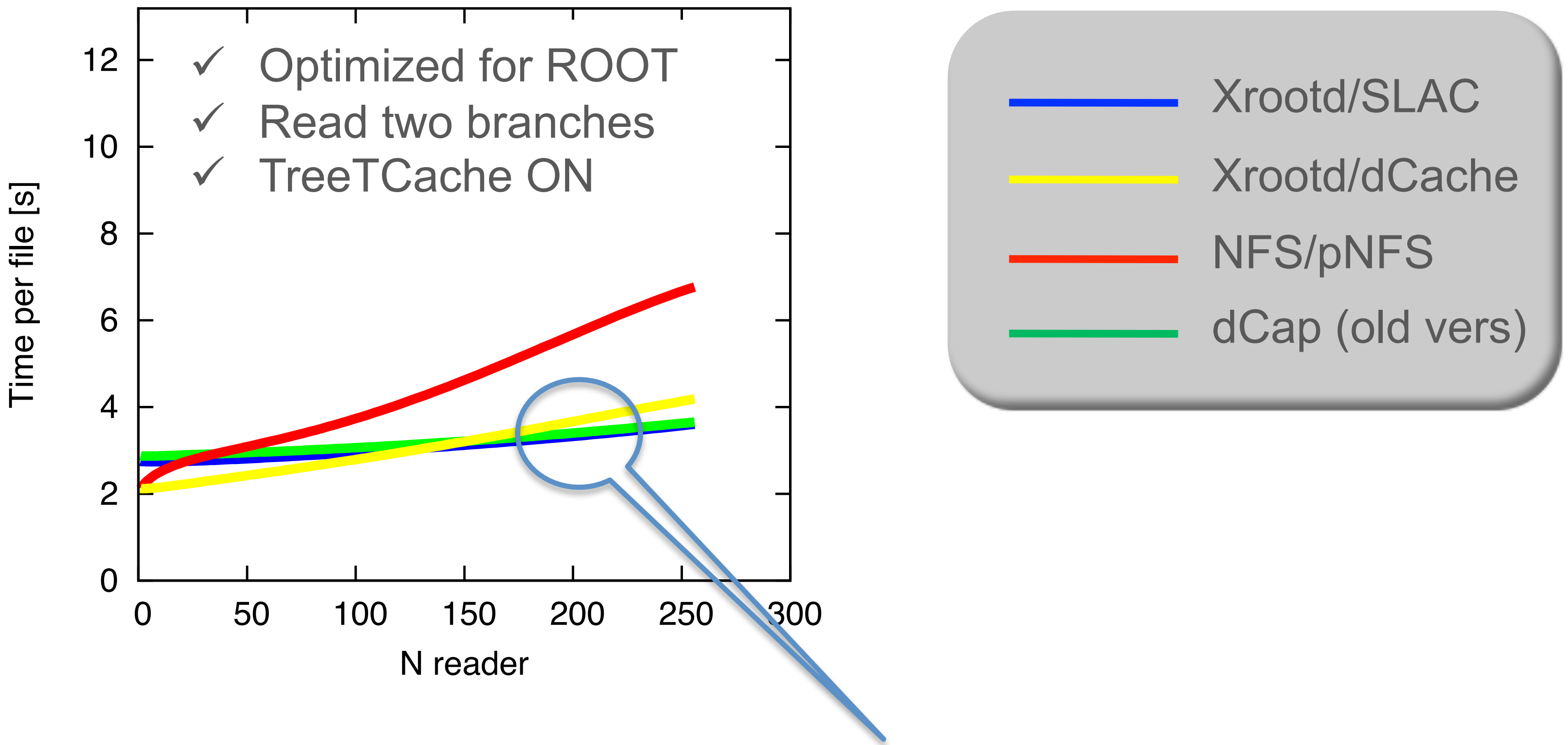
- ✓ Optimized for ROOT
- ✓ Read entire file
- ✓ TreeTCache ON

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

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

xRoot / NFS 4.1

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



- ✓ Optimized for ROOT
- ✓ Read two branches
- ✓ TreeTCache ON

Vector read effect. The ROOT driver is not doing vector read for plain file systems but for dCap/xRoot,

Message

Life is difficult

Conclusion

Protocol verification

- For new protocols, it is not clear *per se* that client and server developers understand all the details.
- But dCache people keep in touch with the experts :
  - dCache.org is member of the CITI group which is coordinating the NFS 4.1 efforts.
  - Three times a year dCache.org is participating the Connecathons resp. Bakaethons to verify compatibility.

Conclusion

Performance

- Reliable and reproducible performance measurements are extremely difficult. The results highly depend on
  - the way the file (ROOT) was written
  - the access profile (ROOT script)
- BUT : Based on our massive testing we are convinced that we and the Linux pNFS kernel developers understand the protocol and that we are running a professional implementation.
- The performance exceeds expectations.

Server configuration (Stolen from Oleg)

```
[nfsv41Domain]
[nfsv41Domain/nfsv41]
```

Start the services:

* nfsv41Domain

and mount /pnfs :

```
# mount -t nfs4 -o minorversion=1,rsize=32768,wsize=32768 localhost:/pnfs /pnfs
```

References

Some references

References

Center for Technology Integration

http://www.citi.umich.edu/

NFS

http://www.nfsv4.org/nfsv4techinfo.html

PNFS

http://www.pnfs.com/

RFC 5661

http://tools.ietf.org/html/rfc5661

NFS 4.1 in first dCache Golden Release (1.9.5)

http://www.dcache.org/downloads/1.9/release-notes-1.9.5-1.html

EMI, The European Middleware Initiate

http://www.eu-emi.eu/en/

EMI, The European Grid Infrastructure

http://www.egi.eu

WLCG Collaboration Workshop, July 20, 2010, Patrick Fuhrmann

http://www.dcache.org/manuals/2010/20100707-2-NFS4_demonstrator.pdf

Grid Deployment Board, Oct 13, 2010, Patrick Fuhrmann

http://www.dcache.org/manuals/2010/NFS41-demonstrator-milestone-2.pdf

11 Reasons you should care, June 16, 2010, Gerd Behrmann

http://www.dcache.org/manuals/2010/20100617-gerd-nfs.pdf

References

CHEP 2010, Oct 20, 2010, Yves Kemp :

http://www.dcache.org/manuals/2010/CHEP2010-NFS41-kemp.pdf

Hepix Fall 2010, Nov 2, 2010, Patrick Fuhrmann

http://www.dcache.org/manuals/2010/20101102-hepix-patrick-nfs41.pdf

Linux Kernel : www.kernel.org

http://www.kernel.org/pub/linux/kernel/v2.6/ChangeLog-2.6.37

NetApp : www.netapp.com

http://media.netapp.com/documents/wp-7057.pdf

BlueArch : www.bluearc.com

http://www.bluearc.com/storage-news/press-releases/101112-bluearc-demos-pnfs-at-supercomputing-2010.shtml

Scientific Linux

http://www.scientificlinux.org

FERMIlab

http://www.fnal.gov

pNFS enabled SL5 Kernel

http://www.dcache.org/chimera/x86_64; dcache-www01.desy.de/yum/nfs4.1/el5/nfsv41.repo

EUROPEAN MIDDLEWARE INITIATIVE

Thank you

EMI is partially funded by the European Commission under Grant Agreement INFSO-RI-261611