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)

Report on the NFS 4.1 pNFS activities

Patrick Fuhrmann
EMI data area lead

Content

- Why should you be interested in pNFS.
  - Experiment, user, infrastructure
- What is the status and the timeline for 2011 ?
  - Availability of the different components !
  - Protocol verification and performance evaluation !
- What is the pNFS funding model for deployment ?
- What is the funding model for support in production beyond 2012 ?

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

Please note

Disclaimer :

In this presentation the term pNFS (Parallel NFS) will be used instead of the correct term : “NFS 4.1 (pNFS)”.

pNFS has nothing at all to do with the dCache PNFS file system.

Why should you be interested

Simplicity

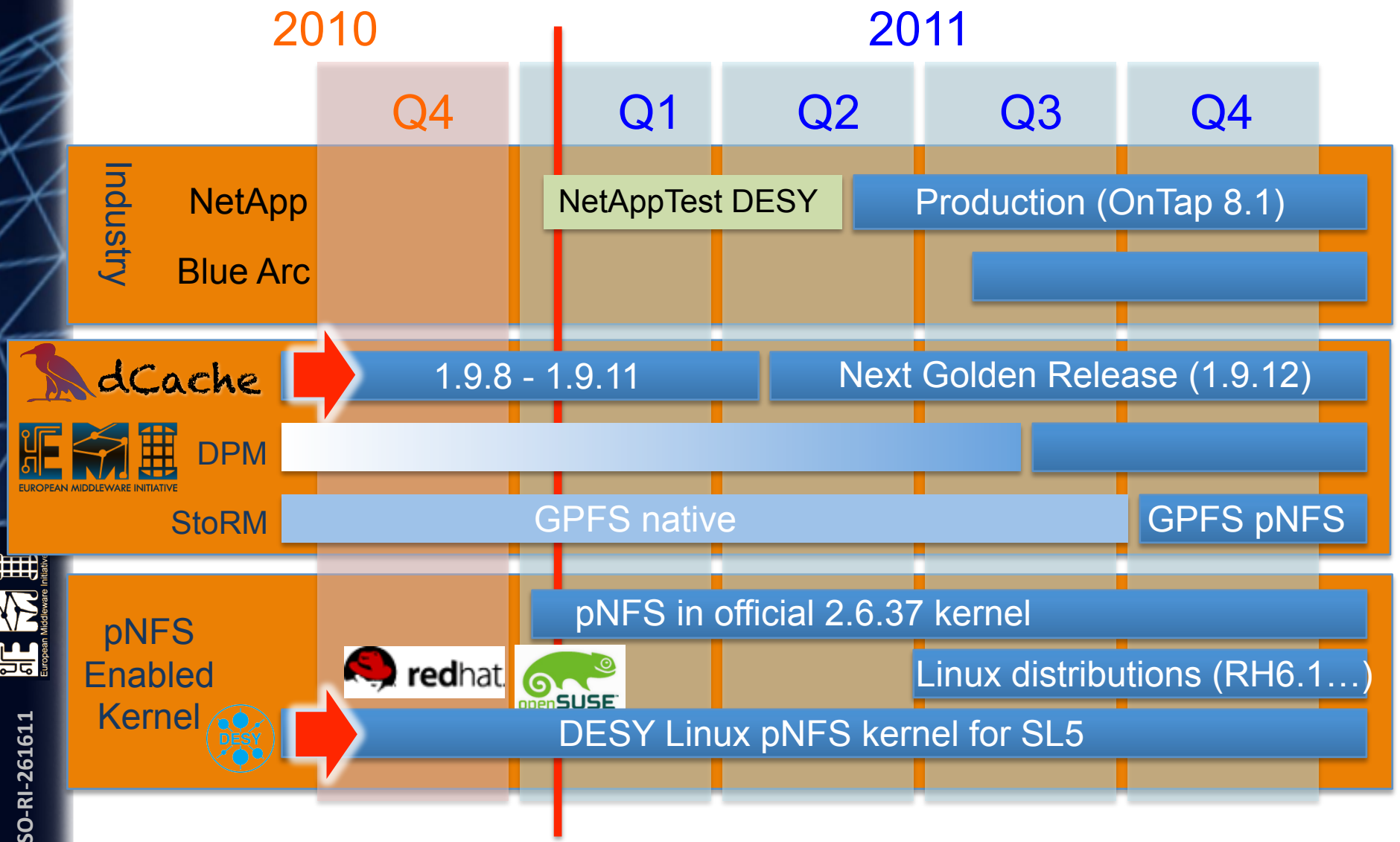
- Regular mount-point and real POSIX I/O
- Can be used by unmodified applications (e.g. Mathematica..)

Cost reduction for software providers and Infrastructures

- Less components to maintain and deploy
  - Data clients provided by the OS
  - No additional server needed
- Smart caching (block caching) development done by OS vendors
- Only single FS driver in ROOT

With EMI-1 dCache and StoRM already provide production ready POSIX I/O, DPM will catch-up later 2011.

Availability for production use



pNFS support in SL5

A pNFS enabled kernel for SL5 is available from the dCache.org web pages.

dCache.org is trying to convince our FERMIlab SL5 colleagues to make an pNFS enabled kernel available in SL5 and possibly support it.

At this point, the help of CERN-IT resp. WLCG would be very much appreciated.

Test stands

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

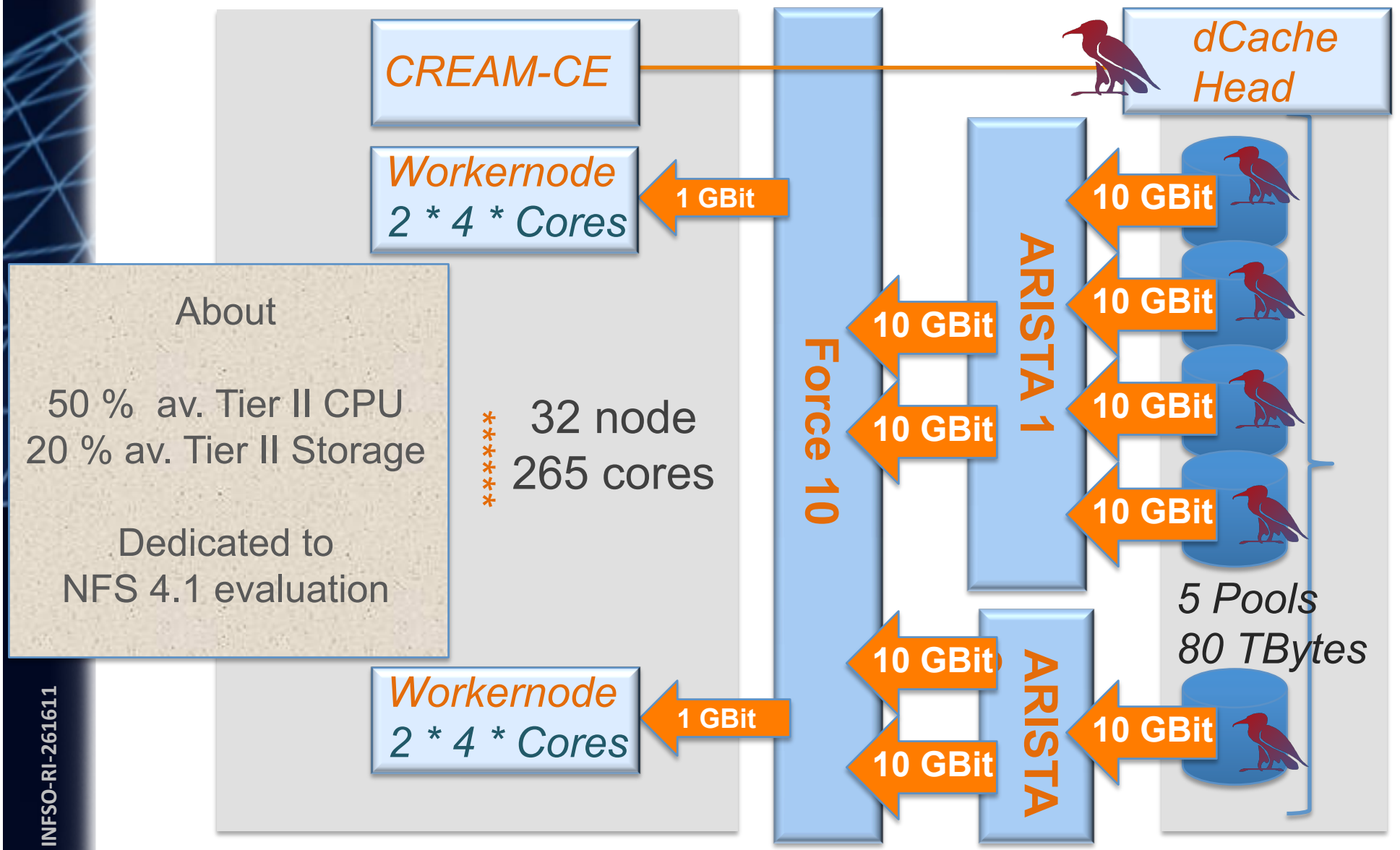
This test stand is REAL and not paperwork and is available for everybody who wants to verify his client/ framework against pNFS. (NFS 4.1)

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.

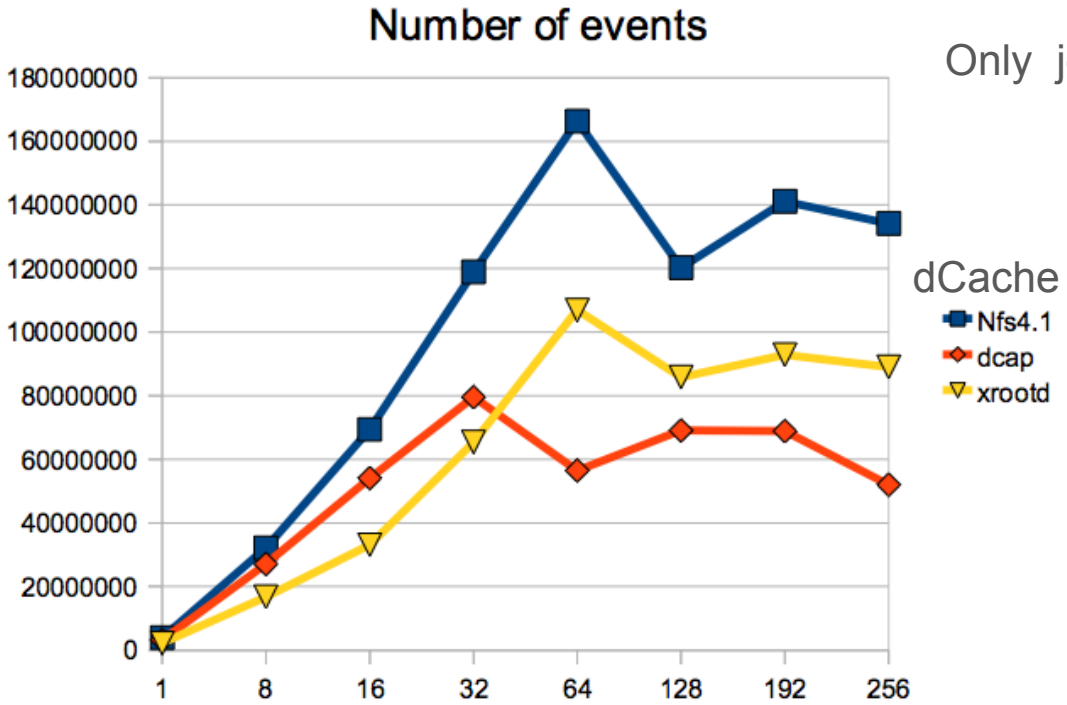
Bottom line : pNFS performance at least as any good as other protocols. However, we have indications that it behaves better.

The DESY pNFS Tier II

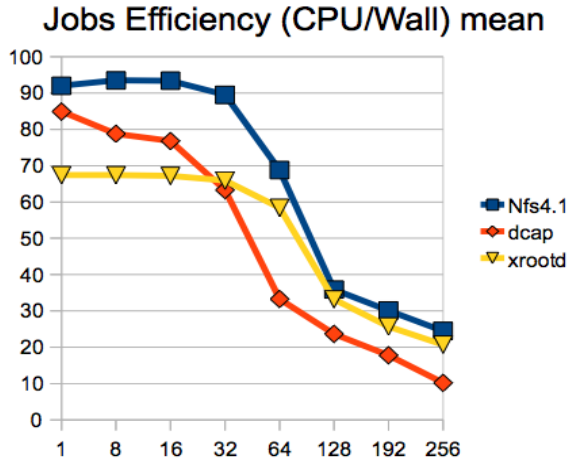


Example : Hammercloud Analysis

With kind permission by Dmitri Ozerov and Yves Kemp (CHEP'10 paper)



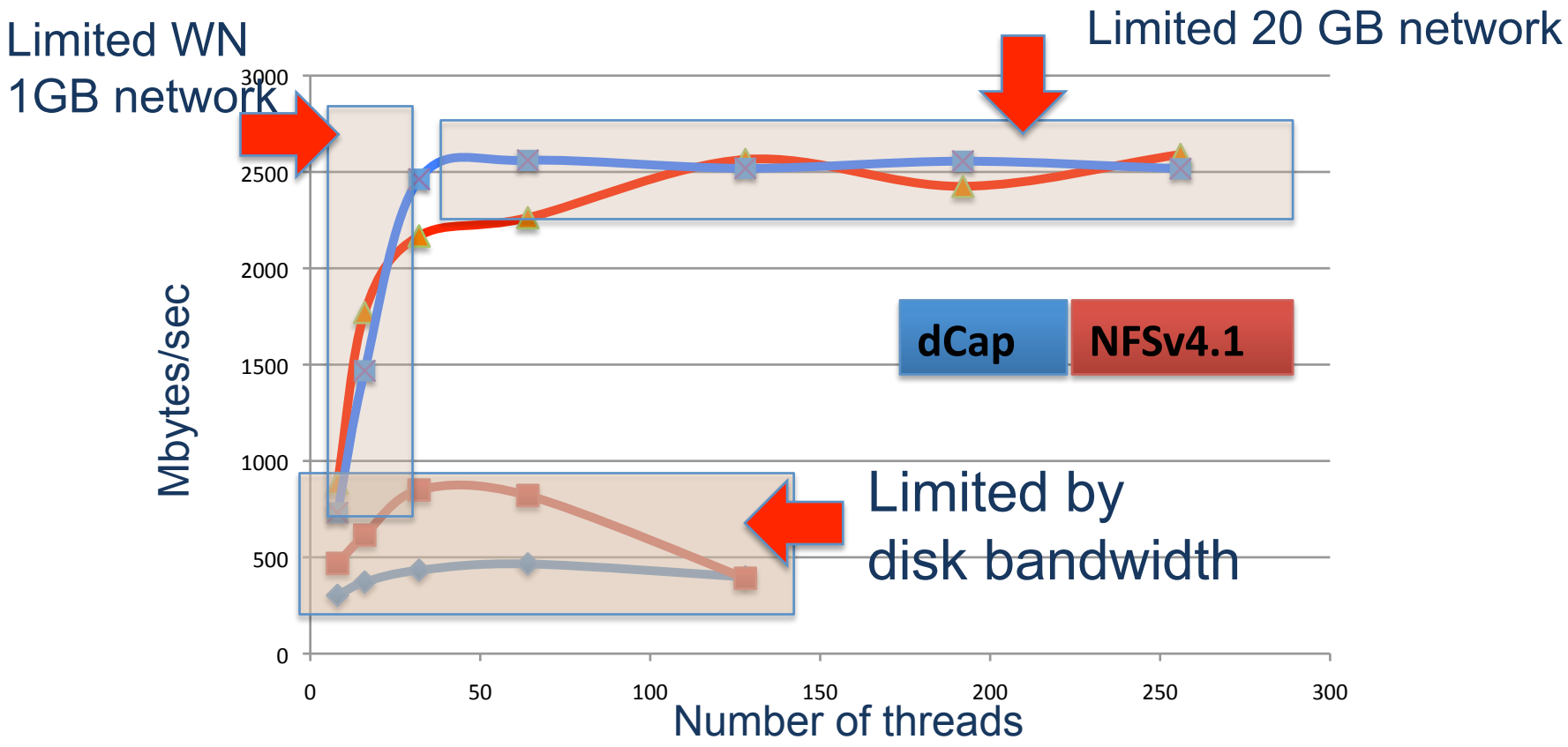
Only jobs finished within 24 hours.



Johannes Elmsheuser :
"The analysis used is a typical ATLAS Athena Monte Carlo analysis analyzing the AOD data format. The analysis is primarily selecting Muons and Muon trigger objects. The Athena version used to process the AODs are 15.6.6 and 16.0.2.3. The Athena versions used to reconstruct and produce the AODs are 15.6.x. The ROOT versions used are 5.22/00h (15.6.6) and 5.26/00e (16.0.2.3)"

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.

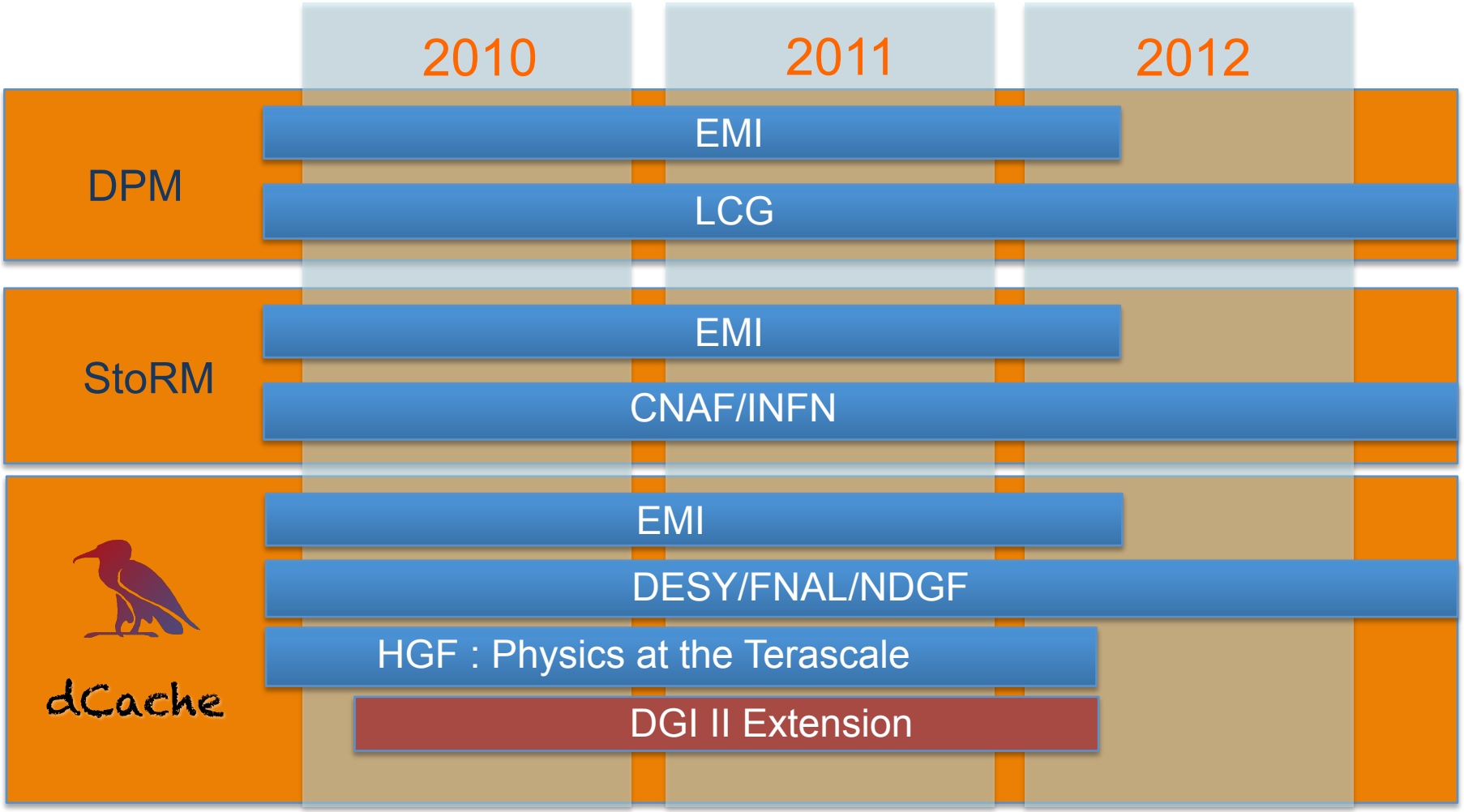
Conclusion

- ✓ Protocol verification
  - 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.
- ✓ Peformance
  - 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.

Funding models

Funding model for deployment and support !

Funding for development and deployment



Funding gives plenty of headroom for pNFS development and deployment.

Funding for support beyond 2011/12

During EMI, development, testing and deployment of pNFS in EMI SE's will be finalized.
After EMI, only low level support is expected, which will can be easily covered by the storage system providers, as :

For DESY, FNAL and NDGF, dCache is a strategic product which they themselves highly depend on. Moreover, dCache is serving more than 50% of the LCG data and as such has sufficient momentum for longterm future support.

For CNAF/INFN, StoRM is a strategic product which they themselves highly depend on.

DPM

Statement Markus Schulz (this morning):
"DPM is a critical data management component for WLCG that is in use at more than 200 sites. CERN is committed to the active support and evolution of the product in the foreseeable future.".

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

Something else

And now for something completely different ….

The global namespace ex. demonstrator

dCache.org is very interested in the global namespace approach of Brian B.

Although NFS 4.1 provides a similar functionality the overlay technique sounds promising.

Therefore we (Gerd, NDGF) successfully started to contribute to Brian's global namespace system.

Replacement of the SLAC xroot deamon by the dCache implementation with X509 support.

However

Gerd (NDGF) would be willing to implement the missing CMSD as well, but only under some conditions :

In order to guarantee interoperability in the future the MB must make sure that :

- The CMSD protocol has to be clearly specified, following common rules, e.g. from IETF RFC's or OGF working groups.
- A change in the protocol needs agreement between the implementers. The process has to be supervised by an independent board composed of developers of the different implementing parties.

Which is normal procedure everywhere else in the world.

Thank you

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