NFS 4.1 / pNFS
The final steps

Patrick Fuhrmann, EMI, DESY, dCache.org

Content

- Who contributed to this presentation ?
- What's the issue ?
- How it works.
- Benefits.
- Who is involved ?
- Performance
- Some last words

Contributions to this presentation

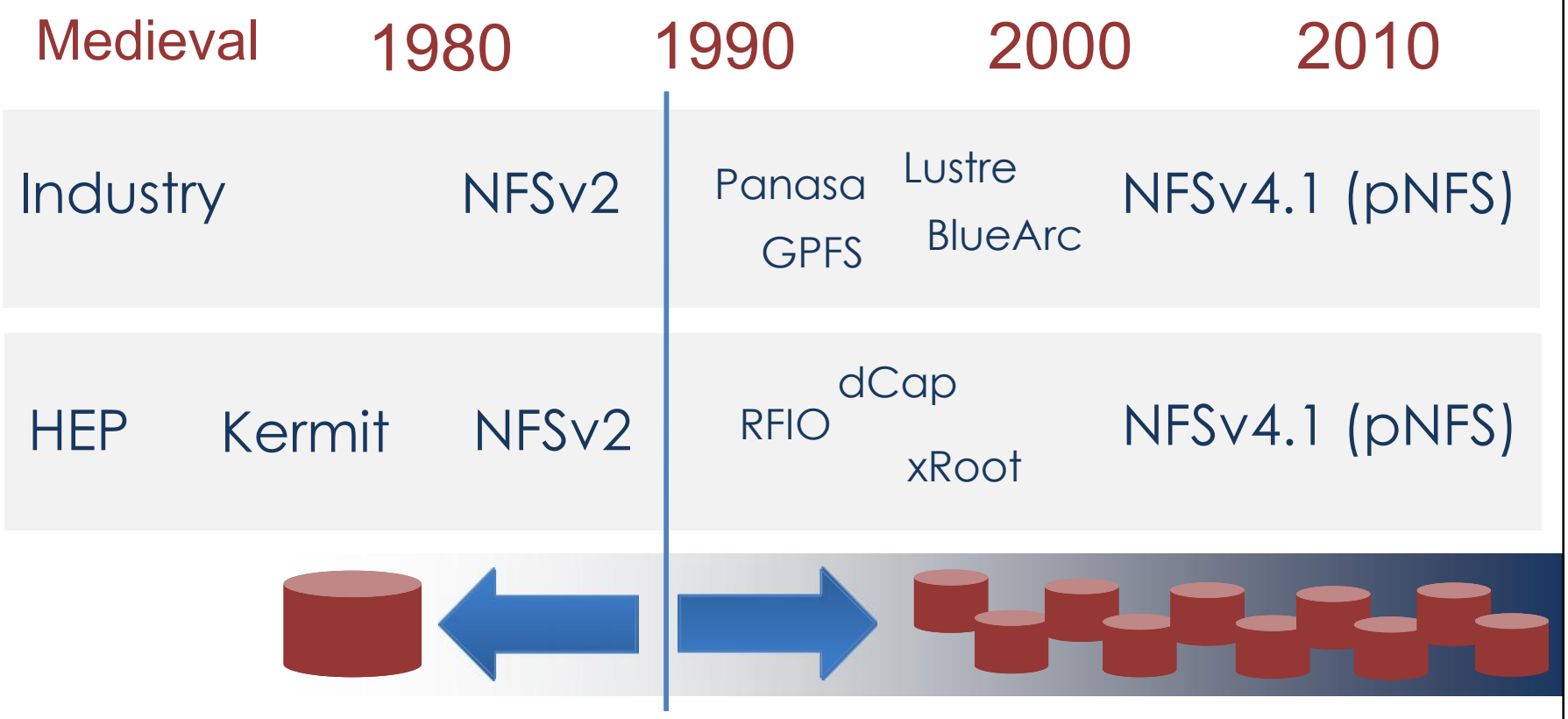
- Technical Background
  - Tigran Mkrtchyan, dCache.org, DESY (dCache, pNFS impl.)
- Evaluation results, gridLab, DESY
  - Yves Kemp, gridLab, DESY
  - Dmitri Ozerov, gridLab, DESY
  - Federica Legger, gridLab, University Münich
  - Sergey Kalinin, Uni Wuppertal
- Slides and more from
  - Brent Welch, Panasas, Inc.
  - Geoffrey Noer, Panasas, Inc.

Motivation

What's the issue ?

Where are we coming from

‘Local’ network data access



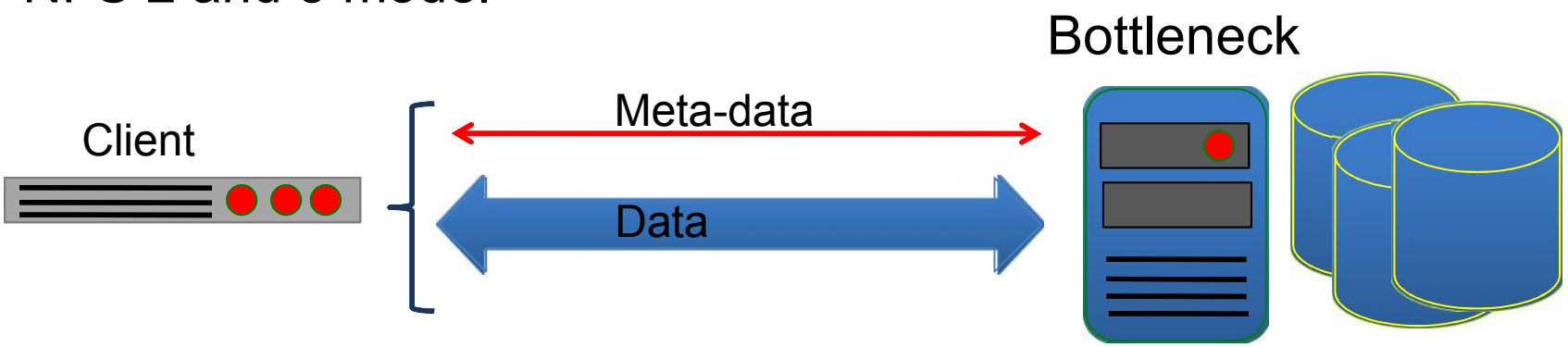
Motivation

Details

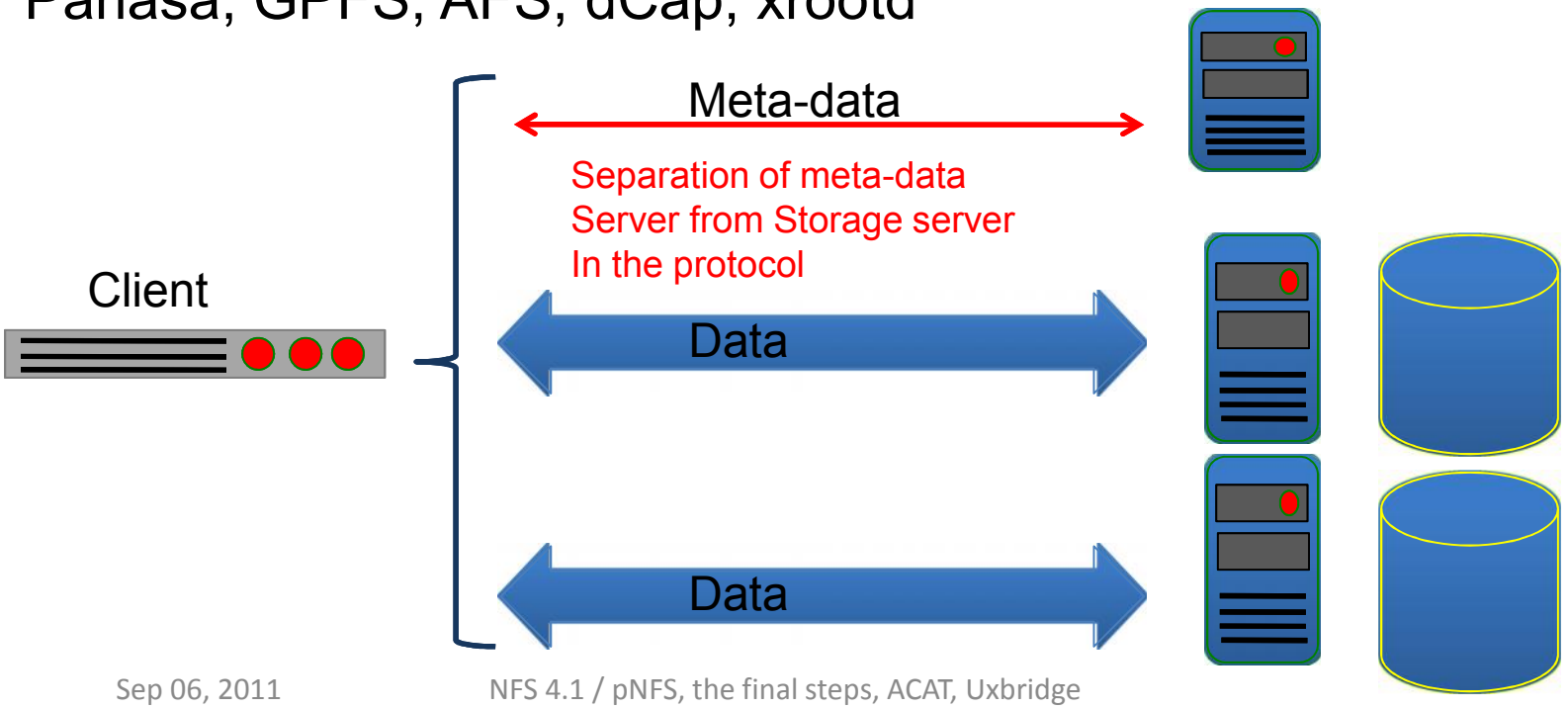
Some information we need to understand the rest.

Some more details on that

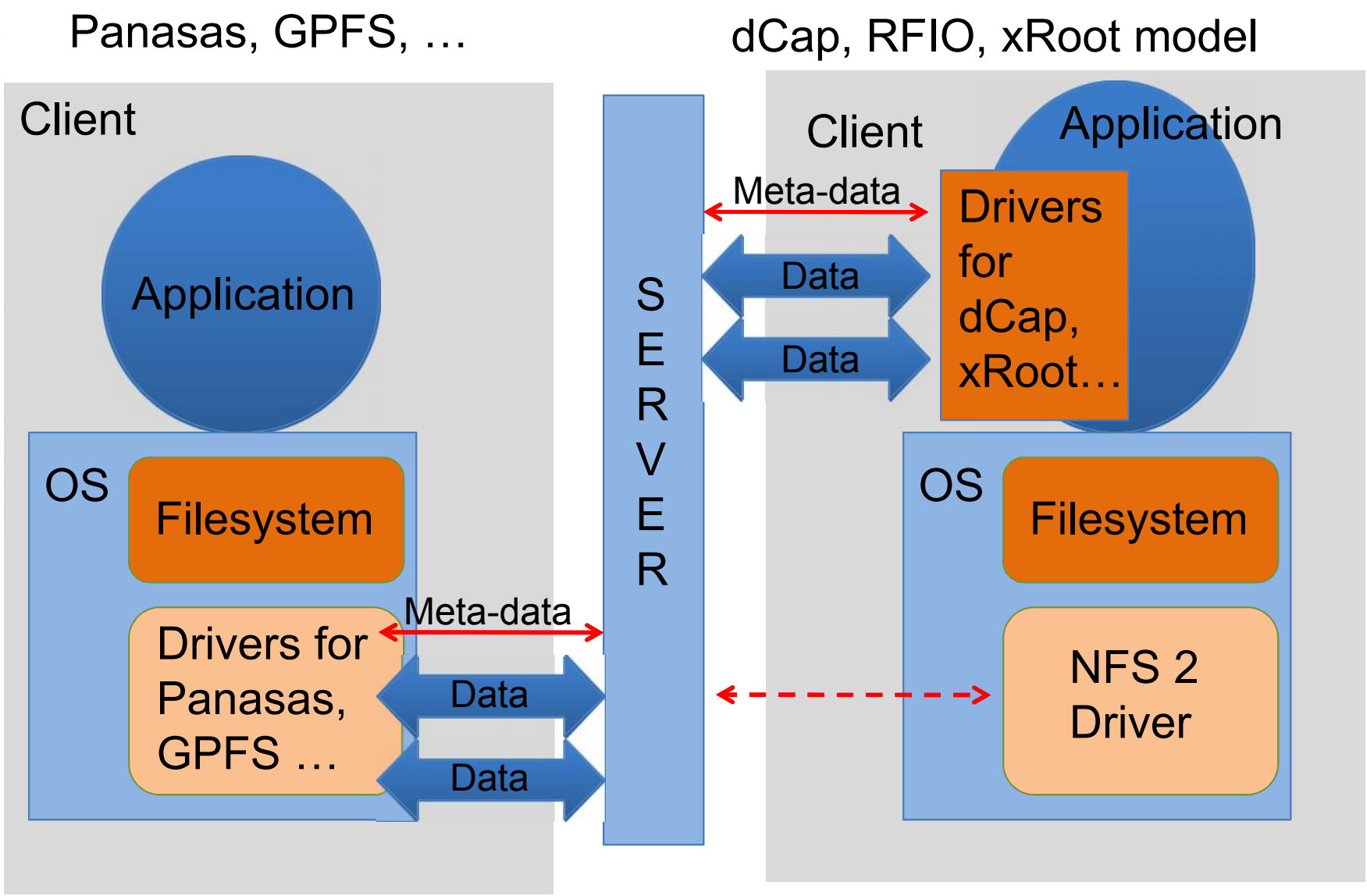
NFS 2 and 3 model



Panasa, GPFS, AFS, dCap, xrootd



And more …


Panasas, GPFS, …
dCap, RFIO, xRoot model
Client
Application
OS
Filesystem
Drivers for Panasas, GPFS …
Meta-data
Data
Data
SERVER
Client
Application
Meta-data
Drivers for dCap, xRoot…
Data
Data
OS
Filesystem
NFS 2 Driver

What's bad with that ?

- What's good with Lustre, GPFS, AFS, BlueArc, Panasas, xrootd, dCap..
  - Client is highly tuned to capabilities of the corresponding server.
- What's so bad with Lustre, GPFS, AFS, BlueArc, Panasas
  - You need to maintain one client kernel driver for each of them.
  - Keep track of all the different versions and dependencies.
  - You are stuck with a kernel version if vendor is late with updates.
  - Some vendors charge you for per client.
- What's so bad with xrootd, dCap, rfio ...
  - Not a mountable file system, you need to link a library to the application, which is not always possible.
  - You have to maintain all those client libraries.

How it works

History and status on one slide

Inevitable

What happened next

- Although proprietary solutions gave companies advantages over their competitors, customers started to suffer.
- A solution for the dilemma was needed.
- As a consequence : 2004 Garth Gibson, Brent Welch (Panasas) and Peter Corbett (NetApp) submitted first pNFS draft to IETF.
- Later CITI (UNI Michigan) coordinated the efforts and SUN, EMC, IBM and others joined. (dCache joined 2006 after I met PH in Sardina).
- Dec 2008 IETF approved internet draft
- Jan 2010 IETF approved pNFS with Objects and Blocks

- Two reference implementations exist. One Open Source (Linux) and at least one private.
- “We assume, all major vendors are working on their servers”

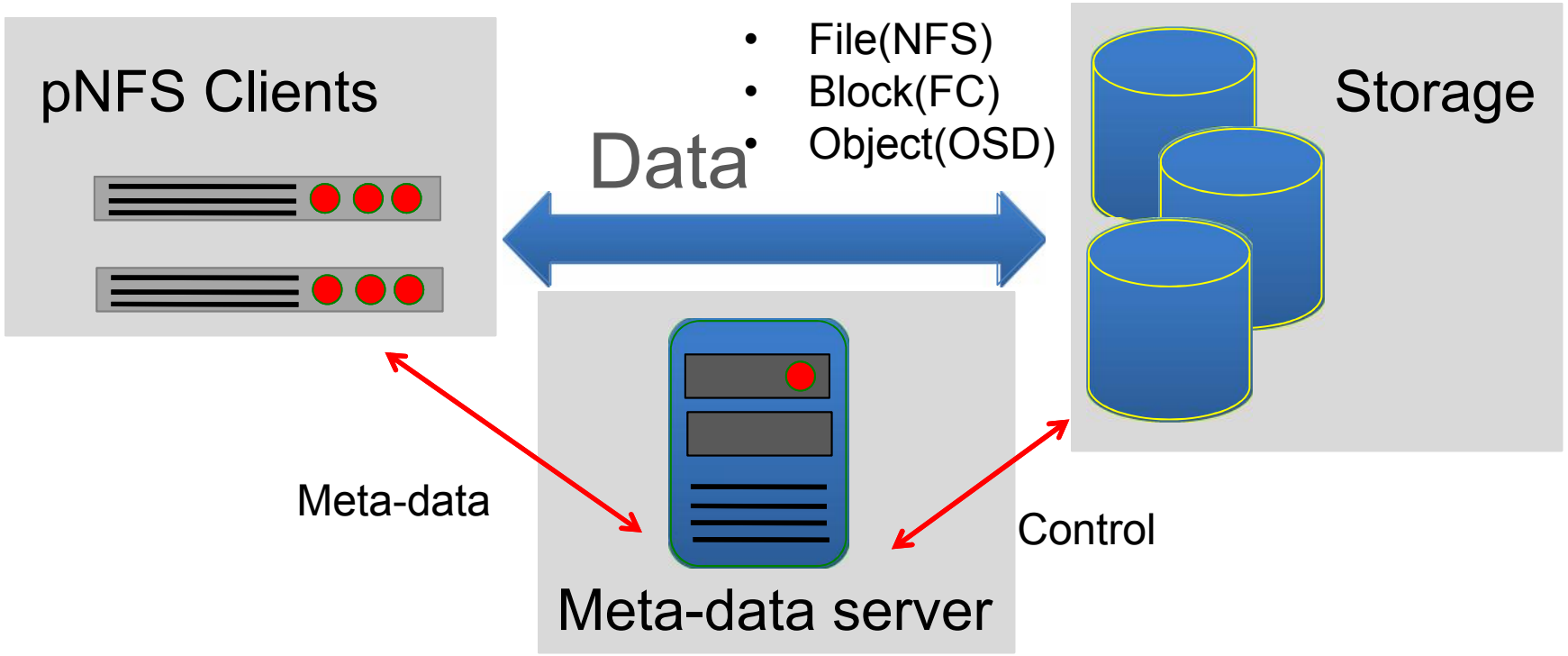
How it works

How it works

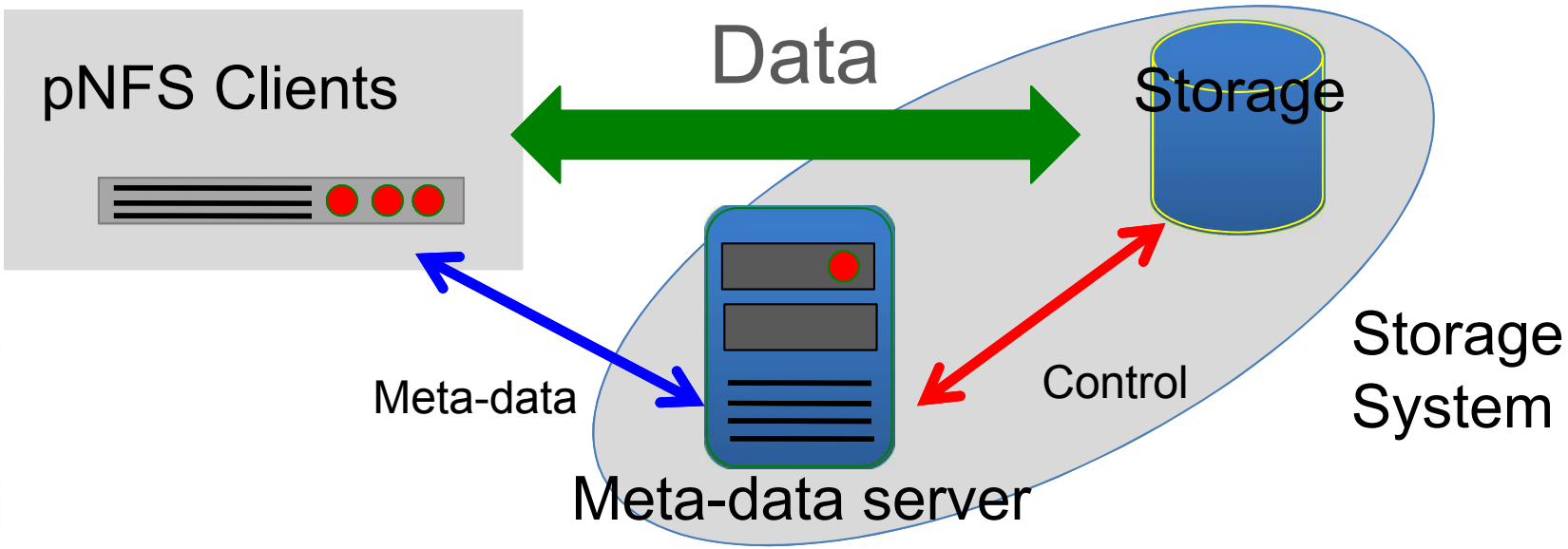
Take a deep breath

pNFS , how it works

- pNFS is an extension to the Network File System v4 protocol standard
- It allows for parallel and direct access
  - From Parallel Network File System clients
  - To Storage Devices over multiple storage protocols
  - Moves the NFS (metadata) server out of the data path.



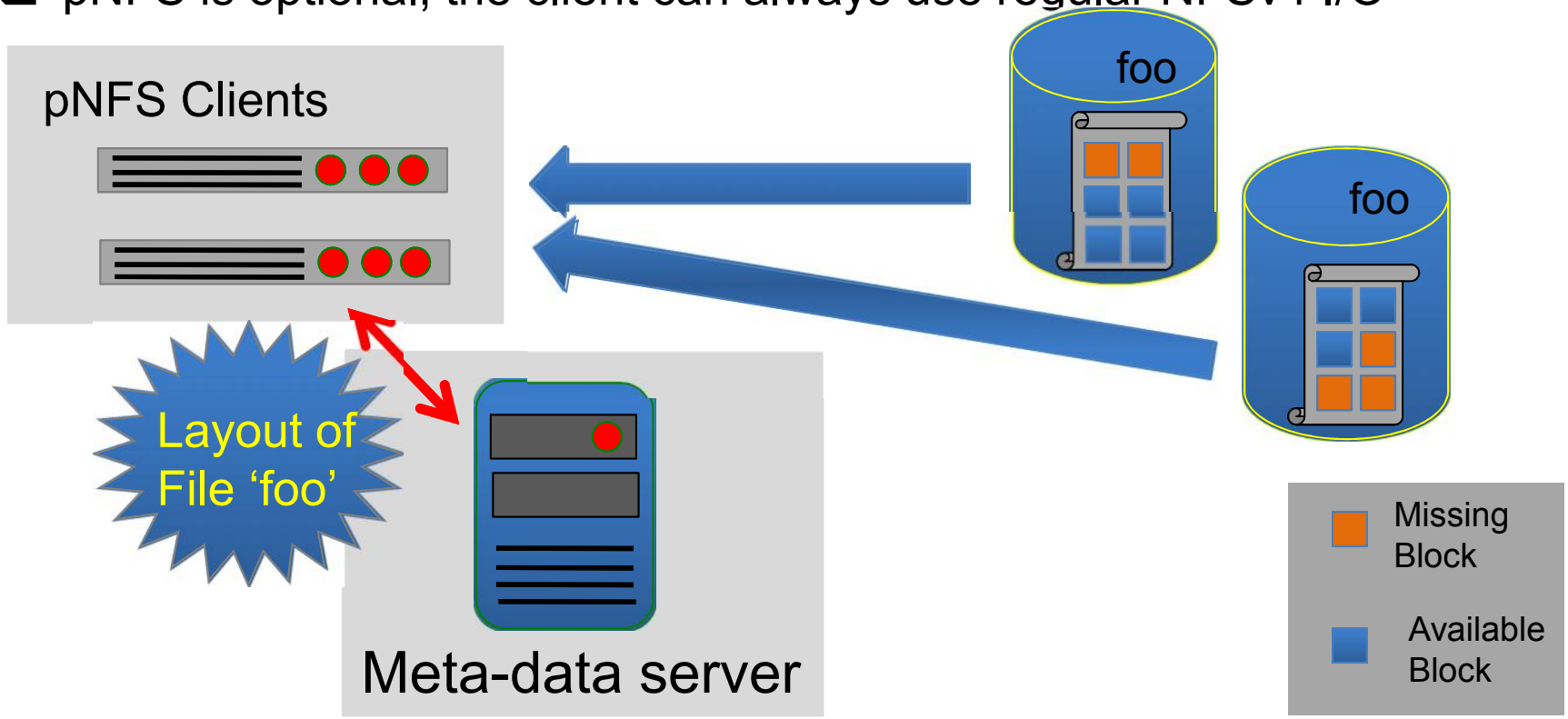
Where is the standard ?



- The pNFS standard defines the NFS 4.1 protocol extension between the (meta-data) server' and the client.
- The I/O protocol between client and storage is defined elsewhere, e.g.
  - SCSI **Block** commands over Fibre Channel
  - SCSI **Object** based storage (OSD) over iSCSI
  - Network **File** System (NFS)
- The control protocol between the server and storage is also specified elsewhere.

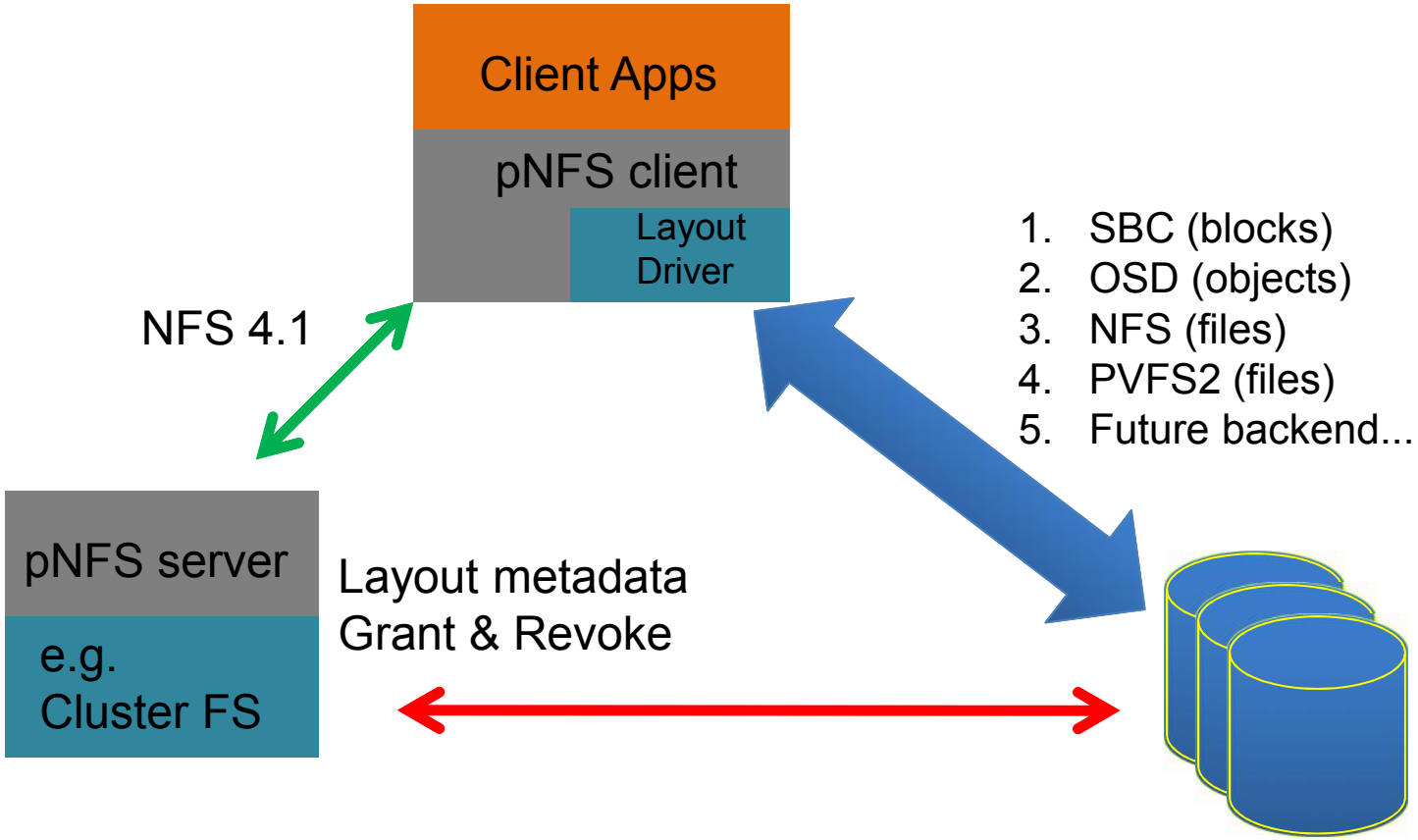
The pNFS layout

- Client gets a *layout from the NFS Server*
- The layout maps the file onto storage devices and addresses
- The client uses the layout to perform direct I/O to storage
- With the layout the client can decide which blocks of the file to fetch in parallel
- At any time the server can recall the layout
- Client commits changes and returns the layout when it's done
- pNFS is optional, the client can always use regular NFSv4 I/O



pNFS clients

- Common client for different storage back ends.
- Wider availability across operating systems.
- Fewer support issues for storage vendors.



Benefits

Benefits

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
- ✓ Security is part of the definition, not an add-on (GSS: Kerberos)
- ✓ Provides POSIXS ACL"s

Performance

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

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)

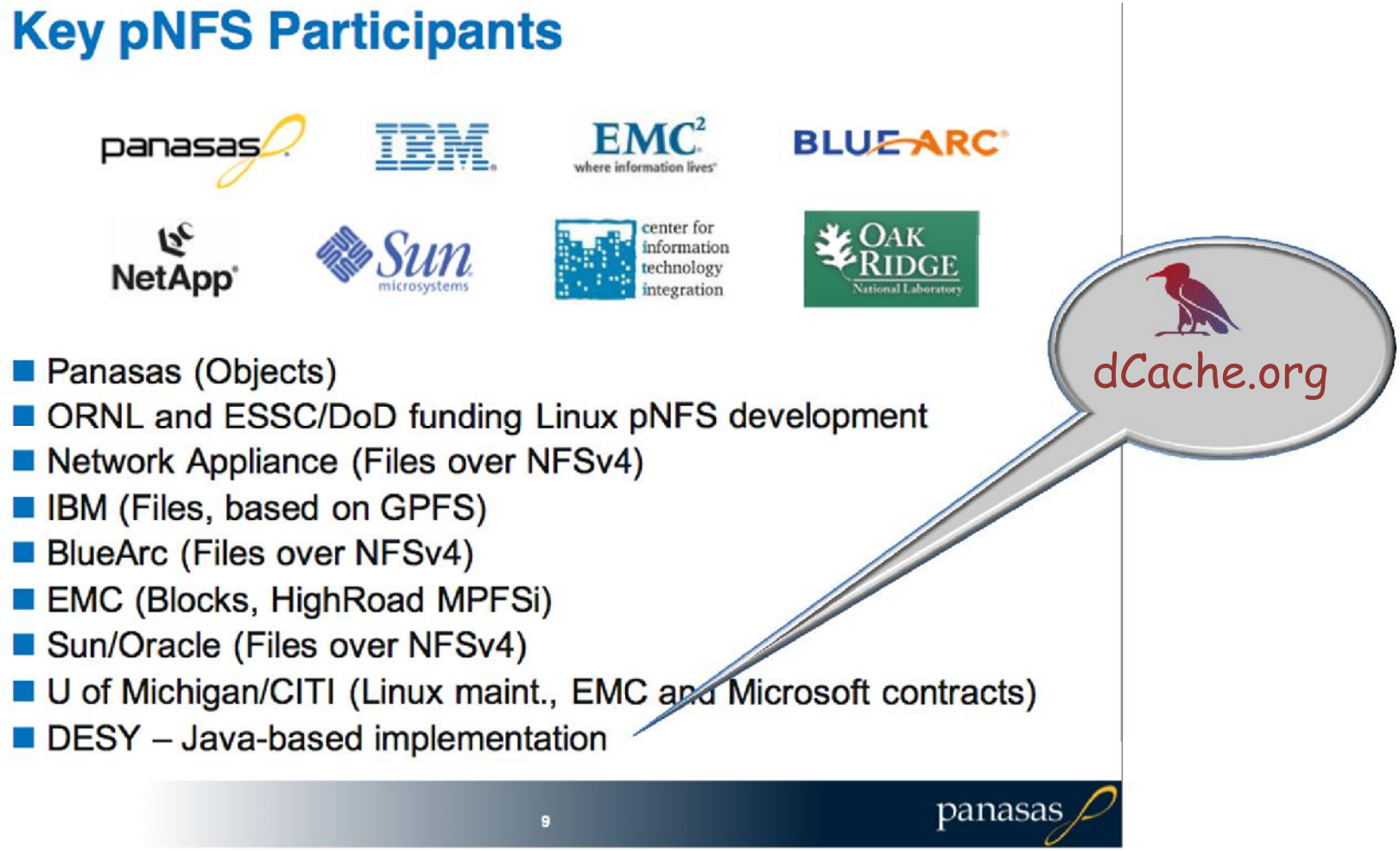
- ✓ Ensures Interoperability among vendor solutions
- ✓ Allows Choice of best-of-breed products
- ✓ Eliminates Risks of deploying proprietary technology

Involvement

Who is involved ?

Active Contribution by Industry

Stolen from
Brent Welch, Panasas, Inc, at the HPC Advisory Council, Lugano, Mar 2011



The European Middleware Initiative

European Middleware Inititative

EMI Factsheet

Budget : about 24 Million Euros over 3 years

Funding : about 50% by EU-FP7, rest by partners

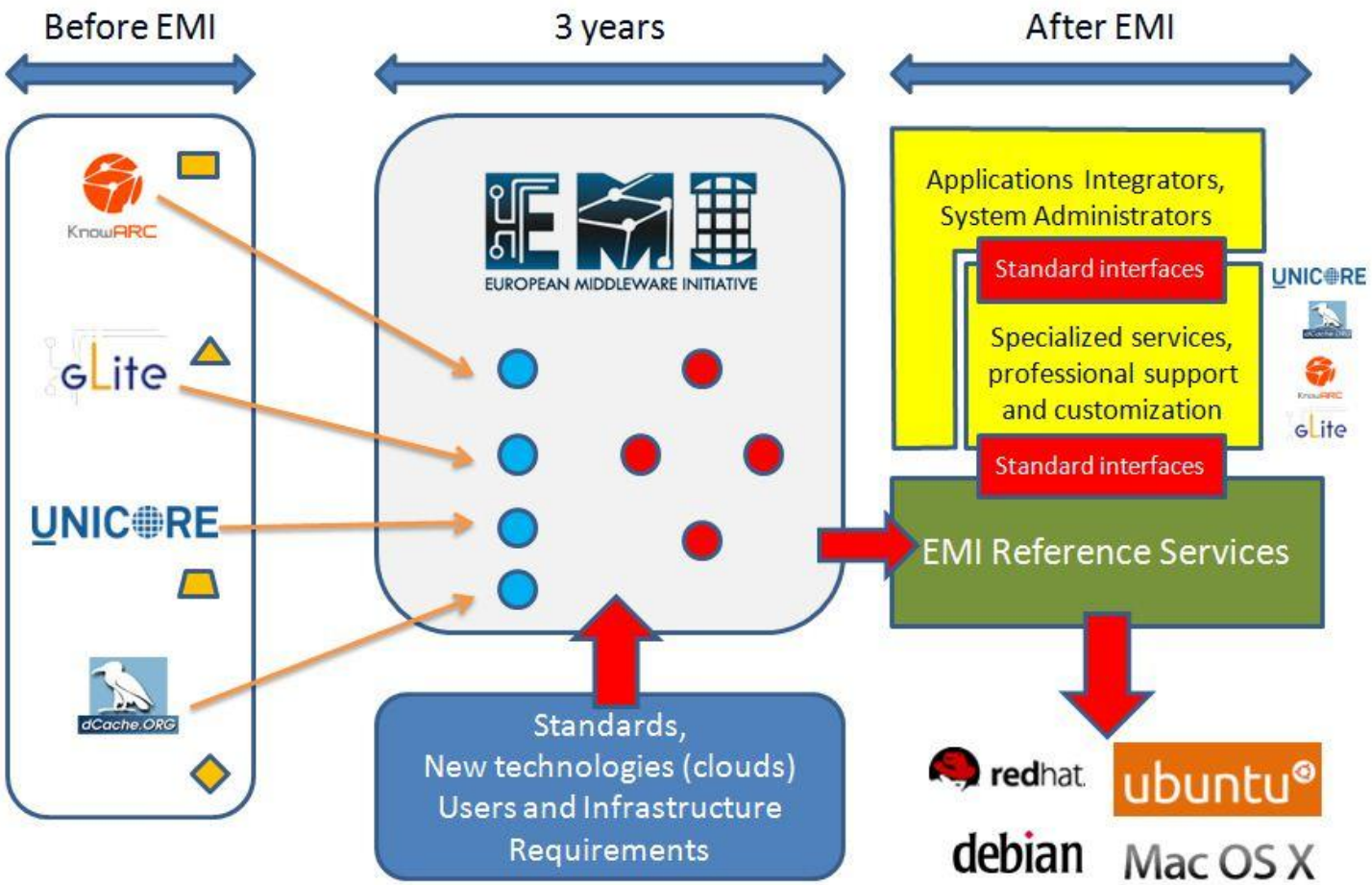
Covers : JRA, SA and NA

Partners : 22

Middlewares: Arc, gLite, UNICORE and dCache

EMI INFSO-RI-261611

European Middleware Initiative



EMI and standards

- Encouraged by the EC, **EMI is strictly committed to standards**.
- EMI supports 3 storage systems
  - DPM (CERN)
  - StoRM (INFN,CNAF)
  - dCache (DESY, NDGF, FERMIlab)
- EMI **is funding the support of standards** in all 3 SE's
  - http, https and WebDAV
  - **NFS 4.1 / pNFS**
  - SRM, Storage Resource Manager
  - Common Storage Accounting Record
  - Common Storage Delegation Service

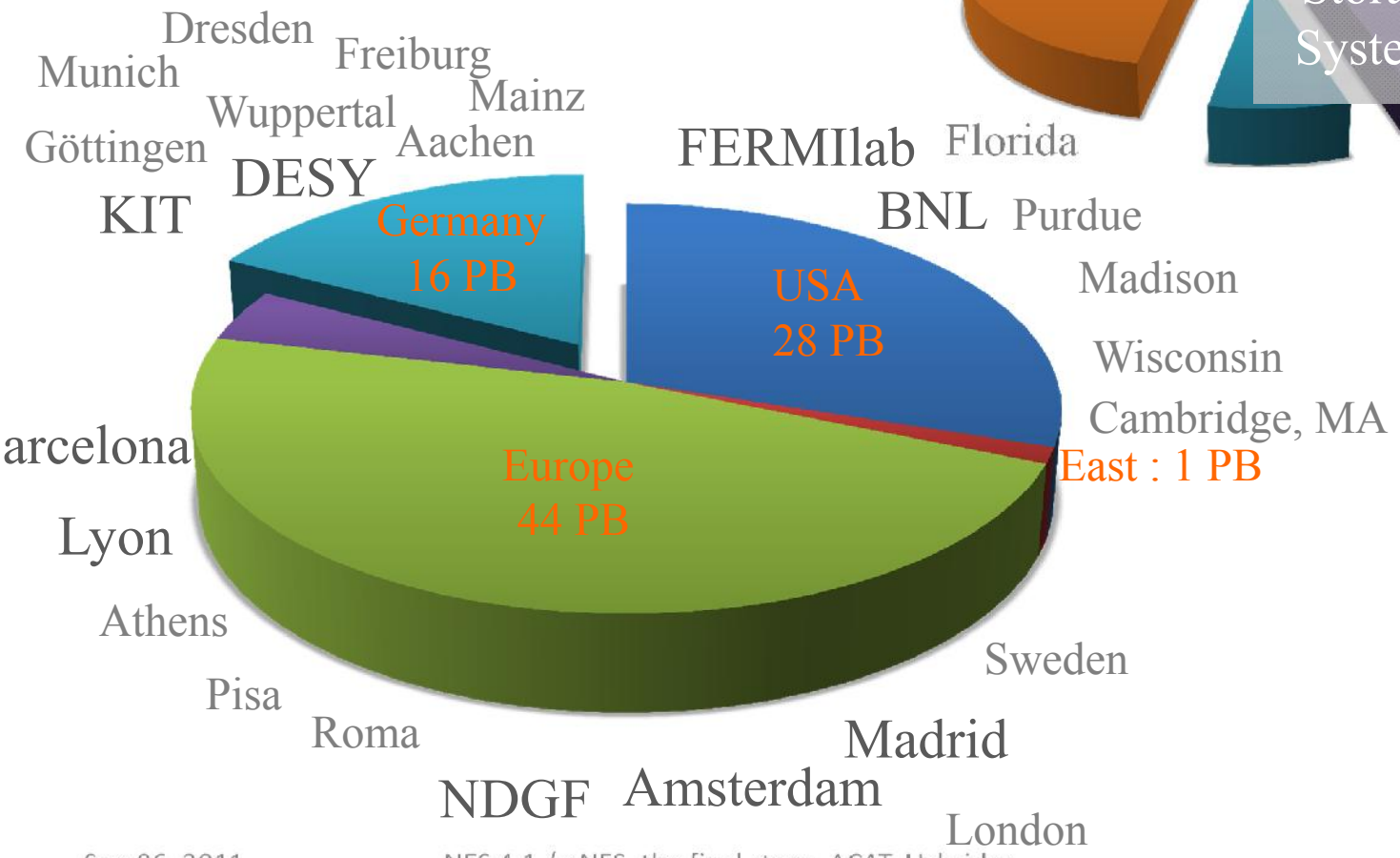
dCache.org

- dCache.org is a collaboration between
  - DESY (Headquarters)
  - The Nordic Data Grid Facility, NDGF
  - FERMIlab
- dCache.org provide the dCache storage element
- dCache is committed to standards
  - **First Storage System running NFS 4.1 / pNFS in production**
  - Http(s)
  - WebDAV
- Participates the regular pNFS Bakethons with all other pNFS vendors

dCache.org

dCache deployment

- 94 PB in total
- 7 Tier I's
- 40 Tier II's



WLCG Storage per SE type

EMI INFSO-RI-261611

Performance

performance

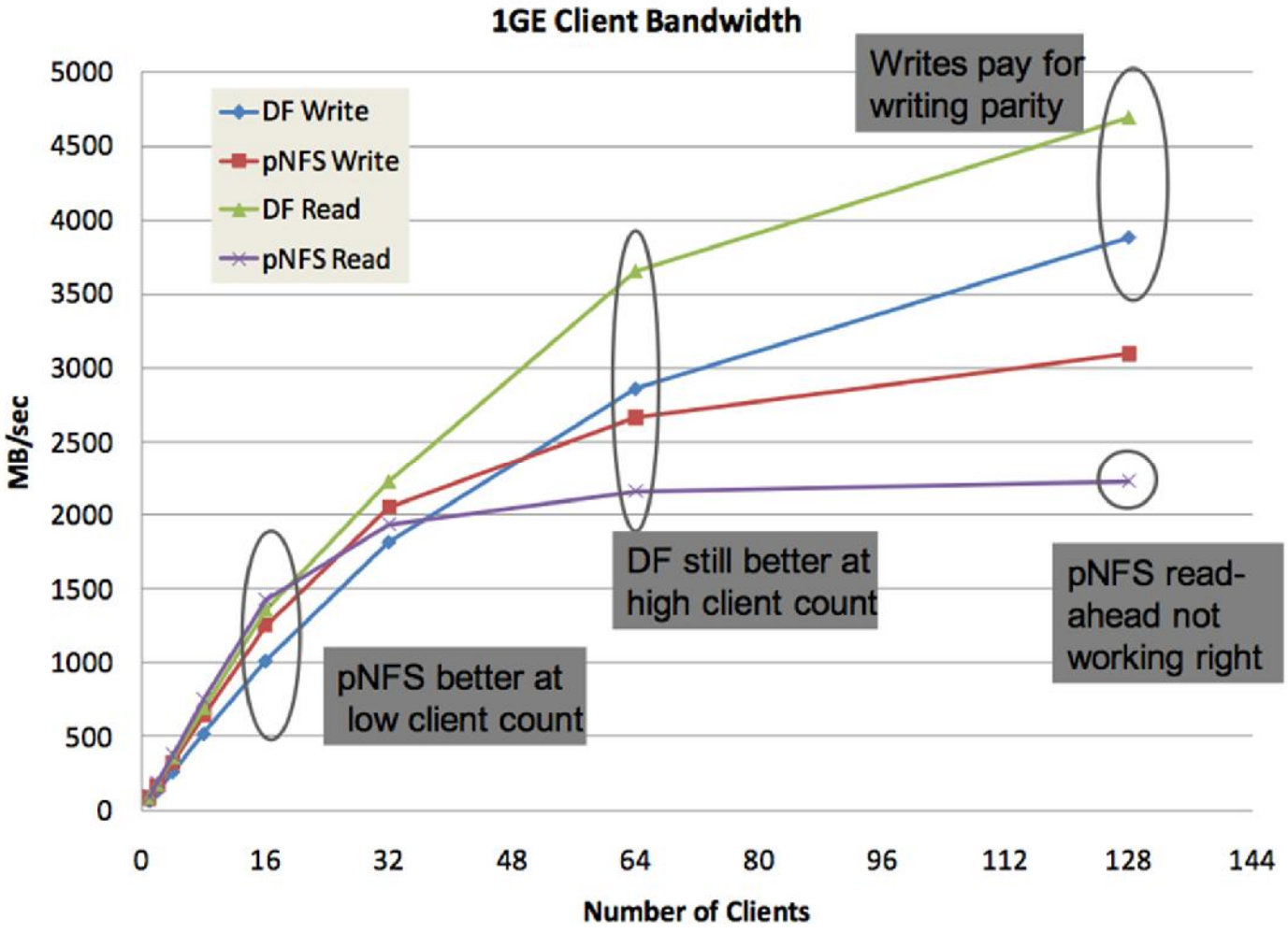
Performance

Panasas

- Iozone benchmark
- DirectFlow versus pNFS
- 1GE files
- Per-file Object RAID
  - Client writes data and parity in RAID-5 pattern
  - Feature of object-based pNFS layout

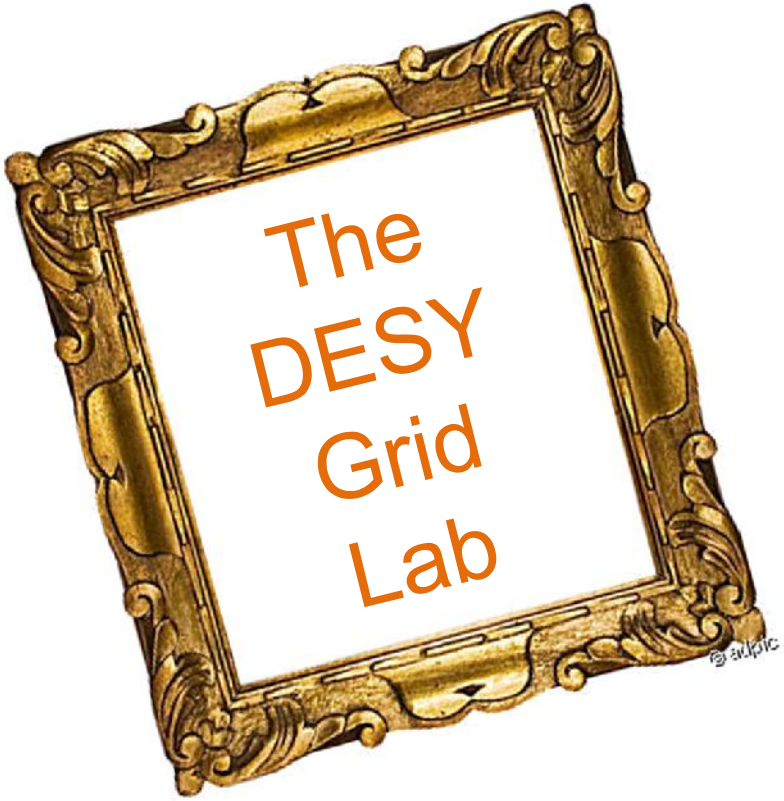
Panasas Performance

Stolen from
Brent Welch, Panasas, Inc, at the HPC Advisory Council, Lugano, Mar 2011



Performance

DESY / gridLab



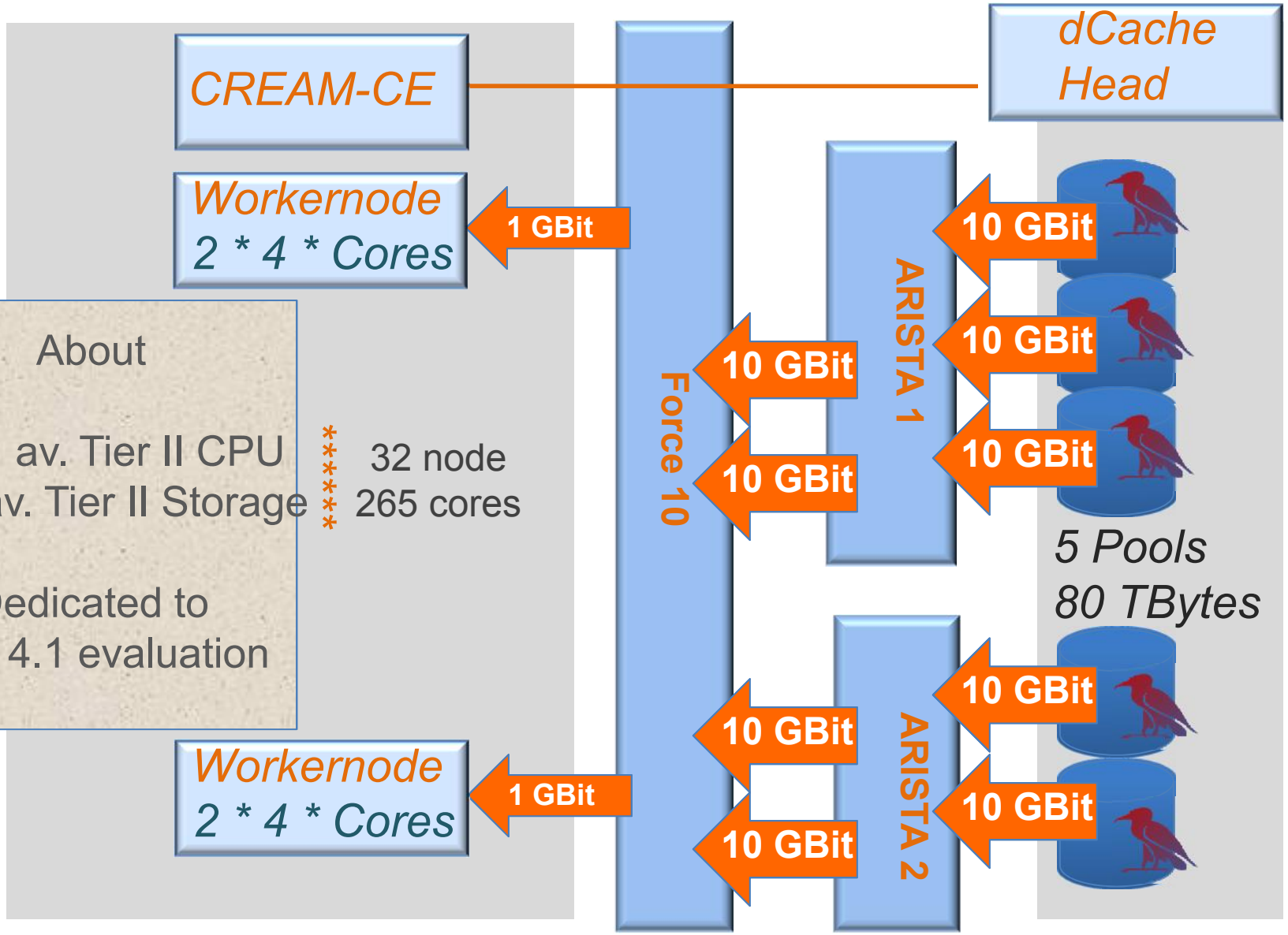
Operated by

Yves Kemp

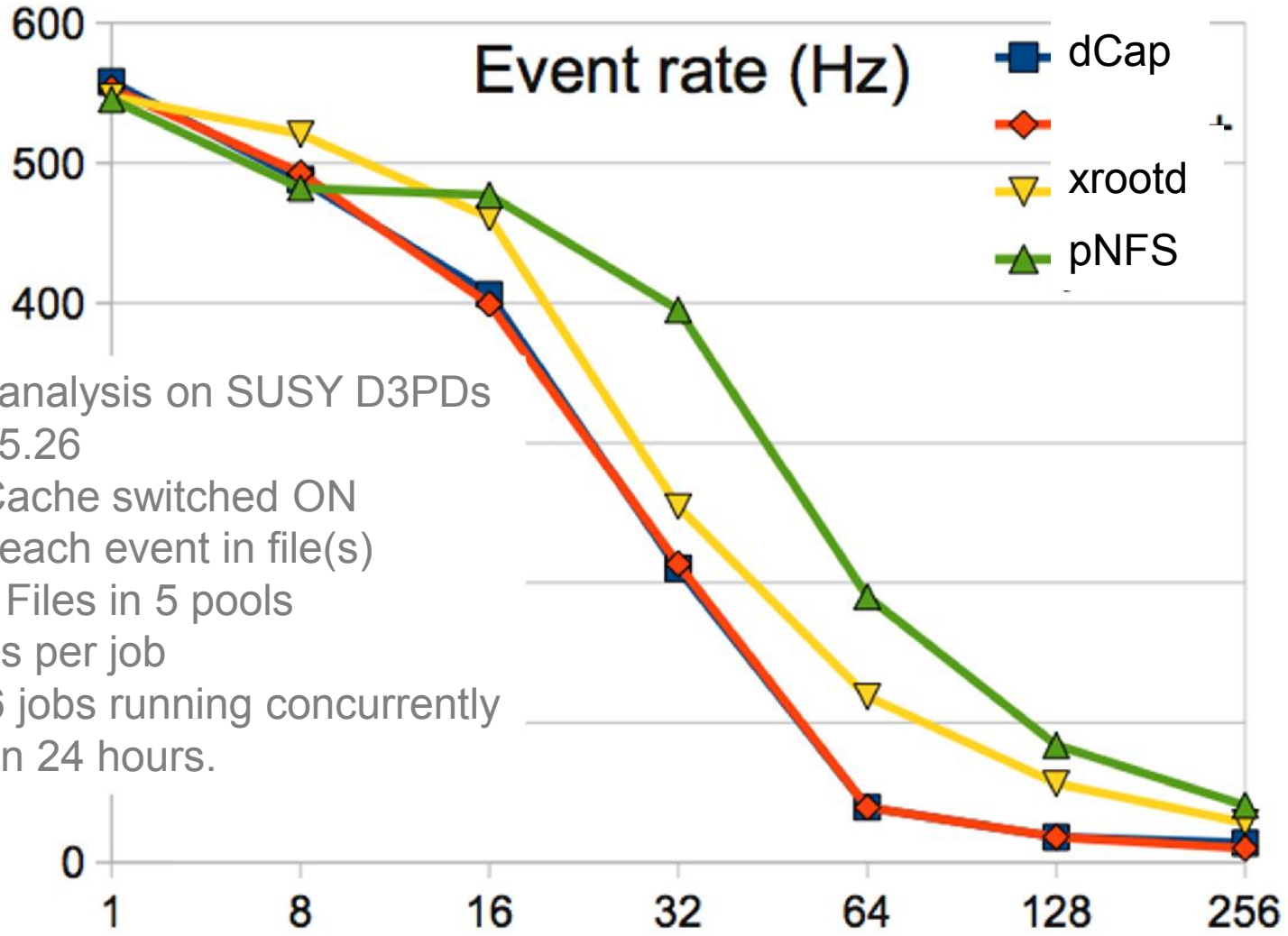
Dmitri Ozerov

But available for
Everyone who wants to
Evaluate pNFS with his/her
Application.

The DESY gridLab



ROOT analysis

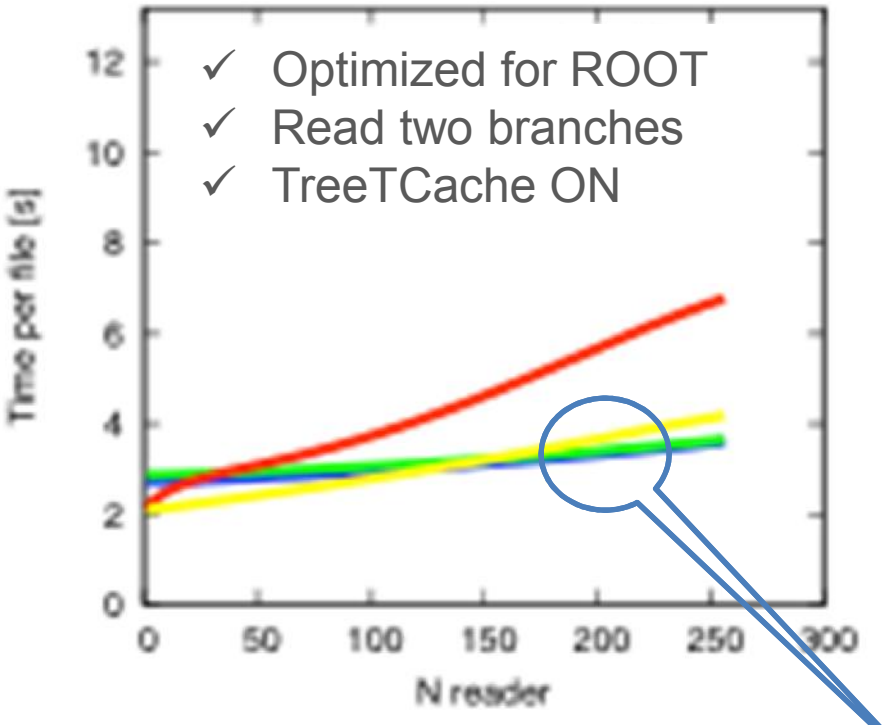


- ✓ROOT analysis on SUSY D3PDs
- ✓ROOT 5.26
- ✓TTreeCache switched ON
- ✓Reads each event in file(s)
- ✓25.000 Files in 5 pools
- ✓100 files per job
- ✓1 – 256 jobs running concurrently
- ✓Duration 24 hours.

Measurements done at DESY/gridLab by Federica Legger

pNFS bad

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

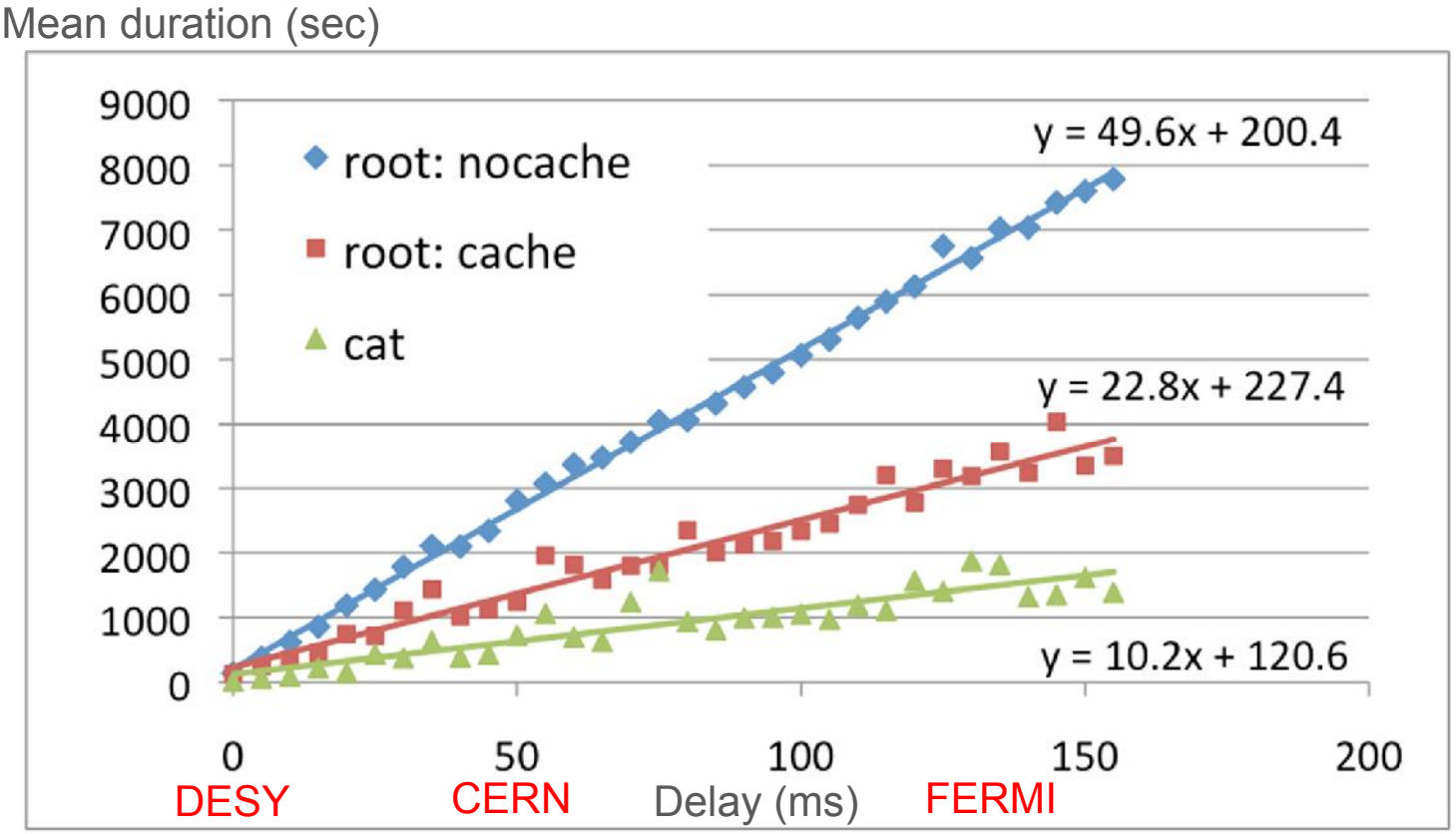


Xrootd/SLAC
Xrootd/dCache
NFS/pNFS
dCap (old vers)

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

Wide area transfers (simulation)

Simulation of wide area transfers with
- ✓constant latencies
- ✓no packet losses.



Measurements done at DESY/gridLab by Yves Kemp

Availability

Availability

- Industry vendor solutions
  - Vendors are still careful. Nobody wants to be the first.
  - NetApp promised something for end of this year (already two times postponed)
  - IBM likely pNFS on GPFS end of 2012
  - BlueArc about beginning of next year.
  - ...
- EMI server
  - DPM in beta
  - StoRM with availability in GPFS
  - dCache : production
- Clients (Linux)
  - With kernel 2.6.39
  - Fedora 16
  - Expected in RH 6.2

Some last words

- pNFS significantly simplifies the current protocol zoo by providing a
  - authenticated, authorized,
  - Parallel and
  - Highly scalable **standard** way of accessing data.
- Proprietary protocols clearly have their advantages, none of which prevails having a common high performance data access standard.
- Future (by Geoffrey Noer, Panasas) “pNFS will be in production use in 2012, fully supported by major Linux distributions, by Panasas and other leading storage vendors”
- Science is well prepared with EMI-Data supporting pNFS, with DPM and dCache.
- A first pNFS system is in production at DESY for the Photon Science community.

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

Thank you

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