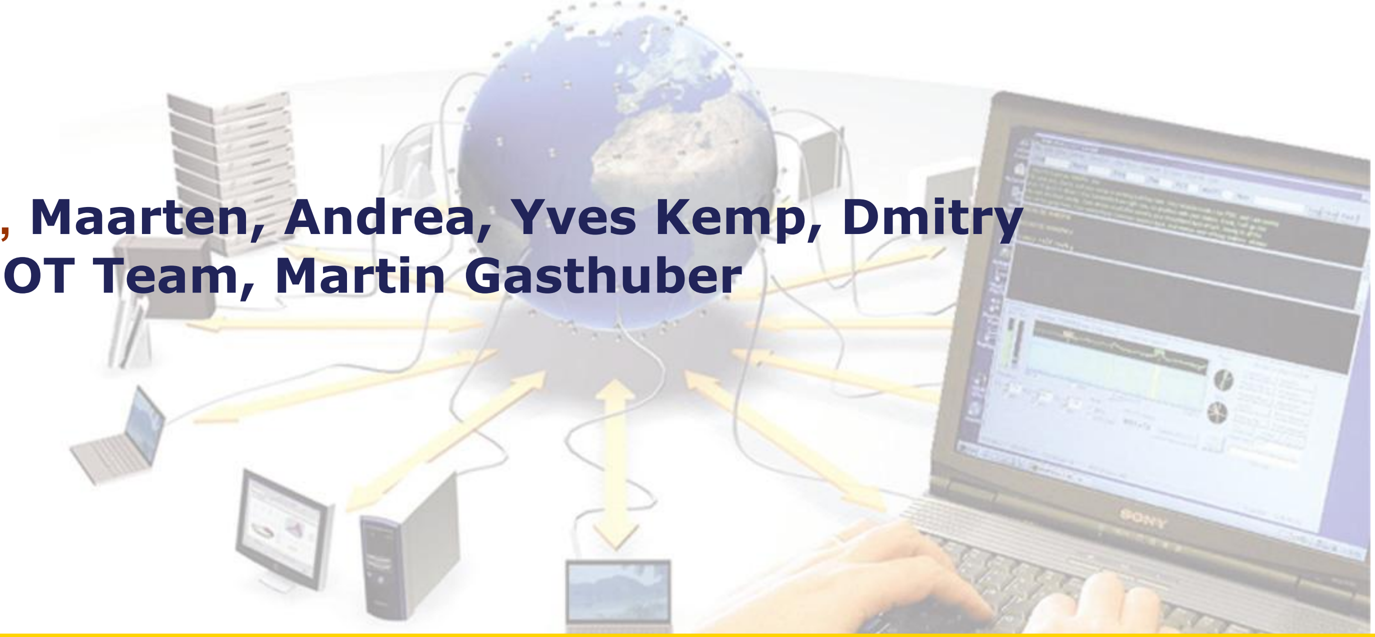
NFS 4.1 demonstrator

Jean-Philippe Baud, IT-GT, CERN
Patrick Fuhrmann, DESY

July 2010



Tigran, Ricardo, **Maarten, Andrea, Yves Kemp, Dmitry Ozerov, desy DOT Team, Martin Gasthuber**

Quick reminder on why we are doing this !

The next chapter is actually the summary of Gerd’s presentation available from dCache.org

Problem: data access

- Proprietary protocols zoo: dependent on Storage Element instance
  - Client libraries conflict with user application libraries
  - Some of the access libraries lack authentication
  - Posix-like, not Posix
  - Need to modify the application code
  - Need to re-invent the wheel when implementing caching
  - Need to install special software on all nodes: UIs, WNs, Metadata servers and disk nodes
  - Problem of portability when going from one OS to another or even from one OS flavor/version to another
  - Publishing the protocols in the Information System and prioritizing the protocols is not easy

NFS 4.1

- Is an IETF standard
- Transparent to applications:
  - True Posix interface, no code change needed
  - No conflicting client libraries
- Strong security: Kerberos 5, (X509 coming), ACLs
- Data flows directly between disk server and client machine
- Common client for different storage back-ends
- Fewer support issues as almost no in-house development
- Well adopted by
  - Storage providers: EMC, IBM, dCache, (DPM)
  - OS providers: Linux, OpenSolaris, Windows
- Can run on existing setups: no data or metadata migration, keep operational knowledge
- No SRM for user data access

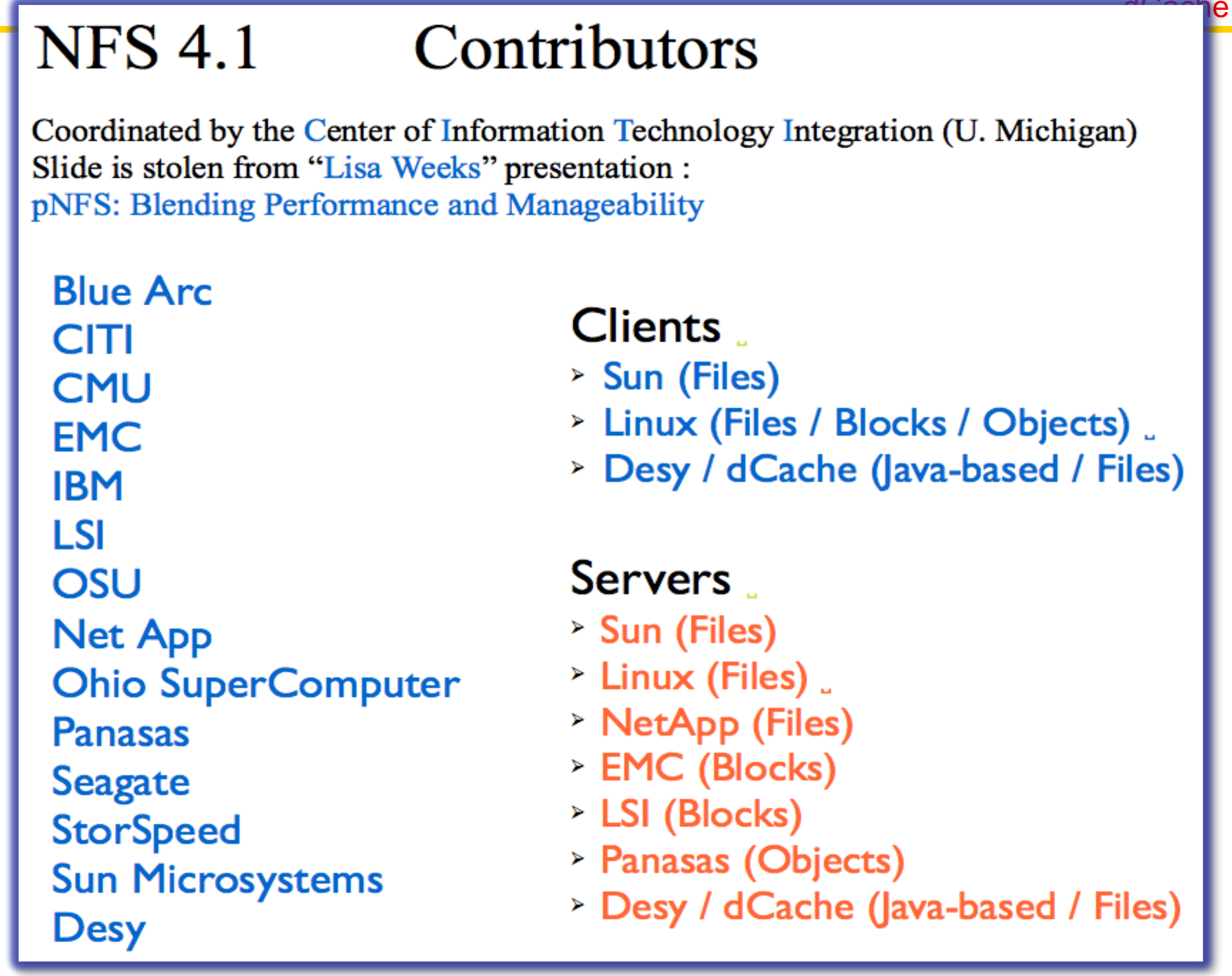
NFS 4.1 installation

- UIs and WNs:
  - kernel coming with standard OS distribution
  - User space daemon for authentication (Kerberos 5 supported off the shelf)
- Disk servers
  - kernel coming with standard OS distribution
- Metadata server
  - User space daemon for handling the name space.
  - This can interact with existing implementations like Chimera or DPNS.
  - This is the only glue to be developed

Availability of NFS4.1 clients

- NFS 4.1 and the linux kernel
  - NFS 4 already in SL5
  - NFS 4.1 in 2.6.32
  - NFS 4.1 plus pNFS in 2.6.33/34
- Kernel 2.6.34 will be in Fedora 13 and RH6 Enterprise (summer)
- NFS 4.1 (pNFS) Kernel available in Fedora 12 (NOW)
- Windows Client expected 4Q10.
- DESY grid-lab is testing with :
  - SL5 and 2.6.33 kernel plus some special RPM. (mount tools)
  - See dCache.org wiki for further information

NFS4.1 industry contributers



NFS 4.1 Contributors

Coordinated by the Center of Information Technology Integration (U. Michigan)
Slide is stolen from “Lisa Weeks” presentation :
pNFS: Blending Performance and Manageability

Blue Arc
CITI
CMU
EMC
IBM
LSI
OSU
Net App
Ohio SuperComputer
Panasas
Seagate
StorSpeed
Sun Microsystems
Desy

Clients

- Sun (Files)
- Linux (Files / Blocks / Objects)
- Desy / dCache (Java-based / Files)

Servers

- Sun (Files)
- Linux (Files)
- NetApp (Files)
- EMC (Blocks)
- LSI (Blocks)
- Panasas (Objects)
- Desy / dCache (Java-based / Files)

NFS 4.1 Demonstrators

dCache.org

Sites

CERN, DESY

Funding through

CERN, WLCG, DESY, dCache.org, EMI

People

Jean-Philippe, Tigran, Patrick, Maarten, Andrea, Ricardo, Yves Kemp, Dmitry Ozerov, Desy DOT Team, Martin Gasthuber

Demonstrator Goals

Goal

We will demonstrate that an industry standard protocol is as useful for WLCG analysis as any other proprietary protocol.

The key point in data access performance is the client side caching and not the protocol used to communicate to the data servers.

Milestones

Phase 1 : Mid of August

Hardware and 'test suite' setup

Phase 2: Chep 10

First performance presentations with 'real' analysis applications and performance matrix.

Phase 3&4 : Beyond CHEP 10

Extends testing beyond 'demonstrator' partners. E.g. HEPIX storage WG.

Phase 1: starting now

Compose a sustainable evaluation test suite, with input from

- ROOT people. (Rene B.)
- Altas Hammer Cloud. (Johannes Elmsheuser)
- Realistic CMS and Altas analysis jobs (N.A.)

Our requirements

- All ‘data access’ demonstrators need to use the same ‘test suite’.
- Changes in the ‘test suites’ need to be communicated.
- Collaboration between the ‘demonstrators’ is essential. Otherwise we, once again, ending up proving nothing.

Phase 1

- Install NFS 4.1 on top of a standard file-system
  - See how easy or difficult it is to install, configure, tune
  - What is the overhead introduced by NFS4.1
  - Investigate how security works
  - Look at performance
- We will have one setup at DESY and one setup at CERN
- We will run tests locally as well as remotely from the 2 sites
- Rene Brun (ROOT) will be involved from the beginning to make sure that the tests are real Physics use cases
- Compare performance results with existing protocols

Phase 2

- Install NFS 4.1 on top of dCache at DESY
  - Look at possible installation/configuration problems
  - Look at performance (local access and remote access)
  - Compare with results of Phase 1 using the same test suites
- Install NFS 4.1 (prototype) on top of DPM
  - Look at possible installation/configuration problems

Milestone : CHEP 10

- DPM
  - Prototype
  - Test results on functionality.
  - Test system at CERN (details later slide)
- dCache
  - Production system
  - Performance evaluation, (comparison with other protocols)
  - Test system at DESY (details later slides)
- COMMON
  - Wide area functionality between CERN/DESY (DPM,dCache)
  - Security : Unix, Kerberos
  - Share test suites and test installations
  - Presentations at CHEP 10 from DPM and dCache (see agenda)

Phase 3: until end of year

- Run the same test suite on DPM as on dCache to investigate
  - Installation, configuration, performance problems
- Run jobs accessing data residing in dCache and DPM cross sites

Phase 4: January-June 2011

- Run production readiness (reliability and scalability) tests using.
  - Setup at DESY
  - Setup at CERN
  - HEPiX setup at KIT in collaboration with Andrei
  - + any site interested in the testing
- Investigate the possibility to have a global filesystem using a dCache instance at one site and a DPM instance at another site.
- If StoRM provides an NFS 4.1 interface, we propose to involve them as well.
- Investigate the use of ZFS on OpenSolaris disk servers.
- Prove the transparency with existing infrastructures.

Hardware configuration @ CERN

At CERN (Available mid of August)

- Workernodes
  - 10 batch nodes
- Data servers
  - 5 disk servers (200 TB)

Hardware configuration @ DESY

Available mid of August at DESY

- Worker Nodes
  - 2 * 16 Blades a 8 cores
  - 1 GB Ether each
  - SL5 with NFS 4.1 enabled kernel
  - Cream CE with PBS
  - VO setup as provided by DESY-HH production grid
  - Available on short notice (some days)
- DATA server
  - 5 * R510 DELL with 24 Tbytes raw each
  - Results in 60 – 100 Tbytes depending on RAID setup
  - 10 GB Ether each
  - 2 Headnodes
  - Available within 25 days

Running regular Experiment Analysis Jobs

NFS 4.1 is cool

People from all areas

(Sites, experiments ...)

are more than invited to join.

BTW : dCache.org is hiring