NFS 4.1 / pNFS activities and progress in EMI-Data

by

Patrick Fuhrmann

With contributions by

Ricardo Rocha
Andrea Sciabà
Credits/Partners

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 Sciabà (CERN)
Hartmut Stadie (DESY, CMS)

The contribution of the EU (EMI) is rather moderate but the return is great.
Outline

✓ Why is pNFS interesting?
✓ What is the status and the timeline for pNFS deployment?
  ✓ Availability of the different components
  ✓ Protocol verification and performance evaluation
✓ Issues to be considered
  ✓ Security
  ✓ Wide area transfers
✓ Some results from the dCache NFS 4.1 / pNFS evaluation at Grid-Lab
✓ Status of DPM NFS 4.1
✓ Conclusions
Why is pNFS interesting

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

(Stolen from : http://www.pnfs.com/ )
Benefits for EMI Customers

Simplicity

✓ Regular mount point and real POSIX I/O
✓ Can be used by unmodified applications
✓ 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. compound requests
### Availability for production use

<table>
<thead>
<tr>
<th>Year</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Industry**
  - NetApp
  - Blue Arc

- **dCache**
  - 1.9.8 - 1.9.11
  - Next Golden Release (1.9.12)

- **DPM**
  - Beta Read-only
  - Beta Read/Write

- **StoRM**
  - GPFS native
  - GPFS pNFS

- **pNFS**
  - Enabled Kernel
  - pNFS in official 2.6.38 kernel

- **DESY**
  - Linux distributions (RH6.2…)

---

**Here we are facing a problem**

**Delayed due to man power issues.**

- NetAppTest DESY
- Production (OnTap 8.1)

**See next slide**
pNFS client support in SL5/6

- Full NFS 4.1/pNFS client **available in 2.6.38**
- RH already back-ported pNFS into 2.6.32 ... but only available for business partners right now
- **Official back port into RH6.x** (kernel rh-2.6.32) expected with RH6.2, maybe earlier
- Based on RH6.x it will be made in SL6.x

Right now it seems that there is a problem with 2.6.38. Don’t know yet if that can be fixed.
pNFS EMI server support

• dCache
  • Full implementation available starting from 1.9.11
  • Plain and Kerberos authentication

• DPM
  • Is ready to testing.
  • Plain and Kerberos authentication

• StoRM
  • Will become available with GPFS pNFS availability
Non-EMI server implementations

• GFS2
  • Available with 2.6.38 (now)
  • Uses pNFS for reading and only NFS 4.1 for writing

• NetApp
  • Available with ONTAP® 8.1 (Now’ish)
  • DESY waiting for a test system

• Others
  • IBM, BlueArc, EMC … are all members of the relevant CITI projects and have server software ready but seem to be reluctant to provide products just yet
Open issues
**Issues to be considered (Security)**

- All **tests** (up to now) done **w/o strong authentication**
- **NFS4.1/pNFS clients** are only available with Kerberos **authentication**, as this is the industry standard
- How about **X509 certificates** (proxies)
  - **Theoretically possible** as NFS4.1 RFC only refers to the GSSAPI
  - However kernel developers are only interested in Kerberos
  - There is an **existing solution** : Creating a Kerberos token from a **X509 certificate** (modified voms-proxy-init talking to kdc)
  - Other solutions would need closer contact with kernel developers
Issues to be considered (Wide Area)

- On wide area performance:
  - No tests done yet
  - Planned between DESY and CERN after CERN test system will become available
Possible wide area NFS 4.1 deployment

- Central configuration of Referral Hosts through FedFS
- Possible parallel trees can be built
DESY Grid Lab

✓ Since mid of last year, DESY provides a **Tier 2 testbed** with dCache/pNFS server and pNFS-enabled SL5 worker nodes

✓ **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 evaluations**

✓ Results have been presented at CHEP’10 and at 2010 Spring HEPIX
The DESY pNFS “Tier 2”

CREAM-CE

Worker node
2 * 4 cores

32 nodes
256 cores

ARISTA

Force 10

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

10 GBit

About

50 % av. Tier 2 CPU
20 % av. Tier 2 Storage

Dedicated to
NFS 4.1 evaluation

Worker node
2 * 4 cores

dCache Head

5 Pools
80 TBytes

Apr 11, 2011

EGI UF 2011, NFS 4.1/pNFS

15
Performance limitations

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

Total throughput does not depend on the protocol
ROOT I/O Framework

File://
Fadvice=async

Application
ROOT

Block Read

TTreeCache
Event Based Caching

Vector Read

Async=no
T-DCACHE

Async=yes
T-XNET

File system cache

Simple read
SMART Block caching

Vector read

dCap-Client

xrootd-Client

NFS 4.1

dCap

xRoot

dCache

SLAC

Apr 11, 2011
EGI UF 2011, NFS 4.1/pNFS
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 implementations behave the same.
xRoot / NFS 4.1 in dCache (II)

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
DPM NFS 4.1 – status

• First full feature prototype in use internally
  – Integrated into an internal DPM testbed at CERN
  – Mounted on a client machine running a patched 2.6.37 kernel
  – Standard functional tests from the NFS Connectathon running daily against this setup

• Features
  – Namespace management (mkdir, ls, rm, stat, …)
  – Permissions (with KRB authentication)
  – Space usage reports (du, df, …)

• Ongoing
  – Extend tests to real world use cases from experiments (like CMS!)
  – Performance improvements in the DPM to cope with NFS access patterns (heavy on metadata queries)
Conclusions (I)

Protocol verification

- For new protocols, it is not clear *per se* that client and server developers understand all the details
- But dCache and DPM people keep in touch with the experts:
  - dCache.org is a *member of the CITI group* which is *coordinating the NFS 4.1 efforts*
  - Compatibility regularly verified for both dCache and DPM
Conclusions (II)

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 extensive testing, the protocol seems to be well understood by dCache and the Linux pNFS kernel developers
- The performance already exceeds expectations
- Performance tests in DPM to be performed using real applications
References

Center for Technology Integration
http://www.citi.umich.edu/

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

PNFS
http://www.pnfs.com/

RFC 5661

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

EMI, The European Grid Infrastructure
http://www.egi.eu

WLCG Collaboration Workshop, July 20, 2010, Patrick Fuhrmann

Grid Deployment Board, Oct 13, 2010, Patrick Fuhrmann

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

CHEP 2010, Oct 20, 2010, Yves Kemp:

Hepix Fall 2010, Nov 2, 2010, Patrick Fuhrmann

Linux Kernel: www.kernel.org

NetApp: www.netapp.com

BlueArch: www.bluearc.com

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

Federated File System
https://datatracker.ietf.org/doc/rfc5716/