NFS 4.1 Working Group

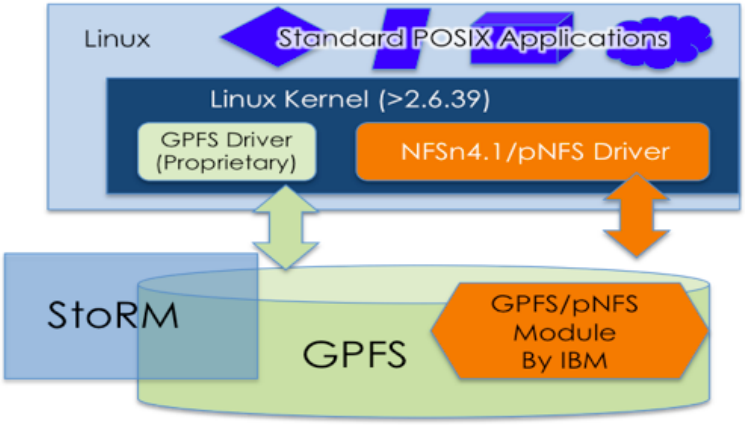
This working group is evaluating the open NFS 4.1 (pNFS) protocol as the generic POSIX access for EMI data sources and applications. Partners involved are DESY, CERN and CNAF/INFN.

Components

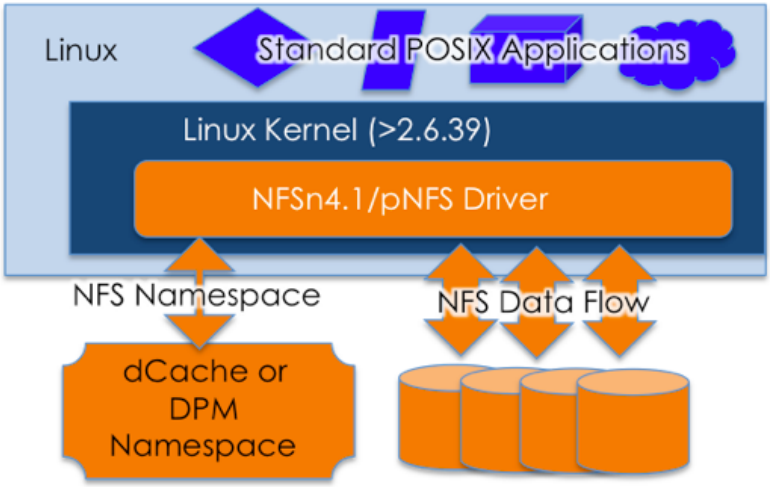
All EMI storage elements are involved. While for dCache and DPM the deliverable requires to implement NFSv4.1/pNFS, for StoRM nothing has to be implemented, as by design, StoRM is build on top of mountable file systems.

Motivation and technical description

The goal of the objective is to allow customers to mount the data repository of EMI storage elements into their local file system space, allowing seamless access without additional client software. This is achieved differently for the different storage systems.



As StoRM is built on top of a file system (GPFS, Lustre, etc)[See Figure 4 1], the goal is achieved by design. The backend file system can be mounted on worker nodes and with that gives direct POSIX access to the data.



DPM and dCache are providing their own file name space engines as well as the different protocol engines [See Figure 4 2]. Therefore, both systems had to find an appropriate network protocol allowing the systems to be mounted into the clients name space. The natural solution was to implement a standard distributed file system protocol, which is NFSv4.1/pNFS. The advantage of the approach is that no proprietary code has to be installed on the client hosts (worker nodes) as the operating system vendors are providing and maintaining the

NFSv4.1/pNFS drivers. Another advantage of pNFS over previous NFS protocols is the fact that pNFS is aware that data can be highly distributed and the name space node is not necessary the node from there the data is delivered. The pNFS protocol redirects requests to an appropriate data server. A temporary disadvantage is the fact the Linux NFSv4.1/pNFS drivers are only fully available with Linux kernel version 2.9.39, which is not officially offered for SL5. Solutions for pNFS in SL5 are under discussion. For SL6 we expect pNFS to be back-ported to the 2.6.32 kernel by the upstream vendor and being available with SL6.2. During the remaining time of the objective the following tasks have to be completed:

- DPM has to upgrade the prototype to a production version. A DPM test-bed has already being setup and sufficient stress testing has to be performed.
- As with the currently available pNFS kernel modules only Kerberos authentication is available, the pNFS working-group has to investigate how X509 certificate can be used for authentication.
- As soon as the DPM test-bed is ready the pNFS group is planning to perform wide area transfer tests between CERN and the GridLab facility at DESY.

Common Documents

Where	What	When	Who	What
CERN	GDB	Jan 2011 GDB	Patrick	NFS 4.1 at GDB; End of demonstrator effort
Cornell	Hepix, Fall	Nov 2010	Patrick	NFS4 at HEPIX, Fall, in Cornell
Taipei	CHEP'10	2010	Yves Kemp	NFS 4.1 at CHEP'10
CERN	Oct GDB	Oct 2010	Patrick	NFS4 at the Oct GDB 2010. Milestone II report
Amsterdam	Jamboree	July 2010	Gerd Behrmann / dCache.org	NFS 4.1, 11 Reasons you should care
London	WLCG Collaboration Workshop	June 2010	Patrick, Jean-Philippe for WLCG and EMI	Introduction of Demonstrator

DESY Test System setup

Staff

Task	People
Testbed	Dima and Yves
pNFS kernel and driver	Tigran
NFS4.1 dCache server	Tigran and Tanja
Hammercloud support	Johannes Elmsheuser, Atlas, Munich

Code

- **dCache server** : In order to allow fast turnover, the recent NFS4.1 code tested here, is not yet checked into the official trunk of the dCache code management system. Therefore, there is no official dCache version available yet, with the features tested here. The plan is to make such a system publicly available before CHEP 10.
- **pNFS client** : We are running the 2.6.35 kernel with pNFS driver, prepared for SL5 plus the corresponding mount tools on SL5 workernodes.
- **Security** :
    - All tests are done w/o integrity and encryption.
    - Beta version of Kerberos code available but not yet sufficiently tested.

Storage and CPU Power

Amount	Type	CPU	RAM	Cores	Network	Disk
1	dCache Headnode	Intel(R) Xeon(R) CPU 5160 @ 3.00GHz	8GB	4	1 GBit	0
5	dCache Pool	Intel(R) Xeon(R) CPU 5520 @ 2.27GHz	12 GB	16	10 GBit	12 * 2 Tbytes
16 or 32	Workernodes	Intel(R) Xeon(R) CPU 5150 @ 2.66GHz	4 GB	8	1 GBit	0

Network

WN	1Gbit	Force 10	4 * 10 GBIT	Arista	10 GBit	dCache pools

Tests

Test type	What	Time / Amount	Result	Detail
Stability	CFEL data transfers	10 days with 13 TBytes sustained writing with 100GB av. filesize	Passed	OK
Stability	CFEL checkusm	13 TBytes; one machine	Slow	Very old client machine
Performance	Hammercloud	128 cores against 100 TBytes	Still ongoing	Performance results are evaluated

-- PatrickFuhrmann - 02-Sep-2010

This topic: EMI > EmiJra1DataDetailsNFS41
Topic revision: r6 - 2011-08-03 - PatrickFuhrmann