NFS frontend for DPM

Shu-Ting Liao

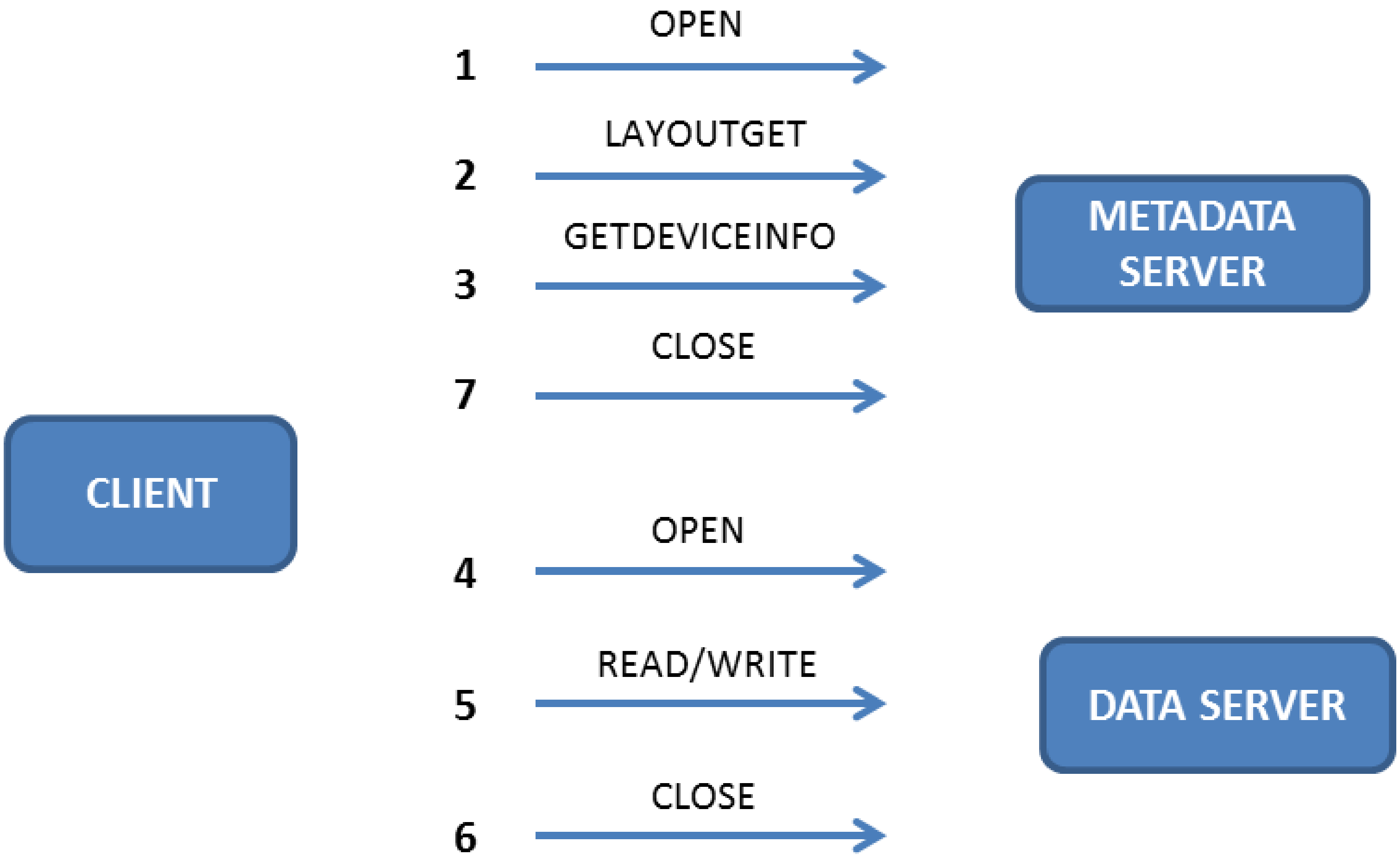
ASGC

Edinburgh DPM Workshop 2013

Reminder

- Main Goal: To allow mount DPM as a regular NFS server providing standard POSIX files access.
- Why pNFS?
  - Direct access to the data, with a standard NFS client
  - Parallel data access
  - No vendor lock-in
  - …

pNFS IO Operations



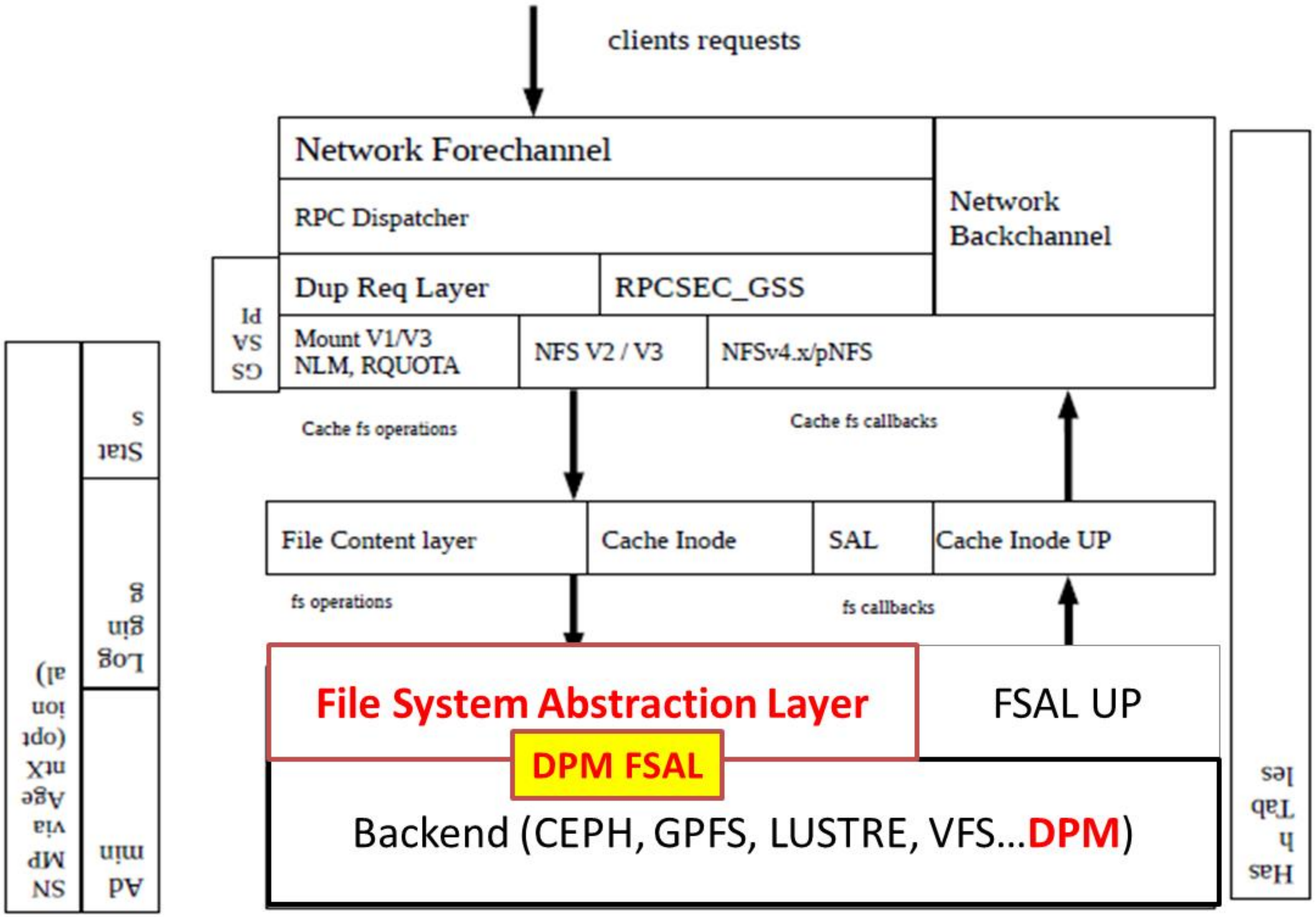
History

- The implementation was based on Ganesha version 1.5
  - A userspace NFS daemon
  - http://sourceforge.net/apps/trac/nfs-ganesha/
  - Read-only available with DPM 1.8.3
  - Not yet based on DMLite
  - Not fully supports pNFS I/O operations
  - Genesha server not stable
  - Performance issue
  - Lack of issues fix and support from Ganesha for 1.5 , we started to work based on new Ganesha 2.0

Ganesha Version 2.0

- Ganesha version 2.0
  - Just released!
  - https://github.com/nfs-ganesha/nfs-ganesha
  - This version is the result of an 18 month effort by an active developer community. There is a lot of new code, a whole lot of improved code, and lots of new features and capabilities.
  - NFSv4.1 support has been greatly improved and now fully supports pNFS .
  - There has been extensive work done to the core of the server. Multi-threaded scaleability and memory usage is much improved. The protocol correctness and export access controls are much better.
  - …

Ganesha Module - FSAL

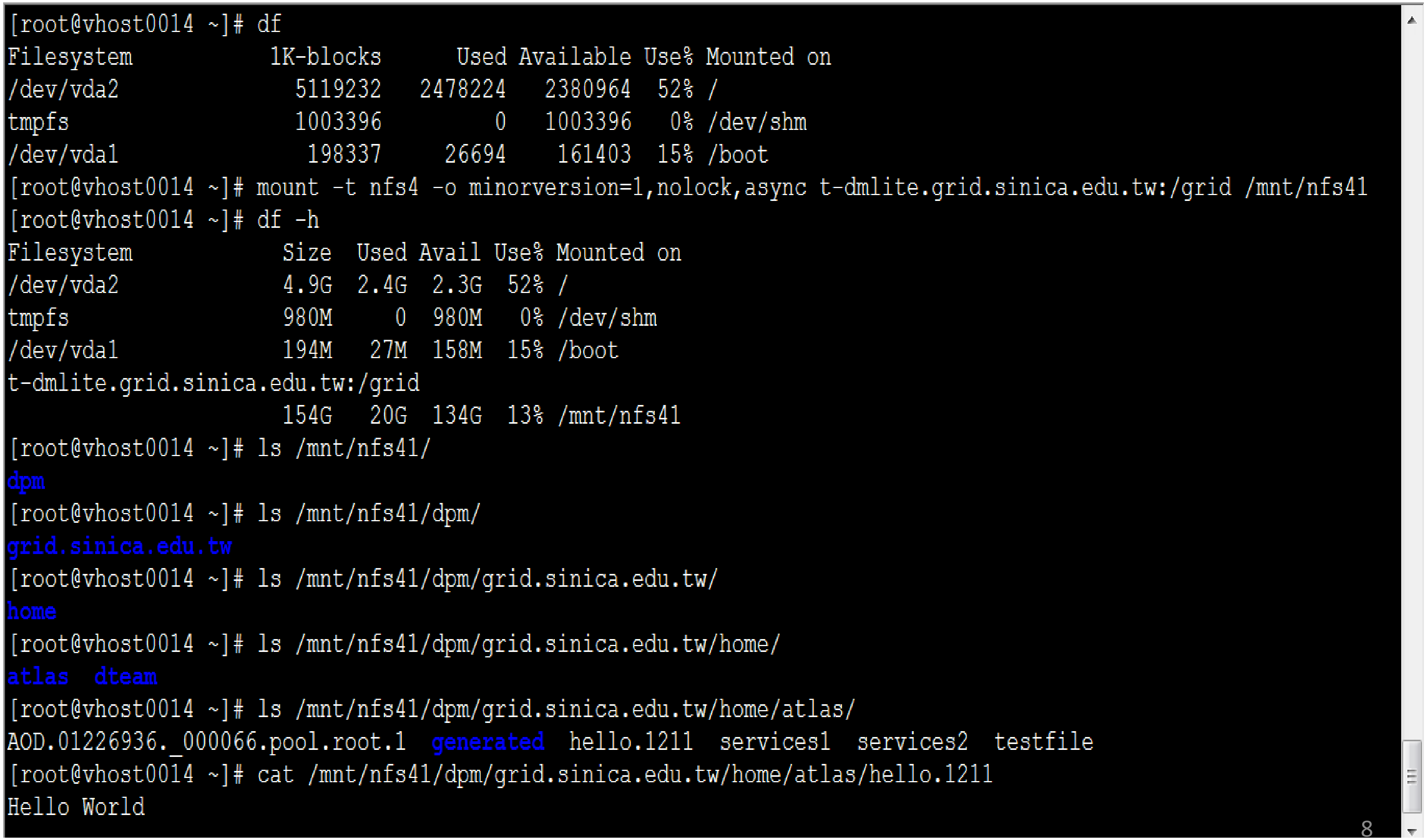


- A FSAL (File System Access Layer) is the interface to a particular filesystem.

Implementation Status

- Completely re-written using DMLite API
- Metadata now working
  - GETATTR
  - LOOKUP
  - READDIR
  - READLINK
  - MKDIR
  - SYMLINK
  - RMDIR
  - RENAME
  - LINK
- Without pnfs layout, read/write goes to head node first.

Testing



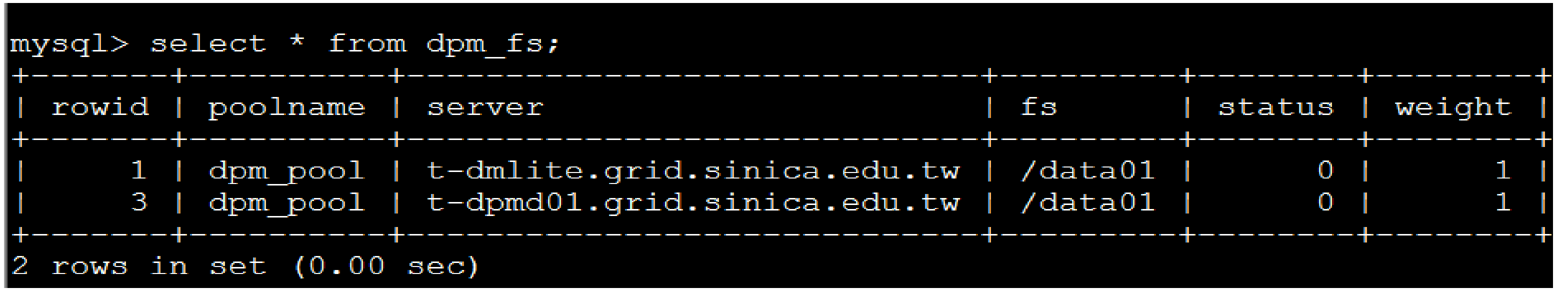
```
[root@vhost0014 ~]# df
Filesystem           1K-blocks      Used Available Use% Mounted on
/dev/vda2              5119232   2478224   2380964  52% /
tmpfs                  1003396         0   1003396   0% /dev/shm
/dev/vda1               198337     26694    161403  15% /boot
[root@vhost0014 ~]# mount -t nfs4 -o minorversion=1,nolock,async t-dmlite.grid.sinica.edu.tw:/grid /mnt/nfs41
[root@vhost0014 ~]# df -h
Filesystem            Size  Used Avail Use% Mounted on
/dev/vda2             4.9G  2.4G  2.3G  52% /
tmpfs                 980M     0  980M   0% /dev/shm
/dev/vda1             194M   27M  158M  15% /boot
t-dmlite.grid.sinica.edu.tw:/grid
                      154G   20G  134G  13% /mnt/nfs41
[root@vhost0014 ~]# ls /mnt/nfs41/
dpm
[root@vhost0014 ~]# ls /mnt/nfs41/dpm/
grid.sinica.edu.tw
[root@vhost0014 ~]# ls /mnt/nfs41/dpm/grid.sinica.edu.tw/
home
[root@vhost0014 ~]# ls /mnt/nfs41/dpm/grid.sinica.edu.tw/home/
atlas  dteam
[root@vhost0014 ~]# ls /mnt/nfs41/dpm/grid.sinica.edu.tw/home/atlas/
AOD.01226936._000066.pool.root.1  generated  hello.1211  services1  services2  testfile
[root@vhost0014 ~]# cat /mnt/nfs41/dpm/grid.sinica.edu.tw/home/atlas/hello.1211
Hello World
```

Ongoing work...

- Moving on pNFS implementations -> implement pNFS operations in DPM FSAL.
- Add proper pNFS access to the disk server -> with the layout going to the client so that it can use it to go directly to the disk server.
- In principle, we do not want to modify DPM to fit pNFS.

Ongoing work...

- Prototyping DPM layout for pNFS
  - Need pnfs device id for disk server
    - rowid -> dpm fsid -> device id -> data server IP address



```
mysql> select * from dpm_fs;
+-------+----------+-----------------------------+---------+--------+--------+
| rowid | poolname | server                      | fs      | status | weight |
+-------+----------+-----------------------------+---------+--------+--------+
|     1 | dpm_pool | t-dmlite.grid.sinica.edu.tw | /data01 |      0 |      1 |
|     3 | dpm_pool | t-dpmd01.grid.sinica.edu.tw | /data01 |      0 |      1 |
+-------+----------+-----------------------------+---------+--------+--------+
2 rows in set (0.00 sec)
```

  - Need striping patterns in the files layout
    - a replica of a file at the begining

Ongoing work…

- pNFS I/O
  - data server handle
  - data server read/write
- Stress testing.
- To deliver by end April 2014.

Thank you!!