

# NFS access to GPFS at DLS, our experience

Frederik Ferner <frederik.ferner@diamond.ac.uk>

Diamond Light Source Ltd

December 5, 2019

# Outline

Diamond

Filesystems and their use

Latest Implementation

CES

CES NFS problems

Cosmetic

Real Issues

Dealbreaker

Solution for DLS

# Diamond

Diamond Light Source (DLS) is the UK's national synchrotron facility. It is located at the Harwell Science and Innovation Campus in Oxfordshire, UK.

- ▶ Synchrotron
- ▶ many beamlines/detectors
- ▶ users/scientists need quick feedback on data quality for experiment planing



# High Performance Filesystems at DLS

## 3 GPFS file systems

- ▶ Shared use with Data Acquisition, Processing and Visualisation in parallel
- ▶ Scientists want to see data on Desktops
- ▶ NFS (and CIFS) access required



## GPFS03 — Initial CES setup

- ▶ GPFS 5.0.2.X
- ▶ 4 nodes
- ▶ CES
- ▶ SMB CES service proved not possible due to POSIX ACLs

# ACLs (POSIX)

```
file Edit View Search Terminal Help
dwxrws--T. 2 gda2          en20287_24      4096 Jul  8 11:37 tmp
dwxrws---. 2 gda2          en20287_24      4096 Jul  8 11:37 xml
[bnh65367@ws386 nnt] $ cd test
[bnh65367@ws386 test] $ ls -l data/2019/en20287-24
total 45060
-rw-rwx---. 1 #06detector en20287_24 1136363 Jul 10 09:51 Atlas_grid2_slot8_1.jpg
-rw-rwx---. 1 #06detector en20287_24 1029629 Jul 10 10:18 Atlas_grid3_slot9_1.jpg
-rw-rwx---. 1 #06detector en20287_24 21883572 Jul 10 09:49 DataAcquisition_grid2_slot8_1.jpg
-rw-rwx---. 1 #06detector en20287_24 21893672 Jul 10 09:55 DataAcquisition_grid2_slot8_2.jpg
dr-xrws---. 3 gda2          en20287_24      4096 Jul 10 15:52 processing
dwxrws---. 13 root            en20287_24      4096 Jul 12 05:40 raw
dwxrws---. 2 gda2          en20287_24      4096 Jul  8 11:37 spool
dwxrwxr-x. 12 #06detector en20287_24      4096 Jul 10 11:09 supervisor_20190710_110010_atlas
dwxrwxr-x. 4 #06detector en20287_24      4096 Jul 10 14:29 supervisor_20190710_111130_data
dwxrws--T. 2 gda2          en20287_24      4096 Jul  8 11:37 tmp
dwxrws---. 2 gda2          en20287_24      4096 Jul  8 11:37 xml
[bnh65367@ws386 test] $ getfacl data/2019/en20287-24
# file: data/2019/en20287-24
# owner: gda2
# group: en20287_24
# flags: -s-
user::rwx
group::rwx
other::---
```

## ACLs over NFS

- ▶ fully enforced (good)
- ▶ not displayed on NFS (CES)

# Impatient Scientist Scenario

- ▶ Detector creates files. File count over NFS lagging behind by initially seconds, increases with number of files in directory, eventually writer completes 10000 files but NFS still listing only 7000-8000 files hours later.
- ▶ expected NFS behaviour???
- ▶ eventually hotfix....

# I/O errors

Some time after restarting everything, user suddenly receive I/O errors over NFS.

- ▶ ganesha process hitting limit of open FDs (No failover trigger?)
- ▶ FD leak?

<https://www.ibm.com/support/pages/ibm-spectrum-scale-nfs-operations-may-fail-io-error>

## random CES IP failovers

With the changes for the FD issue in place, NFS ganesha now so busy that it causes IP failovers? Clients hanging...

## file corruption over NFS

Users reported that their data collection were broken, the data they see is not what they expect!

- ▶ files on native GPFS client OK
- ▶ files on NFS different => not OK
- ▶ only some files affected
- ▶ consistent across all NFS clients
- ▶ persistent for hours (i.e. identified a file at midnight, still same issue next morning)

Still unresolved to date.

# Documentation?

Best practice documents how to set CES up for this? => 404



# Solution for DLS

Giving up on CES (for now??????), replacing it with kernel NFSd solution

Problem solved!

```
File Edit View Search Terminal Help
default:group::r-x
default:group:m06detector:rwx
default:group:m06_data:r-x
default:group:m06_staff:r-x
default:group:dls_bl_cs:r-x
default:group:dls_dasc:rwx
default:group:dls_sysadmin:rwx
default:group:dls-detectors:r-x
default:group:en20287_24:r-x
default:mask::rwx
default:other:----

[bnh65367@cs03r-sc-serv-36 m06] $ ls -l data/2019/en20287-24
total 45060
-rw-rwx---+ 1 m06detector en20287_24 1136263 Jul 10 09:51 Atlas_grid2_slot8_1.jpg
-rw-rwx---+ 1 m06detector en20287_24 1029629 Jul 10 10:10 Atlas_grid2_slot9_1.jpg
-rw-rwx---+ 1 m06detector en20287_24 21883572 Jul 10 09:49 DataAcquisition_grid2_slot8_1.jpg
-rw-rwx---+ 1 m06detector en20287_24 21893672 Jul 10 09:55 DataAcquisition_grid2_slot8_2.jpg
dr-xrws---+ 3 gda2 en20287_24 4096 Jul 10 15:52 processing
drwxrws---+ 13 root en20287_24 4096 Jul 12 05:40 raw
drwxrws---+ 2 gda2 en20287_24 4096 Jul 8 11:37 spool
drwxrwxr-x+ 12 m06detector en20287_24 4096 Jul 10 11:09 supervisor_20190710_110010_atlas
drwxrwxr-x+ 4 m06detector en20287_24 4096 Jul 10 14:29 supervisor_20190710_111130_data
drwxrws-T+ 2 gda2 en20287_24 4096 Jul 8 11:37 tmp
drwxrws---+ 2 gda2 en20287_24 4096 Jul 8 11:37 xnl
[bnh65367@cs03r-sc-serv-36 m06] $
```



# Questions

Questions?

