UNIVERSITY OF BIRMINGHAM

Migrating Spectrum Scale filesystems with minimal downtime

or “Get the DSS-G up and running by Christmas!”

Luke Sudbery
2019-05-09
London/WWSSUG
l.r.sudbery@bham.ac.uk

About me

- University IT for 17 years
- Linux and Windows university sysadmin for 13 years
- Linux HPC Sysadmin for 7 years
- Spectrum Scale user for about 1 year
- Simon’s my boss. So that’s handy.

Challenges

- Large filesystem – 1.5PB data, over 1 billion files.
- No downtime (OK, just a bit…)
- Faulty hardware
- ACLs
- Dodgy code?
- Multiple filesets (>1000)
- Backups – TSM – IBM Spectrum Protect™
- Offline files – HSM migrated – on tape only

Solution! rsync?

```
#!/bin/sh
rsync -a /rds/ /rdsnew
exit 0
```

Solution! rsync?

```
#!/bin/sh
rsync -a /rds/ /rdsnew
exit 0
```

```
sent 13,362,797,459 bytes  received 1,325,431,800 bytes  49,743.98 bytes/sec
total size is 1,042,160,396,855,559  speedup is 70,952.08 (DRY RUN)
rsync warning: some files vanished before they could be transferred (code 24) at
main.c(1178) [sender=3.1.2]

real    4921m16.259s
user    72m6.832s
sys     388m55.421s
```

82 hours – 3 ½ days

Solution! IBM Spectrum Scale AFM – Active File Management?

- "Migration does not pull file system–specific parameters such as quotas, snapshots, file system–level tuning parameters, policies, fileset definitions, encryption keys, and dmapi parameters."

→ **Home cluster errors**

→ **Failure and recovery**

→ **Steps to deal with an IW cache fileset disaster**

For more information about AFM restrictions, see IBM Spectrum Scale™ FAQ in IBM® Knowledge Center.

- Fileset disaster?!

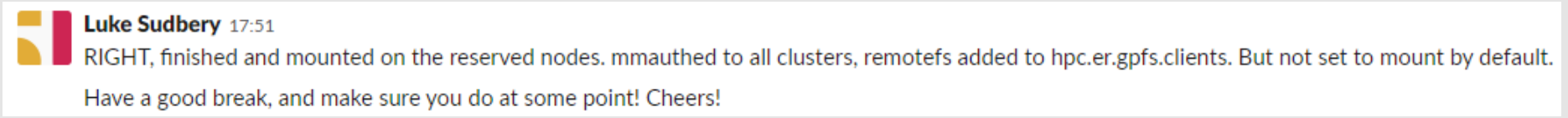
Problem!

See https://lenovopress.com/lp0837-lenovo-dss-g-thinksystem

Lenovo DSS-G241 ships with dodgy ESM!

Distributed Storage Solution for… IBM Spectrum Scale. Don't know where the G comes from…

- Lenovo DSS is like an x86 IBM ESS – bundled hardware and GPFS Native Raid solution
- A nice set of installation tools…
- A little too helpful when you're in a hurry!
- No time to benchmark properly before Christmas



So what did we do?

- Patch `mpifileutils` to support GPFS ACLs and copy data in parallel, over many nodes.
- Use AFM prefetch scripts to generate lists of files to copy and update those.
- Repeat until nearly there.
- Cut over to new filesystem before recalling all offline files – they generally weren't in use anyway (they were migrated offline for a reason).
- Have fun with TSM – backup and HSM recall.

Patching mpifileutils

See https://github.com/hpc/mpifileutils/commit/0df1ad60ca3d8fe584ac09131e6c44b3c21acd15

- Comes down including `gpfs.h` and "get ACL":

```
/* try and get the ACL */
MFU_LOG(MFU_LOG_INFO, "Getting GPFS ACL on %s for %s", src_path, dest_path);
int r = gpfs_getacl(src_path, aclflags, aclbufmem);
```

- And "put ACL":

```
/* Assuming we now have a valid call to an ACL,
* try to place it on dest_path */
if (r == 0) {
  MFU_LOG(MFU_LOG_INFO, "Sucessfully got ACL from %s", src_path);
  r = gpfs_putacl(dest_path, aclflags, aclbufmem);
  ...
} else {
  ...
}
```

(error checking etc omitted here...)

Initial sync of data

- Freeze new storage creation (but users still on the system using existing filesets).
- Mirror existing fileset config – oneoff
- For each fileset:

```
mpirun --np 12 dsync /rds/$fileset/ /rdsnew/$fileset/
```

- And it nearly worked!
- Eventually realised TSM migrated files were causing problem… exclude those files (generate list with a policy scan).

AFM prefetch script

See https://www.ibm.com/developerworks/.../Data Migration Methods using AFM_v2.12.pdf

```
#!/usr/lpp/mmfs/bin/mmksh
#-------------------------------------------------------------------------------------
# Date Written: 25 Oct 2015 Version 1.0.0-01
# Remarks:
# This Script run on the HOME cluster only
# It scans the path provided using mmapplypolify and the policy embedded in this
# script below to generate a files list
# we then locate a previous version of the files list and use tsbuhelper (v4.2 or greater)
# to generate two types of files:
# - changed files list
# - deleted files list
# A file called $migratePath/.mmmigrateCfg/mmmigrate.bootstrap.files is created containing
# the names of the files lists that will then need to be prefetched at the cache
# Once the cache prefetches the bootstrap file, it can then prefetch using the bootstrap
# file to retrieve the files listed.
# At a minimum, the bootstrap file should contain the following:
# - mmmigrate.file.version (contains the current version number)
# - mmmigrate.list.deleted.v${fileVersion}.filelist (filename of list of deleted files)
# - mmmigrate.list.changed.v${fileVersion}.filelist
# - any other files needed to kick off / continue this migration from the cache
#-------------------------------------------------------------------------------------
```

AFM prefetch script

See https://www.ibm.com/developerworks/.../Data Migration Methods using AFM_v2.12.pdf

- Initial run:

```
END PHASE 1: Files scan phase at Sun  5 May 17:24:43 BST 2019 and took 14 seconds. (14 Seconds)

First Time run. So no need for tsbuhelper....


Summary:
Changed Files Version = 0
-------------------------------------------------------------------------------
 # files | File Name
-------------------------------------------------------------------------------
          1228 | /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.v0.filelist
             0 | /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.deleted.v0.filelist
             0 | /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.changed.v0.filelist
-------------------------------------------------------------------------------

JOB ENDED at: Sun  5 May 17:24:43 BST 2019 and took 14 seconds. (14 Seconds)
```

AFM prefetch script

See https://www.ibm.com/developerworks/.../Data Migration Methods using AFM_v2.12.pdf

- Subsequent runs:

```
START PHASE 2: Looking for changed / deleted files and using tsbuhelper Started at Sun  5 May
17:26:05 BST 2019.
::CLUSTERMIGDIFF::0::131::12::0::
END PHASE 2: Looking for changed / deleted files using TSBUHelper Completed at Sun  5 May 17:26:05
BST 2019 and took 0 seconds. (0 Seconds)



Summary:
Changed Files Version = 1
--------------------------------------------------------------------------------
 # files | File Name
--------------------------------------------------------------------------------
          1240 | /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.v1.filelist
            12 | /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.deleted.v1.filelist
           131 | /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.changed.v1.filelist
--------------------------------------------------------------------------------

JOB ENDED at: Sun  5 May 17:26:05 BST 2019 and took 15 seconds. (15 Seconds)
```

AFM prefetch script

See https://www.ibm.com/developerworks/.../Data Migration Methods using AFM_v2.12.pdf

- And it tells you what to do:

```
Next Steps:
------------------------------------------------------------------------------------------------
 (1) If this is the 1st scan for this home, then transfer the file
 /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.v0.filelist) to the cache
 and then, run:
 mmafmctl <filesystem> prefetch -j <fileset> \
 --list-file /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.v0.filenames
 (2) If this is not the 1st Scan, for this home, then transfer these files to cache:
 /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.deleted.v0.filenames
 /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.changed.v0.filenames
 (3) - Run prefetch on deleted files
 mmafmctl <filesystem> prefetch -j <fileset> \
 --list-file /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.deleted.v0.filenames --delete
 (4) - Run prefetch on changed files
 mmafmctl <filesystem> prefetch -j <fileset> \
 --list-file /rds/tools/system-scripts/.mmmigrateCfg/mmmigrate.list.changed.v0.filenames
------------------------------------------------------------------------------------------------
```

Cut over and recall offline files

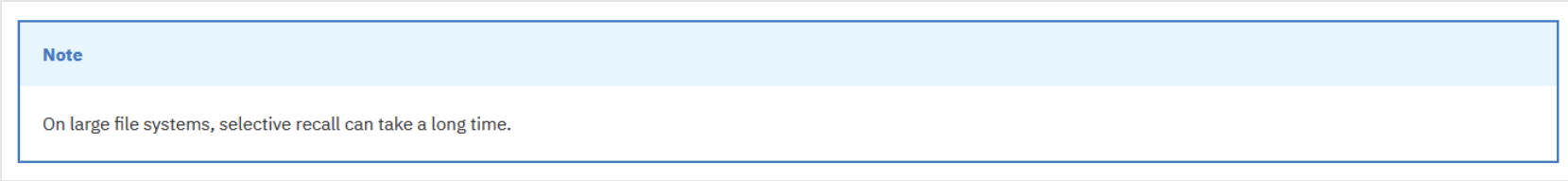
- Change filesystem names and remount… WCGW?!

```
# mmchfs rds –W rdsold –T /rdsold
# mmchfs rdsnew -W rds –T /rds
```

- TSM HSM/migrated files stay with the *filesystem*
- TSM backed up files are associated with the *path*
- No need to back everything up again*
- But it does mean we need to recall and copy all the previously excluded offline files before running backup again or TSM will expire files previously backed up

TSM – recall files

- This was the longest part of the migration…



- But the new filesystem was live now – migrated data was generally not ‘live’ by it’s nature
- Can only *recall* to the same filesystem
- Can *restore* to a different filesystem

TSM – tape optimised recalls

See https://www.ibm.com/support/knowledgecenter/.../dsmrecall.html

- `dsmrecall -h` doesn't mention it, and there is no `man` page…

```
[root@rds-er-gpfs01 ~]# dsmrecall -h
IBM Spectrum Protect
Command Line Space Management Client Interface
  Client Version 8, Release 1, Level 4.1
  Client date/time: 05/05/19   20:03:13
(c) Copyright by IBM Corporation and other(s) 1990, 2018. All Rights Reserved.

Usage: dsmrecall [-Recursive] [-Detail] [-Help] file specs|-FIlelist=file;
   or  dsmrecall [-Detail] -offset=XXXX[kmgKMG] -size=XXXX[kmgKMG] file specs

[root@rds-er-gpfs01 ~]#
```

- But the website goes into a lot more detail:

-PREView

Generate list files that are optimized for tape recalls but do not recall the files. You must also specify **filelist** and a file system. The **preview** option is not valid when **filelist** specifies a collection file.

Where are we?

- Created new filesystem, replicated filesets
- Migrated all the data, incrementally
- Cut over to the new filesystem, done a final sync of changed data and gone live
- Restored everything that was migrated offline – only on tape
- Not done a backup of the new filesystem yet…
- Data partially protected by snapshots (not a backup, we know)

TSM – backing it all up

- 1PB takes a long time to send to tape.
- It's all already there!
- All the metadata is the same – user, group, mode, size, atime, mtime, xattrs and ACLs
- In initial tests TSM wanted to back up every file again…
- But ctime was different on the new filesystem and there is no way to update ctime to a specific value

TSM – 2 useful options

- UPDATECTIME

no

The backup-archive client does not check the change time (ctime attribute) during a backup operation. This value is the default.

yes

The backup-archive client checks the change time (ctime attribute) during a backup operation. If the ctime attribute changed since the last backup operation, the ctime attribute is updated on the IBM Spectrum Protect™ server. The object is not backed up unless it has either ACLs or extended attributes. The client checks files and directories.

- SKIPACLUPDATECHECK

No

If you specify No, the client performs checksum and size comparisons of the ACL data, before and after backup and during incremental processing. This is the default.

Yes

If you specify Yes, the client does not perform checksum and size comparisons of the ACL data.

TSM – first run after migration…

- UPDATECTIME yes

no

The backup-archive client does not check the change time (ctime attribute) during a backup operation. This value is the default.

yes

The backup-archive client checks the change time (ctime attribute) during a backup operation. If the ctime attribute changed since the last backup operation, the ctime attribute is updated on the IBM Spectrum Protect™ server. The object is not backed up unless it has either ACLs or extended attributes. The client checks files and directories.

- SKIPACLUPDATECHECK yes

No

If you specify No, the client performs checksum and size comparisons of the ACL data, before and after backup and during incremental processing. This is the default.

Yes

If you specify Yes, the client does not perform checksum and size comparisons of the ACL data.

TSM – subsequent normal runs

- UPDATECTIME no

no

The backup-archive client does not check the change time (ctime attribute) during a backup operation. This value is the default.

yes

The backup-archive client checks the change time (ctime attribute) during a backup operation. If the ctime attribute changed since the last backup operation, the ctime attribute is updated on the IBM Spectrum Protect™ server. The object is not backed up unless it has either ACLs or extended attributes. The client checks files and directories.

- SKIPACLUPDATECHECK no

No

If you specify No, the client performs checksum and size comparisons of the ACL data, before and after backup and during incremental processing. This is the default.

Yes

If you specify Yes, the client does not perform checksum and size comparisons of the ACL data.

Done!

- Created new filesystem, replicated filesets
- Migrated all the data, incrementally
- Cut over to the new filesystem, done a final sync of changed data and gone live
- Restored everything that was migrated offline – only on tape
- Done a new incremental TSM backup without touching every file
- Not enabled HSM in the new filesystem yet but we will do for SOBAR in due course…

UNIVERSITY OF BIRMINGHAM

Thank you. Questions?

Luke Sudbery
2019-05-09
London/WWSSUG
l.r.sudbery@bham.ac.uk