Working with Large DMF Files

Terabyte-sized files need special treatment

Peter Edwards | Sr Systems & Data Administrator

DMF Users Group| February 2017

The problem

CSIRO has advised DMF users to keep their files under 100 GB where possible to prevent them becoming unwieldy.

“Unwieldy” means:

- Files remain online for a fairly short period of time since their last access. On a previous system, this could at times be measured in just hours. Users would prefer a week or two at a minimum.
- Once offline, the recalls required to reinstate normal access also take a long time, especially when the system is busy.

These unanticipated delays can affect batch jobs, whose time limits are expressed in terms of elapsed time, not CPU time.

The problem (cont'd)

A suggested maximum file size of 100 GB was acceptable for a long time, as files rarely got that long.

This is no longer true; we now see 1 TB files and larger.

Tests show that on our current system, 1 TB is unusable:

- Files go offline at the next *dmfsfree* run, which at a minimum is once per day at 8:00.
- Recalls take an hour, provided tape drives are available.

Moving to STK T10kD drives does not help, as the transfer rate is similar to that of older models.

Proposed “solution”

- **Experiments only – not in production yet**

- Change DMF policies to force a lower weighting for files accessed in the last week, which does not prevent them going offline, but still makes them less likely to do so during that week.

  Which means someone else’s files would have to be chosen instead.

- At migration time, deliberately send different parts of large files to different tapes. These *chunks* could subsequently be recalled in parallel provided enough tape drives were available.

  If they weren’t available, this would make recalls a bit slower, so an appropriate number of drives should be provided.

  Other recalls and migrations could be impacted by any reduced availability of drives.

File weighting - configuration

```
define weight_policy
TYPE policy

# Special treatment for files used within last 2 days
SPACE_WEIGHT            0       5.268e-10 when age <= 2 \
                                and uid in (edw192) and sitetag = 21

# Following weights give an identical weighting factor to a 2MB file
# not accessed for 6 months, and a 32GB file not accessed for 1 day.
AGE_WEIGHT              0       1.0
SPACE_WEIGHT            0       5.268e-9
FREE_DUALSTATE_FIRST on

enddef
```

File weighting - tests

Files subject to the extra weighting parameter – number of daily dmfsfree runs needed to go offline:

Starting:	10 Nov	25 Dec	2 Jan	16 Jan
25 GB	4	4	11	15
50 GB	4	4	11	15
100 GB	4	4	10	14
200 GB	4	1	1	1
400 GB	1	1	1	1
800 GB	1	1	1	1

By comparison, an equivalent set of files not subject to the new weighting all went offline at the next dmfsfree run.

Chunk “striping” - configuration

The Volume Group MAX_CHUNK_SIZE parameters forces files to be divided into multiple pieces (*chunks*) at migration time.

If MAX_CHUNK_SIZE < ZONE_SIZE there is nothing forcing chunks to be sent to different tapes or drives. They will probably end up on the same tape, adjacent to each other.

But as our ZONE_SIZE is 50 GB, this is not an issue for us as we are considering chunk sizes of 100’s of GB.

(If using parallel DMF, use of the MULTITAPE_NODES parameter should be considered to avoid a performance degradation.)

Chunk “striping” – explanation

Where possible, DMF schedules one tape drive per zone (or part thereof) of accumulated data, which means that for large files we will get one drive per chunk.

Full zones will be allocated a drive immediately and migrate concurrently. Partial zones may have to wait a while, satisfying the normal conditions for being written to tape.

At recall time, where possible a number of drives equalling the number of chunks will be normally assigned and the transfers to disk will happen concurrently.

Chunks do not have to arrive on disk in the correct order, according to their position in the file.

Chunk “striping” - tests

Recall times:

chunks	file	mins	mins	mins	mins	mins	mins	mins
1	1024GiB-1chunk	62	62	62	61	61	62	62
2	1024GiB-600GB	62	67	65	58	61	68	64
3	1024GiB-500GB	82	80	79	76	77	82	79
4	1024GiB-330GB	89	84	83	74	43!	86	85

Possible causes of lack of scaling:

- Insufficient filesystem or FC performance to keep multiple drives streaming
- Contention at mount time if tapes serviced by the one robot
- Lack of drives (not applicable to these tests though)

Conclusion

- Changing the weighting policy to increase disk residency for a period works well, as long as other files being forced offline is acceptable.

- Chunk striping to increase effective transfer rates of recalls is problematic, as it carries the risk of overloading I/O paths and therefore not scaling linearly with the number of drives.

Thank you

CSIRO IMT Scientific Computing
Peter Edwards
Sr Systems & Data Administrator

t +61 3 9545 2377
e peter.edwards@csiro.au
w https://wiki.csiro.au/display/ASC/Scientific+Computing+Homepage