




COSMA File systems

Best practice with parallel storage

ICC Theory Lunch
July 2019
Alastair Basden



DiRAC call for proposals

- Now open:
 - www.dirac.ac.uk/callforproposals

Storage allocation

- You have several storage allocations on COSMA
 - /cosma/home
 - ~10GB quota
 - Backed up
 - Source files, information, etc.
 - 37TB XFS / NFS file system
 - /cosma567/data/ [PROJECT] /USERNAME
 - ~10TB quota – can be increased
 - Parallel file systems
 - Lustre or GPFS
 - Multiple redundancy, no backup
 - /snap7/scratch/PROJECT/USERNAME
 - Fast IO for COSMA7. No redundancy, no backup
 - NOT FOR ANY LONG-TERM STORAGE
 - Cleared out regularly
 - Only use if you know what you are doing!
 - Various other locations depending on project
 - e.g. /madfs, /data/dega1

Which file system?

- Use the /cosmaN relevant to the queue to which you will submit jobs
 - e.g. If accessing data on /cosma5, usually best to copy to /cosma6 before running jobs
 - Unless you'll only access it once
 - /cosma5 not available to COSMA7 compute nodes
- Use /cosma/home/ for small files, source code, etc
- Use /snap7/scratch for temporary COSMA7 files
 - e.g. restart files
 - Fast access

Groups

- You might be part of several projects
 - Your Unix group(s) is (are) named after your project
- The “id” command will tell you which groups you are part of, and your current, effective, group:
 - `uid=20957(dc-basd1) gid=64528(durham) groups=64528(durham),1210(dphsprog),20140(dr004),64526(dp004),64532(lg),64603(cosma7),64605(mad),64607(cosma5),64610(madtesters),1295600001(clusterusers)`
- File permissions/ownership is based on your UID and GID:
 - `ls -al:`
 - `-rw-r--r-- 1 dc-basd1 durham 4110 Sep 21 2018 users.txt`
 - User and project disk quotas are derived from file ownership
 - Scattered files are not always easy to find
- `newgrp` command can be used to change your effective group
 - You will then write files as part of that group
- COSMA5 users will be members of group “durham” (DiRAC calls this `hpcicc`)
- To use COSMA6/7, you need to be part of a `dp` group (e.g. `dp004`)

Quotas

- If you get a quota warning email:
 - Don't ignore it!
 - You have 7 days...
 - Clean up some files
 - Use the “quota” command
 - If you want to keep files “just in case...”, we can archive to tape

Parallel file systems

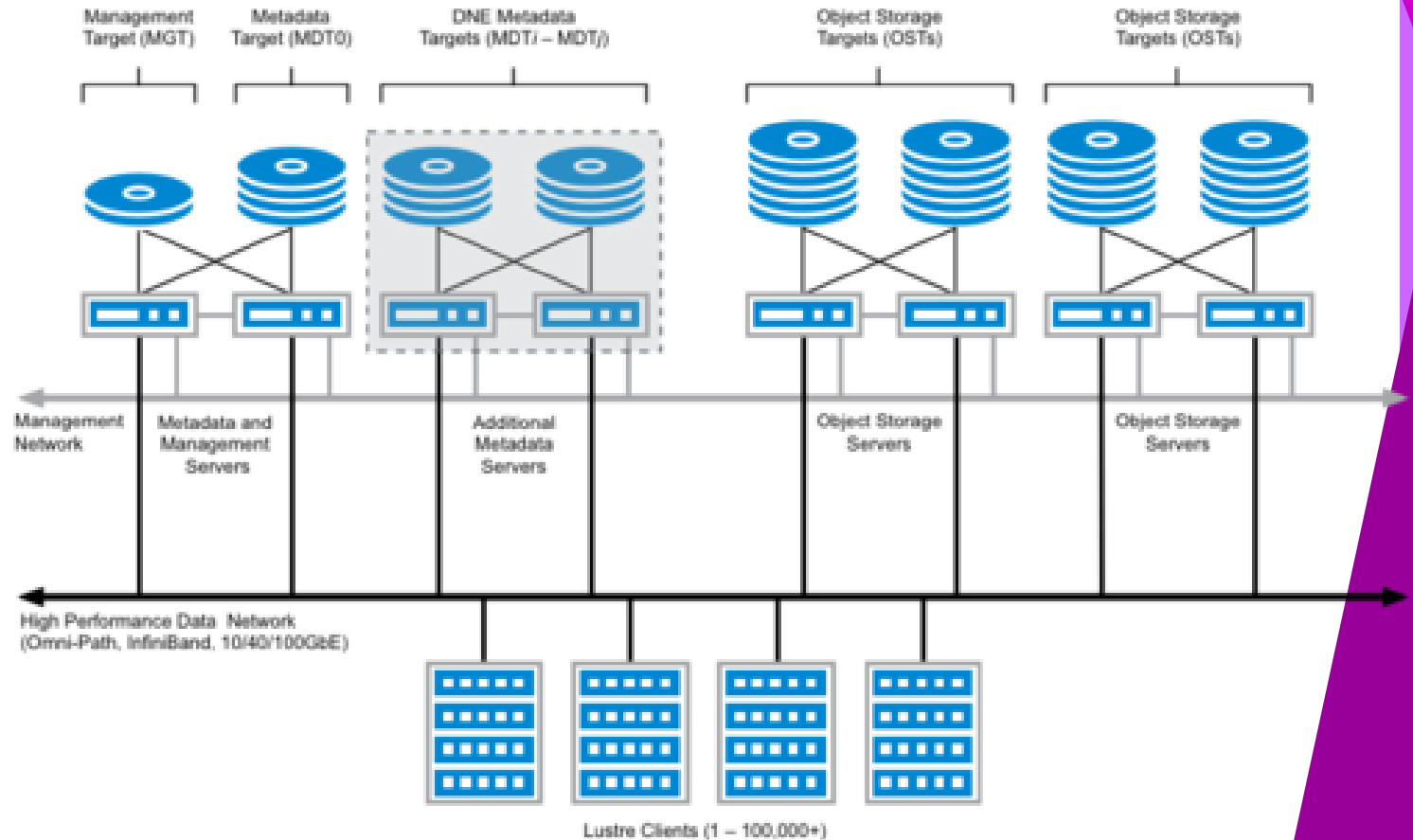
- Understanding general principals is key
- Many small files are not a good idea

Lustre

- The parallel file system used on COSMA6 and COSMA7
- /cosma7:
 - 16 object servers and 1(2) metadata servers
 - 660 disks (8TB)
 - Raid 6 configuration
 - Groups of Logical units
- /snap7:
 - 20 object servers and 1(2) metadata servers
 - 160 disks (3TB SSD)
 - No redundancy
 - Fast
- Files can be written in parallel (i.e. no benefit to writing sequentially)

Lustre overview

- MDS
 - MDT
- OSS
 - OST



Striping files

- Spreading files across multiple disks and servers
 - Provides benefit for parallel access
- Don't stripe small files
 - e.g. a few tens of MB
- Don't stripe if you read/write lots of files simultaneously
 - (e.g. approx equal to the number of Logical disks = 54 on /cosma7)

How to stripe

- Existing files cannot be restriped
 - (though they can be migrated – see later)
- Use the Lustre API
 - Requires code modification
- OR:
 - Add striping to directories/empty files:
 - `lfs setstripe -c <stripe count/-1> -S <size> /directory/or/file`
 - (other options available – `man lfs-setstripe`)
 - Everything within this directory will then have this striping
 - `lfs getstripe /directory/or/file`
 - If a file doesn't exist, it will be “touched” with this striping

Striping recommendations

- For ease of use:
 - Have a striped directory for large files
 - Have a non-striped directory for small files
- Where possible, write files in parallel
 - Each MPI task/thread writing at the same time

Progressive file layouts

- Not yet on /cosma7
 - Requires newest version of Lustre
 - /cosma8
 - Different striping depending on file size
 - Up to size N1 is striped S1
 - from N1 to N2 striped S2
 - etc
 - Entirely user configurable
 - lfs setstripe ...

Small file handling

- /cosma8 will allow small files to be stored on metadata
 - Ifs setstripe...
- Lower latency access for small files
 - Guideline size typically no more than 100k-1MB

lfs migrate

- To move files to different OSTs
 - e.g. change the striping on an existing file
 - `lfs migrate -c <stripe count> -i <start index> -S <stripe size> -o <list of OSTs> filename`
- Useful if you get the striping wrong first...
 - (or you want to restripe files after this talk!)
 - Also if you want to pre-stage files onto particular disks
 - (perhaps more useful for /snap8)

Parallel HDF5 and Lustre

- Several parameters can be tuned to improve performance
 - Chunk size and dimensions
 - Metadata cache size
 - Avoid small file writes
 - Alignment properties
 - Multiple of disk block size
 - MPI-IO tuning
 - block size, buffer size, number of nodes for collective buffering

/snap7

- Use /snap7 for temporary restart files
- This is backed up approximately once per day
 - This may change without warning
- Inconsistency if restart files are backed up whilst being written
 - If important, consider how you write restart files
 - And how often

/cosma8

- Lustre
- ~5-6PB
- ~2x throughput performance improvement
 - From hardware and software improvements

/snap8

- Fast scratch space
- ~1PB
- Storage distributed within compute nodes
- Some optional redundancy
- Metadata server on Optane memory

Summary

- Use the correct storage location
- Stripe files if needed