Lustre IO

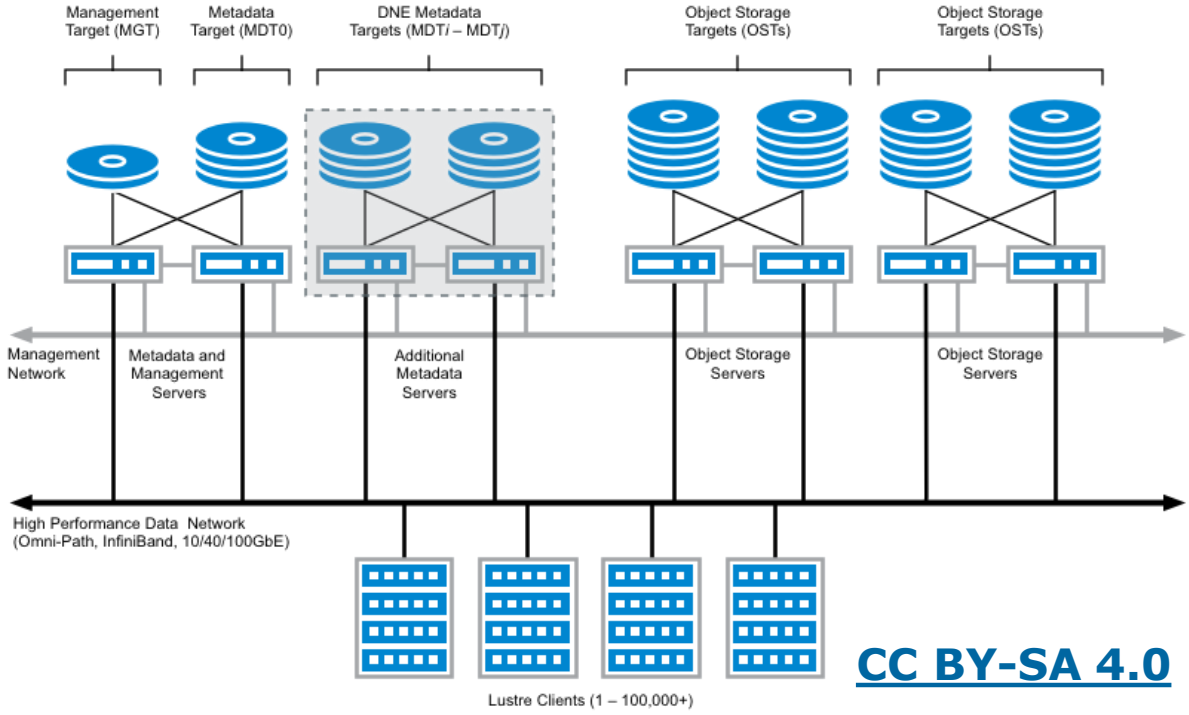
2020

Lustre architecture

Lustre Architecture

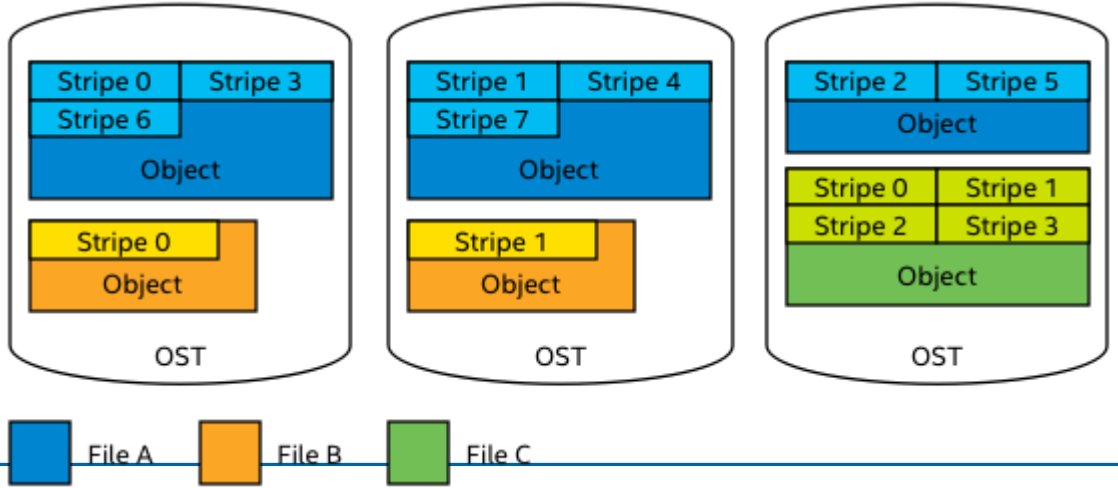
Berlin (Lise):
28 OSTs
85 GB/s

Göttingen (Emmy):
100 OSTs
65 GB/s



Lustre stripe count, size and index

- Lustre stores files in stripes on OSTs
- By default, each stripe is 1MB (stripe size=1MiB)
- By default, all stripes for a file are stored on 1 OST (stripe count=1)
- Lustre chooses the stripe index, on what OST the first stripe is written.



Lustre user settings

Changing the lustre stripe count

- This works well for most cases, but performance can be improved for large files (~GB) by using a larger stripe count:

```
lfs setstripe -c [count] [file]
```

- Stripe count can only be set if the file is not created yet.

Changing the lustre stripe count (2)

Increasing the stripe count for files of ~GBs in size

- Offers benefits:
  - Increases the bandwidth
  - Very large files do not fill up a single OST and lower overall performance.

- And some disadvantages
  - Increases overhead due to network operations and server contention
  - But it also has a higher risk of data loss (data corruption on 1 of the OSTs with stripes corrupts the file.

Changing the lustre stripe count (3)

- Only increase the stripe count for files of 1 GB and larger.
- Recommended maximum stripe count
  - Emmy: 32
  - Lise: 28 (or -1 to use all)

Checking the lustre stripe count

lfs getstripe [file]

```
lmm_stripe_count:  1
lmm_stripe_size:   1048576
lmm_pattern:       raid0
lmm_layout_gen:    0
lmm_stripe_offset: 12
        obdidx           objid           objid           group
            12         4363475       0x4294d3                0
```

Changing the default lustre stripe count of a directory

```
lfs setstripe –c [count] [dir]
```

- Files that are newly created will automatically get the default stripe count.
- Note that files smaller than the stripe size will still be stored on 1 OST only.

- Use a count of -1 to stripe over all available OSTs
  - This probably gives the best performance,
  - But it also has a higher risk of data loss (data corruption on any OST corrupts the file)

What about changing the stripe size or index?

- Please don‘t change the stripe index
  - Lustre automatically distributes files across OSTs.

- In general, you don‘t need to change the stripe size.
  - However, for very large files on **Lise** the best performance can be reached with a stripe size of 16 MiB

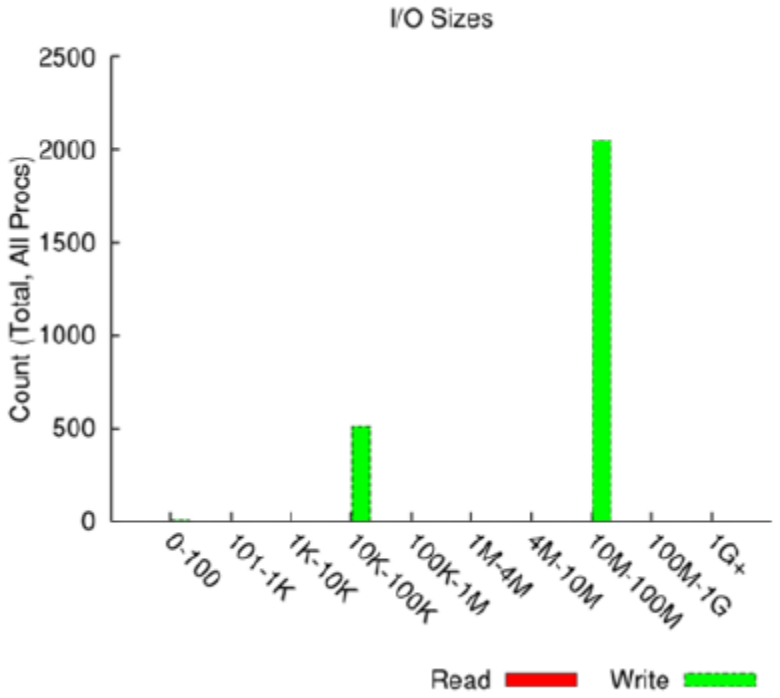
Lustre profiling

Darshan (not yet installed)

- Open source HPC IO Characterization tool
- Shows how efficient your MPI I/O is
- Runtime instrumentation (no recompiling)
- Latest version also shows POSIX I/O

File Count Summary
(estimated by I/O access offsets)

type	number of files	avg. size	max size
total opened	1299	1.1G	8.0G
read-only files	1187	1.1G	8.0G
write-only files	112	418M	2.6G
read/write files	0	0	0
created files	112	418M	2.6G



Other IO hints

Besides Lustre

- If your application is IO-bound, it could be interesting to look at more basic settings.
  - Fortran Formatted IO: export FORTRAN_BUFFERED=yes (try with Nastran)
  - Reduce open & close of files

- Use /tmp for your temporary files
  - It is a RAM disk, so be aware that it lowers available memory for your application.
  - It is about 100Gb in size
  - Also interesting for input files that are accessed a lot during the run

Debugging IO errors

- It is sometimes not clear what is the cause of IO errors
  - srun -l strace -etrace=%file [app]

Thanks for your attention

john.donners@atos.net

Atos