Lustre* ZFS Snapshot Overview

Fan Yong; Zhang Jinghai

High Performance Data Division

How Can Lustre* Snapshots Be Used?

Undo/undelete/recover file(s) from the snapshot

- Removed file by mistake, application failure causes data invalid

Quickly backup the filesystem before system upgrade

- Upgrade Lustre/kernel may hit some trouble and need to roll back

Prepare a consistent frozen data view for backup tools

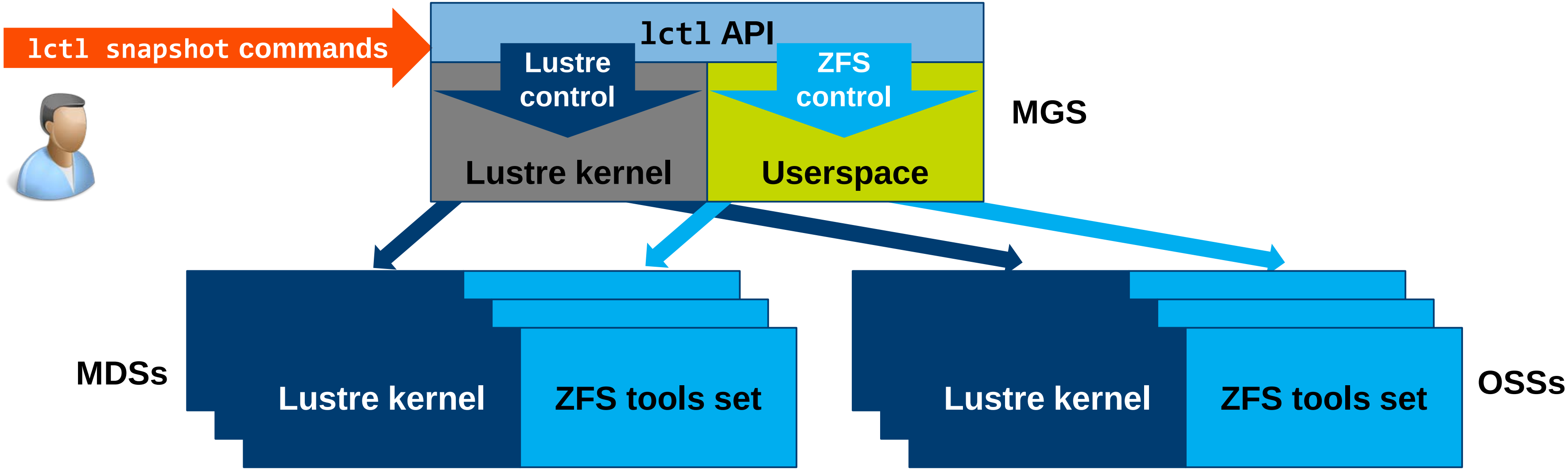
- Ensure system is consistent for the whole backup

Phase I: ZFS-based Lustre* Snapshot

Targeted for Community Lustre 2.10 release

ZFS-based Lustre* Snapshot Overview

- ZFS snapshot created on each target with a new fsname
- Mount as separate read-only Lustre filesystem on client(s)
- Architecture details: http://wiki.lustre.org/Lustre_Snapshots



Global Write Barrier

“Freeze” the system during creating snapshot pieces on every target.

Write barrier on MDTs only

- No orphans, no dangling references

New `lctl` commands for the global write barrier

- `lctl barrier_freeze <`*`fsname`*`>` [*timeout* (seconds)]
- `lctl barrier_thaw <`*`fsname`*`>`
- `lctl barrier_stat <`*`fsname`*`>`

Two Phase Global Write Barrier Setup

User action

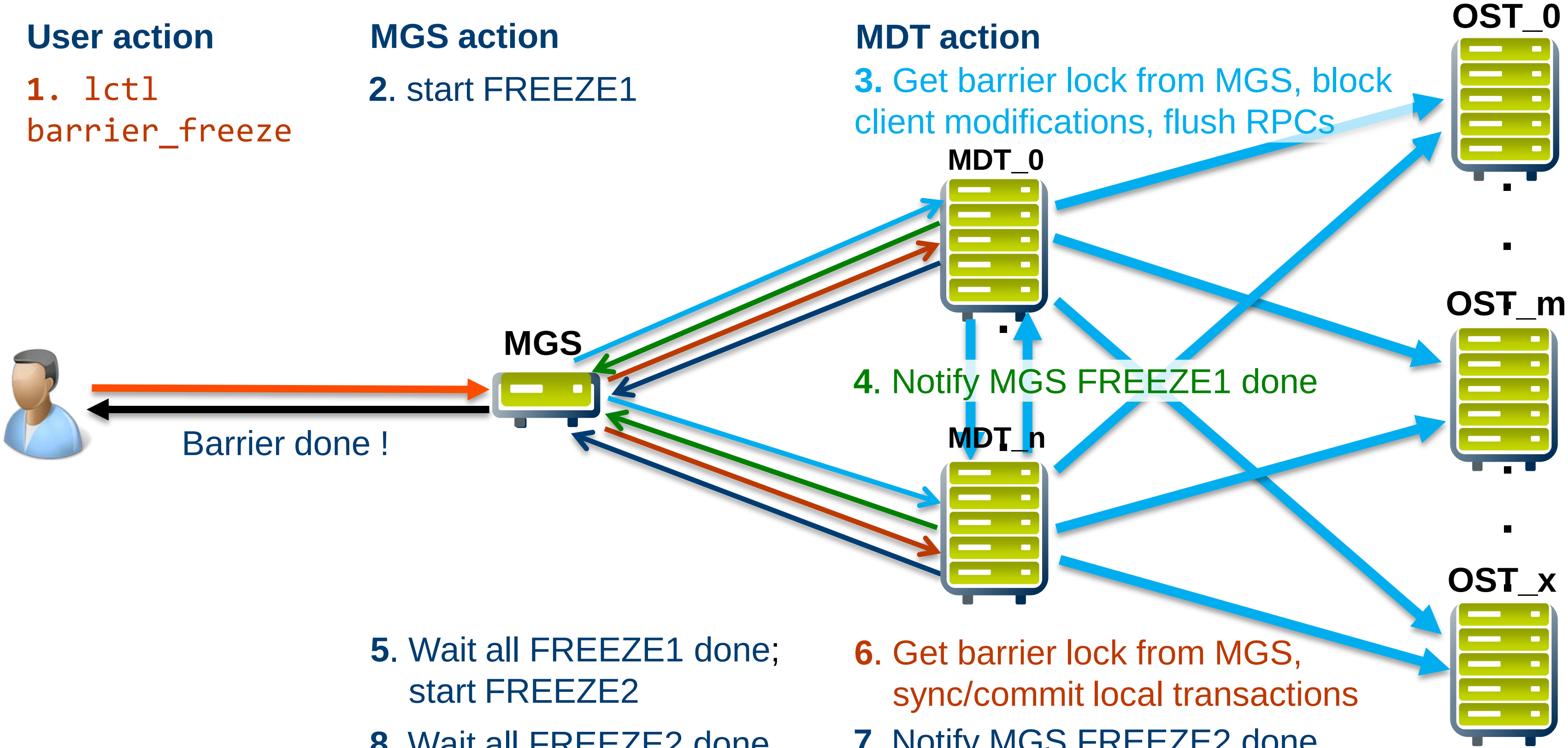
1. `lctl barrier_freeze`

MGS action

2. start FREEZE1

MDT action

3. Get barrier lock from MGS, block client modifications, flush RPCs



Barrier done !

4. Notify MGS FREEZE1 done

5. Wait all FREEZE1 done; start FREEZE2

6. Get barrier lock from MGS, sync/commit local transactions

7. Notify MGS FREEZE2 done

8. Wait all FREEZE2 done

Fork/Erase Configuration Logs

Snapshot is independent from the original filesystem

- New filesystem name (`fsname`) is assigned to the snapshot
- Fsname is part of the configuration logs names
- Fsname exists in the configuration logs entries

New `lctl` commands for fork/erase configuration logs

- `lctl fork_lcfg <fsname> <new_fsname>`
- `lctl erase_lcfg <fsname>`

Mount Snapshot Read-only – Not only “-o ro”

Any modification of ZFS snapshot can trigger backend failure/assertion

- Open ZFS dataset as readonly mode
- NOT start cross-servers sync thread, pre-create thread, quota thread
- Skip sequence file initialization, orphan cleanup, recovery

- Ignore `last_rcvd` modification
- Deny to create transaction
- Forbid LFSCK
- …

Userspace Interfaces – `lctl snapshot_xxx`

Functionality	Usage
Create snapshot	`lctl snapshot_create [-b | --barrier] [-c | --comment comment] <-F | --fsname fsname] [-h | --help] <-n | --name ssname> [-r | --rsh remote_shell][-t | --timeout timeout]`
Destroy snapshot	`lctl snapshot_destroy [-f | --force] <-F | --fsname fsname> [-h | --help] <-n | --name ssname> [-r | --rsh remote_shell]`
Modify snapshot attributes	`lctl snapshot_modify [-c | --comment comment] <-F | --fsname fsname> [-h | --help] <-n | --name ssname> [-N | --new new_ssname] [-r | --rsh remote_shell]`
List the snapshots	`lctl snapshot_list [-d | --detail] <-F | --fsname fsname> [-h | --help] [-n | --name ssname] [-r | --rsh remote_shell]`
Mount snapshot	`lctl snapshot_mount <-F | --fsname fsname> [-h | --help] <-n | --name ssname> [-r | --rsh remote_shell]`
Umount snapshot	`lctl snapshot_umount <-F | --fsname fsname> [-h | --help] <-n | --name ssname> [-r | --rsh remote_shell]`

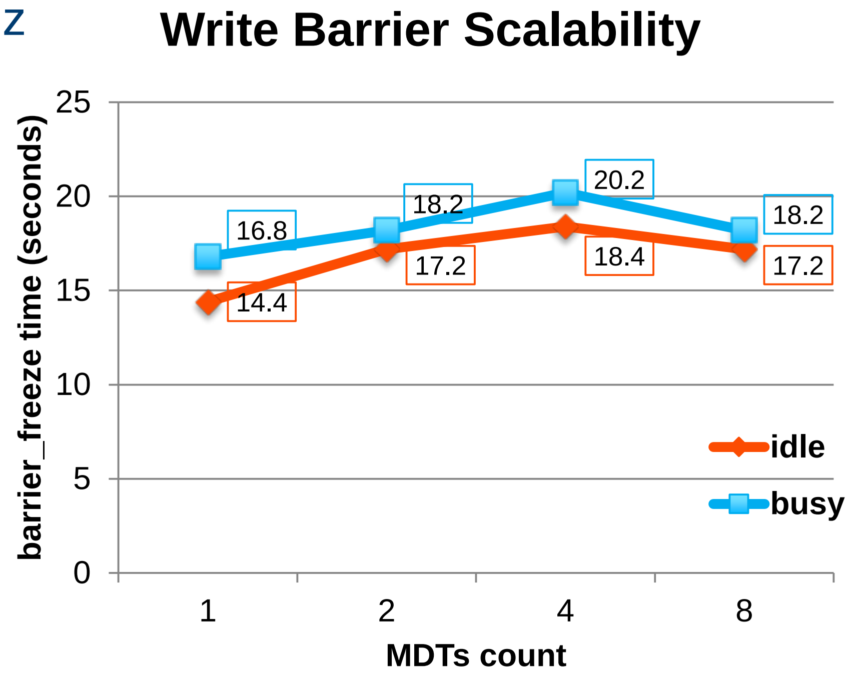
Snapshot tracking – `lsnapshot.log`

- Snapshot logs: **`/var/log/lsnapshot.log`**

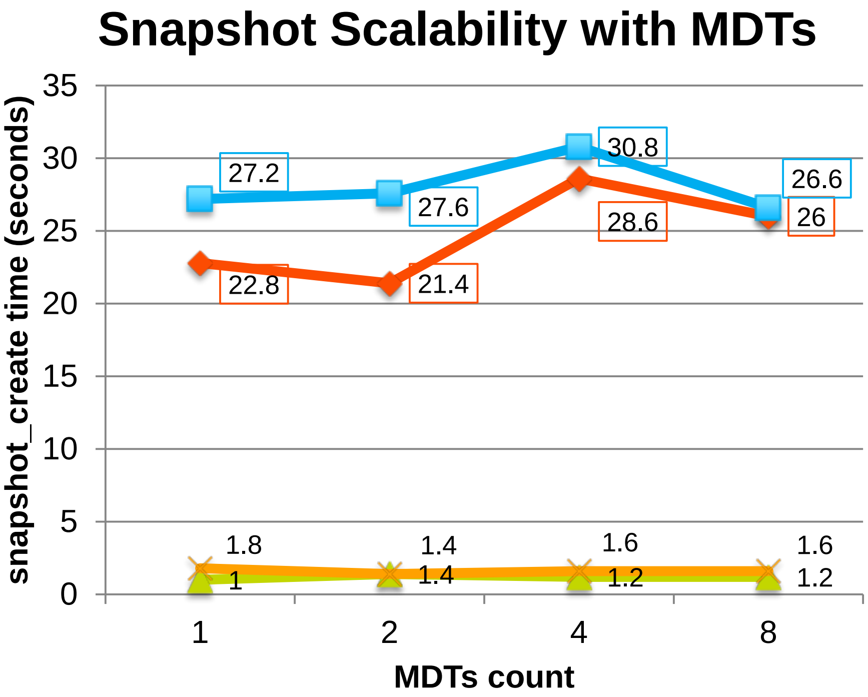
```
# cat /var/log/lsnapshot.log
Sun Mar 13 14:46:05 2016 (32688:jt_snapshot_create:1138:lustre:ssh): Create snapshot mysnapshot
successfully with comment <This is a test>, barrier <enable>, timeout <60>
Sun Mar 13 14:48:27 2016 (515:jt_snapshot_modify:1521:lustre:ssh): Modify snapshot mysnapshot
successfully with name <newsnapshot>, comment <The old name is mysnapshot>
Sun Mar 13 14:49:13 2016 (632:jt_snapshot_mount:2013:lustre:ssh): The snapshot newsnapshot is
mounted
Sun Mar 13 14:53:03 2016 (894:jt_snapshot_modify:1521:lustre:ssh): Modify snapshot newsnapshot
successfully with name <(null)>, comment <Change comment online>
Sun Mar 13 14:53:20 2016 (973:jt_snapshot_umount:2167:lustre:ssh): the snapshot newsnapshot have
been umounted
```

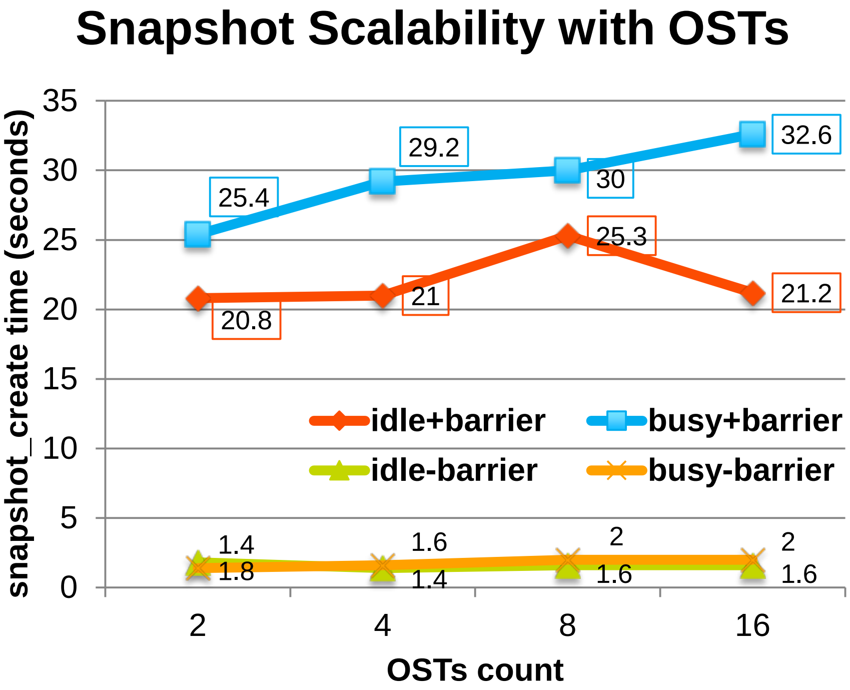
Write Barrier Scalability

- CPU: Intel® Xeon® E5620 @2.40GHz
  - 4 cores * 2, HT
- RAM: 64GB DDR3
- Network: InfiniBand QDR
- Storage: SATA disk arrays
- 2 MDTs per MDS
- 4 OSTs per OSS



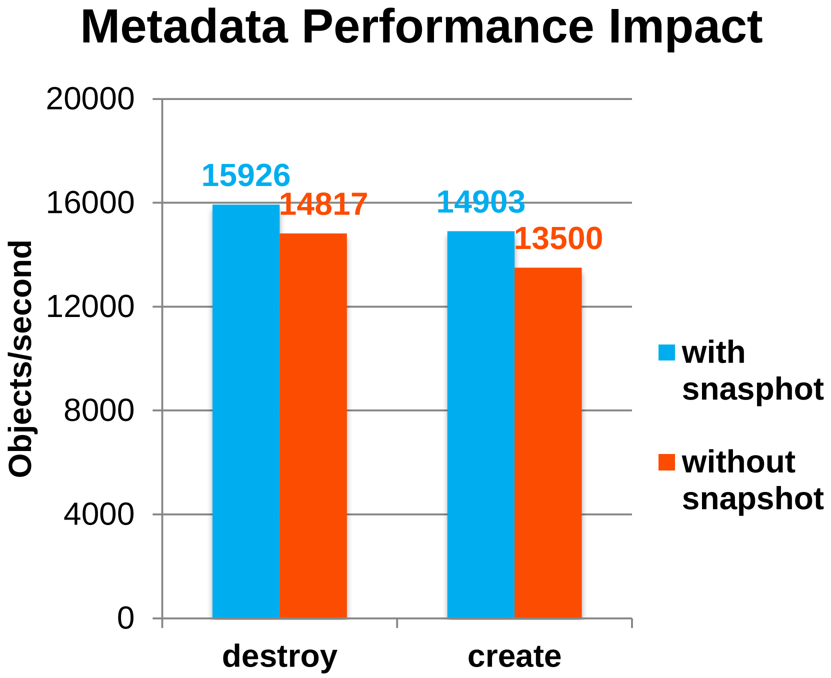
Snapshot I/O Scalability


Snapshot Scalability with MDTs


Snapshot Scalability with OSTs

I/O Performance With Snapshots

- Limited impact on metadata performance
  - Measured via mds-survey on single MDT
  - Slight benefit as changed blocks not freed
- No significant impact on I/O performance
  - Measure via obdfilter-survey on one OST
- Not Lustre* specific, ZFS is COW based



Next Steps for Snapshot Feature

- Phase I: targeted for Community Lustre 2.10 release landing
- Phase II: Lustre* integrated snapshot
  - Depends on users' requirements vs. other Lustre features, performance, etc.
  - More controllable and relatively independent solution
  - Reuse Phase I global write barrier
  - Integrate snapshot creation/mount/unmount into OSD
  - Identify files/objects in each snapshot as part of File Identifier (FID)

Legal Notices and Disclaimers

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit **http://www.intel.com/performance**.

Intel disclaims all express and implied warranties, including, without limitation, the implied warranties and merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing or usage in trade.

Intel, the Intel logo and others are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.

intel®