Lustre* HSM Scalability Work

John L. Hammond

Intel HPDD

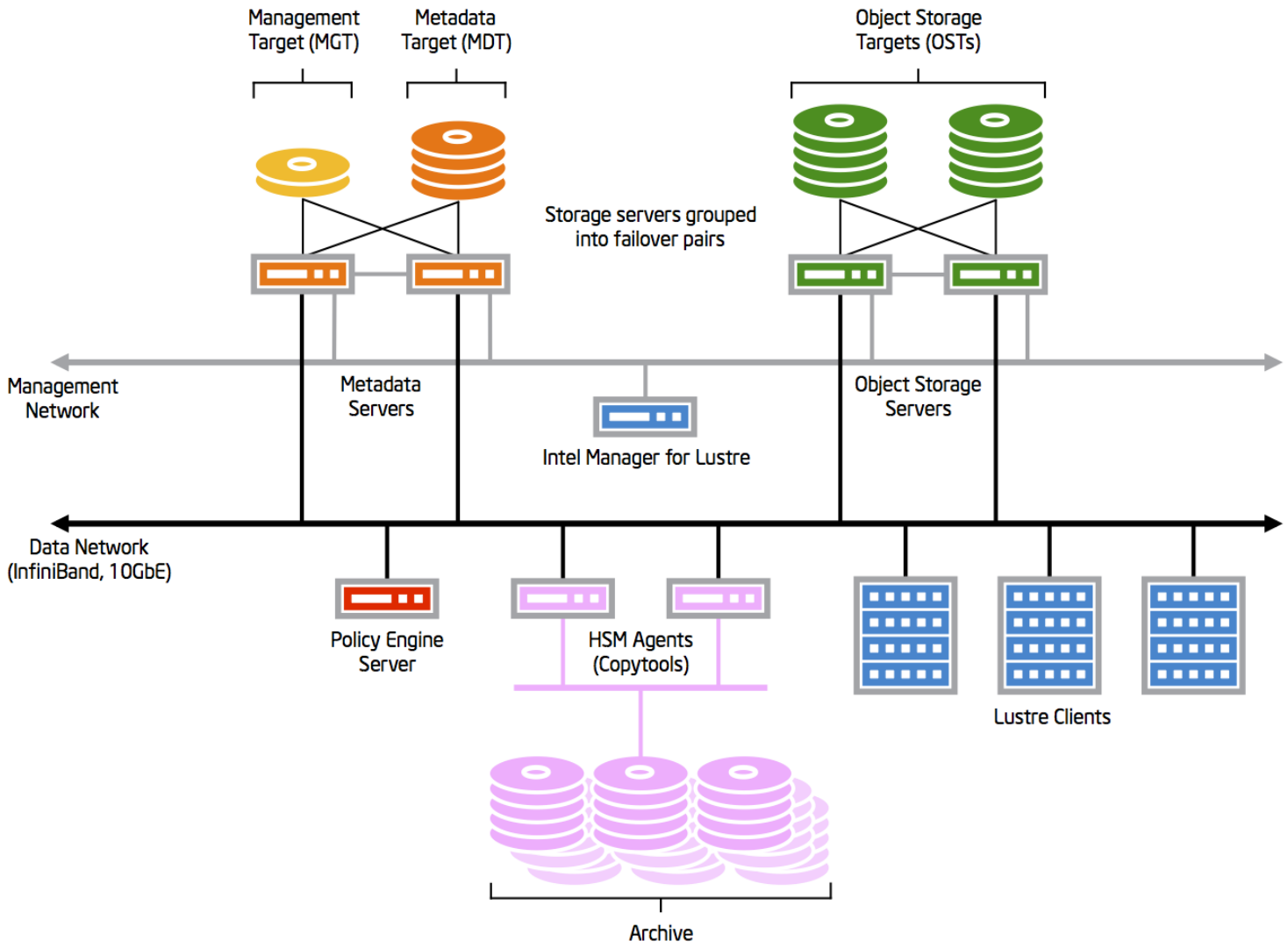
April 23 2018

*Other names and brands may be claimed as the property of others.

Outline

- Review of Hierarchical Storage Management
- Coordinator Issues
- Coordinator Scalability Work

HSM Architecture



HSM Coordinator (CDT)

- Part of a Lustre MDT
- Maintains a log on disk of HSM actions (waiting, started, done…)
- Manages layout locking (LDLM) for restores
- Maintains list of registered copytools
- Sends HSM actions (archive, restore, …) to copytools
- Receives progress RPCs from copytools

Coordinator Issue 1: Action Log Scalability

- Submitting new actions is slow when log is already large
  - Adding an action involves linear scan of log for existing compatible action
  - TODO Put in memory hash in front of log (easy)
  - TODO Use a better data structure like index (harder)
- Log update requires linear scan
  - Cache action record locations (LU-9312)
- Compound ID prevents batching
  - Same compound ID assigned to all requests in a list at submit
  - Only requests with same Compound ID submitted together to CT

Coordinator Issue 2: LDLM Locking

LDLM locking used unnecessarily (and inconsistently):

- HSM state needs protection but local locking would suffice
- Avoid interactions between HSM and xattr caching (for example)
- Clients access HSM state through specialized RPCs (not xattrs)
- LDLM Layout Locking still needed for release and restore

Benefits:

- Faster updates (archive start, archive complete, set dirty)
- Updates not blocked by unresponsive clients

(MDT and) HSM Coordinator Progress (1)

- LU-7988 CDT scalability and stability (2 changes in 2.10, 8 in progress)
- LU-9803 handle early requests vs CT registering
- LU-9927 bypass quota enforcement for HSM release (in 2.10)
- LU-9312 HSM coordinator llog scalability (3 in 2.10, 1 in progress)
- LU-9338 cache agent record locations (1 in progress)
- LU-9385 remove XATTR locking from mdt_add_dirty_flag() (in 2.10)
- LU-9403 prevent HSM leak on re-archive (in 2.10)
- LU-9404 set HSM xattr only when needed (in 2.10)
- LU-9464 use OBD_ALLOC_LARGE() for hsm_scan_data (in 2.10)
- LU-9540 cache HSM actions in memory by FID (in progress)

(MDT and) HSM Coordinator Progress (2)

- LU-10383 deprecate HSM compound id
  - More HSM actions per CDT to CT request
  - Landed to 2.12 (6 change series)
- LU-10699 replace HSM actions llog with an index
  - In progress for 2.12
  - Some DT index work required
    - In-place record update support
    - DT index API does not support variably sized records
    - Minor issues

Future Proofing (around llog to index)

- Archive id (1,2,3…31 to u32/u64/UUID/…)
  - UUID stored in '.' file in root of archive
- Identifier of file in archive (FID to ?)
  - Using FID creates issues for issues for import, disaster recovery, MDT migration, …


intel
Software