Whamcloud

HSM Coordinator Bypass

John Hammond, Whamcloud

LAD'18

DDN® STORAGE

Motivation: HSM Coordinator Scalability

Whamcloud

Old news/Good news/Bad news:

0. HSM Coordinator uses llog as a persistent queue for requested/ongoing/completed HSM actions (archive, restore, remove, cancel). Note that coordinator is not involved in release.

1. HSM Coordinator Scalability much improved since initial HSM feature landings
   - Actions log entries updated in place
   - Mapping of HSM action cookie to actions log entry location cached in RAM

2. Adding N actions to coordinator log is O(N^2) in worst case
   - Linear scan of actions required to determine if a given file already has an action queued
   - Actions llog is the only “thing” that knows this information

Scalability Issues in Production

Whamcloud

Too many archive requests in queue

- Initiated by Robinhood (say)
- No task or user waiting for completion
- Not latency critical (usually)

Compute job or interactive user tries to access released file triggering restore

- Blocks reads, writes waiting on restore
- O(#{queued archive requests}) to insert in llog
- Goes to back of actions llog (no QOS implemented in coordinator, see LU-8324)
- HSM unaware jobs might access file1 (block on restore), access file2 (block on restore), …

Options for HSM Coordinator Scalability

1. Replace HSM Actions llog with Index
2. Re-implement HSM Coordinator in Userspace

2b. HSM Coordinator Bypass for Archive, Remove

3. Store HSM Action State in File Extended Attribute

2. Re-implement HSM Coordinator in Userspace

Whamcloud

"Punt to userspace"

MDT side HSM RPC handlers become thin wrappers.

- Upcall to userspace or write to udev style event

Benefits:

- Flexible (e.g. use SLURM to schedule low and high priority actions)
- Can be thin for some actions
- Easier to modify/debug/… coordinator
- Leverage existing DBs/DBCs
- Everybody gets to write their own coordinator

2. Re-implement HSM Coordinator … (continued)

Whamcloud

Challenges:

- Layout locks for HSM restore require special handling
- Lots of new moving parts (things that admins can forget to turn on, … forget to monitor, …)
- Lots of new interfaces (places for things to get lost in the mail, …)
- Lots of Lustre knowledge copied to Userspace (maybe)
- Extensive discussion about proper language to use
- May require multiple message queues plus database
- Everybody gets to write their own coordinator

2b. Coordinator Bypass for Archive and Remove

LU-10968 hsm: add archive and remove upcall handling

https://review.whamcloud.com/#/c/32212/ +1KLOC Prototype

Offer upcalls for archive and remove to be invoked on the MDT which allow bypassing of the coordinator and better scheduling of archives and removes.

Lots of lower priority (less latency critical) operations (archive and remove) do not degrade performance of higher priority (more latency critical) operations (restore).

Includes new “half copytool” to support upcalled actions.

2b. Coordinator Bypass: MDT Side Upcalls

Whamcloud

```
lctl set_param mdt.*.hsm.upcall_mask=ARCHIVE
lctl set_param mdt.*.hsm.upcall_path=/my/fine/lhsm_mdt_upcall
```

MDS_HSM_REQUEST RPC handler checks to see if requested action type is in upcall mask, if so then invokes upcall and waits for it to complete.

Multiple actions (all of same type) may be passed to upcall. Invocation is of the form:

```
lhsm_mdt_upcall ARCHIVE scratch 1 0 "" [0x200000400:0x1:0x0]...
```

MDT RPC handler is waiting so the upcall shouldn't do too much and definitely shouldn't access Lustre.

- Insert request in DB/persistent queue/... and return.

2b. Coordinator Bypass: Workers

Whamcloud

Coordinator bypass needs modifications to existing copytool or a new copytool:

- Invoked externally (actions not received from Lustre KUC).
- Needs to call modified HSM Progress ioctls & RPCs for begin and end of action.

Divided into two parts:

1. The part I write:

   `lfs hsm_upcall` (new subcommand of existing lfs command)

2. The part I (or you) write:

   `lhsm_worker_posix` (site specific "half copytool", could be a shell script)

2b. … `lfs hsm_upcall`

Whamcloud

Invoked by external action queue consumer.

```
lfs hsm_upcall lhsm_worker_posix ARCHIVE scratch 1 0 "" [0x200000400:0x1:0x0]...
```

- Opens the Lustre file to by archived.
- Sends initial HSM Progress RPC to MDT (bypasses coordinator).
- For each FID invokes

  ```
  lhsm_worker_posix ARCHIVE scratch 1 0 "" [0x200000400:0x1:0x0]
  ```

  with stdin opened to the file to be archived.
- `lhsm_worker_posix` copies file contents from Lustre to archive.
- `lfs hsm_upcall` waits for `lhsm_worker_posix` to complete and sends final HSM Progress RPC to MDT (bypasses coordinator).

Whamcloud

Questions?