Demonstration Milestone Completion for the LFSCK 2 Subproject 3.2 on the Lustre* File System FSCK Project of the SFS-DEV-001 contract.

Revision History

Date	Revision	Author
26/02/14	Original	R. Henwood
13/03/14	Clarifications	R. Henwood

Introduction

The following milestone completion document applies to Subproject 3.2 – LFSCK 2: MDT-OST Consistency. This project is recorded in the OpenSFS Lustre software Development contract SFS-DEV-001 agreed August 30, 2011.

The LFSCK 2: MDT-OST Consistency code is functionally complete. The purpose of this Milestone is to verify the code performs acceptably in a production-like environment. In addition to completing all the Test Scenarios (demonstrated for the Implementation Milestone,) LFSCK 2: MDT-OST Consistency Performance will be measured as agreed on the ticket LU-3423.

All tests were executed on the OpenSFS Functional Test Cluster. Details of the hardware are available in Appendix A. For all the tests, Lustre software Master with LFSCK 2 patches was used.

Correctness

1. sanity-lfsck.sh
   Test will be executed within Autotest and results automatically recorded in Maloo. Test will be automatically completed, triggered by a Gerrit check-in with commit message "Test-Parameters: envdefinitions=ENABLE_QUOTA=yes mdtcount=2 testlist=sanity-lfsck". All test cases must pass.
2. sanity-scrub.sh
   Test will be executed within Autotest and results automatically recorded in Maloo. Test will be automatically completed, triggered by a Gerrit check-in with commit message "Test-Parameters: envdefinitions=ENABLE_QUOTA=yes testlist=sanity-scrub". All test cases must pass.
3. Standard review tests
   The standard collection of review tests (currently including sanity, sanityn, replay-single, conf-sanity, recovery-small, replay-ost-single, insanity, sanity-quota, sanity-sec, lustre-rsync-test, lnet-selftest, and mmp) will be executed within Autotest and results automatically recorded in Maloo. Tests will be automatically completed, triggered by a Gerrit check-in. All test cases should pass except for some known test failures unrelated to the LFSCK functionality.

Result

This test has been completed successfully and the results are recorded in the Implementation milestone:
http://wiki.opensfs.org/images/e/ee/LFSCK_MDT-OSTConsistency_Implementation.pdf

Demonstration context

All tests require a populated directory on the file system. The directory will be created and populated with the following properties:

1. Create 'L' test root directories. 'L' is equal to the MDT count. The directory in the root 'dir-X' is located MDT-X.
2. Set default stripe size as 64KB. Set the default stripe count as 'M'.
3. Create 'N' sub-directories under each of the test root directories.

4. Under each sub-directory create 80K files. Each file is 64 * 'M' KB in size.
5. Once created, the populated directory structure with N sub-directories (described above) is known in the procedure below as PopD_n.

Measure performance of LFSCK 2 against a single MDT device without inconsistencies.

1. Install Lustre Master with LFSCK patches.
2. Run LFSCK on each PopD_n with:
3. L={1}
4. M={1, 2, 4, 8}
5. N={4, 8, 16, 32}.
6. LFSCK must run at full speed, without any Lustre file system system load.

   ```
   lctl lfsck_start -M ${fsname}-MDT0000 -t namespace
   ```
7. Record the following
   a. Start time, end time
   b. lfsck_namespace statistics before and after the run (lctl get_param -n mdd.${fsname}-MDT0000.lfsck_namespace)

Result

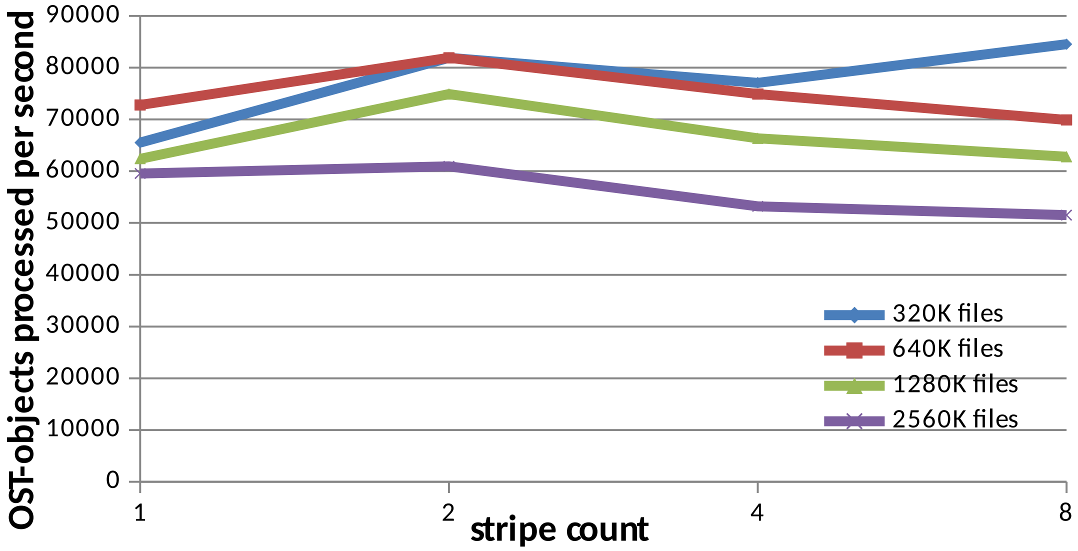

Figure 1: LFSCK 2: MDT-OST consistency running against a single consistent MDT

Checking a completely consistent file system performs well and compares favorably with previous measurements of LFSCK. For each of the varying stripe count measurements consistent performance is observed. Varying the total number of files in the tested directory shows a measurable decrease in performance.

Measure performance of LFSCK 2 against a single MDT device with inconsistencies.

1. Install Lustre Master with LFSCK patches.
2. Run LFSCK on each PopD_n with:
3. L={1}
4. M={1, 2, 4, 8}
5. N={4, 8, 16, 32}.
6. On the OSS, set fail_loc to skip the XATTR_NAME_FID set to simulate the case of MDT-OST inconsistency.
7. LFSCK must run at full speed, without any Lustre file system system load.
   lctl lfsck_start -M ${fsname}-MDT0000 -t namespace
8. Record the following
   a. Start time, end time
9. lfsck_namespace statistics before and after the run (lctl get_param -n mdd.${fsname}-MDT0000.lfsck_namespace)

Result

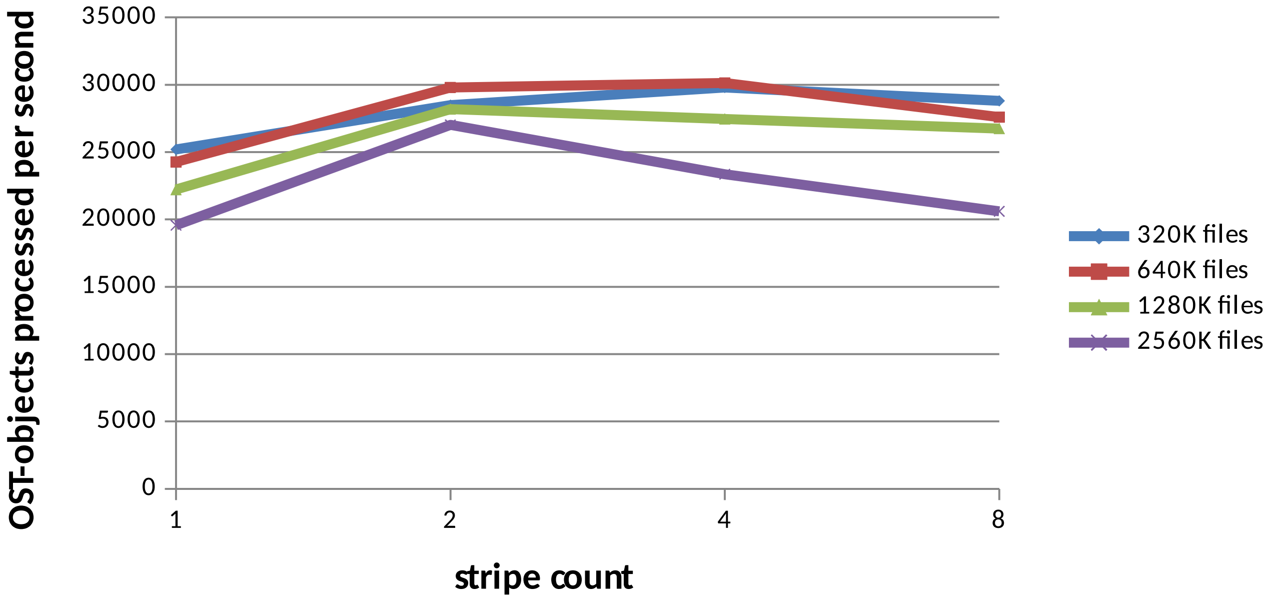

Figure 2: LFSCK 2: MDT-OST consistency repairing dangling references on a single MDT with varying stripe count

Checking and fixing an inconsistent file system performs well. For each of the varying stripe count measurements consistent performance is observed. Varying the total number of files in the tested directory shows a small but measurable decrease in performance. Comparing scanning a consistent file system (previous result) with fixing inconsistencies shows a measurable reduction in performance, as one would expect.

Impact of LFSCK on small file create performance on a single MDT without inconsistencies

1. Install Lustre Master with LFSCK patches.
2. Run LFSCK on each PopD_n with:
3. L={1}.
4. M={8}
5. N={2}
6. Populate each sub-directory with 1M files..
7. Run LFSCK with speed limited to {20%, 40%, 60%, 80%, 100%} of full speed.
8. Simultaneously, use md_test to create an additional 1M files and measure the create performance.

Result

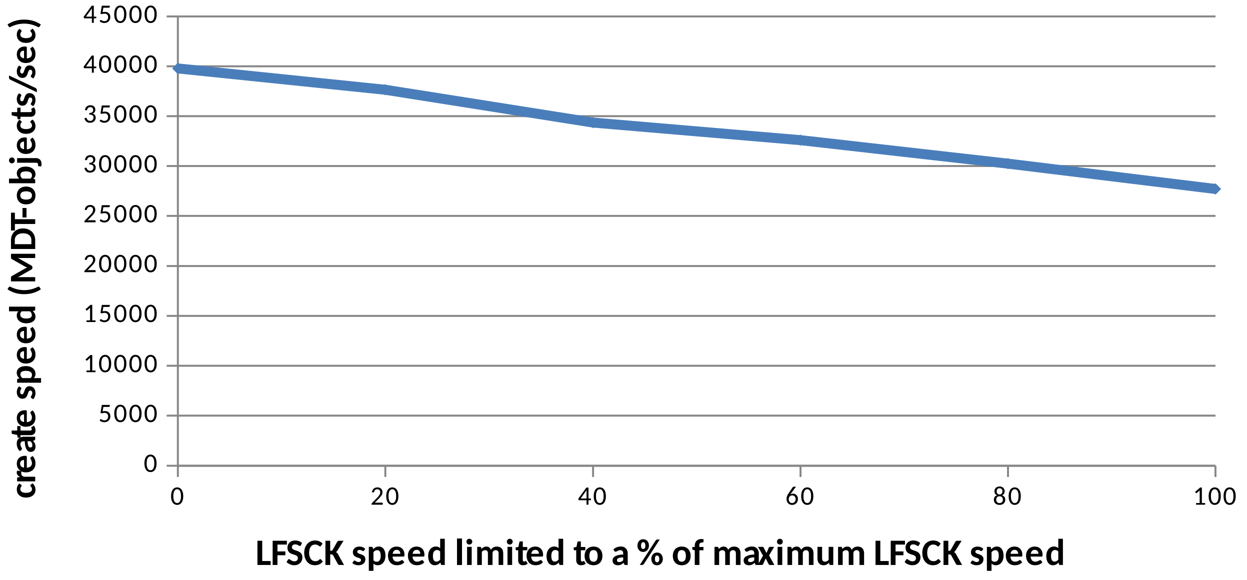

Figure 3: LFSCK 2: MDT-OST consistency impact on create performance.

Creating files on an consistent file system that has LFSCK currently executing performs well. As a general result, as LFSCK scanning speed is increased file creates become slightly slower. This is consistent with expected behavior. It is worth noting that even at 100% LFSCK speed the performance is still a large fraction of the performance without LFSCK running.

Multiple MDT Demonstration

LFSCK 2 is tested against multiple MDTs. Configurations with one, two, three, and four MDTs are created. This is effectively the Demonstration Context above, with values of L={1,2,3,4}.

Impact of LFSCK 2 running against multiple MDT devices without inconsistencies.

1. Install Lustre Master with LFSCK patches.
2. Run LFSCK on each PopD_n with:
3. L={1,2,3,4}
4. M={1, 2, 4, 8}
5. N={4, 8}.
6. LFSCK must run at full speed, without any Lustre file system system load.
   lctl lfsck_start -M ${fsname}-MDT0000 -t namespace
7. Record the following
   a. Start time, end time
8. lfsck_namespace statistics before and after the run (lctl get_param -n mdd.${fsname}-MDT0000.lfsck_namespace)Install Lustre Master with LFSCK patches.

Result

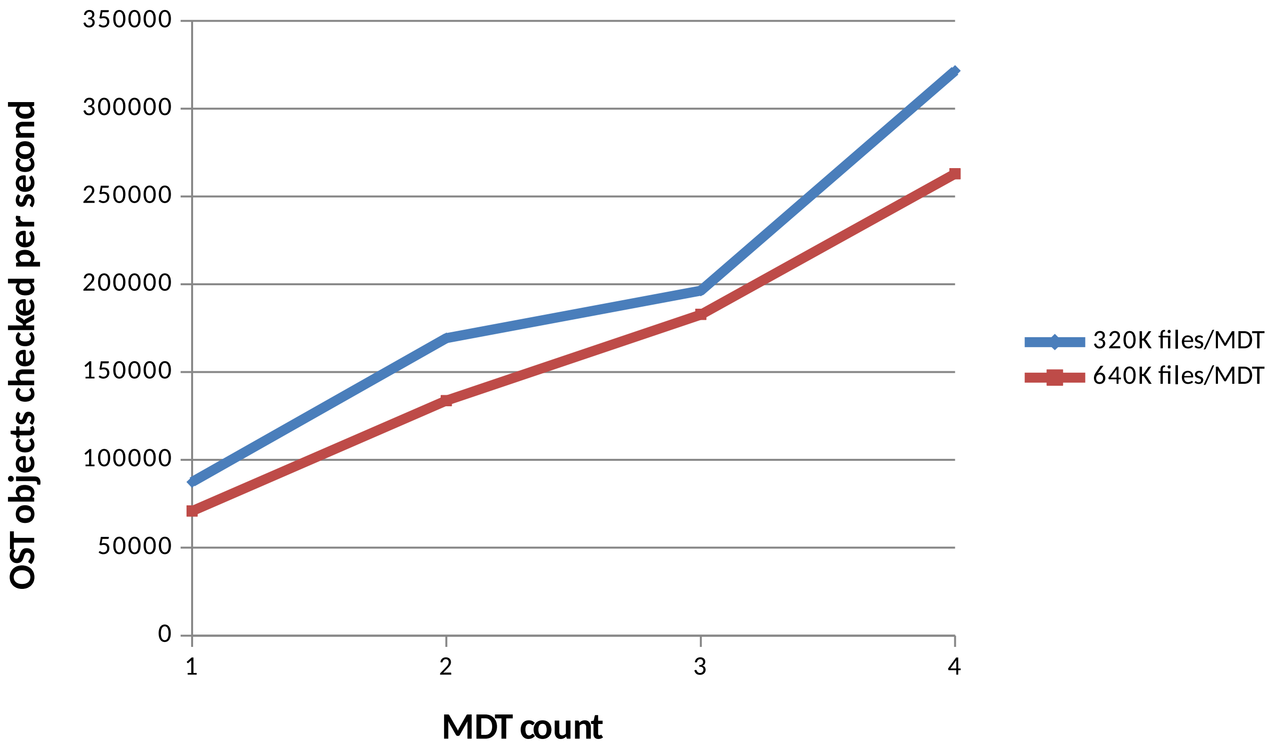

Figure 4: LFSCK 2 MDT-OST consistency performance on multiple consistent MDTs

LFSCK 2 works across multiple consistent MDTs. The scanning performance against a consistent file system across multiple MDTs appears to scale as would be expected. It can also be observed that the result for a single MDT is similar to the results previously recorded in this document.

Impact of LFSCK 2 running against multiple MDT devices during inconsistency repair.

1. Install Lustre Master with LFSCK patches.
2. Run LFSCK on each PopD_n with:
3. L={1,2,3,4}.
4. M={8}
5. N={2,4}
6. Populate each sub-directory with 1M files.
7. Run LFSCK with speed limited to {20%, 40%, 60%, 80%, 100%} of full speed.
8. Simultaneously, use md_test to create an additional 1M files and measure the create performance.

Result

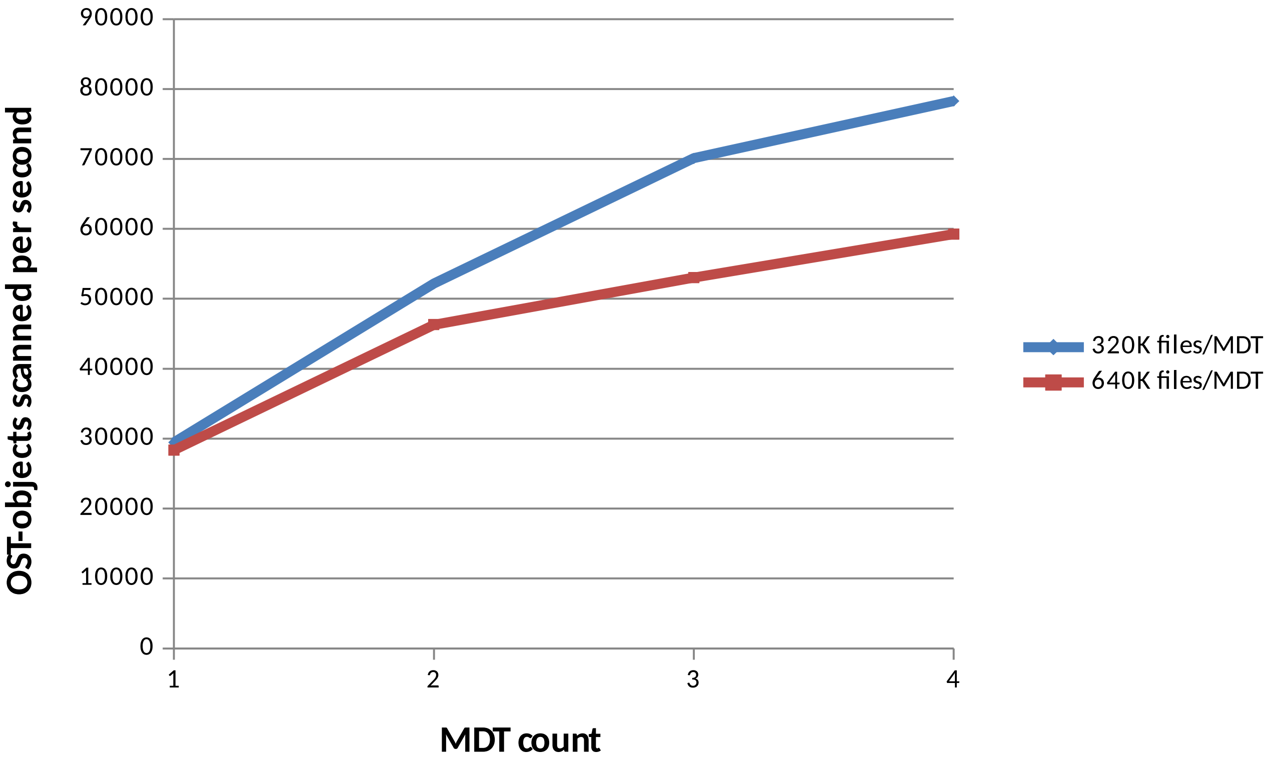

Figure 5: LFSCK 2 MDT-OST repair of dangling references on multiple MDTs with varying stripe count

LFSCK 2 works across multiple inconsistent MDTs. The scanning performance against an inconsistent file system with multiple MDTs does not apparently show linear scaling for 640K files/MDT and does show linear scaling for 320K files/MDT from one to three MDTs. It can also be observed that the result for a single MDT is similar to the results previously recorded in this document.

Conclusion

LFSCK 2: MDT-OST consistency has successfully completed both functional Acceptance and Performance tests. The performance results recorded herein illustrate performance expectations are met or exceeded during online operation and under load. In addition, LFSCK 2 has been shown to meet or exceed expectations running in a single and multiple MDT environment.

Appendix A: OpenSFS Functional Test Cluster specification

client

- (2) Intel E5620 2.4GHz Westmere (Total 8 Cores)
- (1) 64GB DDRIII 1333MHz ECC/REG - (8x8GB Modules Installed) * (1) On Board Dual 10/100/1000T Ports
- (8) Hot Swap Drive Bays for SATA/SAS
- (6) PCi-e Slots 8X
- (3) QDR 40GB QSFP to QSFP iB Cables
- (3) Mellanox QDR 40GB QSFP Single Port

OSS server

- (1) Intel E5620 2.4GHz Westmere (Total 8 Cores)
- (1) 32GB DDRIII 1333MHz ECC/REG - (8x8GB Modules Installed) * (1) On Board Dual 10/100/1000T Ports
- (1) On Board VGA
- (1) On Board IPMI 2.0 Via 3rd. Lan
- (1) 500GB SATA Enterprises 24x7
- (1) 40GB SSD OCZ SATA
- (8) Hot Swap Drive Bays for SATA/SAS
- (6) PCi-e Slots 8X
- (3) QDR 40GB QSFP to QSFP iB Cables
- (3) Mellanox QDR 40GB QSFP Single Port

MDS server

- (1) Intel E5620 2.4GHz Westmere (Total 8 Cores)
- (1) 32GB DDRIII 1333MHz ECC/REG - (8x8GB Modules Installed) * (1) On Board Dual 10/100/1000T Ports
- (1) On Board VGA
- (1) On Board IPMI 2.0 Via 3rd. Lan
- (1) 500GB SATA Enterprises 24x7
- (1) 40GB SSD OCZ SATA
- (8) Hot Swap Drive Bays for SATA/SAS
- (6) PCi-e Slots 8X
- (3) QDR 40GB QSFP to QSFP iB Cables
- (3) Mellanox QDR 40GB QSFP Single Port