Demonstration Report for the Remote Directories Subproject on the DNE Project of the SFS-DEV-001 contract.

Revision History

Date	Revision	Author
04/25/13	Original	R. Henwood
05/06/13	Rev 1: Clarified results, added raw data, described scaling context.	R. Henwood

Contents

Introduction

This document describes demonstration of sub-project 2.1 – Remote Directories – within the OpenSFS Lustre Development contract SFS-DEV-001 signed 7/30/2011. The DNE1: Remote Directories code is functionally complete. The purpose of this Milestone is to verify that the code performs acceptably in a production-like environment. In addition to achieving the Acceptance Criteria (recorded in the DNE1: Remote Directories Solution Architecture), DNE1: Remote Directories Performance will be measured as described below.

Method

Metadata performance scaling

To show metadata performance scaling, the execution of the performance measurements described herein is be repeated with varying MDS and MDT counts. Performance is measured with the following combinations of metadata servers and targets:

- 1 MDS with 1 MDT attached.
- 2 MDSs each with 1 MDT attached.
- 3 MDSs each with 1 MDT attached.
- 4 MDSs each with 1 MDT attached.
- 1 MDS with 2 MDTs attached.
- 2 MDSs each with 2 MDTs attached.
- 3 MDSs each with 2 MDTs attached.
- 4 MDSs each with 2 MDTs attached.

System configuration

Demonstration took place on the LLNL Hyperion testbed. For each test, two OSSs were configured, each with 4 OSTs. A total of 100 clients were used. All machines in the cluster are x86_64 architecture running CentOS 6 and had Lustre 2.3.64 installed. Hardware details are provided in Appendix A.

Metadata Target Performance

Measure the metadata performance of a single underlying MDT using the `mds-survey` tool, which injects a test load directly at the MDD layer on the MDS and isolates the performance of the Lustre MDD/LOD/OSD metadata stack from the network and RPC performance. This provides an upper limit

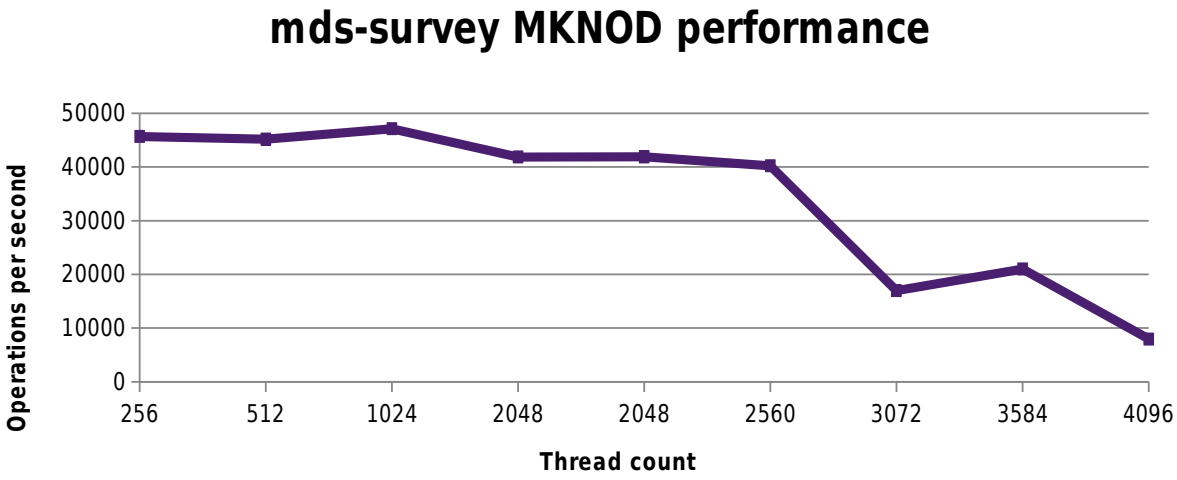
for the metadata operation performance for the lower layers of the MDS code and underlying storage subsystem.

Scaling demonstration

The metadata scaling performance setup was as follows. 100 directories were constructed on MDT0. The performance was measured using mdsrate, with a single thread per directory, and the result recorded as a single MDT. An additional MDT (MDT1) was created. The total of 100 directories were now distributed evenly across both MDTs (50 on each). The performance was measured using mdsrate and the result recorded as two MDTs. On the addition of each MDT, the total of 100 directories are distributed evenly across the pool of MDTs.

From an operations point of view, this experimental design results in a constant load being distributed evenly as more servers are added. The run-time of individual tests declines and the throughput increases with each additional MDT.

MDT Performance Results



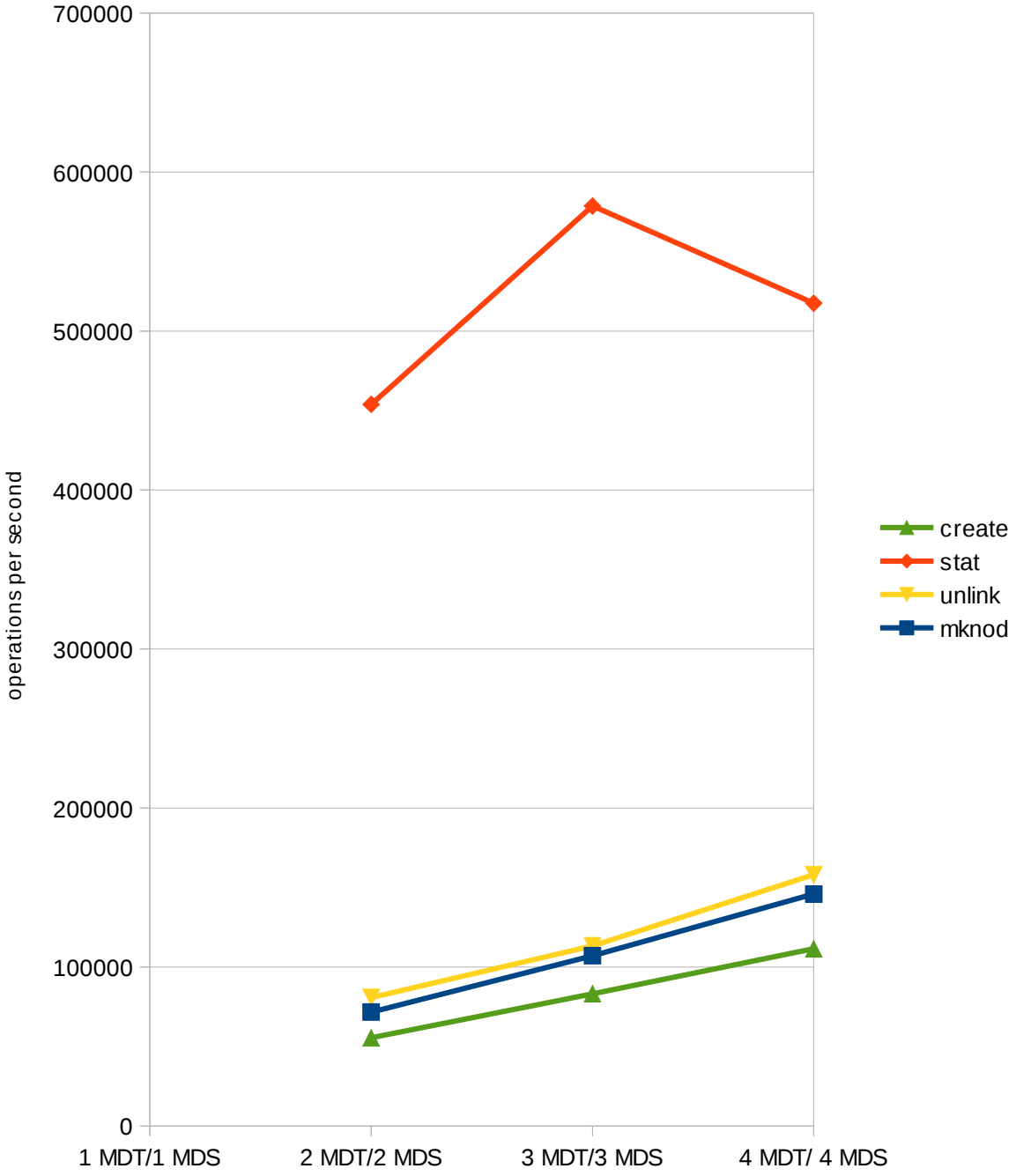
Performance of a single MDT is consistent up to approximately 2500 threads. From 2500 threads and beyond the performance rapidly drops off. MDS nodes are configured with a default maximum of 512 threads, and some are configured with up to 2048 threads, so this falloff should not be visible under normal usage. A rate of 45000 operations per second with mds-survey is acceptable performance on the given hardware.

Scaling performance results

All testing was completed against the tag 2.3.64 on the Hyperion system. The tool mdsrate was used to drive load against the multiple MDSs simultaneously. The mdsrate parameters are as follows:

```
mdsrate parameters: mdsrate --mntfmt='/p/l_wham%d' \
      --mntcount 2 –-{mknod|create|stat|unlink} \
      --mdtcount $MDT --dirfmt='xmds1R%d' \
      --nfiles 20000 --ndirs 100 --filefmt 'g%%d'
```

One metadata target per MDS



The first tests are run with a single MDT per MDS. The measurement of a single MDS with one MDT was reviewed it was concluded the measurement was erroneous. For this reason, this value is omitted from this figure.

The mknod test creates files on the MDT without allocating OST objects. This provides the upper limit of MDT performance for clients and avoids any performance impact from the OSTs. The performance of mknod does not show an increase from one to two MDTs. Adding MDTs after two appears to show linear scaling up to four MDTs.

The create test allocates a single OST object per MDT file and reflects the file creation behavior that would be used by most applications. The performance of create does not show an increase from one to two MDTs. Adding MDTs after two appears to show linear scaling up to four MDTs.

The stat test performs attribute lookups on the client from the MDS. Since stat operations are not modifying the filesystem, clients can send up to eight RPCs per MDT concurrently. The performance of stat does not appear to show linear scaling under this workload.

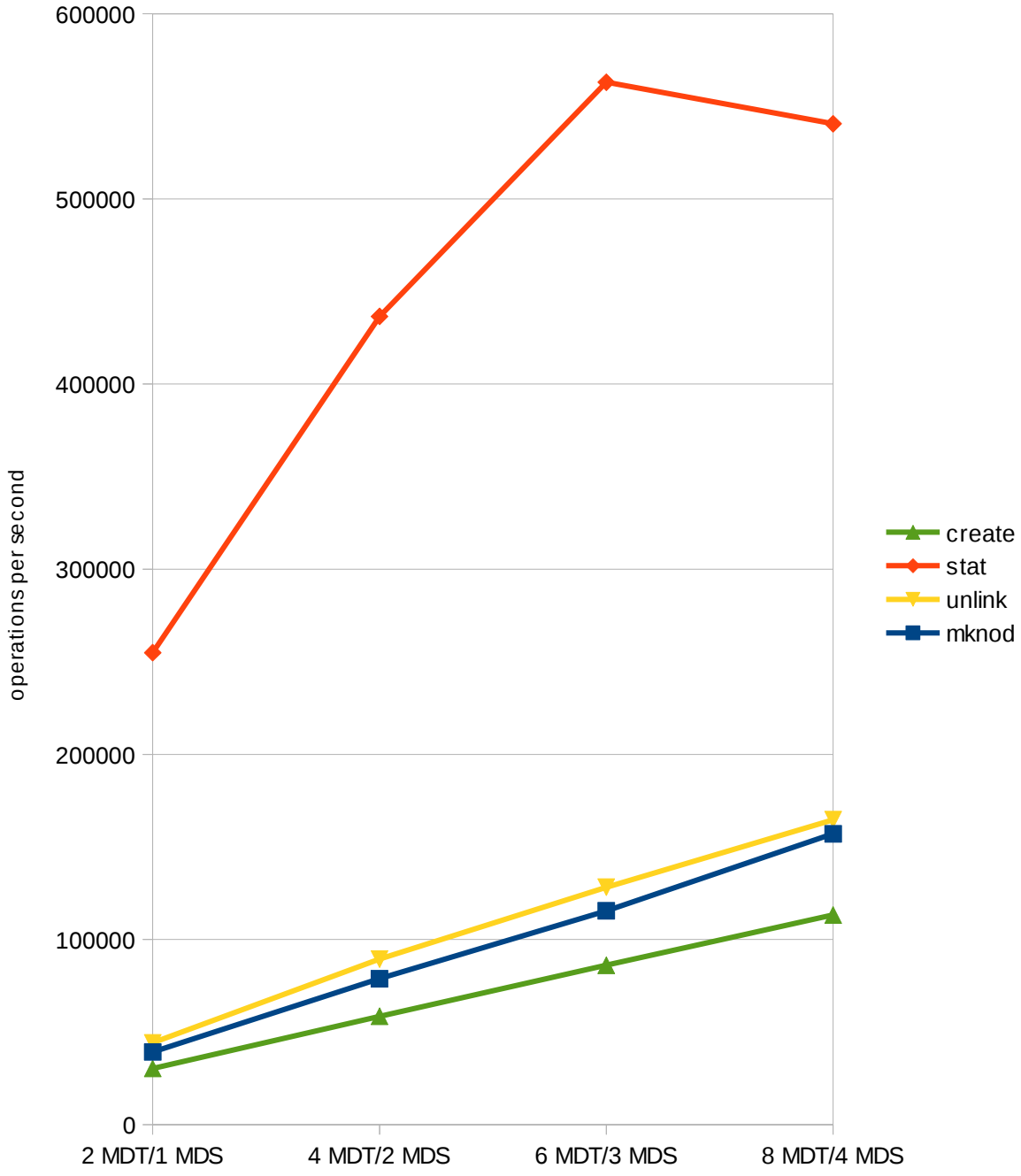
The unlink test deletes files from the MDT and reflects application-visible performance. The performance of unlink does not show an increase from one to two MDTs. Adding MDTs after two appears to show linear scaling up to four MDTs.

Discussion

The overall conclusion of this section of the work is linear scaling is observed with the addition of meta-data servers. There are, however, some issues that could be investigated further:

- The 1MDT/1MDS test results exceeds the mds-survey results on the same single node. This may be caused by the mds-survey tool itself consuming CPU resources on the MDS and negatively impacting the observed performance.
- The flat performance from 1MDT/1MDS to 2MDT/2MDS does not reflect the expected linear scaling that has been observed in prior test runs. One possible explanation is that the defined load for the system was not sufficient. Alternately, there may have been some anomaly in the test configuration during this testing interval.
- The stat performance does not show linear scaling. The large number of stat operations indicates that results were provided from the MDS cache instead of from disk. This is consistent with the experimental design. The high RPC rate may be saturating some component of the test environment, such as the network or client RPC rate.

Two metadata targets per MDS



In this test, each MDS is configured with two MDTs. The performance of create, unlink and mknod all show linear scaling with the addition of MDTs.

The stat performance increases with additional MDTs until six MDTs are present at which point it flattens out.

Discussion

The overall conclusion of this section of the work is linear scaling is observed with the addition of metadata servers. The stat performance does not show linear scaling beyond four MDTs. The large number of stat operations indicates that results were provided from the MDS cache. This is consistent with the experimental design. The flattening out after six MDTs may be a result of saturating the network or client RPC performance.

Conclusions

The Demonstration Milestone for DNE 1: Remote Directories has been successfully completed and linear scaling of metadata requests has been shown. Beyond this important result, a number of additional highlights can be identified:

- The create performance measured with mds-survey of approximately 45K IOPS is close to the performance measured by mdsrate of approximately 55K IOPS. This result increases confidence in the value of mds-survey results, which can be run without the need for a large number of clients to generate testing load.
- The absolute performance of a single metadata server is satisfactory.
- Two MDTs attached to a single MDT performs measurably better than a single MDT attached to a MDS – excluding the case of a single MDS. This effect may be even more noticeable if a large number of disk operations are required (e.g. stat from disk).

Appendix A: System specification of Hyperion DNE Demonstration platform

MDS server

- (1) Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz
- InfiniBand QDR network
- 65791756 KB
- Pci bus

MDT storage

- NetApp HBA controller
- RAID-1+0

OSS

- Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz
- Infinband QDR network
- 65791756 KB
- Pci bus
- RAID-6

Clients

- (1) Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz
- Infiniband QDR network
- PCI Bus
- 65791756 KB

Appendix B: One metadata targets per MDS, raw values

mknod

1 MDT/1 MDS	2 MDT/2 MDS	3 MDT/3 MDS	4 MDT/ 4 MDS
	59743	99506	139018
	72339	105143	117812
No valid result available.	75616	104348	158032
	76156	112801	158636
	74568	113838	156114
	71684	107127	145922

stat

1 MDT/1 MDS	2 MDT/2 MDS	3 MDT/3 MDS	4 MDT/ 4 MDS
	460525	573369	526597
	448522	574950	523034
No valid result available.	455557	575921	524111
	454626	584097	505964
	450068	584816	507457
	453860	578630	517433

unlink

1 MDT/1 MDS	2 MDT/2 MDS	3 MDT/3 MDS	4 MDT/ 4 MDS
	77565	91519	146169
	82912	107108	162856
No valid result available.	85491	126720	159247
	83317	128464	160572
	75004	112838	161154
	80858	113330	158000

create

1 MDT/1 MDS	2 MDT/2 MDS	3 MDT/3 MDS	4 MDT/ 4 MDS
	54702	68536	99257
	55698	86515	113263
No valid result available.	55241	87056	113825
	56087	85926	113512
	55754	87883	117413
	55496	83183	111454

The measurement of a single MDS with one MDT was reviewed and was judged to be erroneous. For this reason, this value is omitted from this figure.

Appendix C: Two metadata targets per MDS, raw values

mknod

2 MDT/1 MDS	4 MDT/2 MDS	6 MDT/3 MDS	8 MDT/4 MDS
37835	74719	114585	154482
39412	77988	108171	154668
39427	78022	118476	159256
39646	81667	118497	158621
39514	81429	117586	158078
39167	78765	115463	157021

stat

2 MDT/1 MDS	4 MDT/2 MDS	6 MDT/3 MDS	8 MDT/4 MDS
248523	440517	565252	536762
253228	438052	559807	533623
257601	438244	557132	544516
257425	438003	564927	542684
257728	427932	568146	545492
254901	436550	563053	540615

unlink

2 MDT/1 MDS	4 MDT/2 MDS	6 MDT/3 MDS	8 MDT/4 MDS
45817	87958	129788	163064
43895	88052	123917	163892
40946	89777	129077	163961
43672	90113	128904	165875
46228	90991	129229	166480
44112	89378	128183	164654

create

2 MDT/1 MDS	4 MDT/2 MDS	6 MDT/3 MDS	8 MDT/4 MDS
30390	57707	85812	106387
30337	58522	85838	113487
29986	58446	85616	115289
30325	58910	87319	115684
30272	58828	85570	115199
30262	58483	86031	113209