Lustre Metadata Performance Testing Best Practice

Cheng Shao
cheng_shao@xyratex.com

xyratex

Outline

- Lustre metadata performance test overview
- Test methodology
- Lustre Metadata performance issue diagnosis

Overview

- Lustre metadata performance was less well-studied than data performance in the past

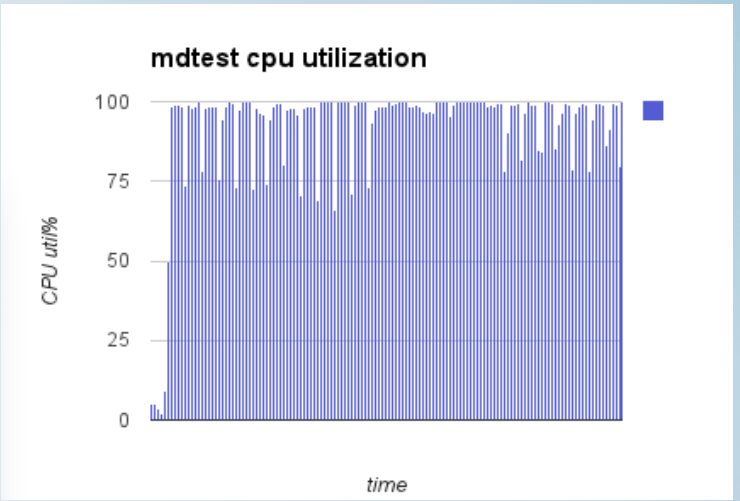
- More attentions on metadata performance lately
  - HPC applications generating huge number of tmp files
  - New application models
    - auditing
    - disaster recovery
    - NFS/CIFS server
  - Competition with other distributed/parallel file system and data storage solutions

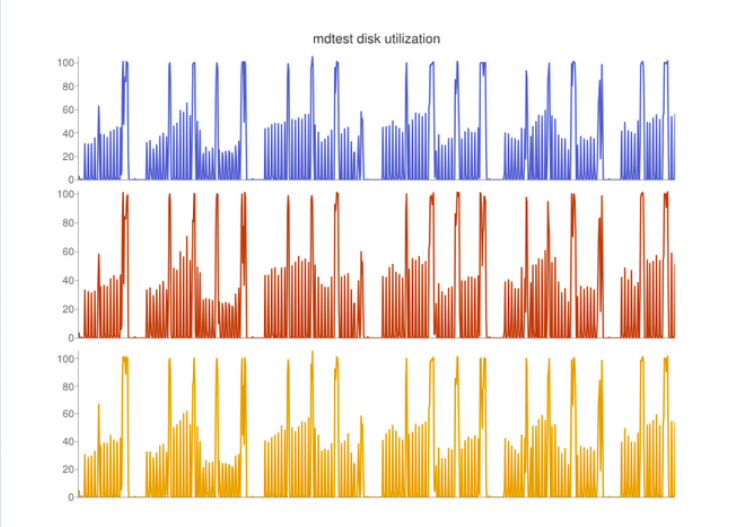
Test Methodology

- Metadata performance benchmarking tools
  - mdtest
  - mds-survey
  - mdsrate
  - metabench
  - bonnie++
  - SPECsfs
  - other home grown tools

Test Methodology

- MDS Hardware Configuration
  - backend storage
  - ldiskfs journal
  - CPU and memory
  - networks (LST)





Test Methodology

- mdtest specifics
  - number of tasks
  - number of files to eliminate caching effect
  - multiple iterations to eliminate outliers
  - shared mount vs. unique mount
  - '-a' and '-o' option to exclude OST impact
  - directory structure

Test Methodology

- Test Design
  - increase number of tasks per client node
  - increase number of client nodes while keeping tasks per node constant
  - monitoring and profiling during the test
  - list test factors and change one factor at a time
  - try to give possible explanation to the difference
  - testing with non-empty filesystems

Metadata Performance Issue Diagnosis

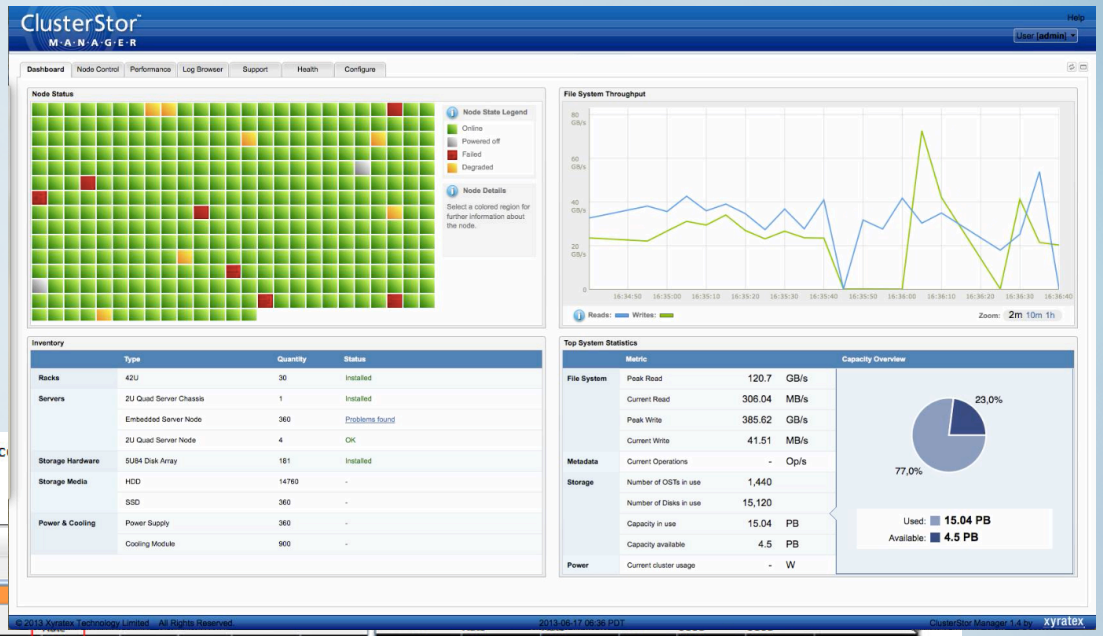
- Issue types
  - performance degradation
  - fail to meet certain performance target
  - end-user complains
- Diagnosis methods
  - profiling
  - Lustre stats
  - crash dump

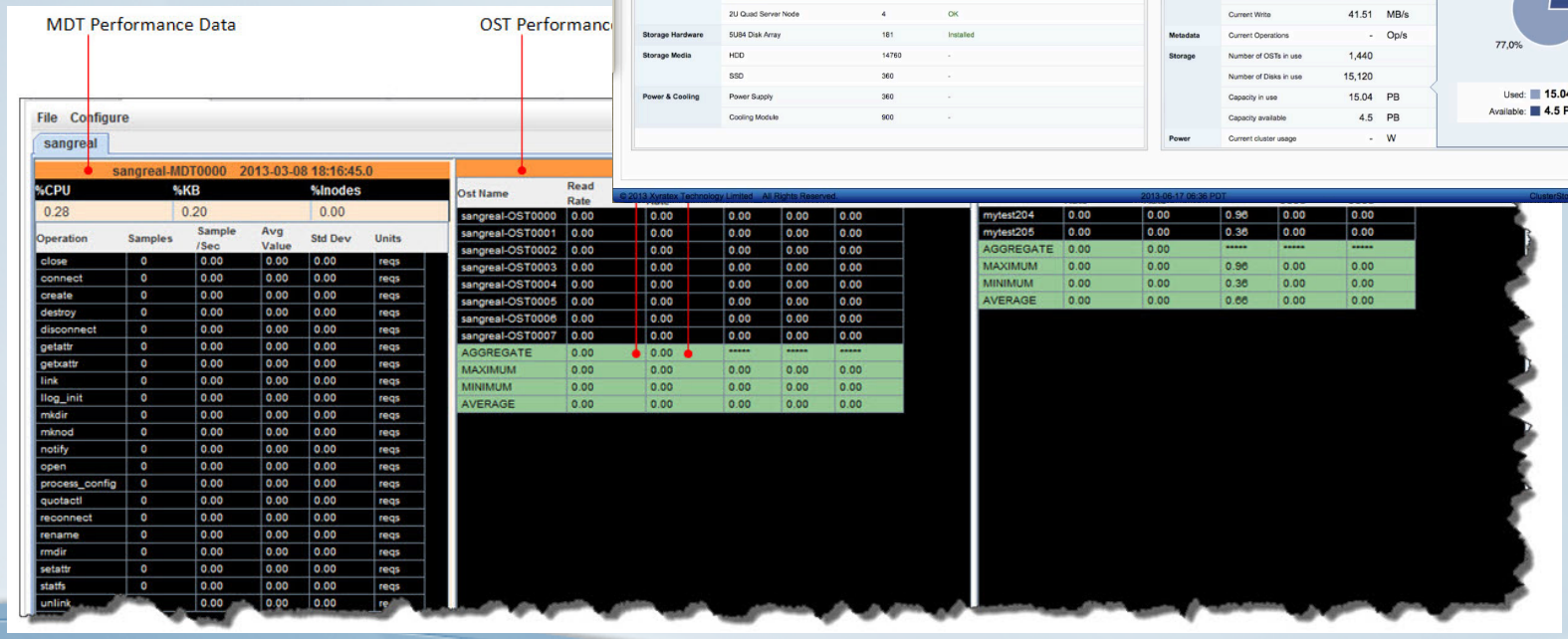
Metadata Performance Issue Diagnosis

- Profiling tools
  - load
    - iostat
    - perf / oprofile
  - latency
    - latencytop
    - systemtap

Metadata Performance Issue Diagnosis

- Lustre stats
  - stats
  - req_history
  - timeouts
  - ldlm stats





Metadata Performance Issue Diagnosis

- One example
  - metadata performance degradation observed in Lustre 2.3 clients, comparing with Lustre 1.8.8 client
  - benchmarking details: create files under a flat directory and then delete
  - creation rate is identical between 2.3 and 1.8.8 clients
  - deletion rate of 2.3 clients is 1/3 of 1.8.8 clients

Metadata Performance Issue Diagnosis

- Investigation
  - lucky enough to replicate in house
  - use systemtap to profile latencies of unlink call stack
  - sample script

```
global times[100]

probe module("lustre").function("ll_unlink").return {
        times["ll_unlink"] <<< (gettimeofday_us()-@entry(gettimeofday_us()))
}

probe module("mdc").function("mdc_unlink").return {
        times["mdc_unlink"] <<< (gettimeofday_us()-@entry(gettimeofday_us()))
}

probe module("mdc").function("mdc_reint").return {
        times["mdc_reint"] <<< (gettimeofday_us()-@entry(gettimeofday_us()))
}

probe end {
        foreach (func in times) {
                printf("\n\n==== %s stat ====\n", func)
                printf("total samples: %d, min %d, max %d, avg %d\n",
                        @count(times[func]),
                        @min(times[func]),
                        @max(times[func]),
                        @avg(times[func]))
                print(@hist_linear(times[func], 0, 45000, 400))
        }

        printf("\nbye now\n")
}
```

Metadata Performance Issue Diagnosis

- Investigation
  - comparing the profiling results of 1.8 and 2.3 clients
  - sample outputs (latency of ll_unlink)

1.8 clients

```
total samples: 8000, min 211, max 144987, avg 506
 value |-------------------------------------------------- count
     0 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@  7989
  2000 |                                                       6
  4000 |                                                       0
  6000 |                                                       0
       ~
 40000 |                                                       0
 42000 |                                                       0
 44000 |                                                       2
 46000 |                                                       0
 48000 |                                                       0
       ~
 54000 |                                                       0
```

2.3 clients

```
total samples: 8000, min 249, max 145504, avg 1539
 value |-------------------------------------------------- count
     0 |@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@  7391
  2000 |                                                      30
  4000 |                                                      47
  6000 |                                                      41
  8000 |                                                      47
 10000 |                                                      75
 12000 |                                                      37
 14000 |                                                      83
 16000 |                                                     104
 18000 |                                                      69
 20000 |                                                      25
 22000 |                                                      19
```

  - most part of the latency comes from clients waiting for responses from MDS

xyratex

Metadata Performance Issue Diagnosis

- Investigation
  - server side code path

```
mdt_reint_unlink
        mdt_object_find_lock // lock parent dir
        mdd_unlink           // actual unlink
```

  - profiling results

function	with 1.8 clients	with 2.3 clients
mdt_reint_unlink	304	1328
mdd_unlink	75	76
mdt_object_find_lock	51	1110

  - why does it take longer to lock the shared parent directory with 2.3 clients

Metadata Performance Issue Diagnosis

- Investigation
  - code review on the benchmark application.
  - file creation work flow
  - file unlink work flow
    - readdir to get next direntry
    - stat direntry to skip everything other than regular files
    - unlink the direntry
- Difference between Lustre 1.8 and 2.3
  - readdir chunk size gets increased from 1 page to 256 pages
  - lock cancellation takes longer on the client side

Metadata Performance Issue Diagnosis

- Conclusion
  - certain workflow in benchmarking tools or user applications may sometimes not be able to achieve the optimal metadata performance Lustre could offer
  - systemtap is proved to be an effective tool for latency profiling and performance diagnosis

Thank You

cheng_shao@xyratex.com