



Metadata Performance Evaluation Effort @ LUG 2014

Sorin Faibish, EMC
Branislav Radovanovic, NetApp
and MD BWG
April 8-10, 2014

OpenBenchmark Metadata Performance Evaluation Effort (MPEE) Team

- Leader: Sorin Faibish - EMC
- Members:
 - Branislav Radovanovic - NetApp
 - Richard Roloff - Cray
 - Cheng Shao, Wang Yibin - Xyratex
 - Keith Mannthey, Bobbie Lind – Intel
 - Gregory Farnum - Inktank

Metadata Performance Evaluation Effort Charter

- Build/select tools that will allow evaluation of File System Metadata performance and scalability
- The tools will help detect pockets of Metadata low performance in cases when users complain of extreme slowness of MD operations
- Benchmark tools will support: POSIX, MPI, and Transactional operations (for CEPH and DAOS)
- Address the very high end HPC as well as small and medium installations benchmark needs
- Tools applicable to Lustre and: CEPH, GPFS...

MPEE Proposed Tools

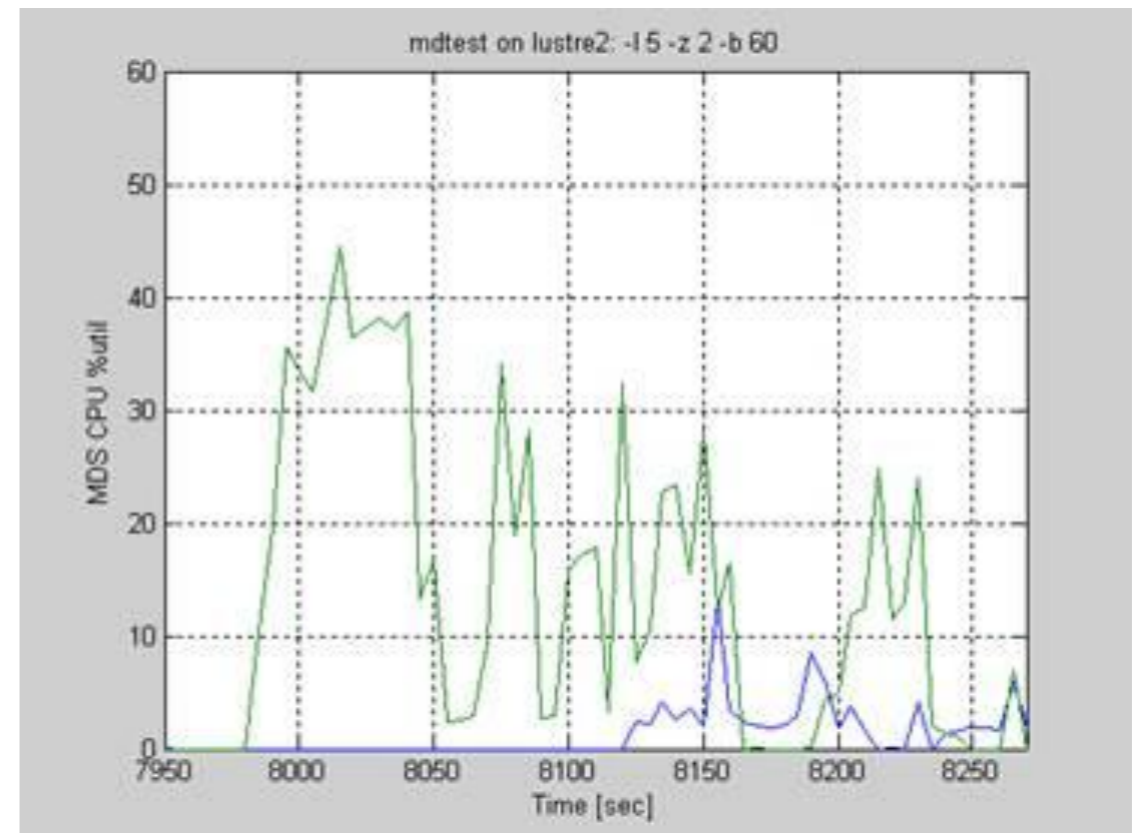
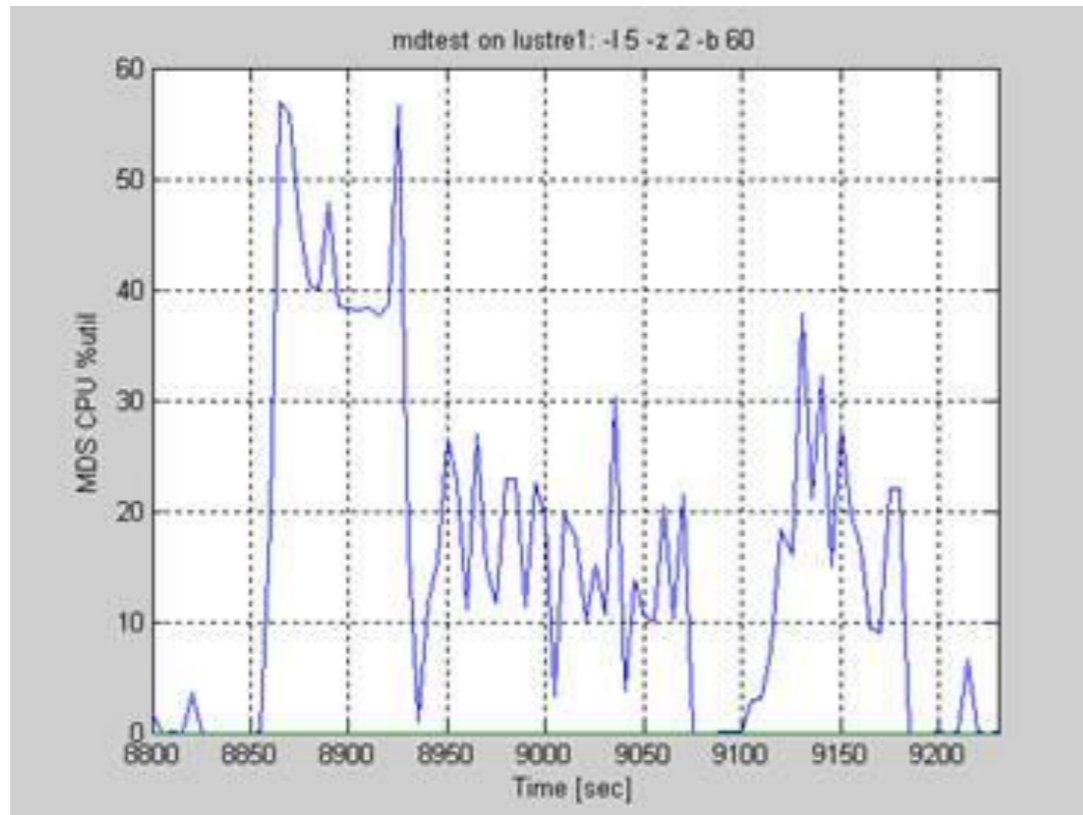
- The current proposed list of benchmarks:
 - **mdtest** – widely used in HPC
 - fstest - used by pvfs/OrangeFS community
 - **Postmark** and MPI version - old NetApp benchmark
 - **Netmist** and MPI version – used by SPECsfs
 - Synthetic tools – used by LANL, ORNL
 - **MDS-Survey** - Intel's metadata workload simulator.
 - Any known open source metadata tools used in HPC
 - Add new Lustre statistics specific to MD operations.

Testbed

- 2 Lustre FS with different characteristics:
 - ✓ Lustre1 using SAS based MDS (4+1) server and disk OST(6x(8+2) SATA
 - ✓ Lustre2 using 400GB SSD devices and SSD MDT/OST
- Monitored just block storage utilization using iostat
 - ✓ Monitored both MDT and OST
- We used mpi based benchmarks: mdtest, postmark, netmist
- Focus on characterization of storage load of MDT and OST for metadata intensive benchmarks.
- New Lustre specific statistic tools?

Mdtest benchmark Low Load

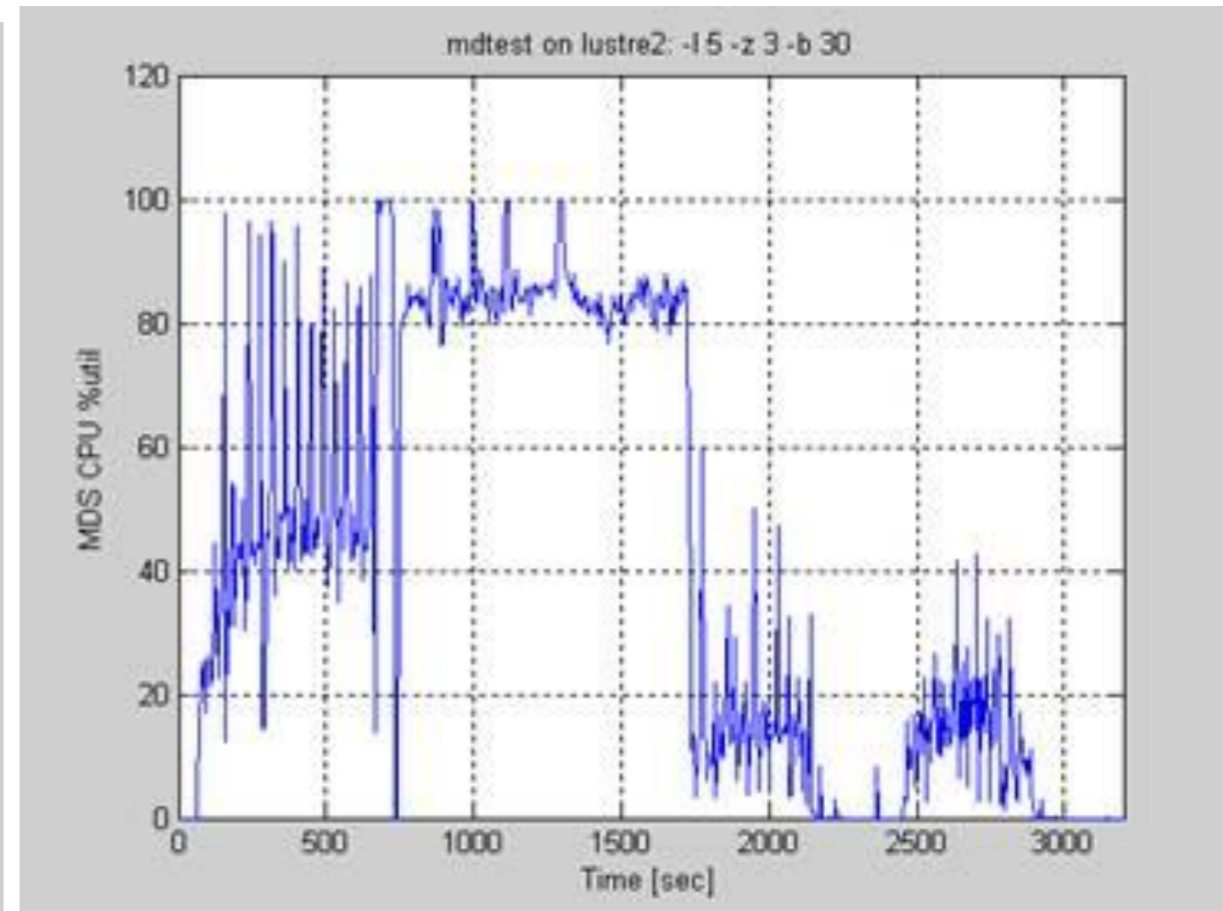
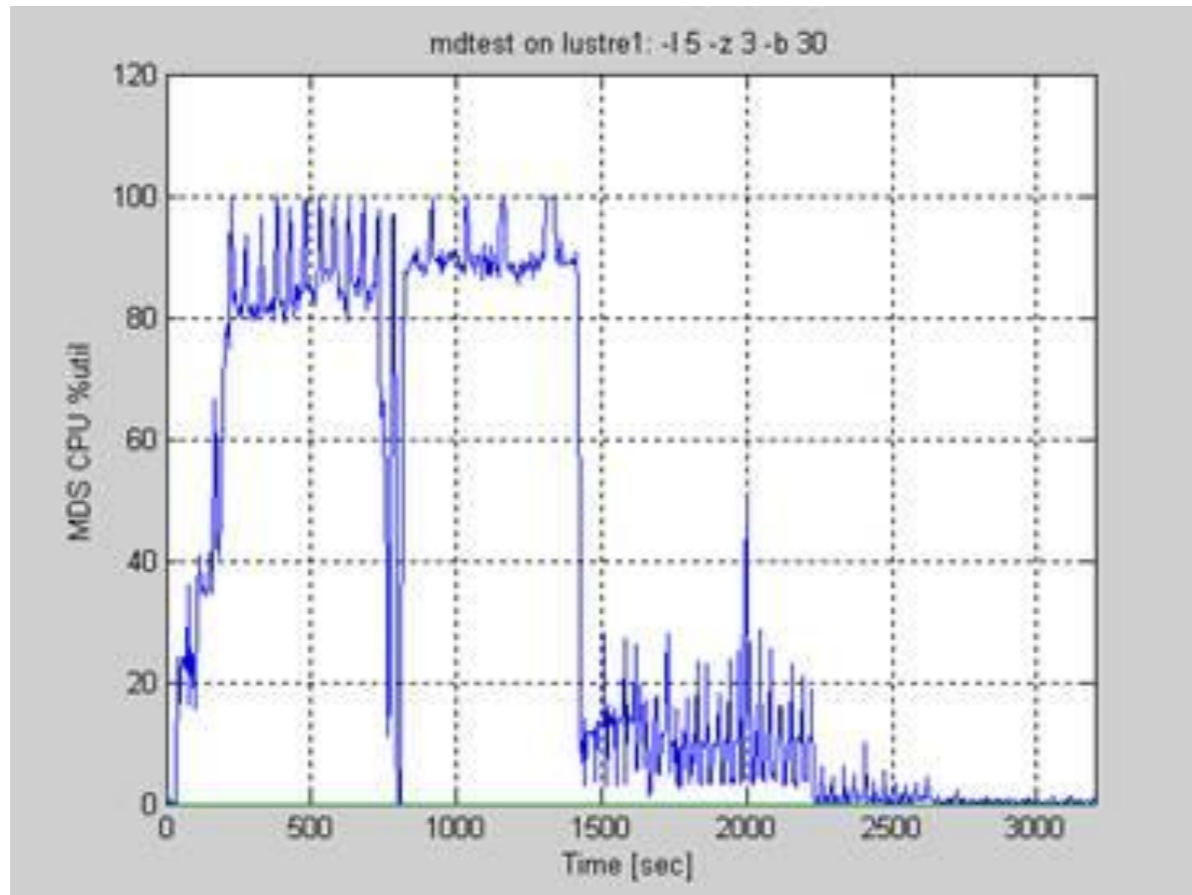
• # mpirun -np 32 -hostfile hostfile /lab/mdtest-1.9.1/mdtest -l 5 -z 2 -b 60 -u -w 1 -d /mnt/lustre1



| | | Directory creation | Directory stat | Directory removal | File creation | File stat | File read | File removal | Tree creation | Tree removal |
|--------|---------|--------------------|----------------|-------------------|---------------|-----------|-----------|--------------|---------------|--------------|
| 585760 | lustre1 | 9206 | 136182 | 8291 | 9270 | 48415 | 23881 | 9877 | 319 | 277 |
| 585760 | lustre2 | 10615 | 119646 | 9271 | 12253 | 91307 | 21311 | 15255 | 461 | 336 |

Mdtest benchmark results high load

•# mpirun -np 32 -hostfile hostfile /root/mdtest-1.9.1/mdtest -l 5 -z 3 -b 30 -u -w 1 -d /mnt/lustre2/



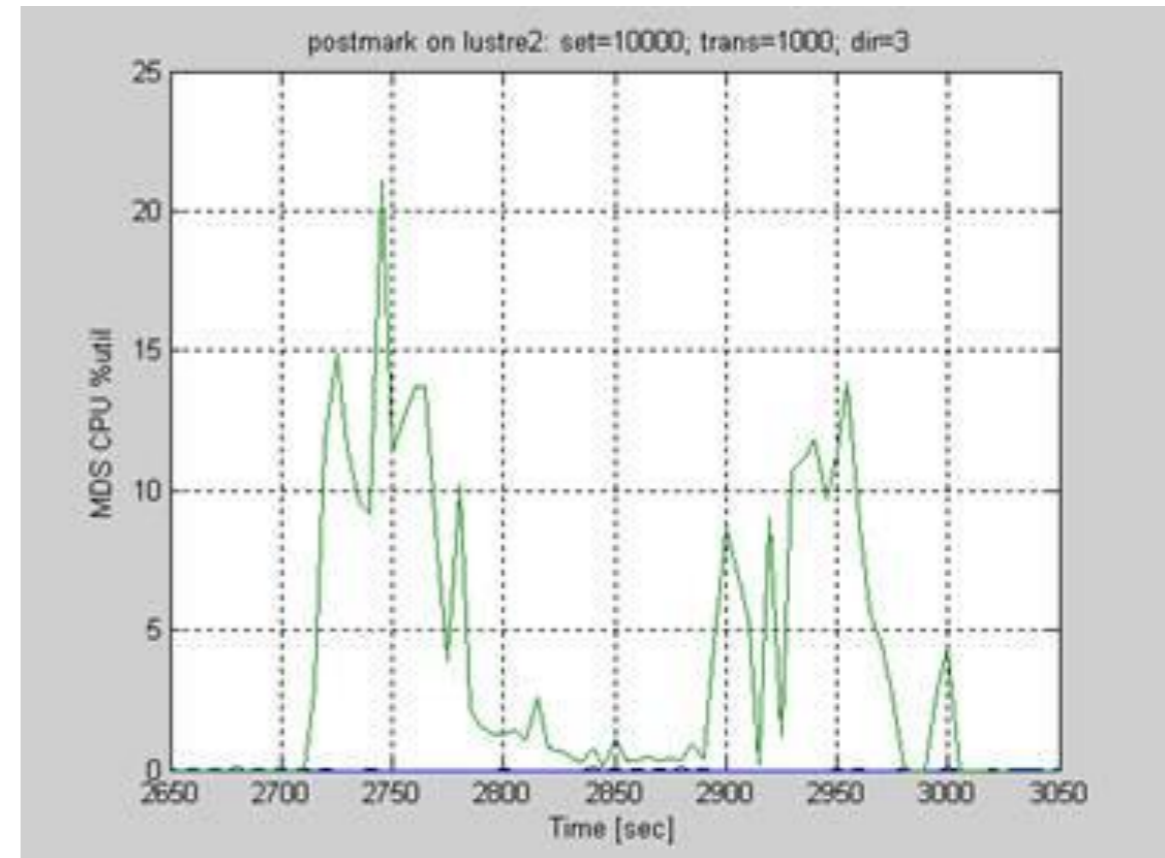
| | | Directory creation | Directory stat | Directory removal | File creation | File stat | File read | File removal | Tree creation | Tree removal |
|---------|---------|--------------------|----------------|-------------------|---------------|-----------|-----------|--------------|---------------|--------------|
| 4468960 | lustre1 | 6960 | 69041 | 4020 | 8091 | 31136 | 22331 | 8169 | 287 | 261 |
| 4468960 | lustre2 | 8025 | 60658 | 4495 | 10694 | 58720 | 19928 | 12617 | 415 | 317 |

Postmark mpi benchmark

```
# mpirun -np 32 -hostfile hostfile /root/mpipostmark-1.5.1/postmark pmtest
```

Pmtest file:

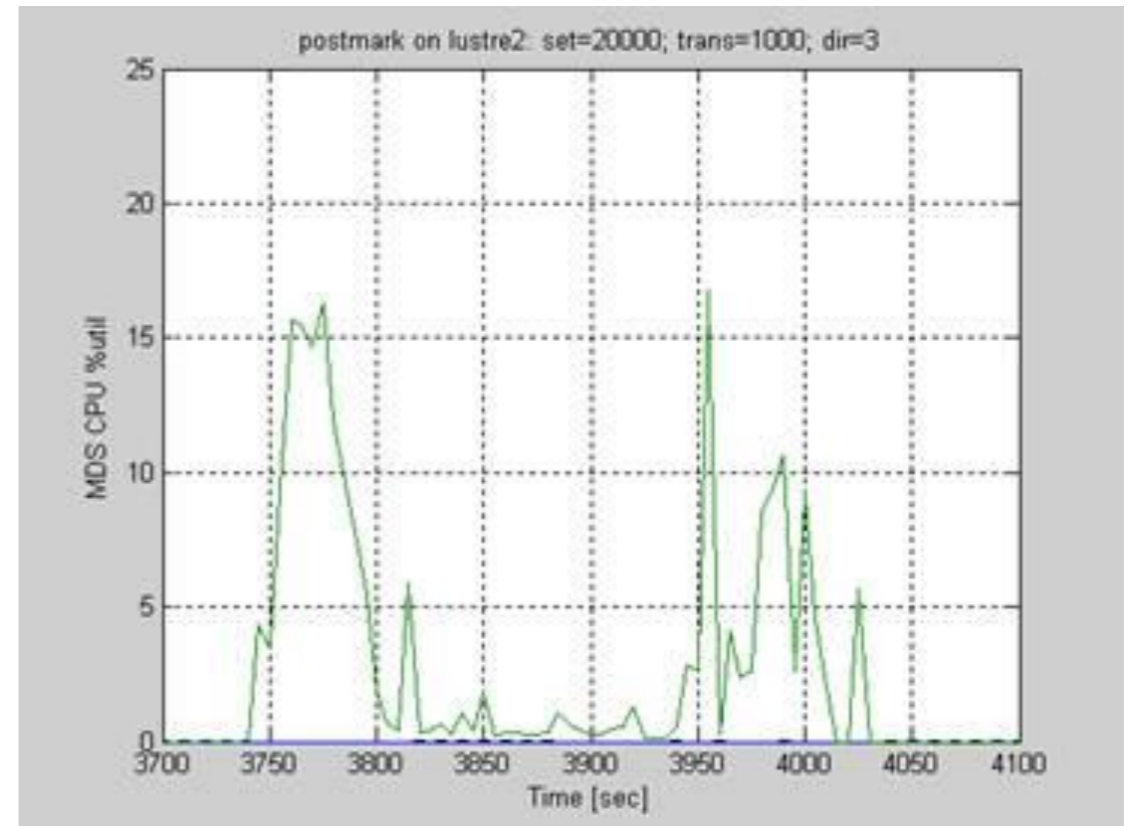
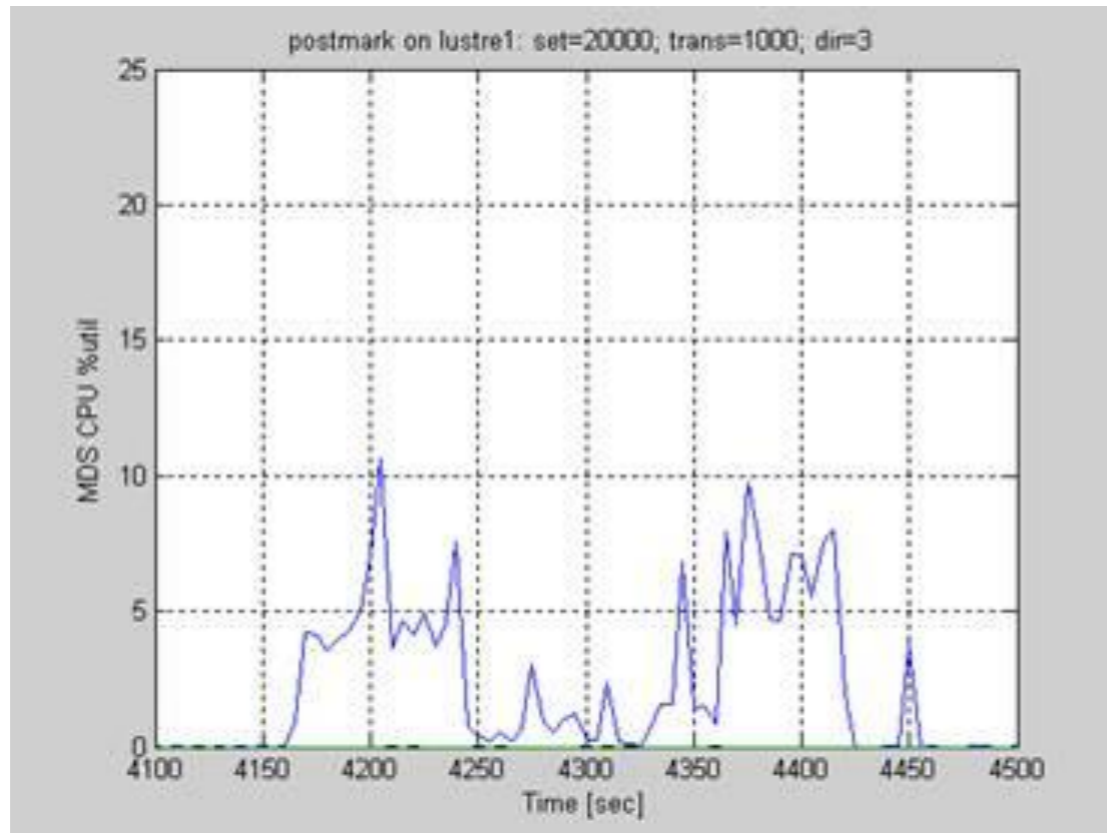
```
set location /mnt/lustre1
set number 10000
set transactions 1000
set subdirectories 3
set size 32768 400000
set bias read 16384
set write 16384
set partition prefix
set report cluster
run
quit
```



| Total files | Tested FS | File Trans/s | File Creates/s | File Reads /s | File Append/s | File Delete/s | Read MB | Write MB | Read MB/sec | Write MB/s |
|-------------|-----------|--------------|----------------|---------------|---------------|---------------|---------|----------|-------------|------------|
| 335840 | lustre2 | 2108 | 4344 | 1085 | 1068 | 4344 | 3338 | 71443 | 44 | 964 |

Postmark benchmark for 2 FS

•20000, 1000, 3



| Total files | Tested FS | File Trans/s | File Creates/s | File Reads/s | File Append/s | File Delete/s | Read MB | Write MB | Read MB/sec | Write MB/s |
|-------------|-----------|--------------|----------------|--------------|---------------|---------------|---------|----------|-------------|------------|
| 656064 | lustre1 | 1066 | 3928 | 549 | 515 | 3927 | 3477 | 137296 | 21 | 841 |
| 656064 | lustre2 | 1777 | 5151 | 1049 | 984 | 5151 | 3477 | 137296 | 28 | 1103 |

Netmist benchmark workload

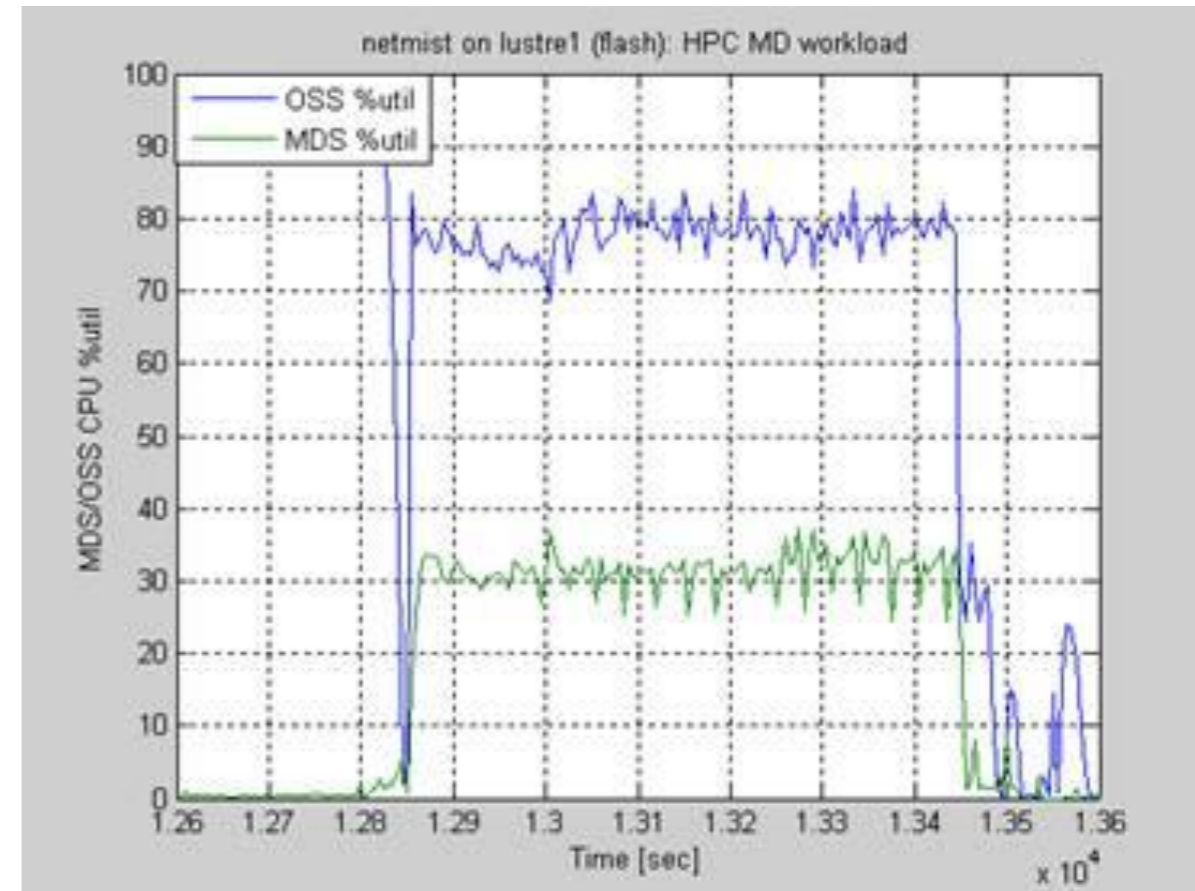
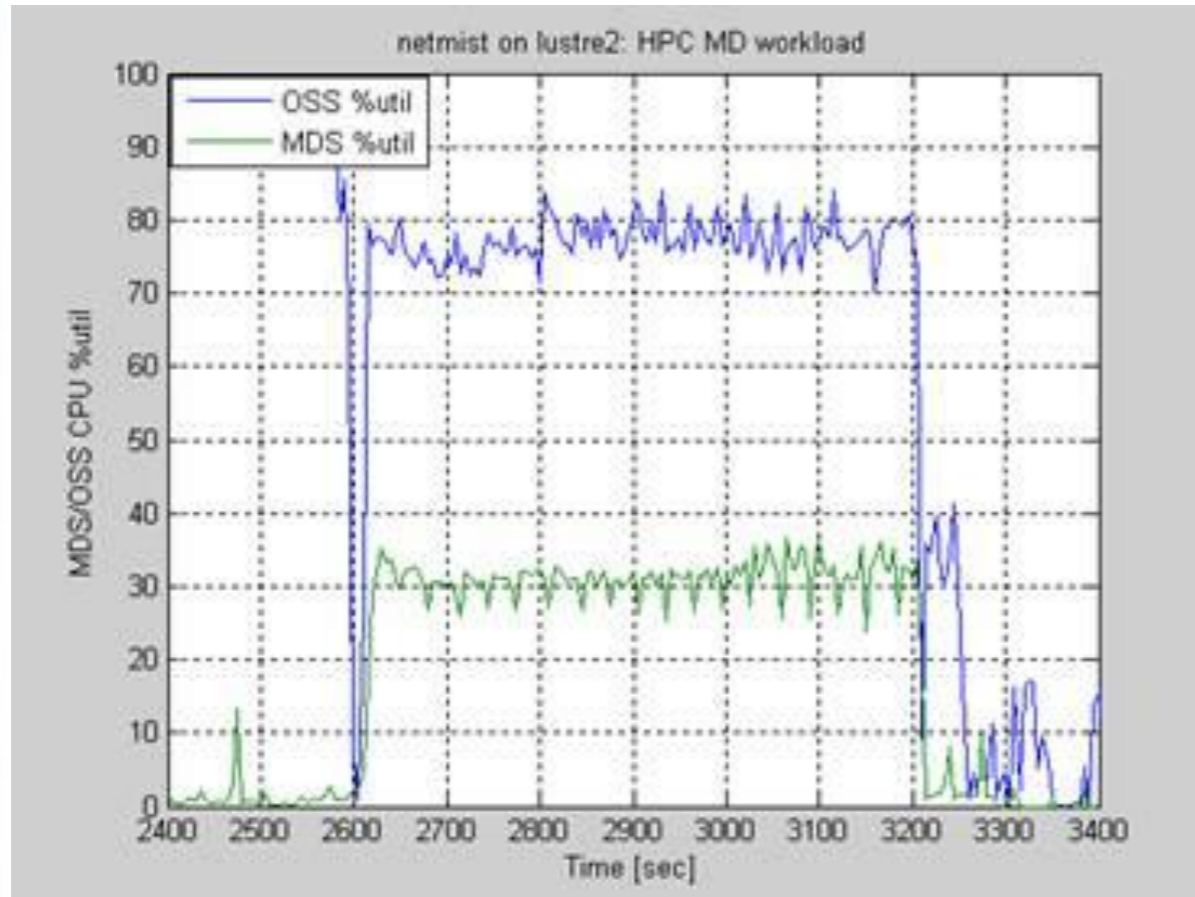
```
# ./netmist -b 128 -d 8 -B 1 -t 300 -w 300 -f client_list
```

| Netmist HPC POSIX on lustre2 (disk) | Netmist HPC POSIX on lustre1 (flash) |
|--|--|
| <p>Test run time = 300 seconds, Warmup = 300 seconds. Running 32 copies of the test on 4 clients Results directory: /lab/demo/netmist_pro_mpi Clients have a total of 512 GiBytes of memory Clients have 16384 MiBytes of memory size per process Clients each have 8 processes Adjustable aggregate data set value set to 1 GiBytes Each process file size = 16800 kbytes Client data set size = 1181250 MiBytes Total starting data set size = 4725000 MiBytes Total initial file space = 4725000 MiBytes Total max file space = 5250000 MiBytes</p> <p>Starting tests: Tue Oct 22 18:35:28 2013</p> <p>Launching 32 processes.</p> <p>-----</p> <p>Overall average latency 3.32 Milli-seconds Overall Netmist_2012 Ops/sec 9628.83 Ops/sec Overall Read_throughput ~ 34558.33 Kbytes/sec Overall Write_throughput ~ 47386.56 Kbytes/sec Overall throughput ~ 81944.90 Kbytes/sec Public Finger</p> <p>Print 841881307</p> <p>-----</p> | <p>Test run time = 300 seconds, Warmup = 300 seconds. Running 32 copies of the test on 4 clients</p> <p>Results directory: /lab/demo/netmist_pro_mpi Clients have a total of 512 GiBytes of memory Clients have 16384 MiBytes of memory size per process Clients each have 8 processes Adjustable aggregate data set value set to 1 GiBytes Each process file size = 16800 kbytes Client data set size = 1181250 MiBytes Total starting data set size = 4725000 MiBytes Total initial file space = 4725000 MiBytes Total max file space = 5250000 MiBytes</p> <p>Starting tests: Tue Oct 22 21:26:13 2013</p> <p>Launching 32 processes.</p> <p>-----</p> <p>Overall average latency 3.29 Milli-seconds Overall Netmist_2012 Ops/sec 9729.73 Ops/sec Overall Read_throughput ~ 34893.23 Kbytes/sec Overall Write_throughput ~ 47683.99 Kbytes/sec Overall throughput ~ 82577.22 Kbytes/sec Public Finger Print 842271196</p> <p>-----</p> |

Netmist benchmark on 2 FS

./netmist -b 128 -d 8 -B 1 -t 300 -w 300 -f client_list

Client_list included 4 clients and 8 threads each or 32 threads



| Total files | Tested FS | Rand read | rand write | rmw | mkdir | unlink | append | access | stat | chmod | readdir | statfs |
|-------------|-----------|-----------|------------|-----|-------|--------|--------|--------|------|-------|---------|--------|
| 650000 | lustre2 | 481 | 481 | 96 | 1540 | 1540 | 1540 | 1540 | 578 | 193 | 1540 | 96 |
| 650000 | lustre1 | 486 | 486 | 97 | 1557 | 1557 | 1557 | 1557 | 584 | 195 | 1557 | 97 |

MDS_survey benchmark

```
# thrlo=64 thrhi=2048 file_count=200000 dir_count=64 tests_str="create lookup md_getattr setxattr  
destroy" mds-survey
```

mdt 1 ,file 200000, dir 64, thr 64

| Create | Lookup | md_getattr | setxattr | destroy |
|----------|---------|------------|----------|----------|
| 70894.61 | 1077702 | 711298.26 | 48365.36 | 67280.78 |

Lustre 2.4, single MDS server, 2 OSS servers, 12 OSTs

mdt 1, file 200000, dir 64, thr 2048

| Create | Lookup | md_getattr | setxattr | destroy |
|----------|---------|------------|----------|----------|
| 60852.58 | 1148586 | 619296.27 | 61283.87 | 59644.49 |

MPEE Proposed Questions

- Should focus on MDS/MDT performance
- Needs to help select best MDS server or help characterize all MD types including directories
- Should we aim to characterize the OST load for small I/O's?
- How important is to test both data and MD access?
- What tools do we use for characterization? Lustre or storage or both. Do we need additional tools?
- We need to help users select the right storage for the MDS/MDT.

MPEE Asks from BWG

- Share any open source synthetic benchmarks code
- Share a list of MD benchmark tools they currently use to allow select the most suitable and used candidates
- Share MD operations tested to allow build Netmist workload objects
- Share the MD workloads that create pain points to Lustre FS
- Share cases of poor MD performance workloads and applications
- Propose Lustre statistic tools to use

Thank you!

