Lustre* Tuning Parameters

High Performance Data Division

Bobbie Lind, Systems Engineer

April 16, 2013

The mystery of tuning Lustre

- Everyone has their own secret sauce
- What parameters are tunable?
- What are the most common tuning parameters?
- How do I monitor those parameters while tuning?
- Where can I find this information?
  - Lustre Operations Manual
  - Online – Web searches
  - Phone a friend
  - HPDD SE's are working on a tuning class

Multiple Layers to Tune

- Underlying OS
- Network (IB, eth, etc.)
- LNet
- Backend Storage
- Lustre Software
- Application I/O

How to find YOUR tuning sweet spot

- Determine what your data looks like
  - Large Files
  - Small Files
  - Video files
  - Other
- Benchmark your current system setup
- Monitor/Change Parameters
- Re-benchmark your system

Baseline Benchmark of Lustre 2.1.5

- Scenario
  - Small Lustre set up
    - 1 MDS
    - 4 OSS (1 OST per)
    - 2 Client
    - Ethernet
  - Sample 300MB files
- Benchmark tools used
  - IOZone
  - DD
  - MDTest

Baseline Benchmark of Lustre 2.1.5

IOZone Baseline Run:

./iozone –w –M –t 1 –s 300m –r 1m –i 0 –i 1 –F /lustre/scratch/iozone/test1 –R

Write: 14184.41 KB/s
Re-Write: 9725.66 KB/s
Read: 2084858.62 KB/s
Re-Read: 2188099

DD Baseline Run:

dd if=/dev/zero of=/lustre/scratch/ddtest/dd1 bs=1M count=300 oflag=direct conv=fdatasync

300+0 records in
300+0 records out
314572800 bytes (315 MB) copied, 29.3717 s, 10.7 MB/s

MDTest Baseline Run:

./mdtest –I 10 –i 5 –z 5 –b 2 –d /lustre/scratch/mdtest

SUMMARY: (of 5 iterations)

Operation	Max	Min	Mean	Std Dev
Directory creation:	472.990	368.132	447.905	40.049
Directory stat :	510.034	425.078	491.362	33.180
Directory removal :	241.112	145.006	219.004	37.092
File creation :	248.578	194.355	235.320	20.586
File stat :	253.255	202.930	239.983	18.683
File read :	260.604	223.014	246.949	14.482
File removal :	458.793	413.149	441.495	16.709
Tree creation :	339.951	325.089	332.853	5.046
Tree removal :	165.150	139.698	156.381	8.903

Common “Lustre Specific” Tuning Parameters

- stripe size
  - lfs getstripe <file or directory>
- max_rpcs_in_flight
  - /proc/fs/lustre/osc/*/max_rpcs_in_flight
- readcache_max_filesize
  - /proc/fs/lustre/obdfileter/*/readcache_max_filesize
- max_dirty_mb
  - /proc/fs/lustre/osc/*/max_dirty_mb
- max_read_ahead_mb
  - /proc/fs/lustre/llite/*/max_read_ahead_mb

Monitoring Those Parameters

- RPCs
  - lctl get_param osc.*.import – realtime
  - lctl get_param osc.*.rpc_stats
- Stripe Size
  - dd if=/dev/zero of=/lustre/test/1stripe/testfile bs=1M count=300 oflag=direct conv=fdatasync
- Monitoring I/O Patterns (for readcache_max_filesize)
  - brw_stats (oss)
  - strace (application)
- Intel Manager for Lustre

Second Benchmark of Lustre 2.1.5

IOZone Second Run:

./iozone –w –M –t 1 –s 300m –r 1m –i 0 –i 1 –F /lustre/scratch/iozone/test1 –R

Write: 19218.00 KB/s
Re-Write: 12180.47 KB/s
Read: 2505728.50 KB/s
Re-Read: 2130480.00

DD Second Run:

dd if=/dev/zero of=/lustre/scratch/ddtest/dd1 bs=1M count=300 oflag=direct conv=fdatasync

300+0 records in
300+0 records out
314572800 bytes (315 MB) copied, 28.5367 s, 11.0 MB

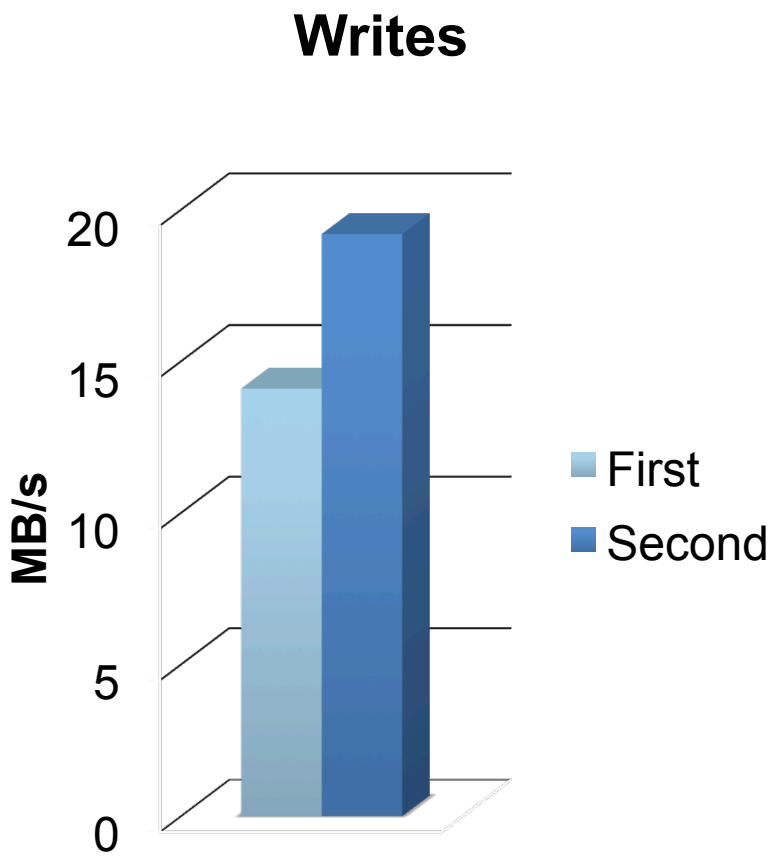
MDTest Baseline Run:

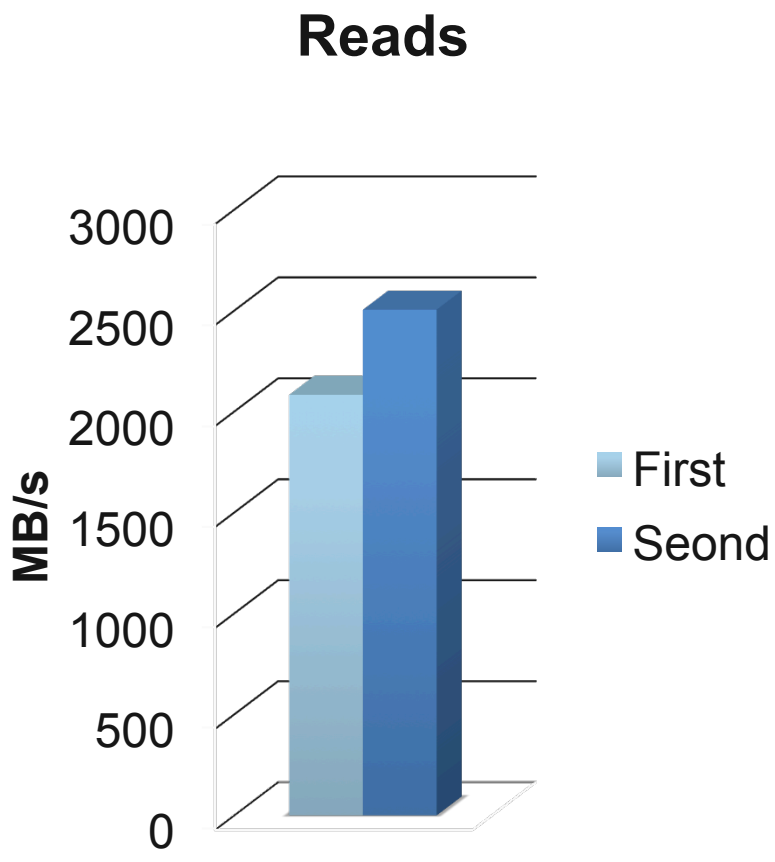
./mdtest –I 10 –i 5 –z 5 –b 2 –d /lustre/scratch/mdtest

SUMMARY: (of 5 iterations)

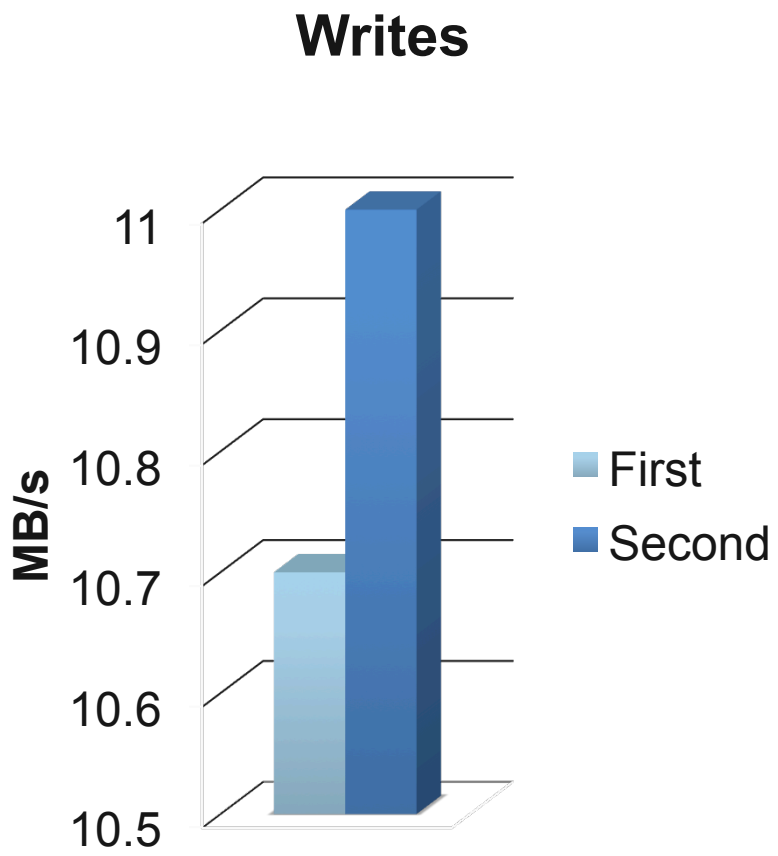
Operation	Max	Min	Mean	Std Dev
Directory creation:	502.790	378.232	483.737	37.049
Directory stat :	510.034	495.063	501.189	20.530
Directory removal :	241.112	217.506	220.543	9.092
File creation :	248.478	230.355	240.026	5.786
File stat :	252.257	239.930	244.783	6.698
File read :	258.444	240.054	249.418	12.483
File removal :	458.623	441.161	450.324	8.506
Tree creation :	346.499	339.089	341.174	3.821
Tree removal :	168.150	152.234	159.508	9.904

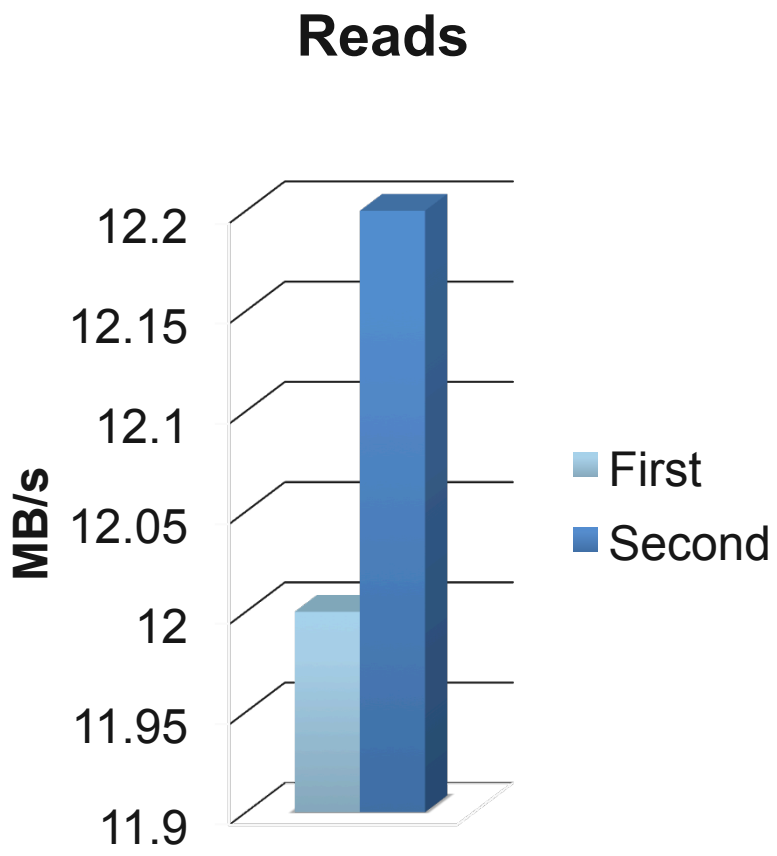
Changes Between Benchmarks - IOZone



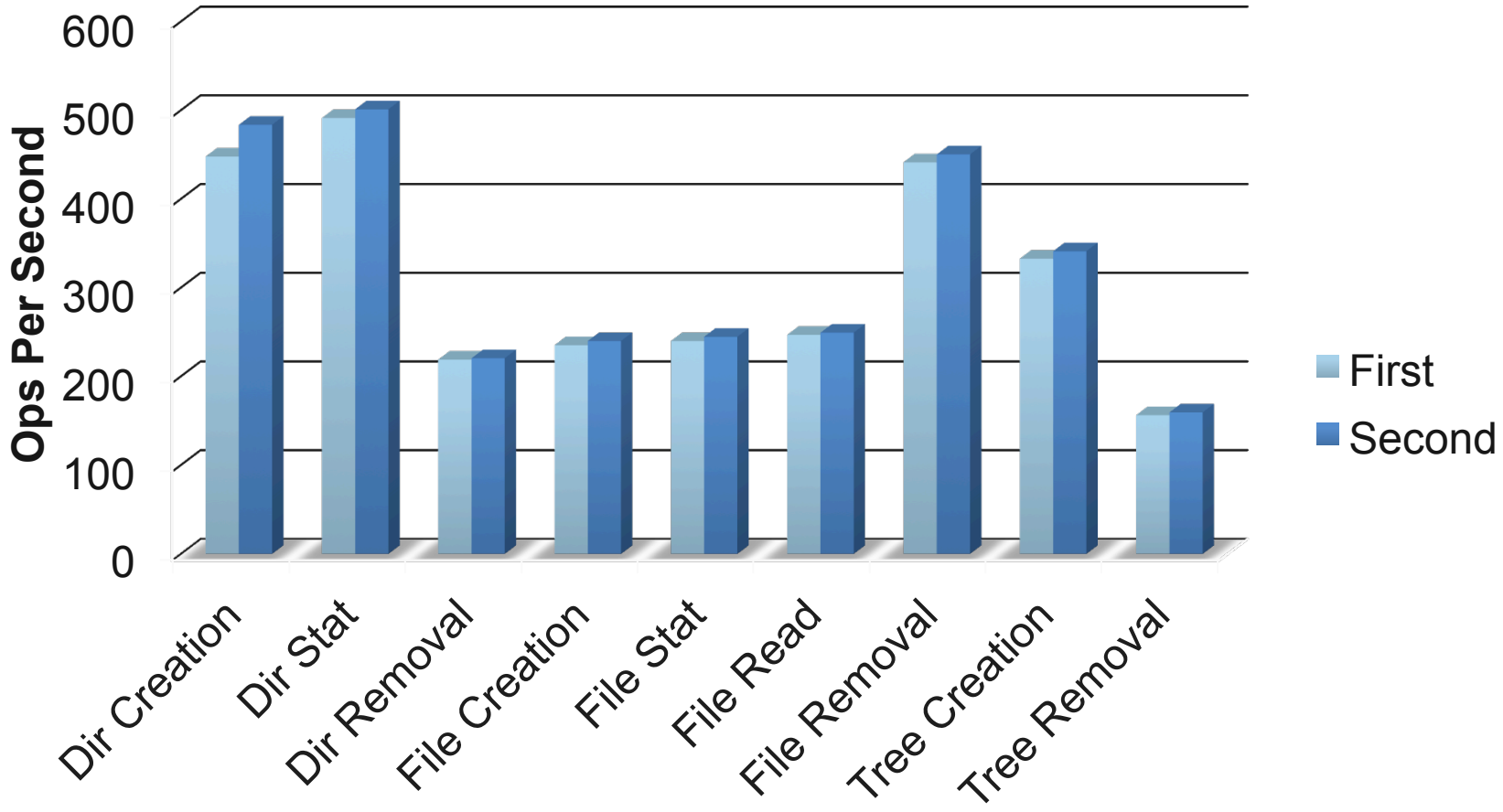


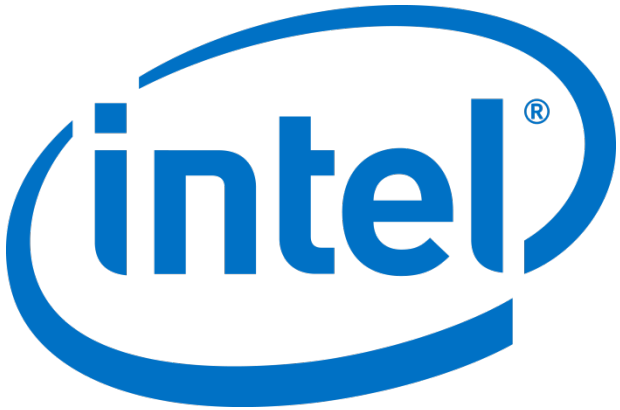
Changes Between Benchmarks - DD





Changes Between Benchmarks - MDTest




intel