

A low-angle photograph of several rows of blue solar panels, reflecting the sky and clouds, extending towards the horizon.

Lustre Benchmarking Tips And Tricks

Atul Vidwansa

Professional Services, Sun Microsystems

Agenda

Why Benchmark?

Benchmarking disks and Lustre OSS

Benchmarking Lustre Network

Measuring Data & Metadata Performance

Tools & Resources

Why Benchmark?

To measure performance

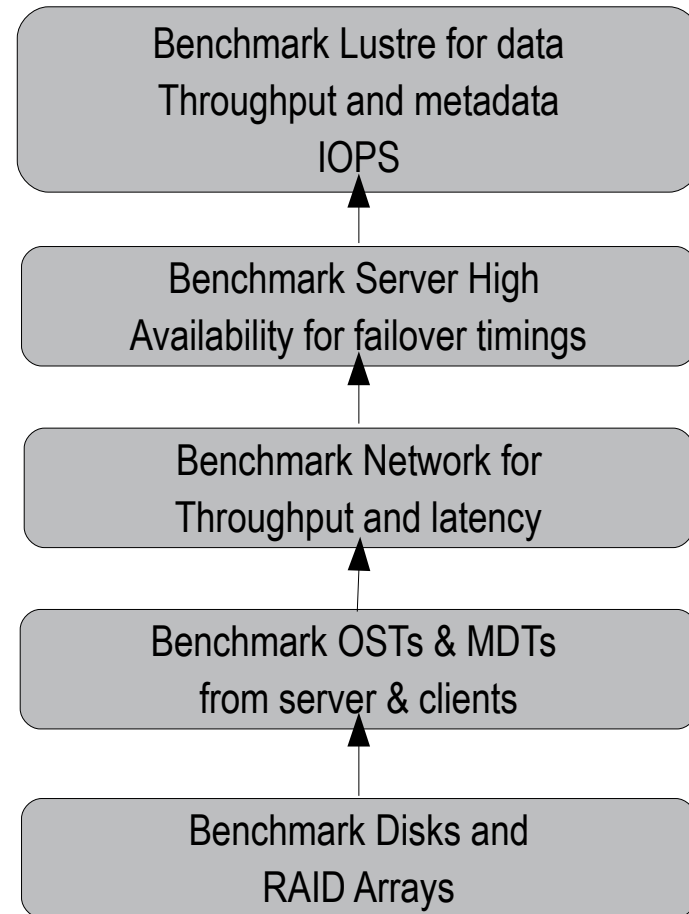
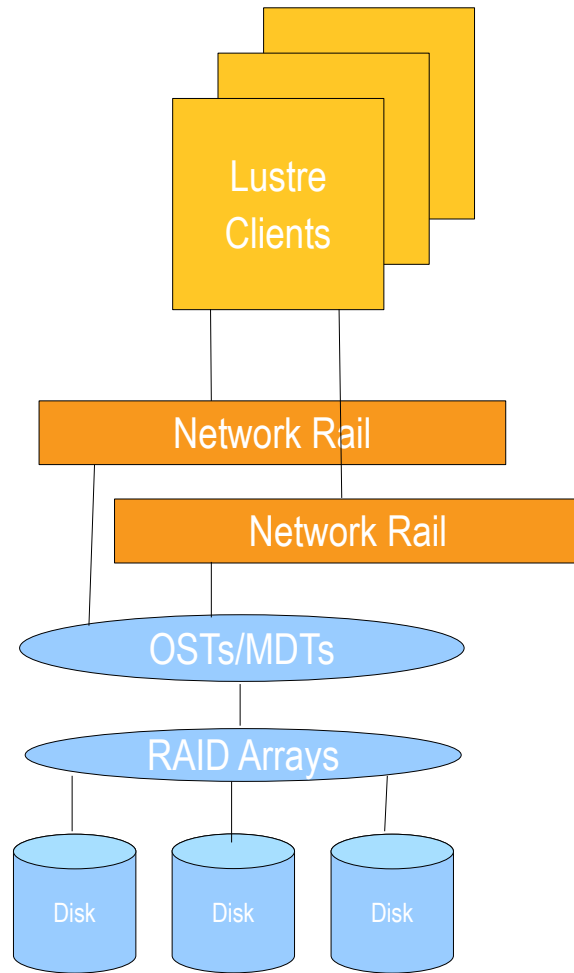
To tune Lustre

To validate acceptance criteria

To debug problems

To verify SLAs

Benchmarking Lustre Stack



Benchmarking Disks and Lustre Object Storage Targets (OSTs)

Benchmarking Disks

Lustre can use any type of storage that provides block device interface.

Disks, LUNs and storage arrays are at the bottom of the IO stack and needs to be benchmark-ed first in order to:

- Verify and tune disk parameters

- Remove slow/non performing disks

- Guarantee consistent performance across set of disks

Lustre provides excellent set of scripts in form of Lustre IOKit available from CVS repository and [Lustre downloads site](http://downloads.lustre.org) (<http://downloads.lustre.org>)

Disk and RAID Tuning

Before benchmarking disks must be tuned to get maximum performance. We recommend following settings:

Deadline scheduler

```
"/sys/block/sd*/queue/scheduler"
```

max_sectors_kb=max_hw_sectors_kb

```
"/sys/block/sd*/queue/max_sectors_kb"
```

```
"/sys/block/sd*/queue/max_hw_sectors_kb"
```

Write through cache for disks/storage arrays without battery backed cache

```
"/sys/block/sd*/device/scsi_disk*/cache_type"
```

Tune RAID5/RAID6 arrays for stripe cache size and read-ahead settings. These may vary for different storage devices

```
"echo 16384 > /sys/block/md$i/md/stripe_cache_size"
```

```
"blockdev --setra 8192 /dev/md$i"
```

Sgpdd Survey for Disks & S/w RAID

Sgpdd-survey is a wrapper script available in Lustre IOKIT for benchmarking disks with `sgp_dd` command, which is part of generic SCSI device utilities (`sg3-utils`) package.

`Sgp_dd` is scalable version of “`dd`” command with options to do multi threaded IO with different block sizes and region sizes.

`Sgpdd-survey` script generates large sequential IO workload on underlying disks

`Sgpdd-survey` can also be used to benchmark software RAID arrays with some modifications.

- Replace `sg_readcap` with s/w raid array size

- Present s/w raid arrays as raw device.

- Refer to modified `sgpdd-survey` script in Bug 17218

Understanding Sgpdd-survey

Generate writes and reads on devices with parameters:

rszlo-rszhi: Record sizes in KB

Affects how many blocks (#bpt) can be transferred in each IO transaction.
Simulates Lustre RPC size.

crglo-crghi: No of Regions

Tells how many separate regions on disk will be read or written.
Simulates multiple Lustre clients per OST. More regions = less performance, as disk needs to do seeking.

thrlo-thrhi: No of Threads

Simulates Lustre OSS threads.

Size: Total size in MB per sgp-dd command

Blocksize: 512 Bytes

Default size is 8GB and blocksize is 512 bytes but, 32GB size (or 2x system memory) and 1MB blocksize recommended to simulate Lustre sequential IO workload

Recommended Parameters

While invoking `sgpdd-survey`, we recommend following parameters to measure performance under large sequential IO workload:

`Rszlo=1024 rszhi=1024`

`Thrlo=1 thrhi=16`

`Crglo=1 crghi=16`

`Size=32768` (or twice that of system memory)

`Dio=1 oflag=direct iflag=direct`

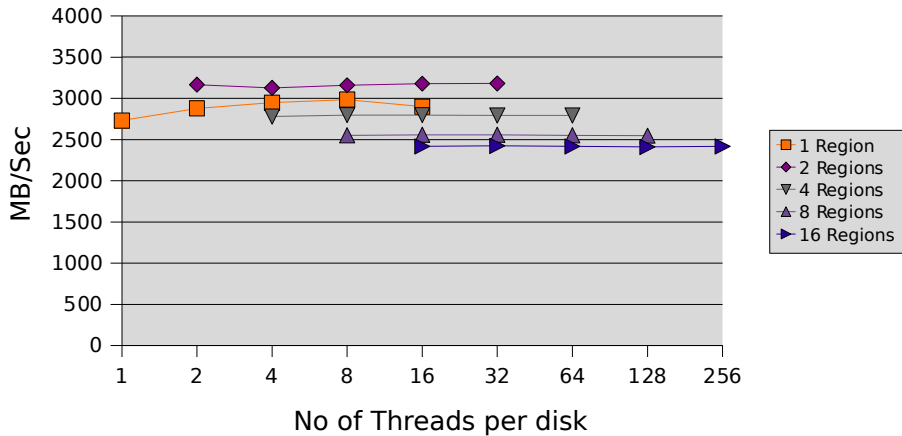
`Bs=1048576`

Note that, “size” parameter also determines amount of seek `sgp_dd` command will do. With `size=32GB`, only outer portions of the disk will be exercised which are usually faster.

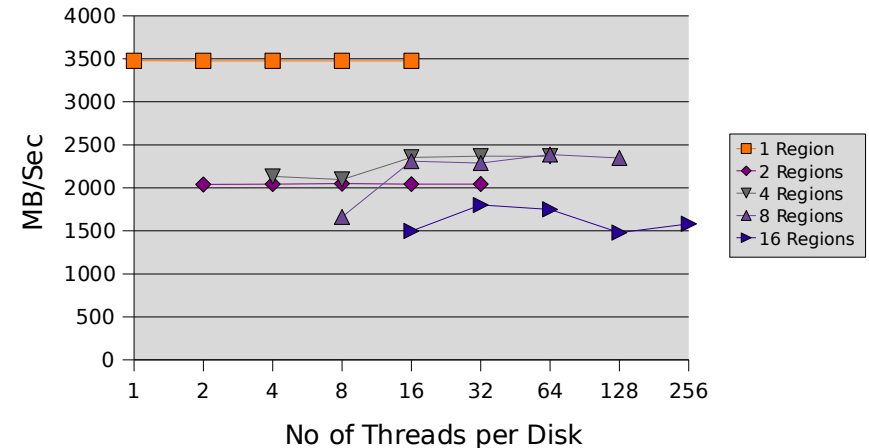
To determine performance of complete disk, use patch in Bug 17218 against `sgpdd-survey`.

Sgpdd-survey on Sun Fire X4540

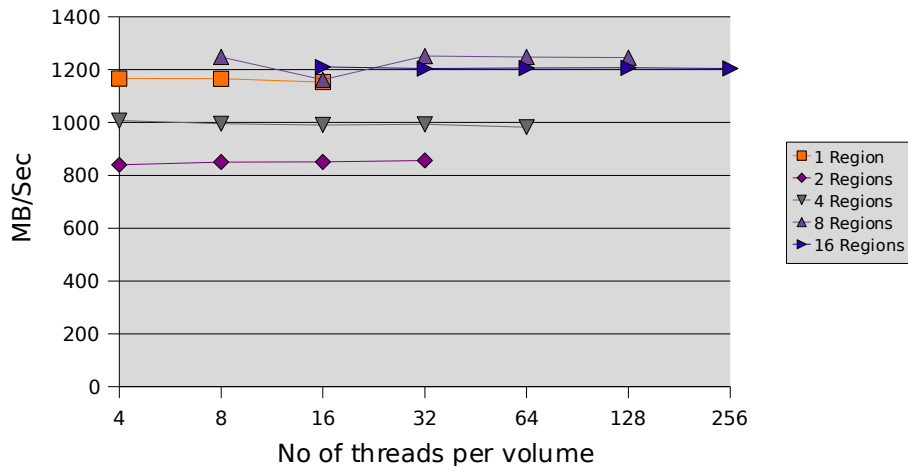
Sgpdd-survey write on 42 disks



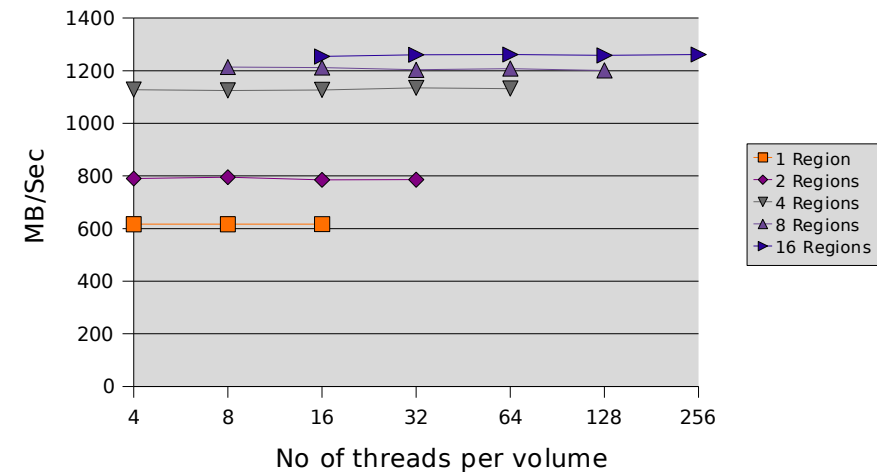
Sgpdd-survey read on 42 Disks



Sgpdd-survey write on 7 RAID6 volumes



Sgpdd-survey read on 7 RAID6 volumes



More Tools for Benchmarking Disks

There exists lots of tools for benchmarking disks other than `sgp_dd`.

XDD is also a good tool and Lustre group provides a similar [xdd-survey](#) script which generates sequential as well as random read/write workload.

LMDD, VDBench, Filebench, bonnie++ are some more tools used for benchmarking storage.

Benchmarking Lustre OSTs

Once disks are verified for performance, next step is to benchmark Lustre OSTs

OSTs could be hardware or software RAID arrays, usually they are raid5 or raid6 arrays.

Lustre provides tools like Sgpdd-survey and Obdfilter-survey for benchmarking OSTs

Lustre uses improved linux RAID5/RAID6 code to get better performance by aligning IO requests and using zero copy mechanism

OBDFilter-survey

Lustre IOKIT provides obdfilter-survey script which exercises obdfilter layer in Lustre IO stack for reading, writing and rewriting Lustre objects.

Obdfilter-survey can be run on Lustre OSTs, both loopback and disk-backed, either from same node or from another node(s) over the network.

Obdfilter-survey is primarily used for sizing OST throughput performance over the network.

Lustre OSS needs to be configured before running survey on OSTs

Check output of “lctl dl” command on OSS nodes to verify existence of obdfilter instances

Understanding Obdfilter-survey

Obdfilter-survey can be invoked with following parameters

case: local-disk, network-echo, network-disk

Run survey on disk-backed local obdfilter instances, network loopback or disk instances

thrl0-thrhi: High - low counts of threads

nobjlo-nobjhi: No of objects to read/write

rszlo-rszhi: High - low record size in KB

size: Total IO size in MB

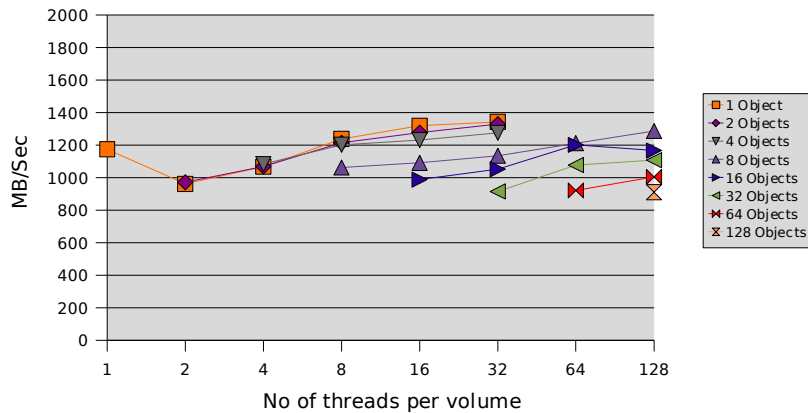
targets: names of obdfilter instances

Recommended parameters are:

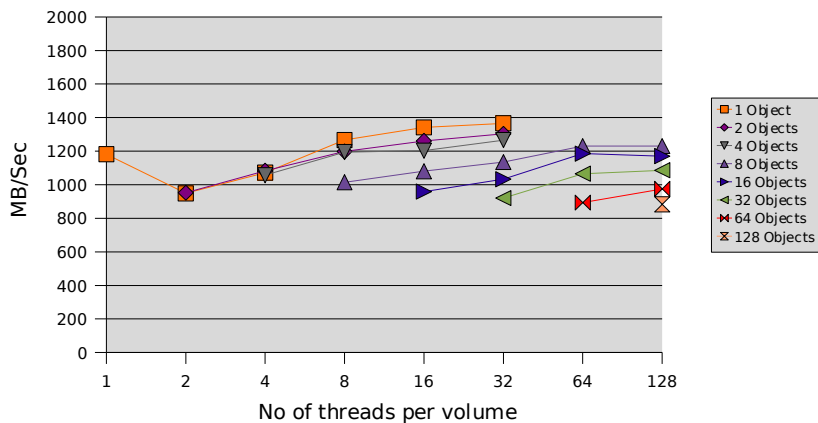
rszlo=rszhi=1024, nobjhi=128, thrhi=128

Obdfilter-survey on 7 RAID6 OSTs

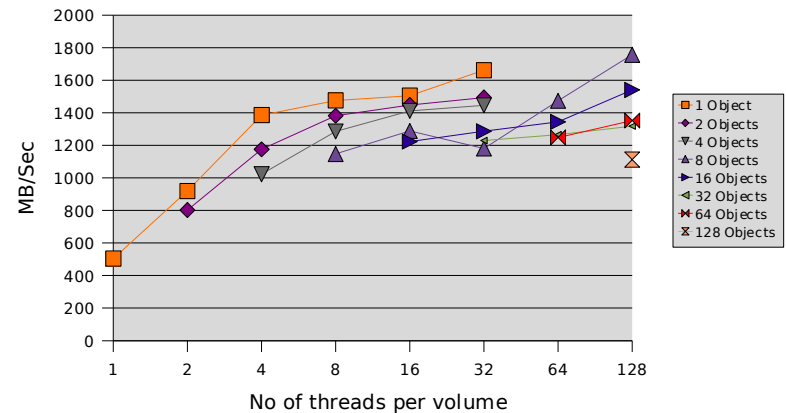
OBDFilter-survey write on 7 RAID6 Volumes



OBDFilter-survey rewrite on 7 RAID6 volumes



OBDFilter-survey read on 7 RAID6 volumes



Benchmarking Lustre Network

Lustre Network Performance

Lustre supports various types of network technologies including TCP/IP over Gbe/10Gbe, Infiniband, Elan, Myrinet MX, and Cray Seastar

Measuring network performance is important as it can be a bottleneck compared to storage performance.

Performance can be measured in terms of throughput and latency.

Lustre provides LNET Selftest (LST) for measuring performance with Lustre networking protocol. Open-source tools like netperf can be used as well.

[Lustre manual](#) provides detailed description of LST

Using LNET Selftest

LNET Selftest should be used to measure network throughput and RPC operations in Lustre environment from:

- Single Client to Single Server

- Multiple clients to Single Server

- Multiple clients to multiple Servers

With parameters like:

- Different message sizes (size=)

- Different concurrency (concurrency=)

- Writes and read operations (brw read/write)

- With and without data checksum (check=)

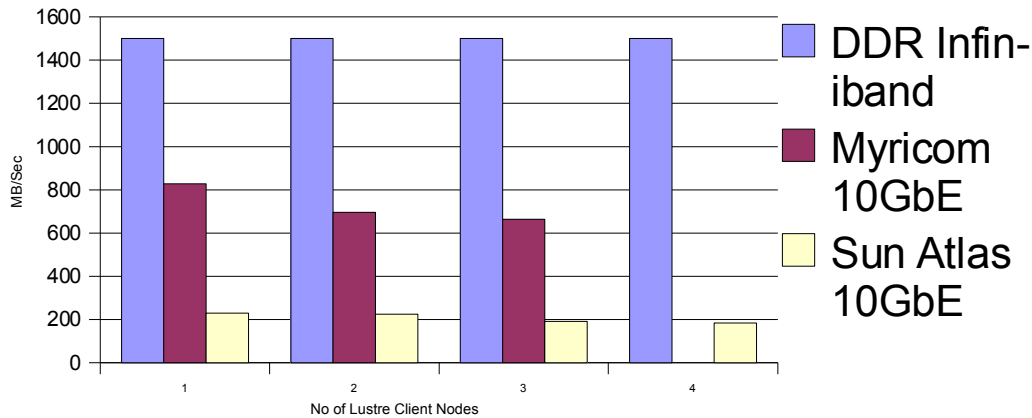
Sample LNET Selftest Script

```
#!/bin/bash
export LST_SESSION=$$
lst new_session read/write
lst add_group servers 5.6.128.233@o2ib
lst add_group readers 5.6.132.[30-37]@o2ib
lst add_group writers 5.6.132.[30-37]@o2ib
lst add_batch bulk_rw
lst add_test --batch bulk_arw --concurrency 8 --from readers --to servers \
brw read size=1M
lst add_test --batch bulk_rw --concurrency 8 --from writers --to servers \
brw write size=1M

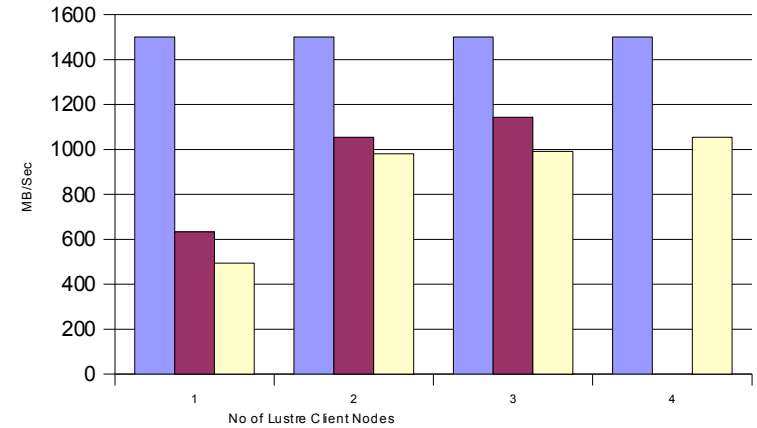
# start running
lst run bulk_rw
# display server stats for 180 seconds
lst stat servers & sleep 180
lst stop bulk_rw
# tear down
lst end_session
pkill lst
```

LST Results for N clients and 1 Server

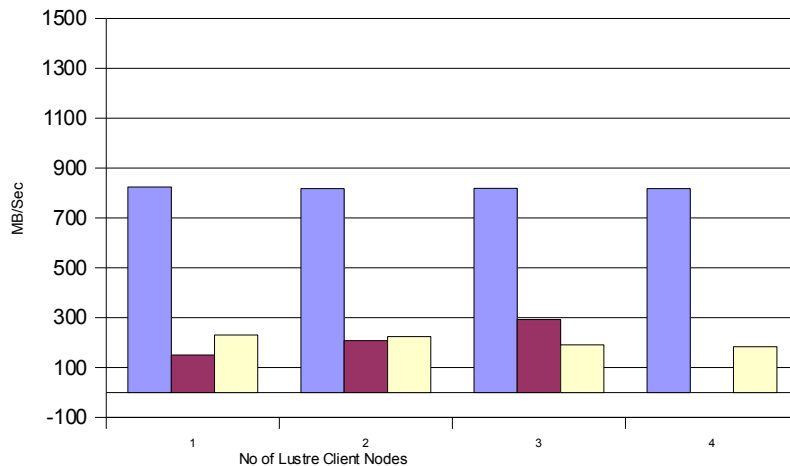
Unidirectional Write



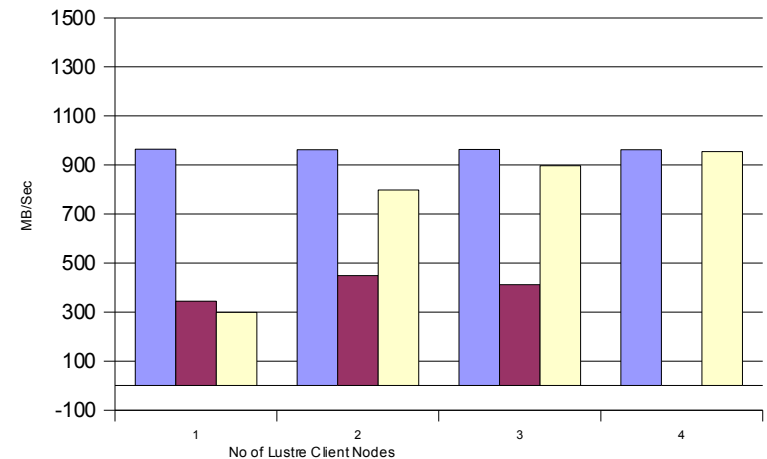
Unidirectional Read



Bidirectional Write



Bidirectional Read



Benchmarking Data and Metadata Performance with Lustre

Measuring read/write performance

After benchmarking disks, RAID arrays, OSTs and network individually, next level is to measure data read/write performance of Lustre filesystem.

Read/write performance should be measured from Lustre client between

- Single client and single server

- Multiple clients and single server

- Multiple clients and multiple servers

Benchmarking scenarios should include:

- Single Shared file read/write

- File per process/client read/write

Open-source tools like IOR and IOZone are most commonly used for measuring read/write performance of clustered file systems.

IOR and IOZone

Both IOR and IOZone are excellent benchmarking tools. Each essentially capable of doing large sequential reads/writes as well as has unique features.

IOZone provides additional data load patterns like re-read, rewrite, read backwards, read strided, random io, aio, mmap io etc.

IOR has support for manipulating Lustre striping and can use POSIX, MPI_IO, HDF5 or NCMPI api for IO.

Use IOR as:

```
mpirun -np 4 ./IOR -a POSIX -r -w -b 32g -t 1m -o /mnt/lustre/IOFile -O lustreStripeCount=-1 -v -i 3
```

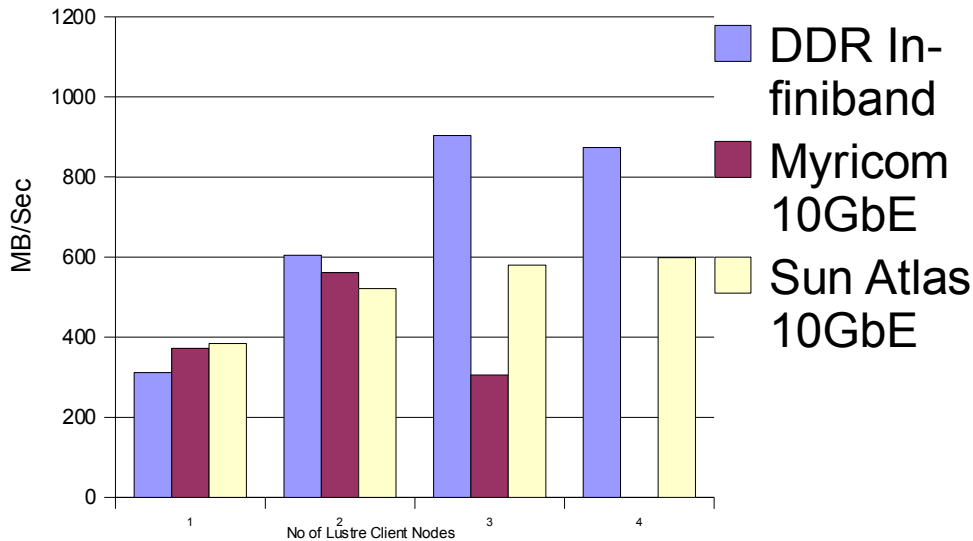
Use IOZone as:

```
./iozone -w -M -t 120 -s 32g -r 1m -i0 -i1 -+m client_list
```

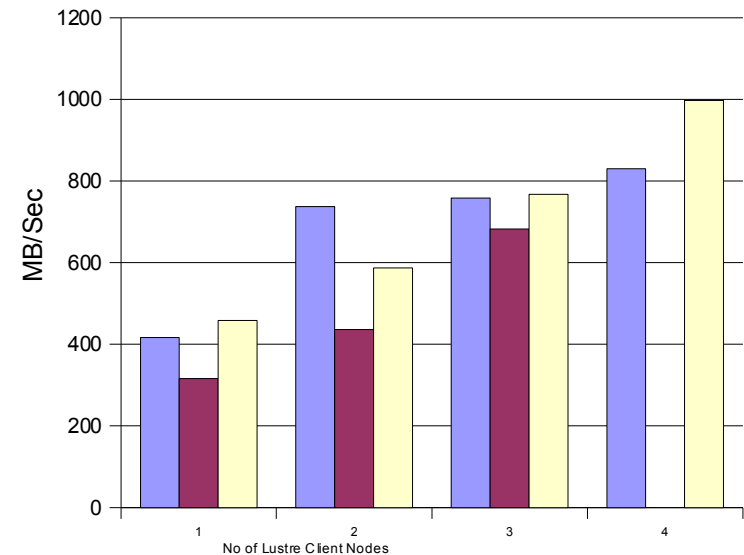
IOR Results

IOR large sequential read/write from 1-4 clients to single OSS connected by DDR Infiniband and TCP/IP over 10GbE

IOR write performance



IOR read performance



Lustre Metadata Measurements

Many HPC applications generate heavy metadata traffic e.g. creating large number of temporary files.

It is important to measure Lustre metadata performance like rate of file and directory creation, stat and delete operations.

Currently Lustre supports single metadata server and hence metadata performance is limited. This is expected to change with Clustered MDS.

MDS Benchmarking Tools

There are excellent metadata benchmarking tools available like Metabench from NERSC, Mdtest from LLNL, Mdsrate from HP (distributed with Lustre) and bonnie++.

These tools allow load generation and OPS measurements of

- File creation in single or multiple directories

- Multiple Directory creation

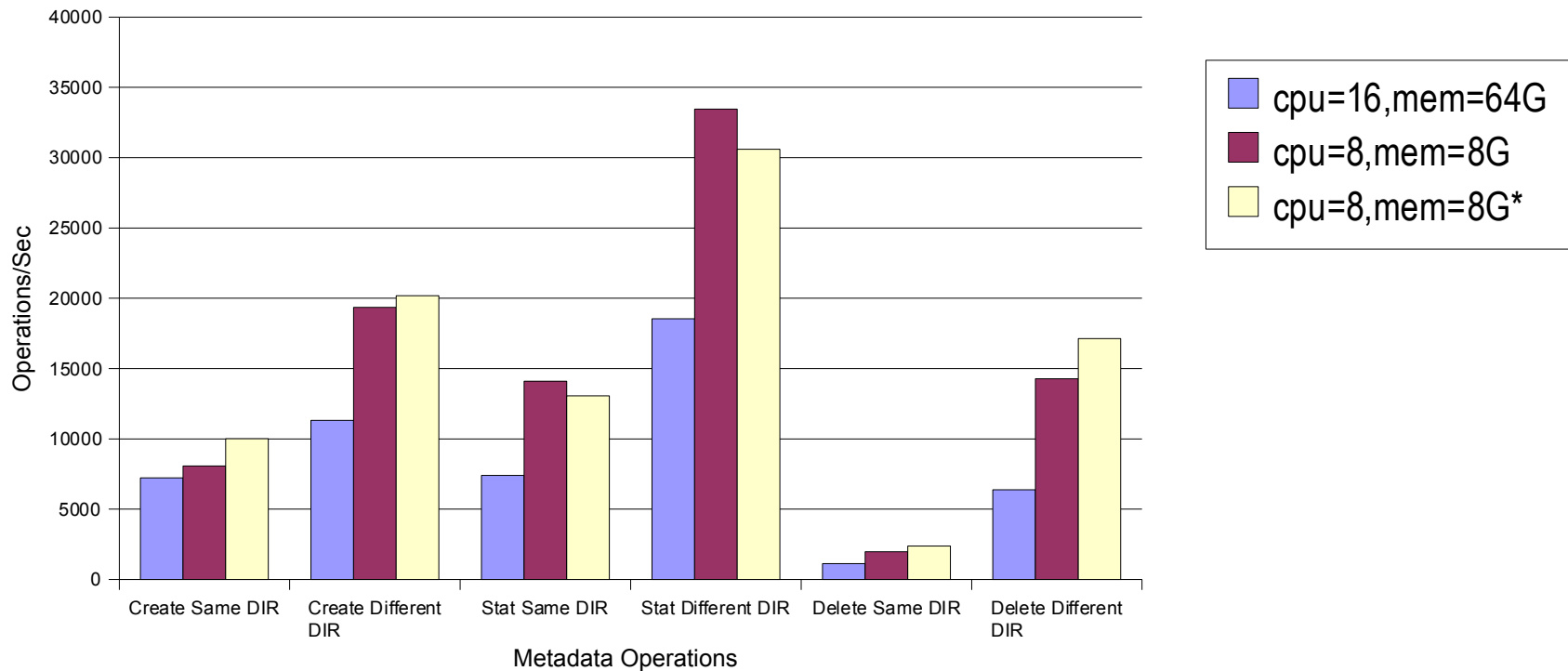
- File stat, unlink/delete in single or multiple directories

- Stat in readdir or random order

All these tools need MPI support in order to run on Lustre.

Lustre Metadata Performance

Lustre Metadata Performance on 24 SAS (15K RPM) Disk RAID10 MDT



Measuring Lustre Performance over period of time

A fully functioning empty filesystem performs at its best. To understand performance of Lustre over period of time, following scenarios should be benchmarked:

- Fully functioning empty fs performance

- Fully functioning, 80 % full and fragmented filesystem

- FS performance with RAID arrays are rebuilding/resyncing

- FS performance with mixture of large sequential and small random IO workload.

- Single and multi streamed IO in half and full duplex mode

- Extreme file creation, stat and deletion rate in empty directory

- Extreme file creation, stat and deletion rate in directory with 1 Million existing files

Tools and Resources

Monitoring

- Collectl and Lustre Monitoring Tool (LMT)

- Custom scripts to pull data from /proc filesystem on MDS, OSS and clients

Performance measurements

- IOR and IOZone for read/write performance

- Metabench, mdtest and mdsrate for metadata performance

- Sgpdd-survey, XDD-survey and Obdfilter-survey for disk, raid, OST level performance

- Netperf, LNET Selftest for network performance

Documentation

- Lustre IOKIT README files

- Benchmarking chapter in Lustre Manual



Questions ?

A close-up photograph of water splashing, with white foam and blue water, set against a dark blue background with a yellow curved line on the right side.

THANK YOU

Atul Vidwansa <atul@sun.com>

Lustre Group, Sun Microsystems