SEAGATE

New and Improved Lustre Performance Monitoring Tool

Torben Kling Petersen, PhD
Principal Engineer

Chris Bloxham
Principal Architect

Lustre® monitoring

- Performance
  - Granular
  - Aggregated
  - Components
  - Subsystem
- Lustre
  - Health
  - Functional
  - Problems
  - Hotspots
- Systems
  - Health
  - FRU components
  - OS and firmware
  - Temperature
  - Power consumption
- Clients
  - I/O patterns
  - Job execution
  - Health status

Lustre Monitoring Tool v 3.x

fs1-MDT0000 2012-02-20 01:12:40.0

%CPU	%KB	%Inodes
1.36	0.03	0.00

Operation	Samples	Sample /Sec	Avg Value	Std Dev	Units
close	0	0.00	0.00	0.00	reqs
connect	0	0.00	0.00	0.00	reqs
create	0	0.00	0.00	0.00	reqs
destroy	0	0.00	0.00	0.00	reqs
disconnect	0	0.00	0.00	0.00	reqs
getattr	0	0.00	0.00	0.00	reqs
getxattr	0	0.00	0.00	0.00	reqs
link	0	0.00	0.00	0.00	reqs
llog_init	0	0.00	0.00	0.00	reqs
mkdir	0	0.00	0.00	0.00	reqs
mknod	0	0.00	0.00	0.00	reqs
notify	0	0.00	0.00	0.00	reqs
open	0	0.00	0.00	0.00	reqs
process_config	0	0.00	0.00	0.00	reqs
quotactl	0	0.00	0.00	0.00	reqs
reconnect	0	0.00	0.00	0.00	reqs
rename	0	0.00	0.00	0.00	reqs
rmdir	0	0.00	0.00	0.00	reqs
setattr	0	0.00	0.00	0.00	reqs
statfs	0	0.00	0.00	0.00	reqs
unlink	0	0.00	0.00	0.00	reqs

OST 2012-02-20 01:12:40.0

Ost Name	Read Rate	Write Rate	%CPU	%KB	%Inodes
fs1-OST0000	0.00	467.60	*****	0.57	0.00
fs1-OST000b	0.00	496.00	*****	0.16	0.00
fs1-OST000a	0.00	478.40	*****	0.15	0.00
fs1-OST0009	0.00	472.20	*****	0.25	0.00
fs1-OST0008	0.00	468.20	*****	0.25	0.00
fs1-OST0007	0.00	450.20	*****	0.50	0.00
fs1-OST0006	0.00	451.20	*****	0.51	0.00
fs1-OST0005	0.00	444.40	*****	0.30	0.00
fs1-OST0001	0.00	489.00	*****	0.36	0.00
fs1-OST0004	0.00	454.80	*****	0.30	0.00
fs1-OST0003	0.00	466.80	*****	0.04	0.00
fs1-OST0002	0.00	465.40	*****	0.04	0.00
AGGREGATE	0.00	5,604.20	*****	*****	*****
MAXIMUM	0.00	496.00	*****	0.57	0.00
MINIMUM	0.00	444.40	*****	0.04	0.00
AVERAGE	0.00	467.02	0.00	0.29	0.00

OSS 2012-02-20 01:12:40.0

Oss Name	Read Rate	Write Rate	%CPU	%Space Used	%Inodes Used
dvtrack202	0.00	956.60	0.42	0.36	0.00
dvtrack207	0.00	974.40	0.34	0.15	0.00
dvtrack206	0.00	940.40	0.10	0.25	0.00
dvtrack205	0.00	901.40	3.58	0.51	0.00
dvtrack204	0.00	899.20	0.80	0.30	0.00
dvtrack203	0.00	932.20	1.15	0.04	0.00
AGGREGATE	0.00	5,604.20	*****	*****	*****
MAXIMUM	0.00	974.40	3.58	0.51	0.00
MINIMUM	0.00	899.20	0.10	0.04	0.00
AVERAGE	0.00	934.03	1.06	0.27	0.00

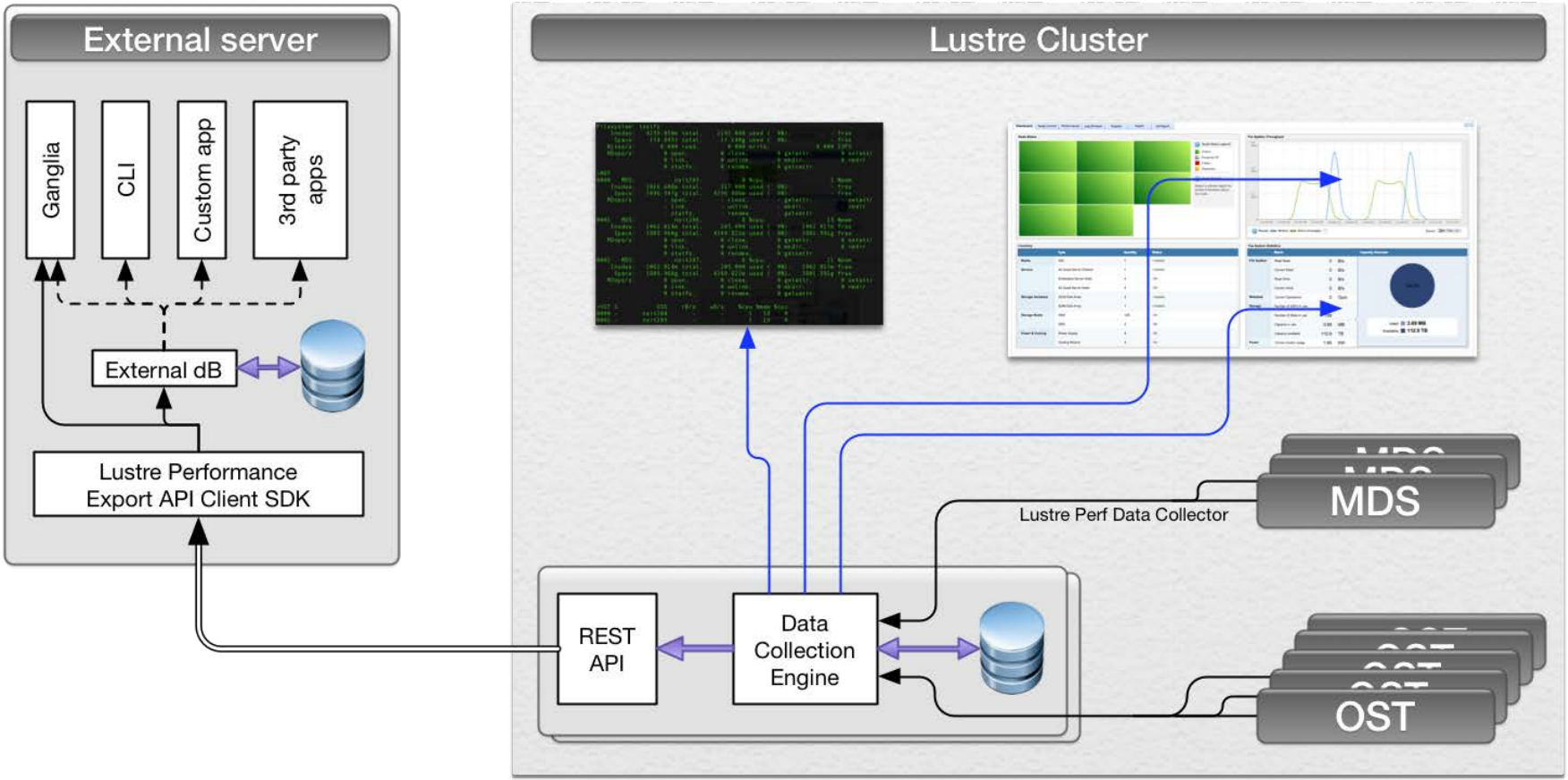
Lustre Performance Monitoring - Background

- Challenges with LMT
  - LMT does not report on extra MDS/MDTs in Lustre 2.5 DNE
  - LMT's trouble reporting data on large, busy clusters (dropped packets over multicast)
  - LMT Performance UI is legacy, Java based and currently no-one seems willing to enhance or maintain the code
  - Customers want Lustre performance data exported to their own tools
  - Customers could make use of better statistics than those from LMT

Lustre Performance Monitoring Goals

- Replace LMT dB and associated data collection with new solution
  - System is highly resilient, self monitoring and highly scalable.
- Data collection is delivered intelligently, for instance status data is updated every few seconds only when recovery is in progress otherwise sampling reduces to changes only.
- File Systems are discovered on the fly, no setup required.
- Introduce a REST based API for customers to obtain the data along with plugins to use the API to work with 3rd party tools such as Ganglia, Zenoss …
- Replace the basic LMT tools (specifically ltop)

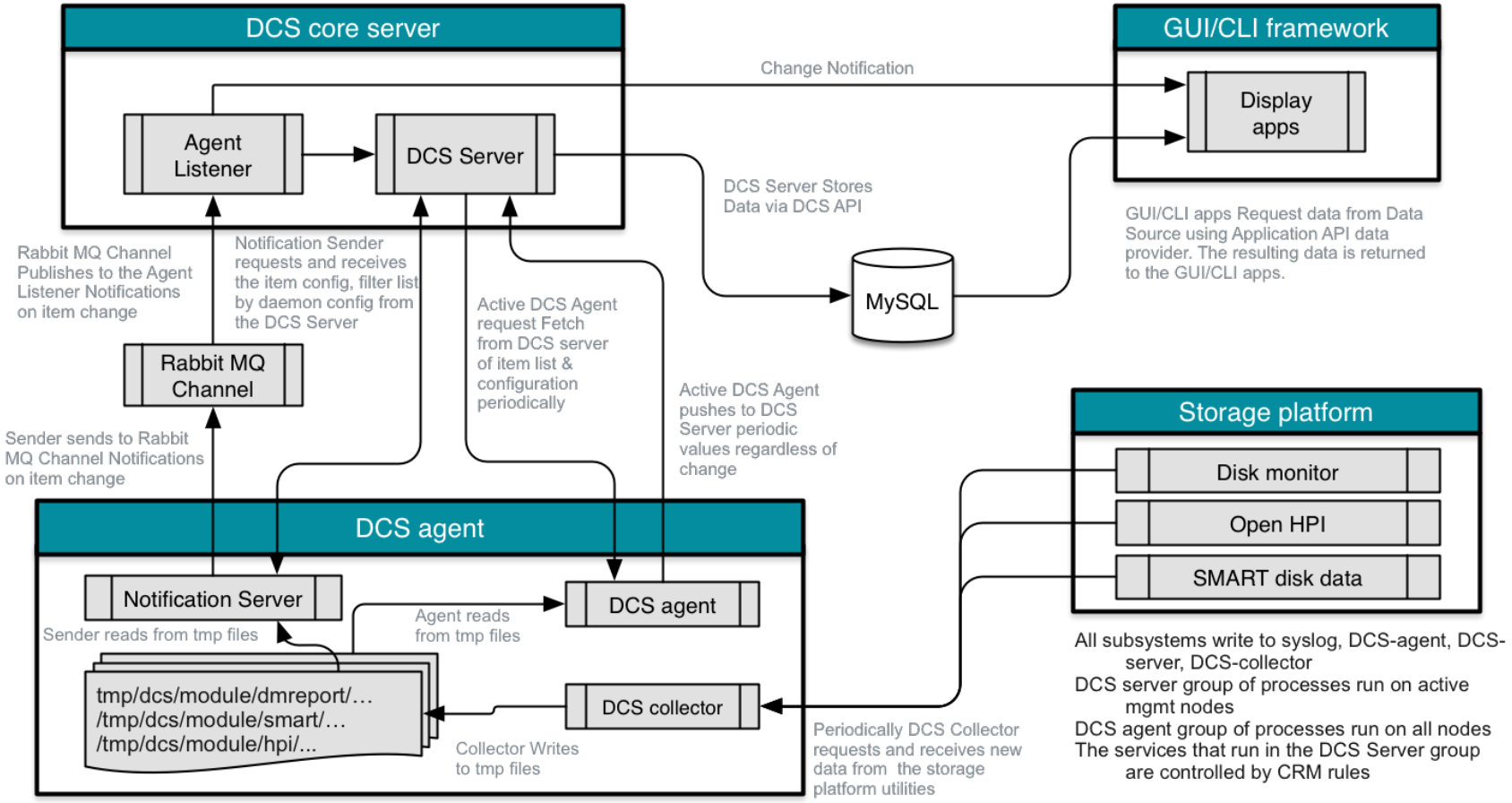
Lustre performance data collection



Details - Data Collection System (DCS)

- All historical data is stored in a standard DB - MySQL
- Python/JSON API for improved integration with Apps
- SNMP monitoring of ethernet switches and PDU’s
- Data collection can/should be event driven using triggers
- Self monitors, internal queues, I/O use, SQL and RAID use etc
- Tested to 20K Samples per second on Storage Hardware, Network and management servers
- Tested to 1 Million individual items monitored
- Tests show network use 100x lower than using Nagios and LMT for the same data monitored.
- Node Data Collectors - to reduce load on file system nodes we created a background data collector which pulls data from utilities such a disk monitor, HPI, SMART.

DCS Architecture



Data Collection

New scheme collects initially the same data that is stored in LMT today*.
Sample rate for some metrics (in bold) is slower,
these statistically cannot change much in 5 sec intervals:

- MDS/DNE hardware data
  - 5 sec samples: CPU Utilization, Memory Utilization
- MDT0 – MDTn data
  - 5 sec samples: close, connect, create, destroy, disconnect, getattr, getxattr, link, llog_init, mkdir, mknod, notify, open, process_config, quotactl, reconnect, rename, rmdir, setattr, statfs, unlink
  - 60 sec samples: **kbytes free, kbytes avail, free inodes, used inodes**
- OSS hardware data
  - 5 sec samples: CPU Utilization, Memory Utilization
- OST Data
  - 5 sec samples: read_bytes, write_bytes
  - 60 sec samples: **kbytes free, kbyte savail, free inodes, used inodes**

* Based on customer feedback

Phase 1 data collection details

Source	Data	Sample Delay
OSS	CPU utilization, Mem Usage	5 sec
MDS *	CPU utilization, Mem Usage	5 sec
OST	read_bytes, write_bytes	5 sec
OST	kbytesfree, kbytesavail, free_inodes, used_inodes	60 sec
MDT*	kbytesfree, kbytesavail, free_inodes, used_inodes	60 sec
MDT*	open, cancel_unused, clear_open_replay_data, close, create, crossdir_rename, done_writing, enqueue, find_cbdata, free_lustre_md, get_lustre_md, get_remote_perm, getattr, getattr_name, getstatus, getxattr, init_ea_size, intent_getattr_async, intent_lock, is_subdir, link, lock_match, mkdir, mknod, null_data, readpage, rename, renew_capa, revalidate_lock, rmdir, samedir_rename, set_lock_data, set_open_replay_data, setattr, setxattr, statfs, sync, unlink, unpack_capa, connect, destroy, llog_init, notify, process_config, quotactl, reconnect	5 sec

* includes DNEs/MDT1 - MDTn

lustre_perf --help

```
Usage: lustre_perf <subcommand> [option]
Options:
      --help

LustrePerf CLI client
Commands:
    fetch    Export historical lustre data between --starttime and --endtime
             to the local filesystem. Use '(cscli) cscli lustre_perf list'
             to find the location of the resulting output.
    ltop     display live information about a Lustre file system
             --help for more detail.
    list     List the full path of any existing log files --help for more detail
    status   Return the status of the last run command (or the currently running
             command if it is non blocking and a process is still running)
             --help for more detail.
    abort    Abort the currently running export job --help for more detail.
    clean    Delete all export files in the export folder
             --help for more detail.
```

lustre_perf --help

```
lustre_perf ltop --help

Usage: lustre_perf ltop
Options:
  -n, --no-summary  Omit summary of filesystem stats in output
  -f, --filter=     Filter by regular expression of target name.
                    [default: ]
      --help

Example Filters:
 •  all MDSs or OSSs with ID 0000 or 0001:   --filter='0000|0001'
 •  OSSs 3-7:                                --filter=OSS000[3-7]
 •  nodes with indices 2-5 or 9:             --filter=000[2-5,9]
 •  all MDSs:                                --filter=MDS
```

lustre_perf fetch sample

```
[admin@host]$lustre_perf fetch -s 2014-01-01t12:00:00+00:00 -e 2015-01-01t12:00:00+00:00
Lustre performance statistics export started
Please use 'lustre_perf status' to monitor progress and 'lustre_perf list' to find the
resulting output.
[admin@host]$lustre_perf status
Status: running - fetch for data from 2014-01-01t12:00:00+00:00 2015-01-01t12:00:00+00:00
[admin@host]$lustre_perf status
Status: completed partially - fetch for data from 2014-01-01t12:00:00+00:00 2015-01-
01t12:00:00+00:00
Filename: 20140101120000000_20140620092927000.csv.gz
Message: Last fetch ran out of disk quota.   Last successfully fetched record had a timestamp
of 2014-06-20t09:29:27+00:00. File contains data from 2014-01-01t12:01:23+00:00 to 2014-06-
20t09:29:27+00:00.

Please copy (scp) the last output file to a safe location off of the cluster and then free up
disk space by calling ‘lustre_perf clean’ (to remove the local copy).

You can export the next chunk of data with ‘lustre_perf fetch -s 2014-06-20t09:29:28+00:00 -e
2015-01-01t12:00:00+00:00’.
```

lustre_perf fetch sample

```
[admin@host]$lustre_perf list
/mnt/mgmt/var/lib/lustre_perf/data/201401011200000000_201406200929270000.csv.gz
[admin@host]$scp 201401011200000000_201406200929270000.csv.gz my-computer.network.com:/.
[admin@host]$lustre_perf clean
Successfully deleted all export files
[admin@host]$lustre_perf fetch -s 2014-06-20t09:29:28+00:00 -e 2015-01-01t12:00:00+00:00
Lustre performance statistics export started
Please use 'lustre_perf status' to monitor progress and 'lustre_perf list' to find the
resulting output.
[admin@host]$lustre_perf status
Status: running - fetch for data from 2014-06-20t09:29:28+00:00 to 2015-01-01t12:00:00+00:00
[admin@host]$lustre_perf status
Status: completed - fetch for data from 2014-06-20t09:29:28+00:00 2015-01-01t12:00:00+00:00
Filename: 201406200929280000_201501011200000000.csv.gz
Message: Successfully completed fetching data from 2014-06-20t09:29:28+00:00 to 2015-01-
01t12:00:00+00:00
[admin@host]$lustre_perf list
/mnt/mgmt/var/lib/lustre_perf/data/201406200929280000_201501011200000000.csv.gz
[admin@host]$scp 201406200929280000_201501011200000000.csv.gz my-computer.network.com:/.
```

DKRZ Mistral system – Phase 1

Compute

- 900+ Bullx Haswell blades
  - > 25,000 cores
  - >1.1 PFLOPs peak
- Total memory: 65 TB
- FDR interconnect

Storage

- ~20 PB usable
- 210 GB/s throughput*
- 27 HA OSS pairs
- 27 expansion unitss
  - 54 OSSes
  - 108 OSTs
  - 4,482 HDDs
- 1 MDS (Active/Passive)
- 4 DNEs (Active/Active)
- 10 Racks with rear cooling doors
  - Est full power usage: ~94 kW

* Estimated. Performance tests under way.

DKRZ Mistral node configuration …

```
admin@ds000 ~]$ show_nodes
-------------------------------------------------------------------------------------------------------------
Hostname          Role              Power State   Service state   Targets     HA Partner    HA Resources
-------------------------------------------------------------------------------------------------------------
ds000             MGMT                   On          -----       0 / 0       ds001             -----
ds001             (MGMT)                 On          -----       0 / 0       ds000             -----
ds002             MGS,(MDS)              On          N/a         0 / 0       ds003             None
ds003             MDS,(MGS)              On          Started     1 / 1       ds002             Local
ds004             OSS                    On          Started     2 / 2       ds005             Local
ds005             OSS                    On          Started     2 / 2       ds004             Local
ds006             OSS                    On          Started     2 / 2       ds007             Local
ds007             OSS                    On          Started     2 / 2       ds006             Local
ds008             OSS                    On          Started     2 / 2       ds009             Local
ds009             OSS                    On          Started     2 / 2       ds008             Local
ds010             OSS                    On          Started     2 / 2       ds011             Local
ds011             OSS                    On          Started     2 / 2       ds010             Local
….
ds057             OSS                    On          Started     2 / 2       ds056             Local
ds058             OSS                    On          Started     2 / 2       ds059             Local
ds059             OSS                    On          Started     2 / 2       ds058             Local
ds060             OSS                    On          Started     2 / 2       ds061             Local
ds061             OSS                    On          Started     2 / 2       ds060             Local
ds062             MDS                    On          Started     1 / 1       ds063             Local
ds063             MDS                    On          Started     1 / 1       ds062             Local
ds064             MDS                    On          Started     1 / 1       ds065             Local
ds065             MDS                    On          Started     1 / 1       ds064             Local
-------------------------------------------------------------------------------------------------------------
```

lustre_perf ltop

```
[admin@ds000 ~]$ lustre_perf ltop
Filesystem: lustre01
    Inodes:     26.870g total,    121.593k used (  0%),     26.870g free
     Space:     21.618p total,   9163.457g used (  0%),           - free
   Bytes/s:       0.000 read,         0.000 write,             0.000 IOPS
   MDops/s:         0 open,           0 close,         0 getattr,         0 setattr
                    0 link,           0 unlink,        0 mkdir,           0 rmdir
                    0 statfs,         0 rename,        0 getxattr
>MDT
0000   MDS:              ds003,            1 %cpu,                   2 %mem
    Inodes:   2016.608m total,     446.000 used (  0%),   2016.608m free
     Space:   3096.947g total,   4400.693m used (  0%),   3092.546g free
   MDops/s:         - open,           - close,         - getattr,         - setattr
                    - link,           - unlink,        - mkdir,           - rmdir
                    - statfs,         - rename,        - getxattr
0001   MDS:              ds062,            6 %cpu,                   3 %mem
    Inodes:   1062.014m total,     433.000 used (  0%),   1062.013m free
     Space:   3805.960g total,   4372.955m used (  0%),   3801.587g free
   MDops/s:         0 open,           0 close,         0 getattr,         0 setattr
                    0 link,           0 unlink,        0 mkdir,           0 rmdir
                    0 statfs,         0 rename,        0 getxattr
0002   MDS:              ds063,            7 %cpu,                   4 %mem
    Inodes:   1062.014m total,     433.000 used (  0%),   1062.013m free
     Space:   3805.960g total,   4372.955m used (  0%),   3801.587g free
   MDops/s:         - open,           - close,         - getattr,         - setattr
                    - link,           - unlink,        - mkdir,           - rmdir
                    - statfs,         - rename,        - getxattr
```

lustre_perf ltop --filter=OST

```
#[admin@ds000 ~]$ lustre_perf ltop –f OST
Filesystem: fs1
    Inodes:   5561.716m total,   1879.536k used (  0%),   5559.837m free
     Space:   1500.344t total,   5622.349g used (  0%),   1494.722t free
   Bytes/s:       0.000 read,         0.000 write,             0.000 IOPS
   MDops/s:        0 open,          0 close,         0 getattr,         0 setattr
                   0 link,          0 unlink,        0 mkdir,           0 rmdir
                   0 statfs,        0 rename,        0 getxattr
>OST S           OSS    rB/s    wB/s    %cpu %mem %spc
0000 -         ds004       0       0       2   27    0
0001 -         ds004       0       0       2   27    0
0002 -         ds005       0       0       2   27    0
0003 -         ds005       0       0       2   27    0
0004 -         ds006       0       0       2   27    0
0005 -         ds006       0       0       2   27    0
0006 -         ds007       0       0       2   26    0
0007 -         ds007       0       0       2   26    0
0008 -         ds008       0       0       2   27    0
0009 -         ds008       0       0       2   27    0
000a -         ds009       0       0       2   26    0
000b -         ds009       0       0       2   26    0
000c -         ds010       0       0       2   27    0
…
0071 -         ds060       0       0       2   27    0
0072 -         ds061       0       0       2   27    0
0073 -         ds061       0       0       2   27    0
--------------------------------------------------------------------------------
```

lustre_perf ltop --filter=MDT0001 --no-summary

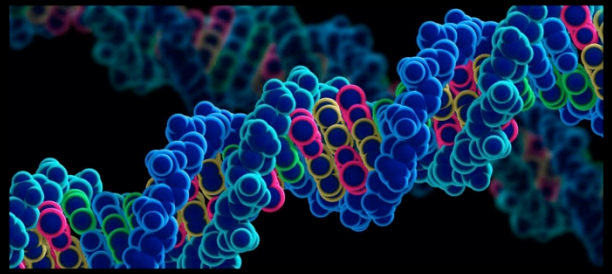
```
#[admin@ds000 ~]$ lustre_perf ltop –f MDT0001 –no-summary
>MDT
0001   MDS:             ds062,           6 %cpu,                  3 %mem
    Inodes:   1062.014m total,     433.000 used (  0%),   1062.013m free
     Space:   3805.960g total,   4372.955m used (  0%),   3801.587g free
   MDops/s:        0 open,         0 close,        0 getattr,        0 setattr
                   0 link,         0 unlink,       0 mkdir,          0 rmdir
                   0 statfs,       0 rename,       0 getxattr
-------------------------------------------------------------------------

#[admin@ds000 ~]$ lustre_perf ltop –f OST004 –no-summary

>OST S           OSS    rB/s    wB/s    %cpu %mem %spc
0004 -         ds006       0       0       2   27    0
```

Live demo - DKRZ Mistral system

Future plans ??

- The Data Collection System toolkit will be expanded to cover other dynamic variables for simplified export such as:
  - System temperature (ambient, disk drives, servers)
  - Power usage (both in system and PDU measurements)
- The LTOP functionality equipped with a customizable GUI
- Long term historical data will be available for export or in GUI visualization (granularity will decrease as data gets older …)
- Health data (Lustre as well as systems) will be available in the toolkit
- More Plugins for more 3rd party tools
  - Please let us know your preference.
- Date for availability is being discussed by management and is TBD.

Seagate is HPC Storage


Unmatched speed and efficiency from the
Trusted Leader in HPC storage

SEAGATE