CRAY SWMG

CARIBOU/SONEXION @ CSCS demo

Alpha release (03/2017)

.

.

.

.

Matteo Chesi / Jg Piccinali

TDS: Cray XC (dom)

Lustre layout Lustre implements a separation of data and metadata:

- The metadata is stored on a Metadata Target (`MDT`)
- The data is stored on a number of Object Storage Targets (`OSTs`)
- A Metadata Server (`MDS`) serves all file system requests for metadata, and it looks after the MDT
- A number of Object Storage Servers (`OSS`) each look after several OSTs and serve requests for data on those OSTs.

```
piccinal@dom101:~$
df -h /scratch/*
Filesystem                                          Size  Used Avail Use% Mounted on
148.187.4.99@o2ib6002:148.187.4.100@o2ib6001:/snx11031  170T  1.4T  166T   1% /scratch/snx1600tds
148.187.4.82@o2ib3002:148.187.4.83@o2ib3001:/snx11104   226T  296G  223T   1% /scratch/snx2000tds
```

```
piccinal@dom101:~$
lfs osts
OBDS::
0: snx11104-OST0000_UUID ACTIVE
1: snx11104-OST0001_UUID ACTIVE
OBDS::
0: snx11031-OST0000_UUID ACTIVE
1: snx11031-OST0001_UUID ACTIVE
```

```
piccinal@dom101:~$
lfs df -h /scratch/snx1600tds/
UUID                       bytes        Used   Available Use% Mounted on
snx11031-MDT0000_UUID        2.1T        4.3G        2.1T   0% /scratch/snx1600tds[MDT:0]
snx11031-OST0000_UUID       84.5T      338.9G       83.2T   0% /scratch/snx1600tds[OST:0]
snx11031-OST0001_UUID       84.5T        1.0T       82.5T   1% /scratch/snx1600tds[OST:1]

filesystem summary:       169.1T        1.3T      165.7T   1% /scratch/snx1600tds
```

TDS: Cray XC (dom)

> Lustre striping Lustre allows the user to have explicit control over how a file is striped over the OSTs: chunks are sent to the different OSTs to improve disk bandwidth.

- export `MPICH_MPIIO_STATS`=1
- srun -n192 ./GNU.DOM

```
lt -h out_1120x720x80.16x12.000*
-rw-r--r-- 1 piccinal csstaff 3.7G Mar 24 14:32 out_1120x720x80.16x12.0000.bin
-rw-r--r-- 1 piccinal csstaff 3.7G Mar 24 14:32 out_1120x720x80.16x12.0001.bin
-rw-r--r-- 1 piccinal csstaff 3.7G Mar 24 14:33 out_1120x720x80.16x12.0002.bin
-rw-r--r-- 1 piccinal csstaff 3.7G Mar 24 14:33 out_1120x720x80.16x12.0003.bin
```

- lfs setstripe `-c 1` and lfs setstripe `-c 2`

```
+-------------------------------------------------------+
| MPIIO write access patterns for out_1120x720x80.16x12.0003.bin
|   independent writes      = 0
|   collective writes       = 1920
|   independent writers     = 0
|   aggregators             = 1
|   stripe count            = 1
|   stripe size             = 1048576
|   system writes           = 3750
|   stripe sized writes     = 3750
|   total bytes for writes  = 3932160000 = 3750 MiB = 3 GiB
|   ave system write size   = 1048576
|   read-modify-write count = 0
|   read-modify-write bytes = 0
|   number of write gaps    = 0
|   ave write gap size      = NA
| See "Optimizing MPI I/O on Cray XE Systems" S-0013-20 for explanations.
+-------------------------------------------------------+
Testing get_procmem... 7516160.000000 45846528.000000 38330368.000000
written grids of 80,80,80
written 4 iterations
MPI Elapsed time: 50.929825 sec
average  0.028762 Gbytes/sec
real 54.23
```

```
+-------------------------------------------------------+
| MPIIO write access patterns for out_1120x720x80.16x12.0003.bin
|   independent writes      = 0
|   collective writes       = 1920
|   independent writers     = 0
|   aggregators             = 2
|   stripe count            = 2
|   stripe size             = 1048576
|   system writes           = 3750
|   stripe sized writes     = 3750
|   total bytes for writes  = 3932160000 = 3750 MiB = 3 GiB
|   ave system write size   = 1048576
|   read-modify-write count = 0
|   read-modify-write bytes = 0
|   number of write gaps    = 0
|   ave write gap size      = NA
| See "Optimizing MPI I/O on Cray XE Systems" S-0013-20 for explanations.
+-------------------------------------------------------+
Testing get_procmem... 7516160.000000 42070016.000000 34553856.000000
written grids of 80,80,80
written 4 iterations
MPI Elapsed time: 25.865780 sec
average  0.056632 Gbytes/sec
real 29.15
```

Cray Caribou

CRAY THE SUPERCOMPUTER COMPANY

admin

- Identity
- Alarms and Notifications
- Statistics
  - Sonexion Status
  - Jobs
  - OSTs
  - OST Performance

Caribou

Last 15 minutes

snx11031

OST I/O	Metadata ops	Capacity
0.00 B/s Average Read	0.00 k/s Requests	98.01 % Available
0.00 B/s Average Write		169.07 TB Total

Jobs	IB network	OST
1 / 1 total / error	26 / 35 switches / HAs	2 / 2 total / warning

0 Health Events

snx11104

OST I/O	Metadata ops	Capacity
0.00 B/s Average Read	0.00 k/s Requests	98.68 % Available
0.00 B/s Average Write		225.80 TB Total

Jobs	IB network	OST
5 / 5 total / error	26 / 35 switches / HAs	2 / 2 total / warning

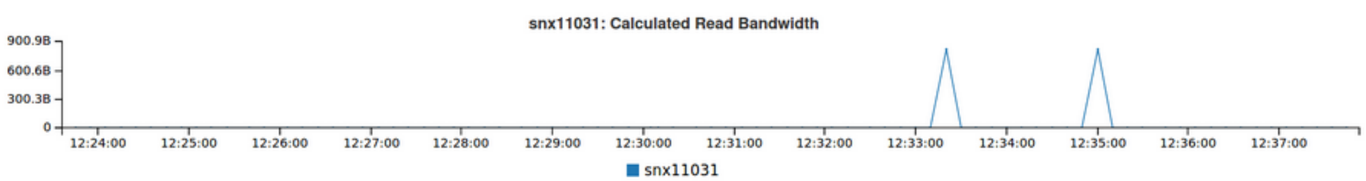
0 Health Events

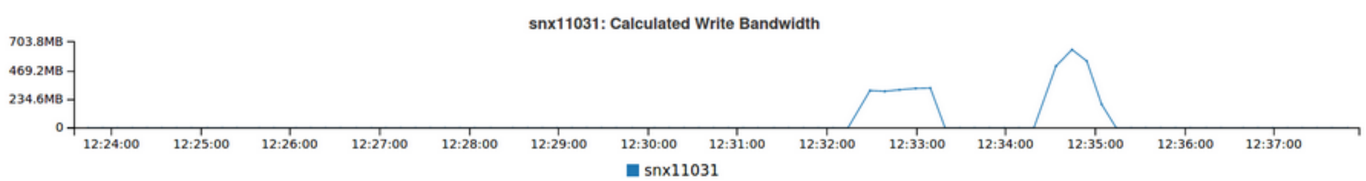
Add Sonexion

Jobs

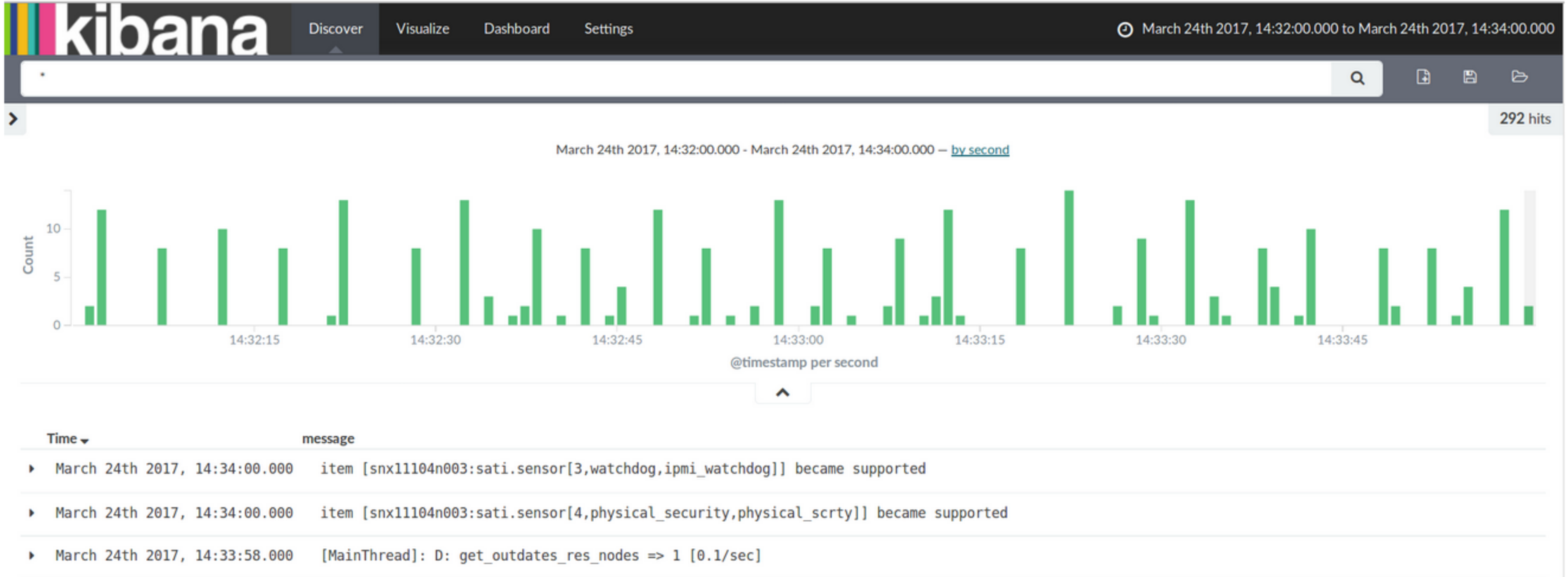
snx11031 — Last 15 minutes

Job	User ID	Application	Start Time	End Time	Avg. I/O Size	Metadata Ratio
649566	21099		2017-03-24T13:32:20Z	2017-03-24T13:33:14Z	1.0MB	4.8MB
649567	21099		2017-03-24T13:34:29Z	2017-03-24T13:34:58Z	1.0MB	3.7MB

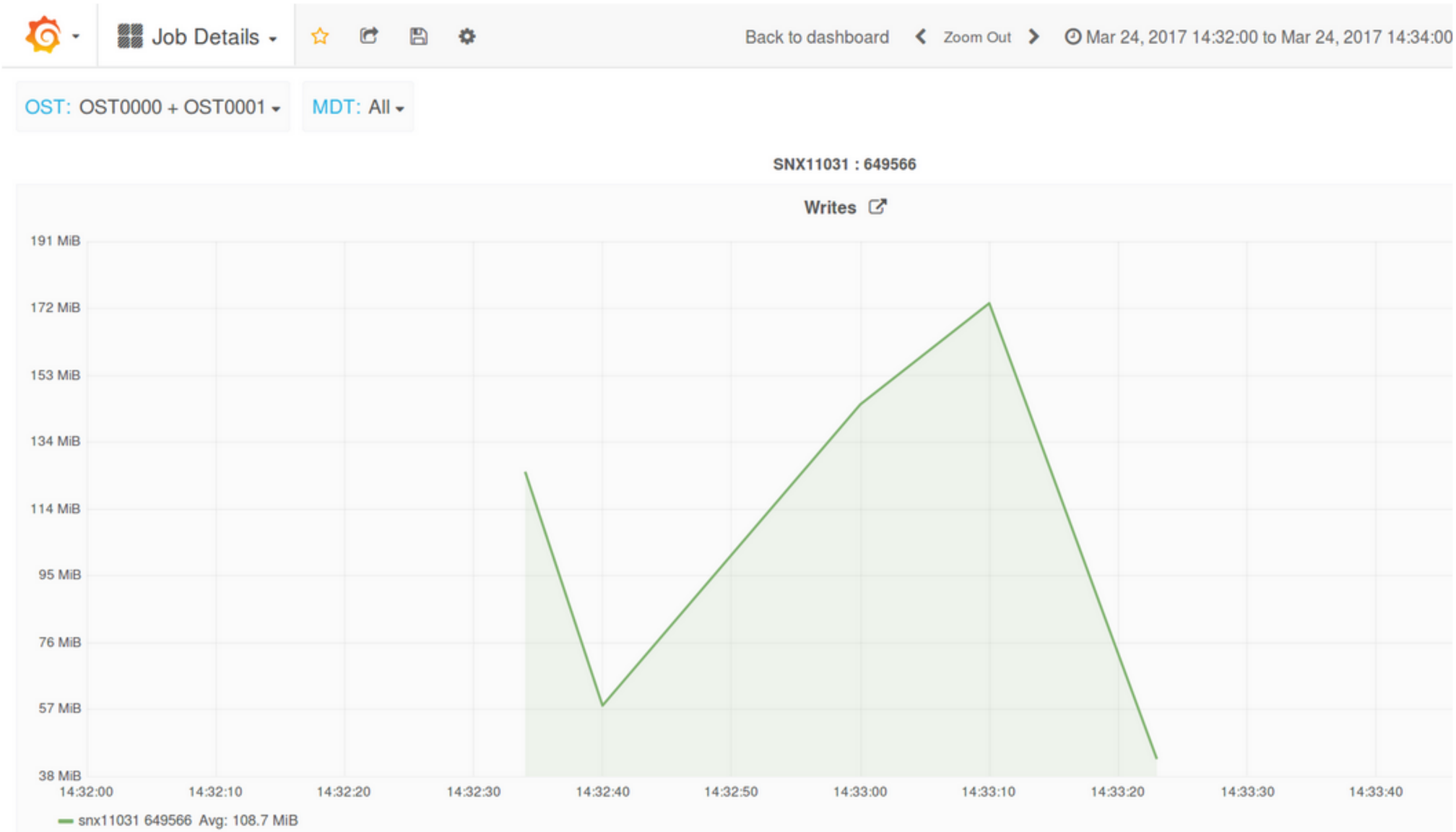


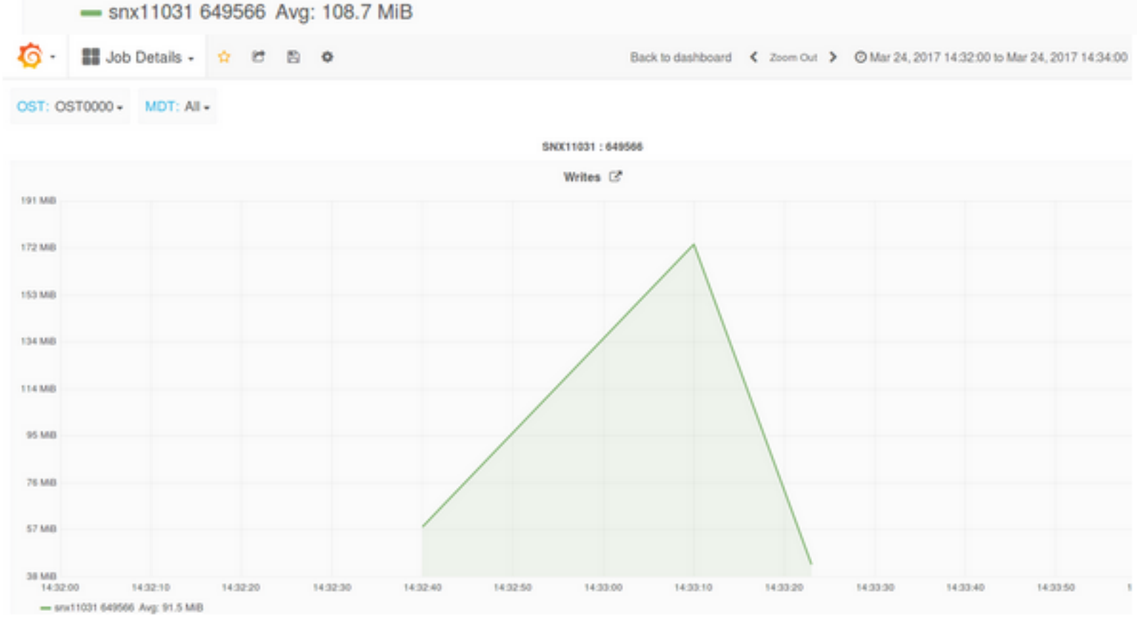


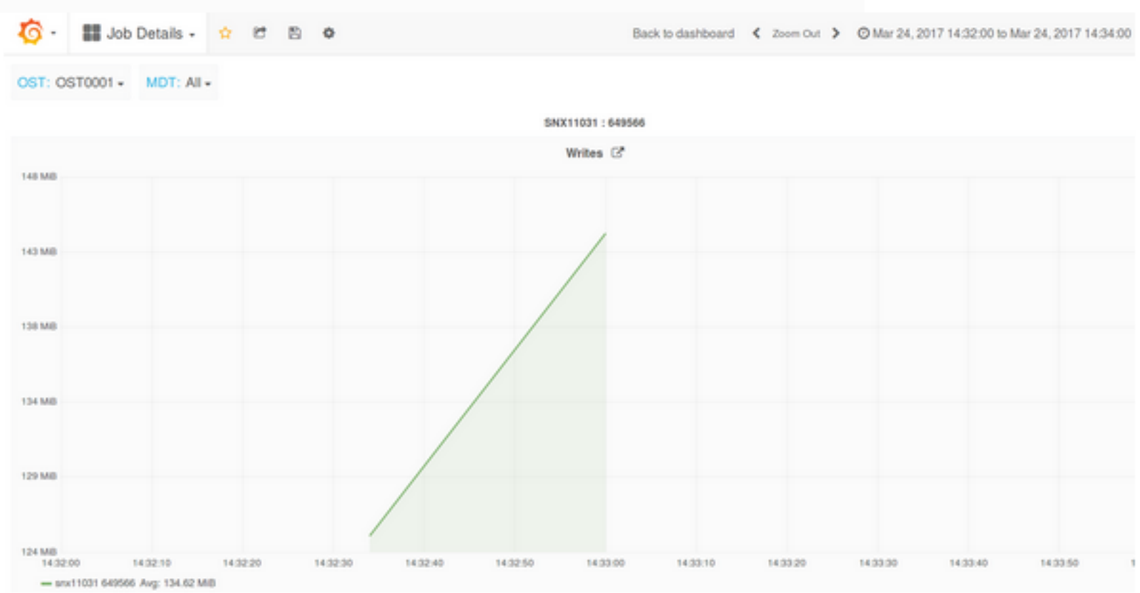
kibana


kibana
Discover
Visualize
Dashboard
Settings
March 24th 2017, 14:32:00.000 to March 24th 2017, 14:34:00.000
*
292 hits
March 24th 2017, 14:32:00.000 - March 24th 2017, 14:34:00.000 — by second
Count
10
5
0
14:32:15
14:32:30
14:32:45
14:33:00
14:33:15
14:33:30
14:33:45
@timestamp per second
Time
message
March 24th 2017, 14:34:00.000 item [snx11104n003:sati.sensor[3,watchdog,ipmi_watchdog]] became supported
March 24th 2017, 14:34:00.000 item [snx11104n003:sati.sensor[4,physical_security,physical_scrty]] became supported
March 24th 2017, 14:33:58.000 [MainThread]: D: get_outdates_res_nodes => 1 [0.1/sec]

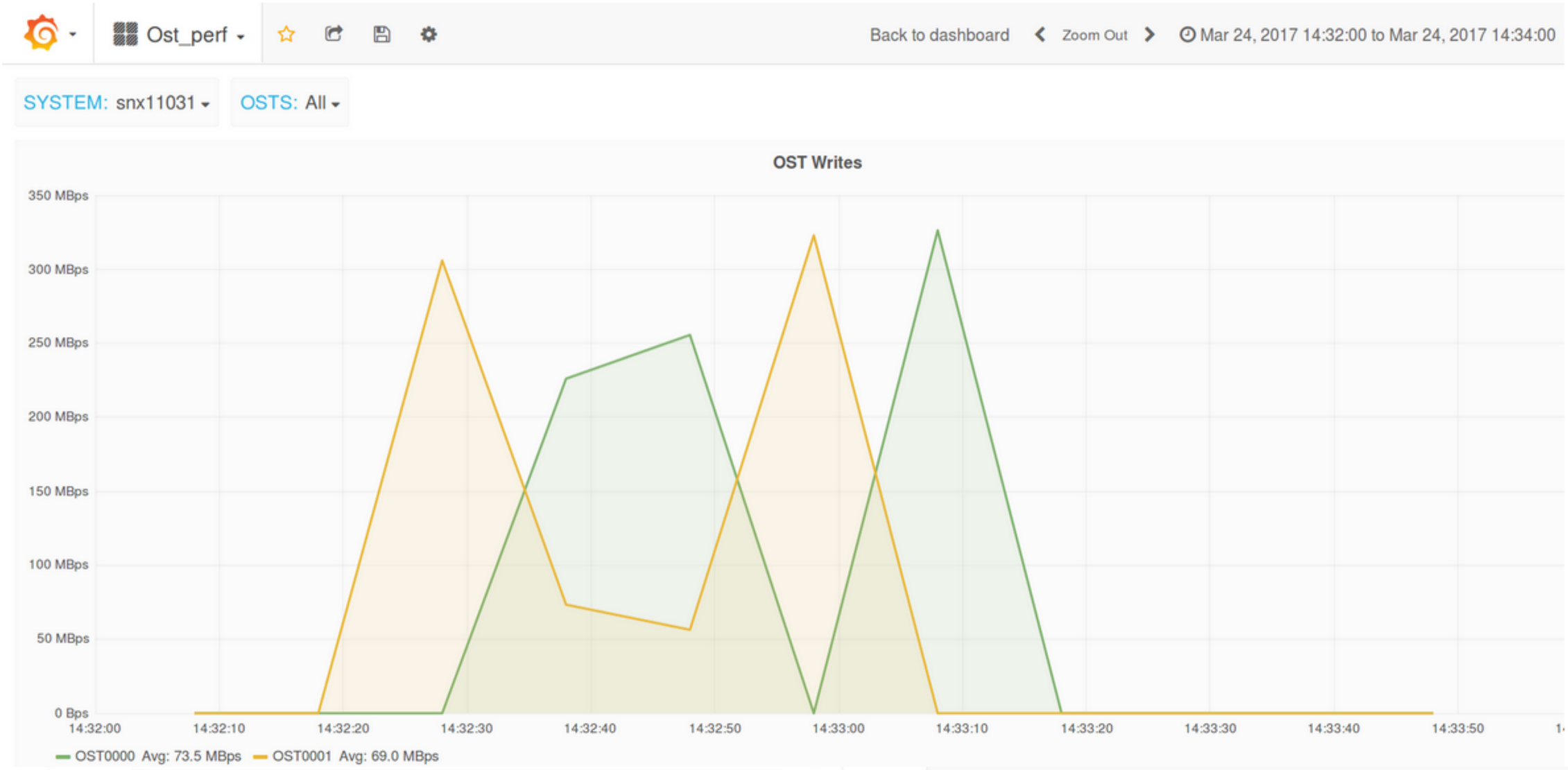
grafana:job

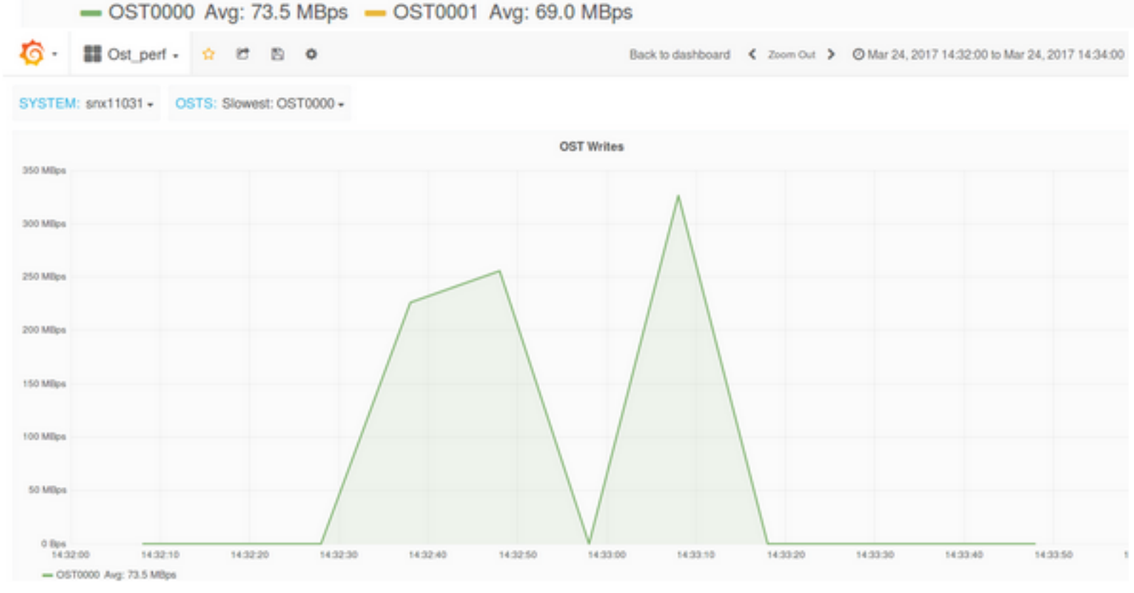

Job Details
Back to dashboard
Zoom Out
Mar 24, 2017 14:32:00 to Mar 24, 2017 14:34:00
OST: OST0000 + OST0001
MDT: All
SNX11031 : 649566
Writes
191 MiB
172 MiB
153 MiB
134 MiB
114 MiB
95 MiB
76 MiB
57 MiB
38 MiB
14:32:00
14:32:10
14:32:20
14:32:30
14:32:40
14:32:50
14:33:00
14:33:10
14:33:20
14:33:30
14:33:40
snx11031 649566 Avg: 108.7 MiB

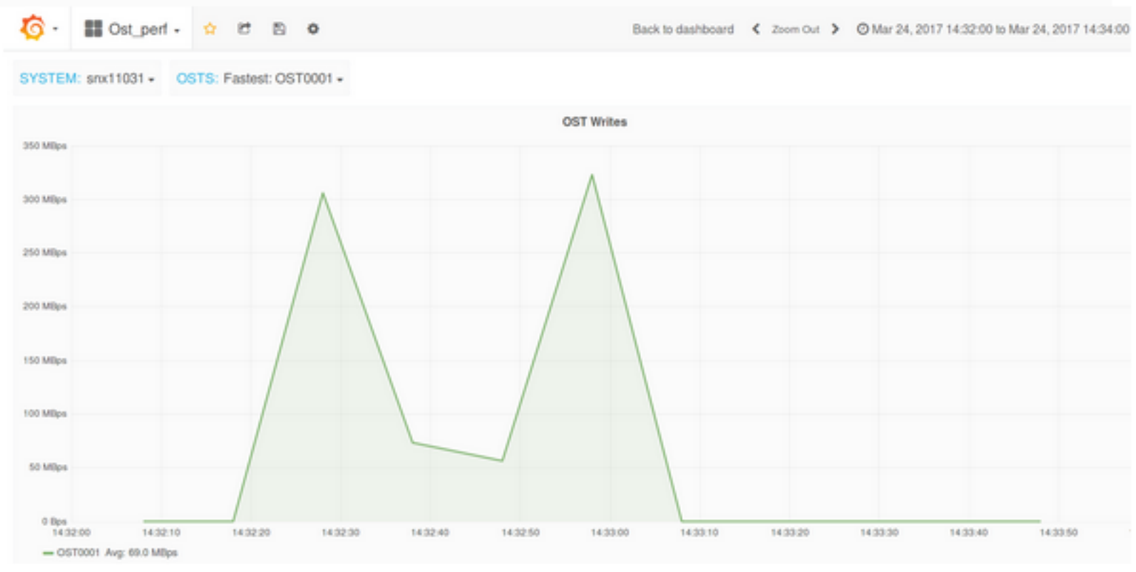

Job Details
Back to dashboard
Zoom Out
Mar 24, 2017 14:32:00 to Mar 24, 2017 14:34:00
OST: OST0000
MDT: All
SNX11031 : 649566
Writes
snx11031 649566 Avg: 91.5 MiB


Job Details
Back to dashboard
Zoom Out
Mar 24, 2017 14:32:00 to Mar 24, 2017 14:34:00
OST: OST0001
MDT: All
SNX11031 : 649566
Writes
snx11031 649566 Avg: 134.62 MiB

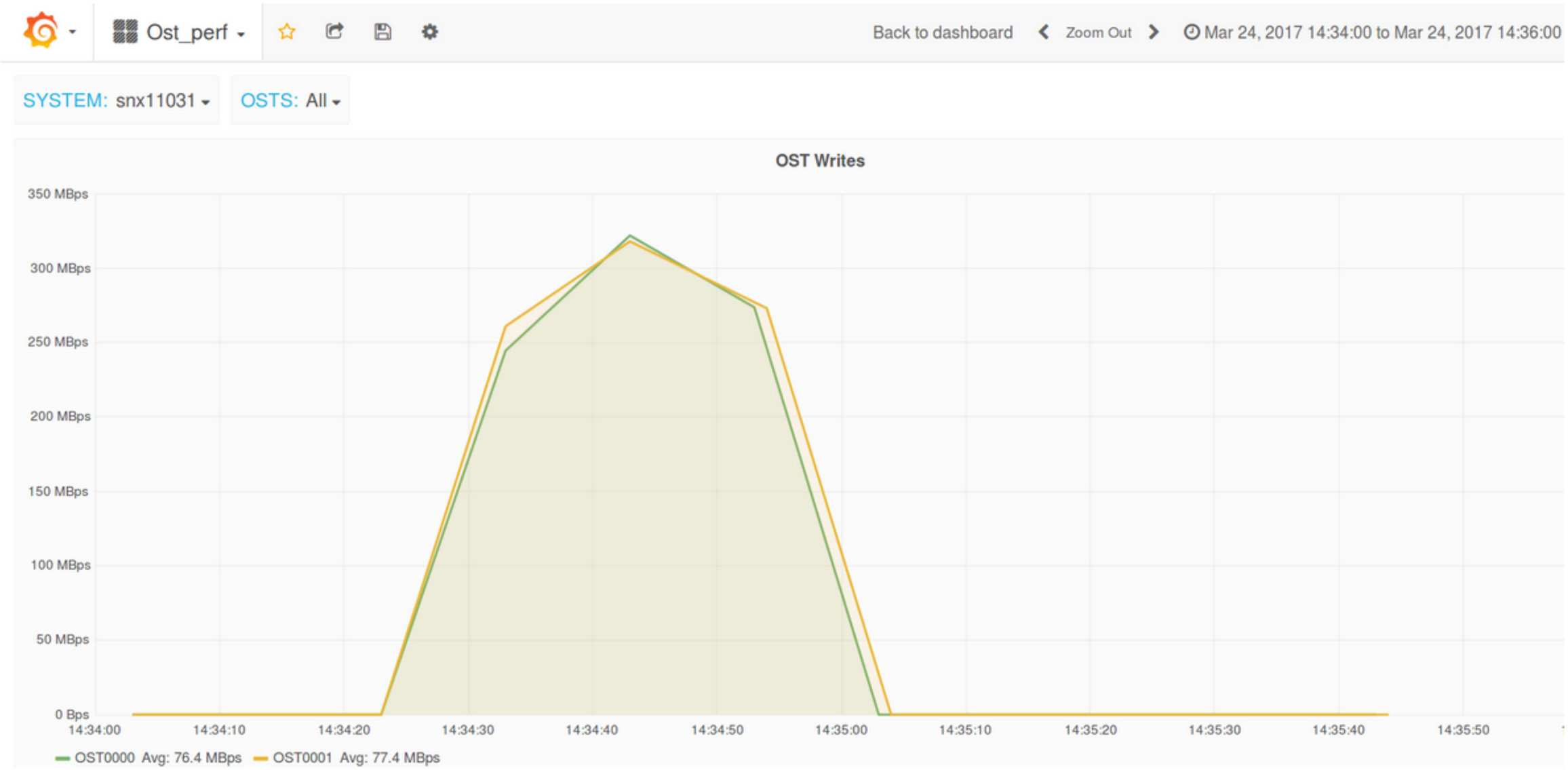
grafana:ost

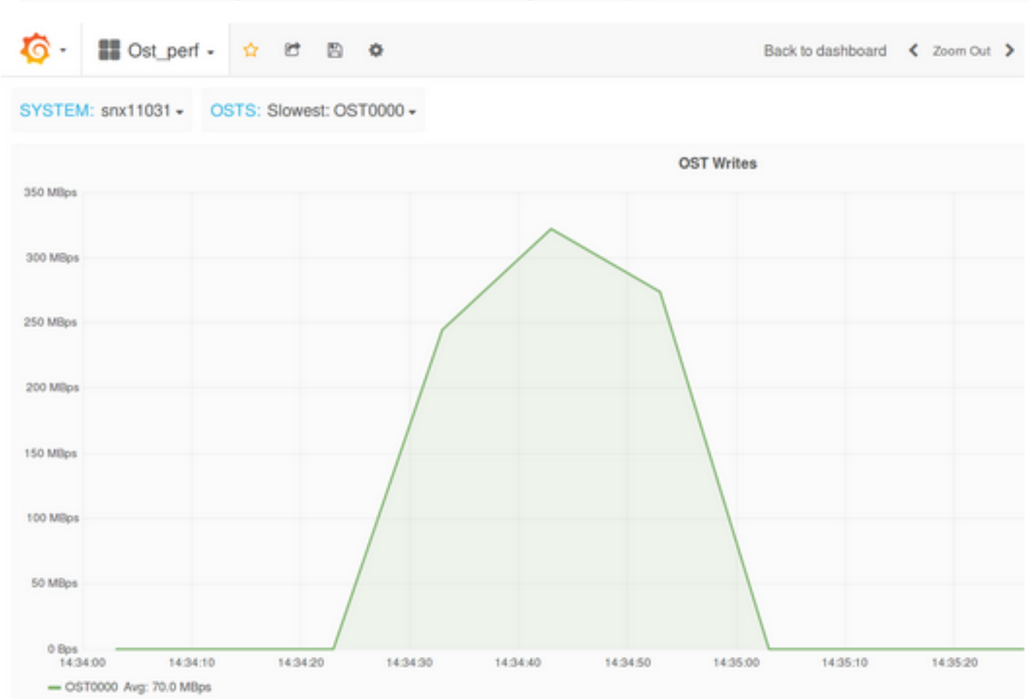

Ost_perf
Back to dashboard
Zoom Out
Mar 24, 2017 14:32:00 to Mar 24, 2017 14:34:00
SYSTEM: snx11031
OSTS: All
OST Writes
350 MBps
300 MBps
250 MBps
200 MBps
150 MBps
100 MBps
50 MBps
0 Bps
14:32:00
14:32:10
14:32:20
14:32:30
14:32:40
14:32:50
14:33:00
14:33:10
14:33:20
14:33:30
14:33:40
14:33:50
OST0000 Avg: 73.5 MBps
OST0001 Avg: 69.0 MBps

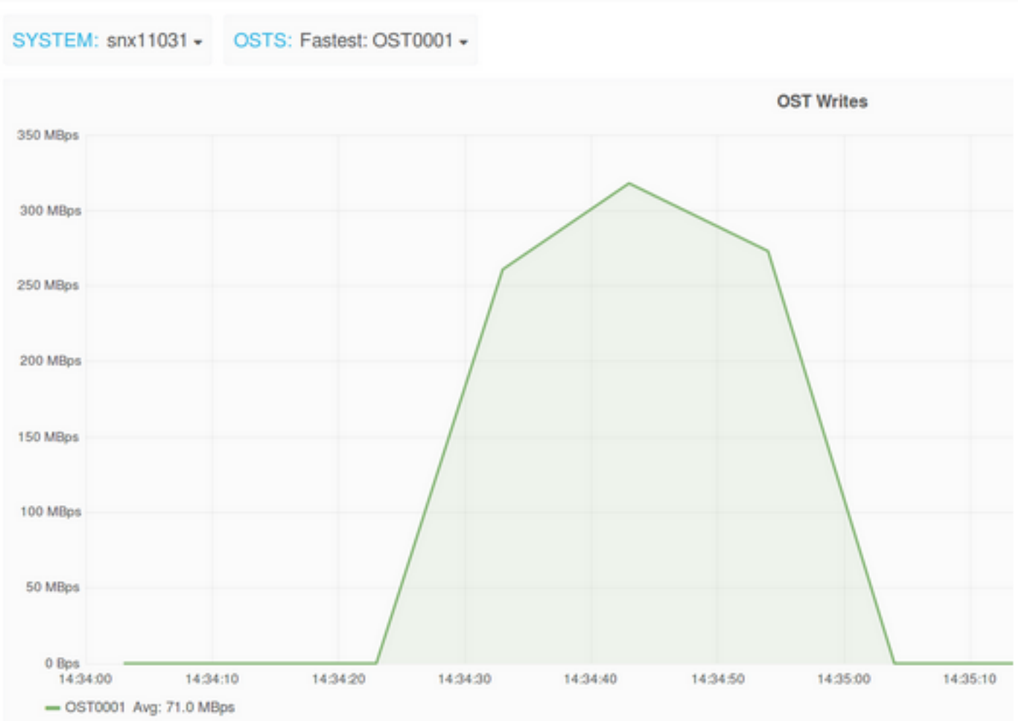

Ost_perf
Back to dashboard
Zoom Out
Mar 24, 2017 14:32:00 to Mar 24, 2017 14:34:00
SYSTEM: snx11031
OSTS: Slowest: OST0000
OST Writes
OST0000 Avg: 73.5 MBps


Ost_perf
Back to dashboard
Zoom Out
Mar 24, 2017 14:32:00 to Mar 24, 2017 14:34:00
SYSTEM: snx11031
OSTS: Fastest: OST0001
OST Writes
OST0001 Avg: 69.0 MBps

grafana:striping on 2 osts







CUG'17

- Tuesday May 9th / 4:40pm

`HPC Storage Operations BOF`

> Matteo Chesi (CSCS), Tina Declerck (NERSC), Oliver Treiber (ECMWF)

- Thursday May 11th / 10:30am

`Caribou : Monitoring and Metrics for Cray Sonexion Storage`

> Patricia Langer, Craig Flaskerud (Cray)