



A Case for Using CephFS

BJ Lougee
Software Engineer
Federal Reserve Bank of Kansas City
Center for the Advancement of Data and Research in Economics

The opinions expressed herein are those of the authors and do not reflect the views of the Federal Reserve Bank of Kansas City or Federal Reserve System



Our Dilemma

- PanFS is used in our HPC environment
- SAS is used for some very important workloads
 - Sporadic workloads but very IO intensive
- Carved out 70TB of PanFS for SAS temp workspace
- Needed to reclaim that space
 - Cheaper and/or faster parallel filesystem
 - It's only for temp/transient data

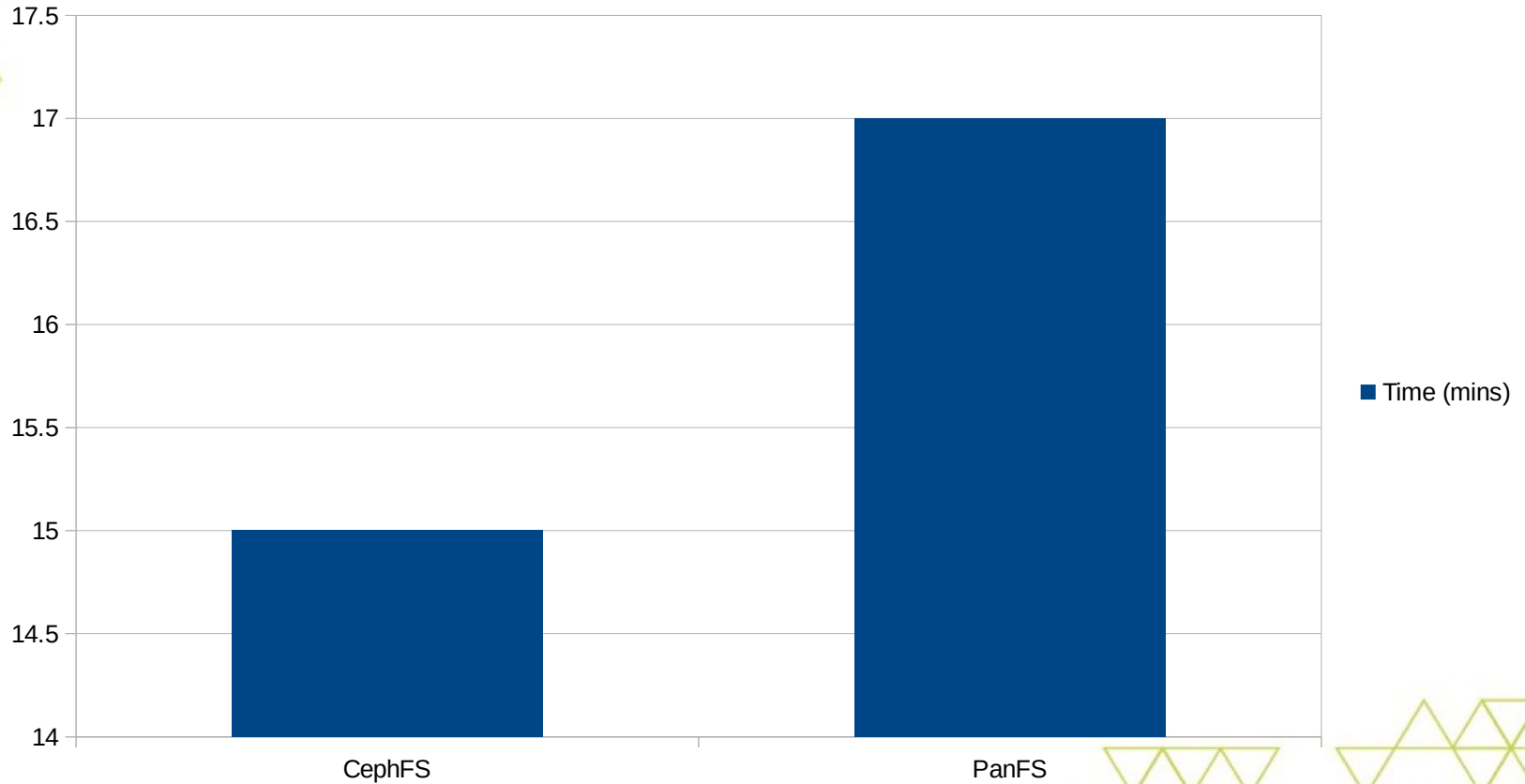
Choices

- Lustre
 - Hardware redundant
 - Open Source
 - Scales
- CephFS
 - **Software redundant**
 - Open Source
 - Scales
 - **Ceph ties in well with our OpenStack plans**

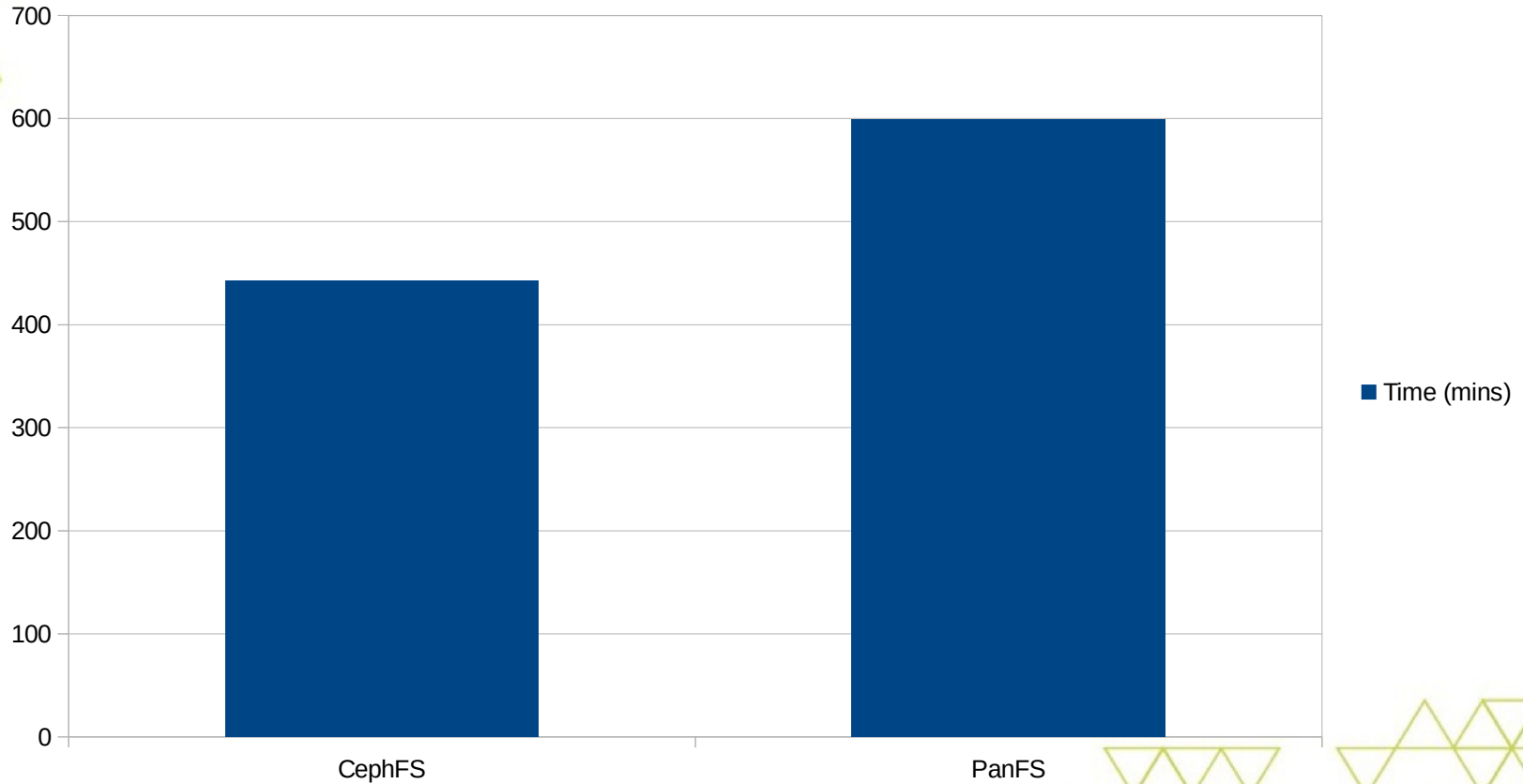
CephFS Testing

- #1 goal : What kind of minimum can I get away with?
- Various types of usage paradigms
- Tested performance between Giant and Hammer
- Tested a lot of different config options in ceph.conf
- Tested Different Journal Setups

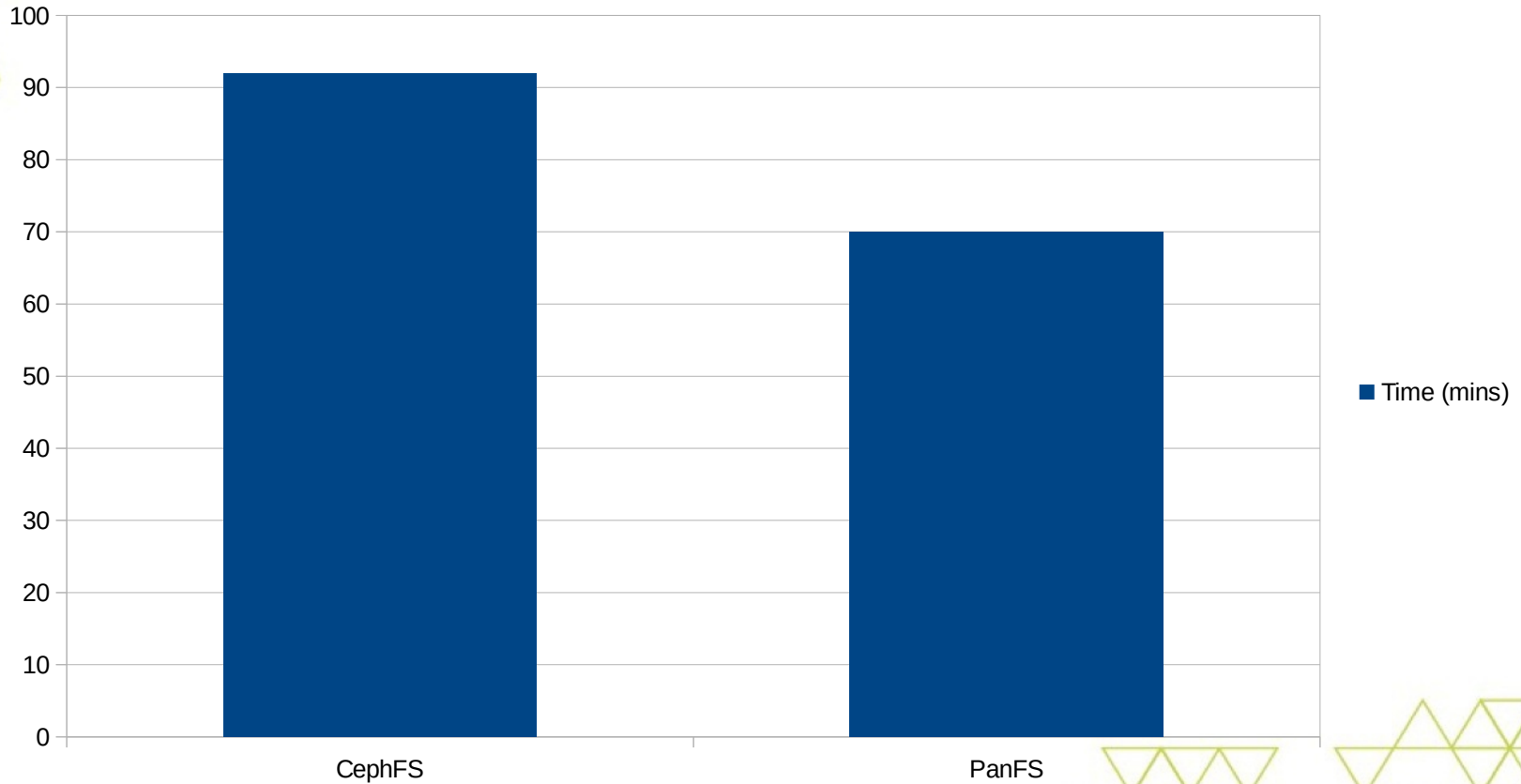
SAS Regression



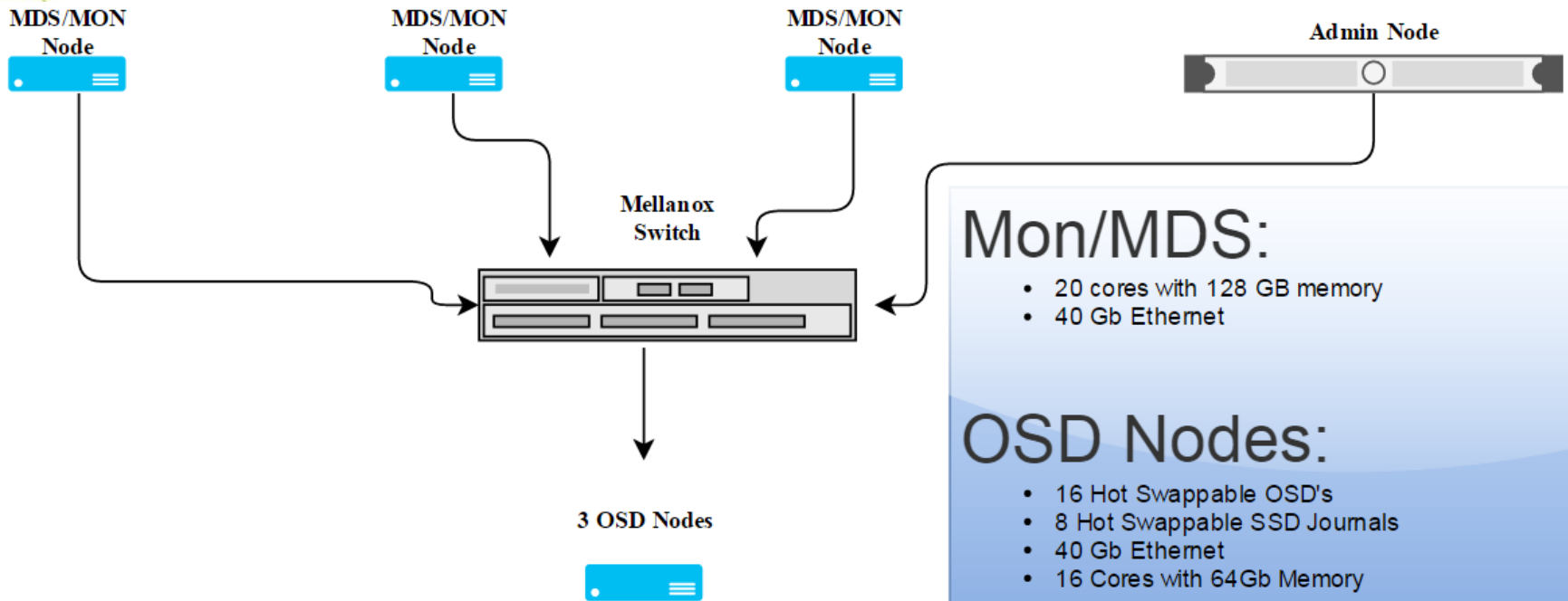
SAS Bench Large



cp 100,000 1MB files



CephFS in Production



Ceph status

```
root@c
Every 0.5s: ceph -s

cluster a21445d2-
health HEALTH_OK
monmap e3: 3 mons at {                01=                :6789/0,
      election epoch 72, quorum 0,1,2      01,
mdsmap e570: 1/1/1 up {0=                02=up:active}, 2 up:standby
osdmap e22042: 52 osds: 48 up, 48 in
pgmap v3925792: 4608 pgs, 2 pools, 2504 GB data, 1328 kobjects
      7555 GB used, 167 TB / 174 TB avail
      4608 active+clean
client io 269 MB/s rd, 1196 MB/s wr, 727 op/s
```

```
root@c
Every 0.5s: ceph osd pool stats

pool cephfs_metadata id 10
  client io 31550 B/s rd, 1 op/s

pool cephfs_data id 11
  client io 269 MB/s rd, 1196 MB/s wr, 726 op/s
```

Collectl

```

## RECORD 3112132 >>>          01 <<< (1446753044.001) (Thu Nov  5 13:50:44 2015) ###
# DISK STATISTICS (/sec)0      0  0      0  0  0  0  0  0  0  0
#Name  <-----reads-----><-----writes-----><-----averages-----> Pct
#Name  KBytes Merged  IOs Size  KBytes Merged  IOs Size  RWSize  QLen  Wait  SvcTim Util
dd      0      0      0  0  221860    5 507 438    437  114  264    1  95
dj      0      0      0  0  154552    5 389 397    397   36   78    2  83
dl      0      0      0  0      0      0  0  0      0      0  0  0  0  0
dk      0      0      0  0  105288    2 276 381    381   14   24    1  46
dm      0      0      0  0   71660    0 190 377    377    9   25    2  48
do      0      0      0  0  239320    6 538 445    444  125 266    1  98
dp      0      0      0  0   57636    0 160 360    360    9   24    2  41
dn      0      0      0  0   90380    0 230 393    392   17   44    2  58
dq      0      0      0  0   93960    4 252 373    372   27   86    2  65
df      8      0      2  4   32972    8  72 458    445    3   14    3  26
dh      8      0      2  4   86224    48 185 466    461   34   23    4  84
da     4116    0  13 317  28924    7  63 459    434    3   12    3  26
db      68     0  16  4  66172    16 143 463    416    4   24    5  87
de     12     0  3  4  43296    10  95 456    441    3   16    4  40
dg      4      0  1  4  61296    14 134 457    454    3   17    4  61
di      0      0      0  0   83976    20 183 459    458    3   16    4  79
dc     4120    0  13 317  41236    10  90 458    440    3   13    4  46
dr      16     0  3  5  99136    27 198 501    493  162 819    4  99
ds     12     0  3  4   8800     2  20 440    383    3   11    3  6
dt     24     0  6  4  55676    13 122 456    435    3   15    4  51
du     36     0  9  4  51596    12 113 457    423    3   16    4  52
dy     24     0  6  4  67780     8 137 495    474   31 329    5  72
dx     32     0  8  4  76132    11 159 479    456   15 176    4  83
dw     24     0  6  4  78412    15 167 470    453    7   59    4  84
dv     8272    0  34 243 41432     8  87 476    410    7   43    5  61
dz      0      0      0  0      0      0  0  0      0      0  0  0  0
daa     0      0      0  0      0      0  0  0      0      0  0  0  0
dab     0      0      0  0      0      0  0  0      0      0  0  0  0
dad     0      0      0  0      0      0  0  0      0      0  0  0  0
dac     0      0      0  0      0      0  0  0      0      0  0  0  0
dm-0    0      0      0  0      0      0  0  0      0      0  0  0  0
dm-1    0      0      0  0      0      0  0  0      0      0  0  0  0

```

```

## RECORD 3042005 >>>          02 <<< (1446753044.001) (Thu Nov  5 13:50:44 2015) ###
# DISK STATISTICS (/sec)vcTim Util  0  0  0
#Name  <-----reads-----><-----writes-----><-----averages-----> Pct
#Name  KBytes Merged  IOs Size  KBytes Merged  IOs Size  RWSize  QLen  Wait  SvcTim Util
sdj      0      0      0  0   72900     0 203 359    359    8   13    1  30
sdi      0      0      0  0   67400     0 192 351    351    5    8    1  28
sdb      0      0      0  0      0      0  0  0      0      0  0  0  0
sdb     12     0  3  4   53528     9 110 487    473   24 332    5  66
sdc      0      0      0  0   16612     4  36 461    461    4   30    6  22
sda     12     0  2  6   83976    14 170 494    488   59 605    5  99
sdd      0      0      0  0   66960    34 147 456    455    6   33    6  97
sde     4108    0  10 411 36864     0  72 512    499   34 270    5  45
sdg      0      0      0  0   65936    16 144 458    457    5   33    6  96
sdf      4      0  1  4   70052    16 152 461    457    5   32    6  96
sdh      0      0      0  0   74048     8 151 490    490   19 194    5  88
sdm      0      0      0  0  110992     2 280 396    396   48 120    2  69
sdl      0      0      0  0   73520     0 196 375    375   11  29    2  49
sdn      0      0      0  0   126236    1 314 402    402   24  58    2  74
sdo      0      0      0  0   176232    4 414 426    425   65 148    2  86
sdp      0      0      0  0   156768    1 413 380    379   15  28    1  68
sdq      0      0      0  0   126452     0 326 388    387   12  23    1  58
sdr     24     0  4  6   64020    16 140 457    444    5   33    6  96
sdt      0      0      0  0   84896    25 174 488    487   55 200    5  98
sdv     12     0  3  4   80704    32 166 486    477   64 130    5  97
sdu     12     0  2  6   64516    16 141 458    451    5   33    6  96
sds      0      0      0  0   53608    12 117 458    458    5   33    6  75
sdw     4096    0  8 512 61976    15 135 459    462    5   36    6  95
sdx      8      0  2  4   70572    59 151 467    461   75  51    6  97
sdy      0      0      0  0   61804    14 135 458    457    5   31    6  82
dm-0    0      0      0  0      0      0  0  0      0      0  0  0  0
dm-1    0      0      0  0      0      0  0  0      0      0  0  0  0
sdz      0      0      0  0      0      0  0  0      0      0  0  0  0
sdaa    0      0      0  0      0      0  0  0      0      0  0  0  0
sdab    0      0      0  0      0      0  0  0      0      0  0  0  0
sdac    0      0      0  0      0      0  0  0      0      0  0  0  0
sdad    0      0      0  0      0      0  0  0      0      0  0  0  0

```

Zabbix

ZABBIX

Help | Get support | Print | Profile | Logo

Monitoring | Inventory | Reports | Configuration | Administration

Nav

Dashboard | Overview | Web | Latest data | Triggers | Events | Graphs | Screens | Maps | Discovery | IT services

Search

History: Configuration of user groups » Dashboard » Custom graphs » Custom screens » Custom graphs

GRAPHS

+ -

Ceph OSD Pool

Group: Ceph | Host: | Graph: Ceph OSD Pool

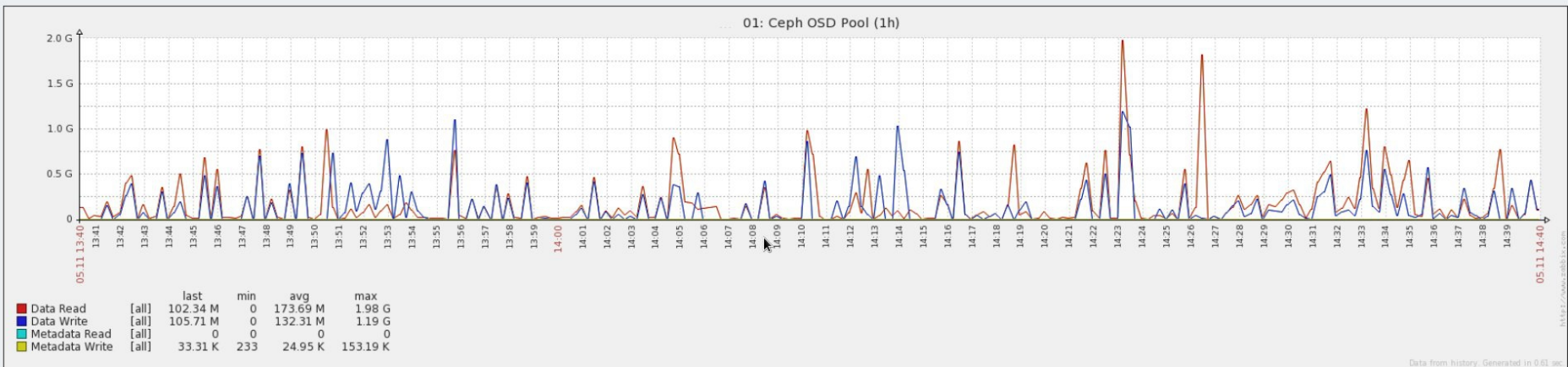
Hide filter

Zoom: 1h 2h 3h 6h 12h 1d 7d All

2015-11-05 14:37 - 2015-11-05 15:37 (now)

7d 1d 12h 1h | 1h 12h 1d 7d >>>

1h (reset)



Data from history. Generated in 0.03 sec



Conclusion

- Did not lose any performance PanFS → CephFS
- Users have been happy with the performance
- There really is a minimum!
- Need MOARRR OSD nodes :)

Questions

?