PRESENTED BY RED HAT

LEARN. NETWORK.
EXPERIENCE OPEN SOURCE.

www.theredhatsummit.com

Tuning the Red Hat Enterprise Linux 6 I/O Subsystem & Using I/O cGroups

Jeff Moyer
Principal Software Engineer, RedHat Inc.

Vivek Goyal
Senior Software Engineer, RedHat Inc.

May 5th, 2011

Agenda

- Characterizing Application Workloads
- Matching Workloads to Storage
- Tuning the I/O Subsystem
- I/O Cgroups
  - IO Throttling
  - Proportional disk time division
  - Demo

Characterizing Application Workloads
Tools of the Trade

vmstat

File Edit View Terminal Help

```
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 1  0      0 2994592  57604 432024    0    0     0 12976 7544 4967 11 24 47 18  0
 0  1      0 2993120  58756 432340    0    0     0 24672 8760 5687 13 29 37 21  0
 3  1      0 2991832  59600 432252    0    0     0 34140 7586 5000 11 27 39 23  0
 1  1      0 2988864  60272 434020    0    0     0 40816 6783 4707  9 23 43 25  0
 2  0      0 2990088  60832 432096    0    0     0 43612 5364 3765  8 22 43 27  0
 0  1      0 2985840  61280 433980    0    0     0 41360 5119 3633  7 20 44 29  0
 2  1      0 2985732  61288 434084    0    0     0    32  783  873  1  1 49 48  0
 0  1      0 2987872  61296 432036    0    0     0    36  822  852  1  2 49 48  0
 0  1      0 2987716  61296 431952    0    0     0  2048 1028 1022  2  4 47 47  0
 0  1      0 2987972  61300 431948    0    0     0    32  778  926  2  1 49 47  0
 1  1      0 2987972  61308 432008    0    0     0    36  820  921  2  2 50 47  0
 0  1      0 2987972  61312 432012    0    0     0  2052  845  893  3  2 49 46  0
 0  1      0 2987972  61316 432008    0    0     0    44  837 1012  1  2 47 49  0
 1  1      0 2986236  61324 433984    0    0     0    32  767  831  1  1 49 48  0
 1  1      0 2989956  61332 429980    0    0     0    44  901  950  2  2 49 47  0
 0  1      0 2988344  61336 432020    0    0     0  2052  747  802  1  1 48 49  0
 0  1      0 2988220  61340 431992    0    0     0    36 1479 1406  3  2 47 48  0
 1  1      0 2984212  61548 436004    0    0     4 43232 4119 3831 25 14 29 33  0
 1  0      0 2986684  61876 433060    0    0     0 50284 4864 4783  7 22 43 28  0
 2  0      0 2986560  62180 433384    0    0     0 54588 4257 3436  6 18 46 30  0
^C
[phro@localhost ~]$
```


iostat

```
File  Edit  View  Terminal  Help

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.00 12165.00    0.00 1429.00     0.00 111512.00    78.03     7.09    4.90   0.51  72.80

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           5.37    0.00   18.54   29.27    0.00   46.83

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.00 11152.00    0.00 1588.00     0.00 105656.00    66.53     7.25    4.67   0.44  70.00

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           2.93    0.00    6.34   42.93    0.00   47.80

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.00  3601.00    0.00  453.00     0.00 32912.00    72.65     3.47    6.77   2.09  94.60

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           2.43    0.00    2.91   45.15    0.00   49.51

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.00     0.00    0.00    4.00     0.00    24.00     6.00     2.30  376.75 249.25  99.70

^C
[phro@localhost ~]$
```


blkparse

```
File  Edit  View  Terminal  Help
  8,16   0        1     0.000000000 31177  Q  WS 29304 + 8 [aio-stress]
  8,16   0        2     0.000000573 31177  M  WS 29304 + 8 [aio-stress]
  8,16   0        3     0.000001528 31177  U   N [aio-stress] 1
  8,16   0        4     0.000002324 31177  D   W 29248 + 64 [aio-stress]
  8,16   2        1     0.000100707     0  C   W 29248 + 64 [0]
  8,16   0        5     0.000121681 31177  Q  WS 29312 + 8 [aio-stress]
  8,16   0        6     0.000122375 31177  G  WS 29312 + 8 [aio-stress]
  8,16   0        7     0.000123034 31177  P   N [aio-stress]
  8,16   0        8     0.000123403 31177  I   W 29312 + 8 [aio-stress]
  8,16   0        9     0.000126002 31177  Q  WS 29320 + 8 [aio-stress]
  8,16   0       10     0.000126350 31177  M  WS 29320 + 8 [aio-stress]
  8,16   0       11     0.000128516 31177  Q  WS 29328 + 8 [aio-stress]
  8,16   0       12     0.000128846 31177  M  WS 29328 + 8 [aio-stress]
  8,16   0       13     0.000130859 31177  Q  WS 29336 + 8 [aio-stress]
  8,16   0       14     0.000131092 31177  M  WS 29336 + 8 [aio-stress]
  8,16   0       15     0.000133085 31177  Q  WS 29344 + 8 [aio-stress]
  8,16   0       16     0.000133388 31177  M  WS 29344 + 8 [aio-stress]
  8,16   0       17     0.000135366 31177  Q  WS 29352 + 8 [aio-stress]
  8,16   0       18     0.000135596 31177  M  WS 29352 + 8 [aio-stress]
  8,16   0       19     0.000137624 31177  Q  WS 29360 + 8 [aio-stress]
  8,16   0       20     0.000137942 31177  M  WS 29360 + 8 [aio-stress]
  8,16   0       21     0.000139972 31177  Q  WS 29368 + 8 [aio-stress]
  8,16   0       22     0.000140190 31177  M  WS 29368 + 8 [aio-stress]
  8,16   0       23     0.000140829 31177  U   N [aio-stress] 1
  8,16   0       24     0.000141549 31177  D   W 29312 + 64 [aio-stress]
  8,16   2        2     0.000233631     0  C   W 29312 + 64 [0]
  8,16   0       25     0.000255604 31177  Q  WS 29376 + 8 [aio-stress]
  8,16   0       26     0.000256288 31177  G  WS 29376 + 8 [aio-stress]
  8,16   0       27     0.000256892 31177  P   N [aio-stress]
  8,16   0       28     0.000257188 31177  I   W 29376 + 8 [aio-stress]
  8,16   0       29     0.000260118 31177  Q  WS 29384 + 8 [aio-stress]
  8,16   0       30     0.000260466 31177  M  WS 29384 + 8 [aio-stress]
  8,16   0       31     0.000262622 31177  Q  WS 29392 + 8 [aio-stress]
  8,16   0       32     0.000262933 31177  M  WS 29392 + 8 [aio-stress]
  8,16   0       33     0.000264923 31177  Q  WS 29400 + 8 [aio-stress]
  8,16   0       34     0.000265143 31177  M  WS 29400 + 8 [aio-stress]
  8,16   0       35     0.000267193 31177  Q  WS 29408 + 8 [aio-stress]
:
```


Blkparse Summary

File Edit View Terminal Help

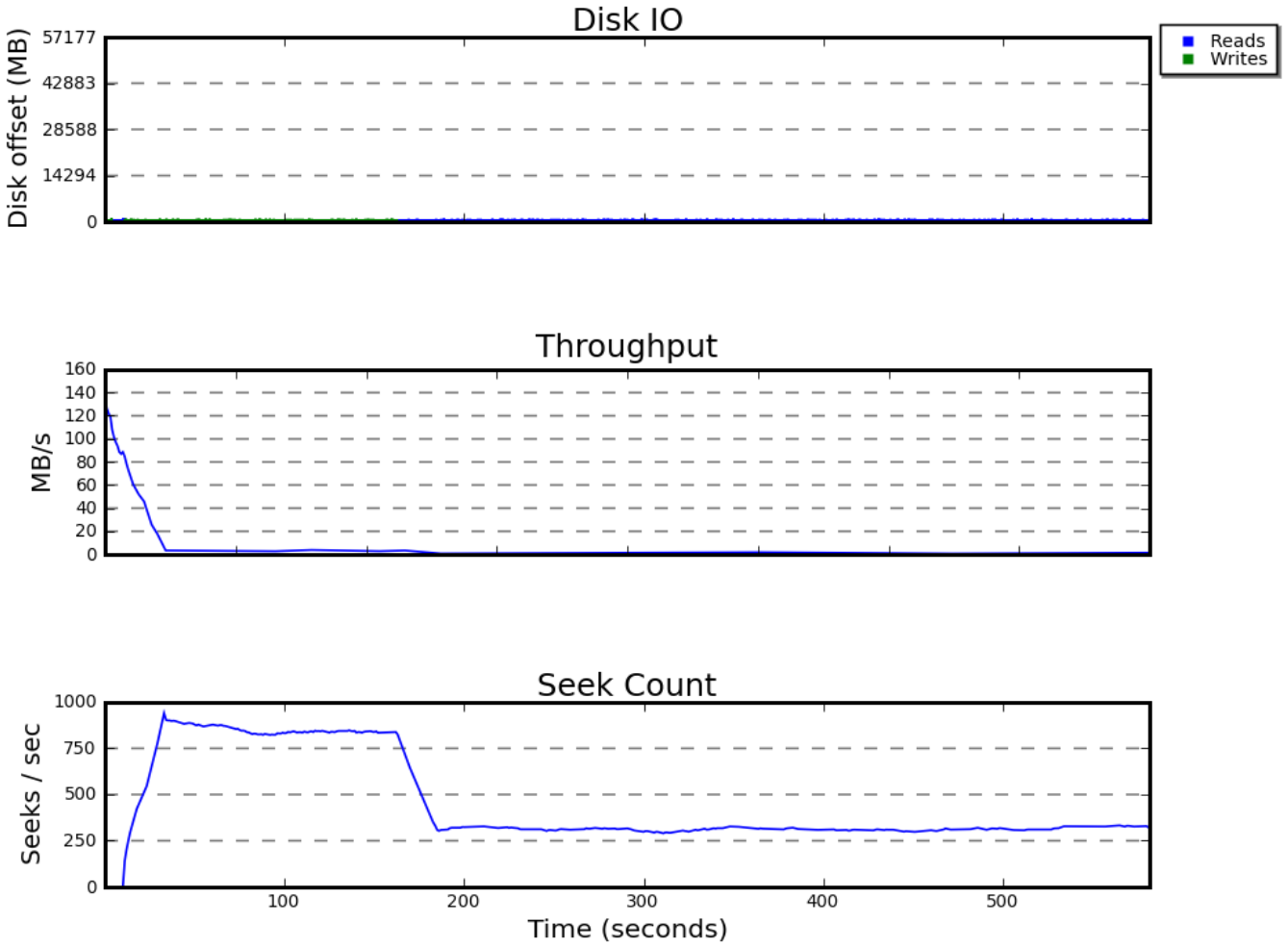
```
Total (sdb):
 Reads Queued:     262,154,    1,048MiB  Writes Queued:     258,485,    1,033MiB
 Read Dispatches:  147,464,    1,048MiB  Write Dispatches:  146,998,    1,033MiB
 Reads Requeued:         0               Writes Requeued:         0
 Reads Completed:  147,463,    1,048MiB  Writes Completed:  147,002,    1,033MiB
 Read Merges:      114,690,  458,760KiB  Write Merges:      111,484,  445,936KiB
 PC Reads Queued:        0,        0KiB  PC Writes Queued:        0,        0KiB
 PC Read Disp.:          4,        0KiB  PC Write Disp.:          0,        0KiB
 PC Reads Req.:          0               PC Writes Req.:          0
 PC Reads Compl.:        4               PC Writes Compl.:  147,002
 IO unplugs:        65,089               Timer unplugs:           0

Throughput (R/W): 1,802KiB/s / 1,777KiB/s
Events (sdb): 2,087,639 entries
Skips: 0 forward (0 -   0.0%)
Input file sdb.blktrace.0 added
Input file sdb.blktrace.1 added
Input file sdb.blktrace.2 added
Input file sdb.blktrace.3 added
[phro@localhost 10krpm-blktrace]$
```


btt

File Edit View Terminal Help

```
==================== All Devices ====================

            ALL           MIN           AVG           MAX           N
--------------- ------------ ------------ ------------ -----------

Q2Q               0.000001566   0.001117258   0.520601700      520629
Q2G               0.000000295   0.000000735   0.000115444      294461
G2I               0.000000182   0.000000736   0.000047225      294461
Q2M               0.000000150   0.000000304   0.000049239      226174
I2D               0.000001454   0.000011665   0.000197490      294461
M2D               0.000000899   0.000008160   0.000189766      226174
D2C               0.000066198   0.004451998   0.520529250      520629
Q2C               0.000071221   0.004463105   0.520545724      520629

==================== Device Overhead ====================

       DEV |       Q2G       G2I       Q2M       I2D       D2C
---------- | --------- --------- --------- --------- ---------
 (  8, 16) |   0.0093%   0.0093%   0.0030%   0.1478%  99.7511%
---------- | --------- --------- --------- --------- ---------
   Overall |   0.0093%   0.0093%   0.0030%   0.1478%  99.7511%
:
```

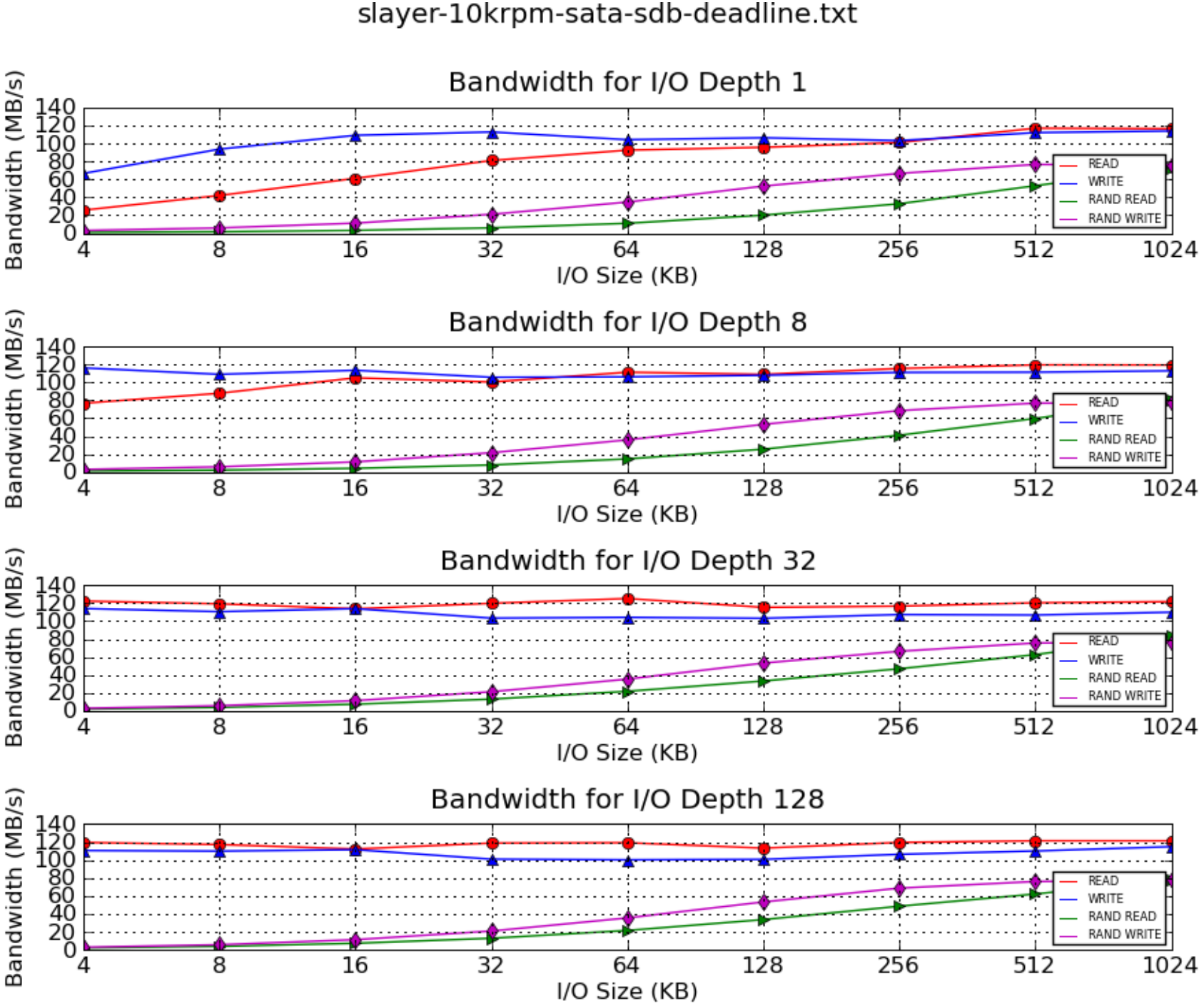

Seekwatcher

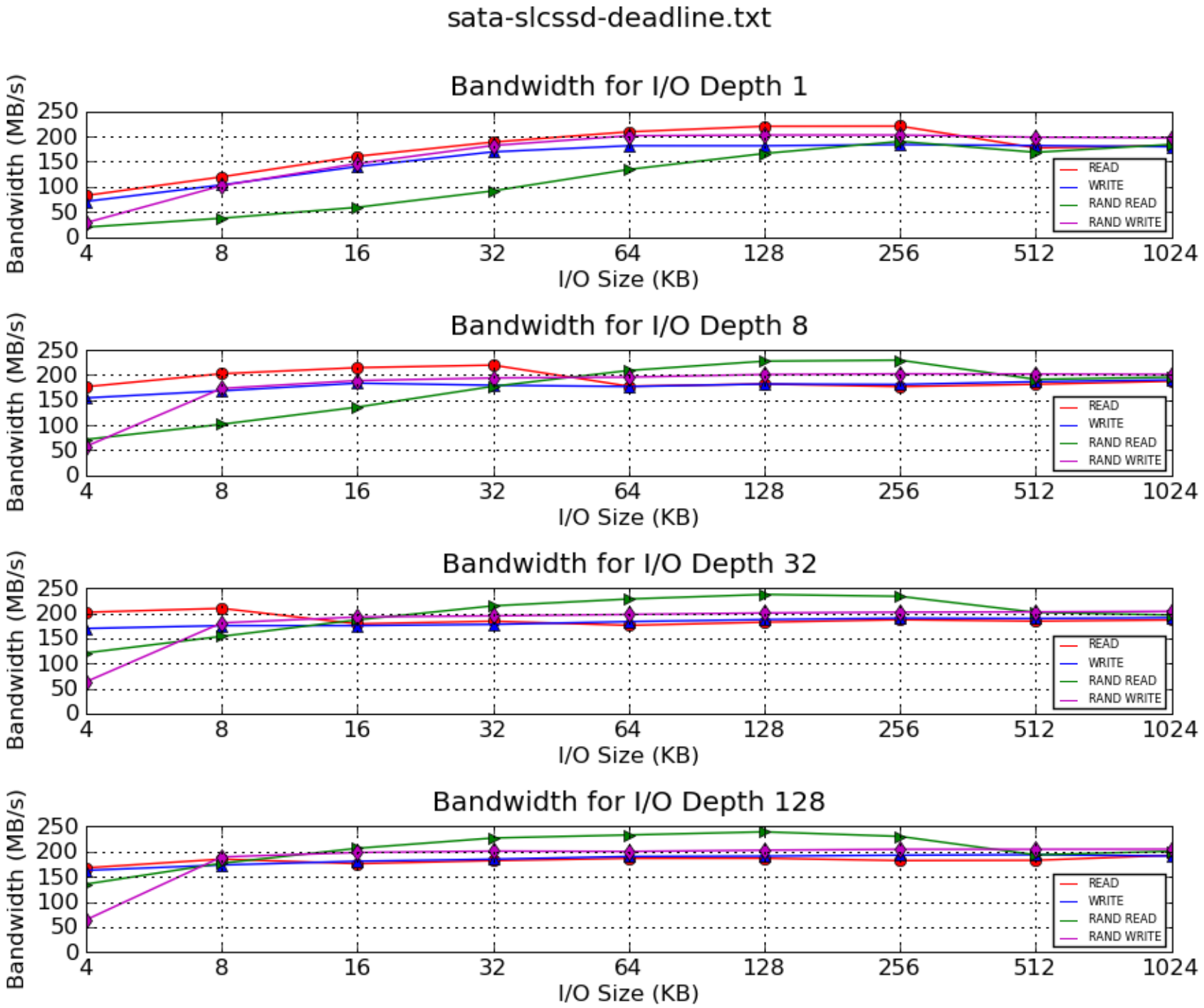


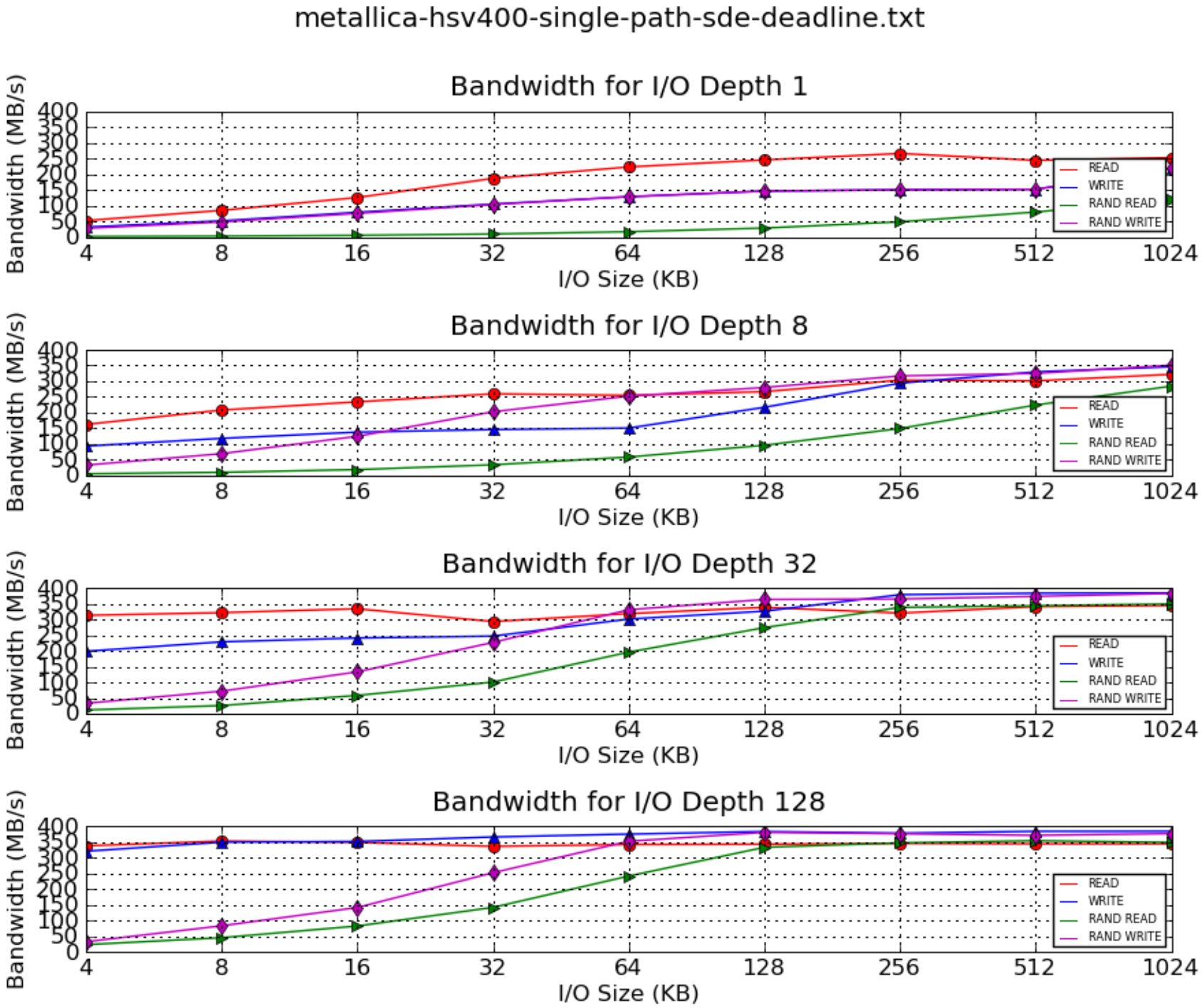
What We Can Measure So Far

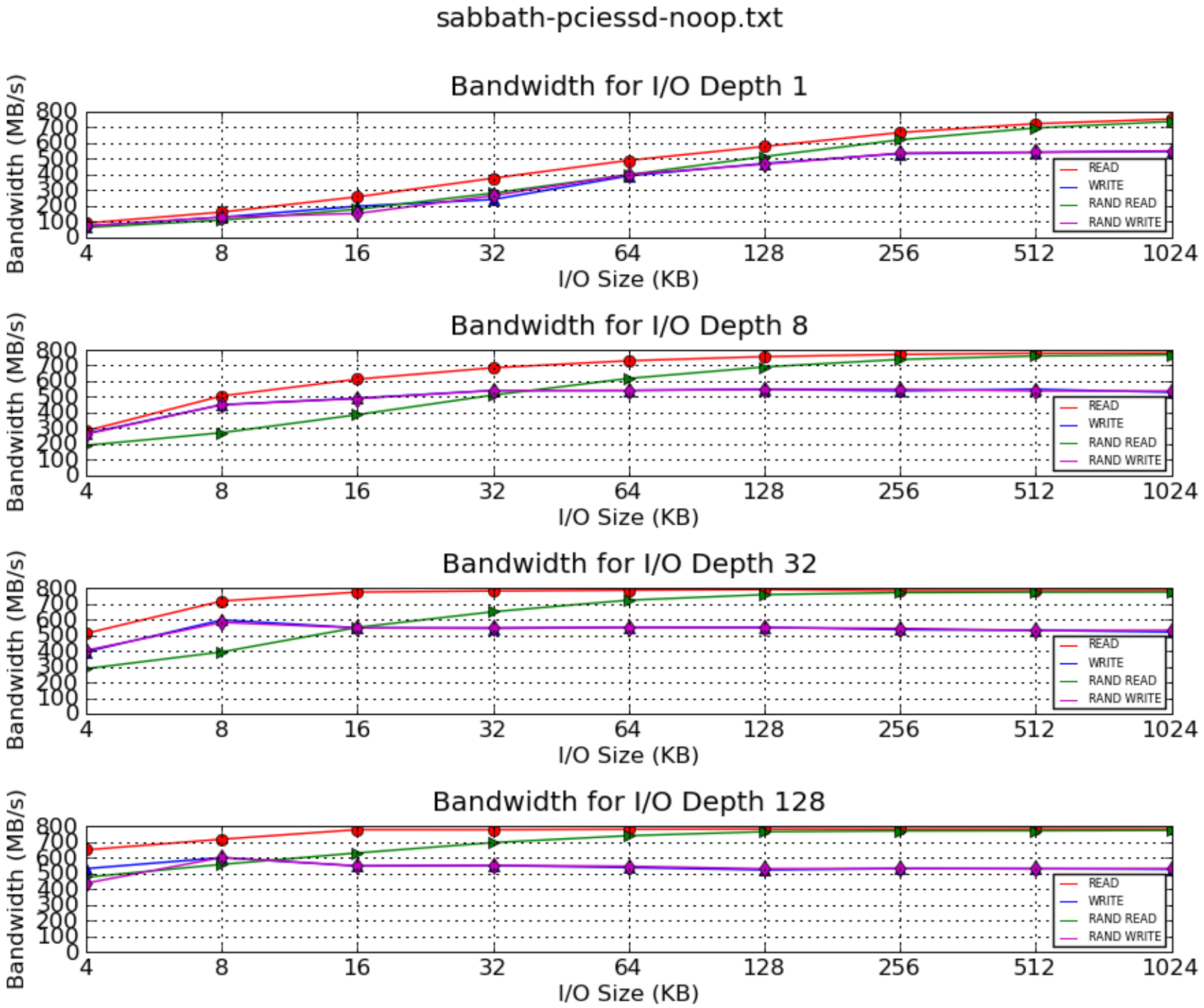
- Type of workload
- Average I/O sizes
- Average bandwidth & IOPS
- Where I/O spends its time
- What applications are issuing I/O

Matching a Workload to a Storage Solution


slayer-10krpm-sata-sdb-deadline.txt
Bandwidth for I/O Depth 1
Bandwidth for I/O Depth 8
Bandwidth for I/O Depth 32
Bandwidth for I/O Depth 128
Bandwidth (MB/s)
I/O Size (KB)
0
20
40
60
80
100
120
140
4
8
16
32
64
128
256
512
1024
READ
WRITE
RAND READ
RAND WRITE


sata-slcssd-deadline.txt
Bandwidth for I/O Depth 1
Bandwidth for I/O Depth 8
Bandwidth for I/O Depth 32
Bandwidth for I/O Depth 128
Bandwidth (MB/s)
I/O Size (KB)
0
50
100
150
200
250
4
8
16
32
64
128
256
512
1024
READ
WRITE
RAND READ
RAND WRITE


metallica-hsv400-single-path-sde-deadline.txt
Bandwidth for I/O Depth 1
Bandwidth for I/O Depth 8
Bandwidth for I/O Depth 32
Bandwidth for I/O Depth 128
Bandwidth (MB/s)
I/O Size (KB)
READ
WRITE
RAND READ
RAND WRITE
0
50
100
150
200
250
300
350
400
4
8
16
32
64
128
256
512
1024


sabbath-pciessd-noop.txt
Bandwidth for I/O Depth 1
Bandwidth for I/O Depth 8
Bandwidth for I/O Depth 32
Bandwidth for I/O Depth 128
Bandwidth (MB/s)
I/O Size (KB)
READ
WRITE
RAND READ
RAND WRITE
0
100
200
300
400
500
600
700
800
4
8
16
32
64
128
256
512
1024

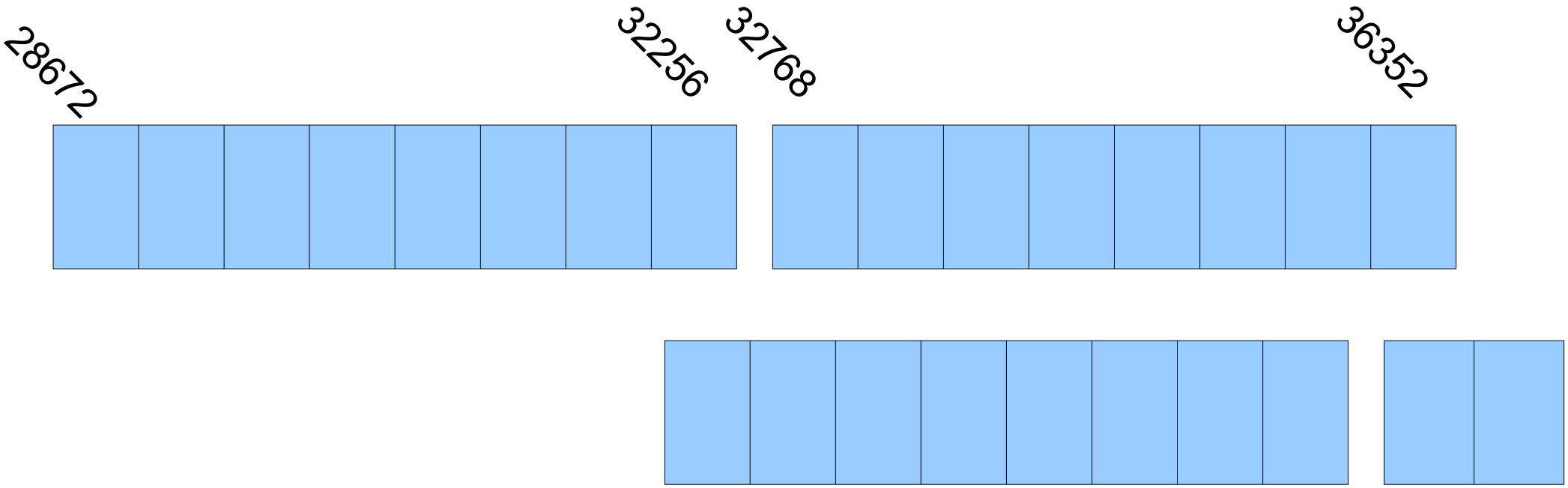
Tuning the I/O Stack

tuned-adm

- Profiles
  - enterprise-storage
  - throughput-performance
  - latency-performance

I did everything right, and it's still slow... What now?

Alignment

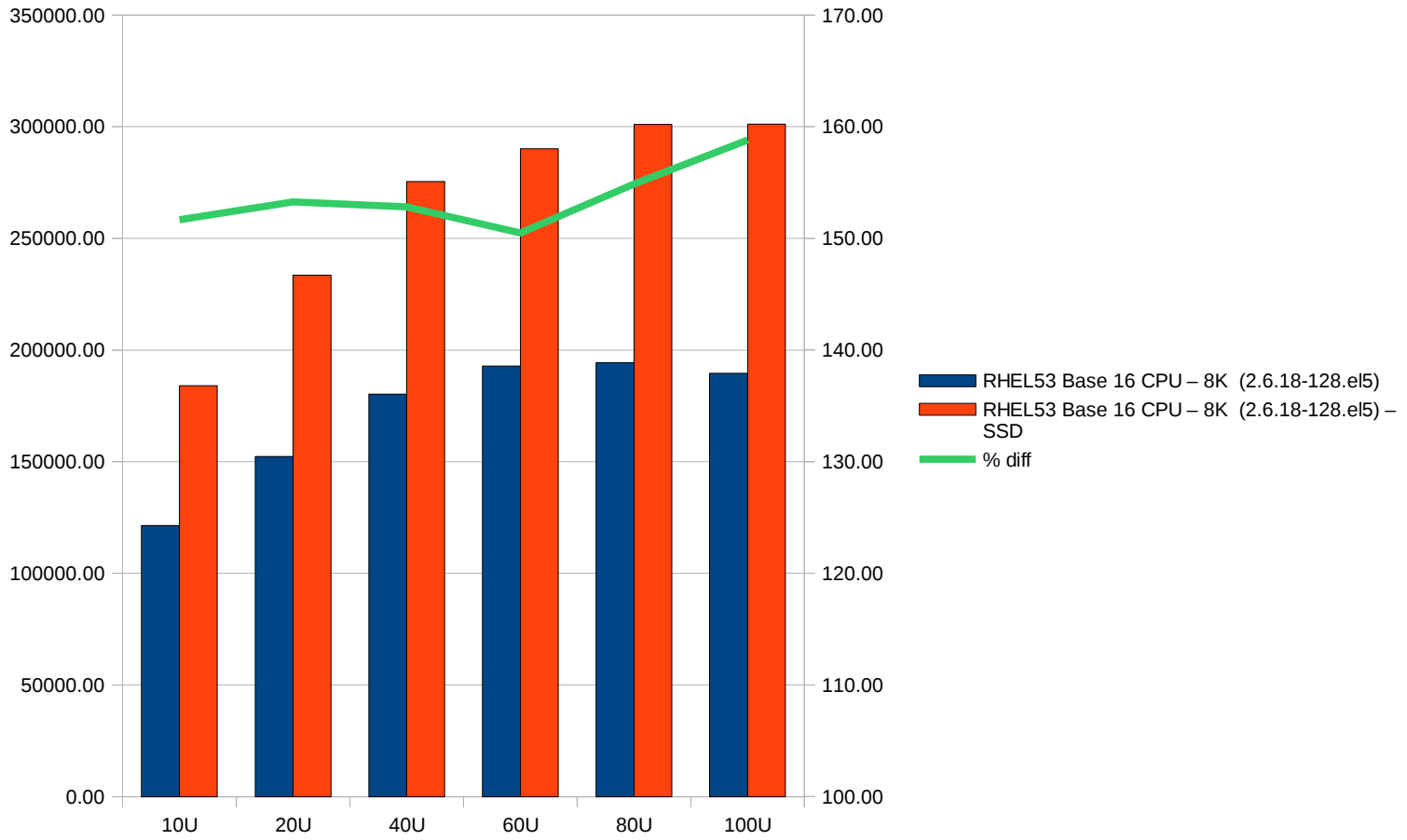


4KB = 8 512 byte blocks
Historically, partition 1 starts on sector 63.
63 * 512 = 32256

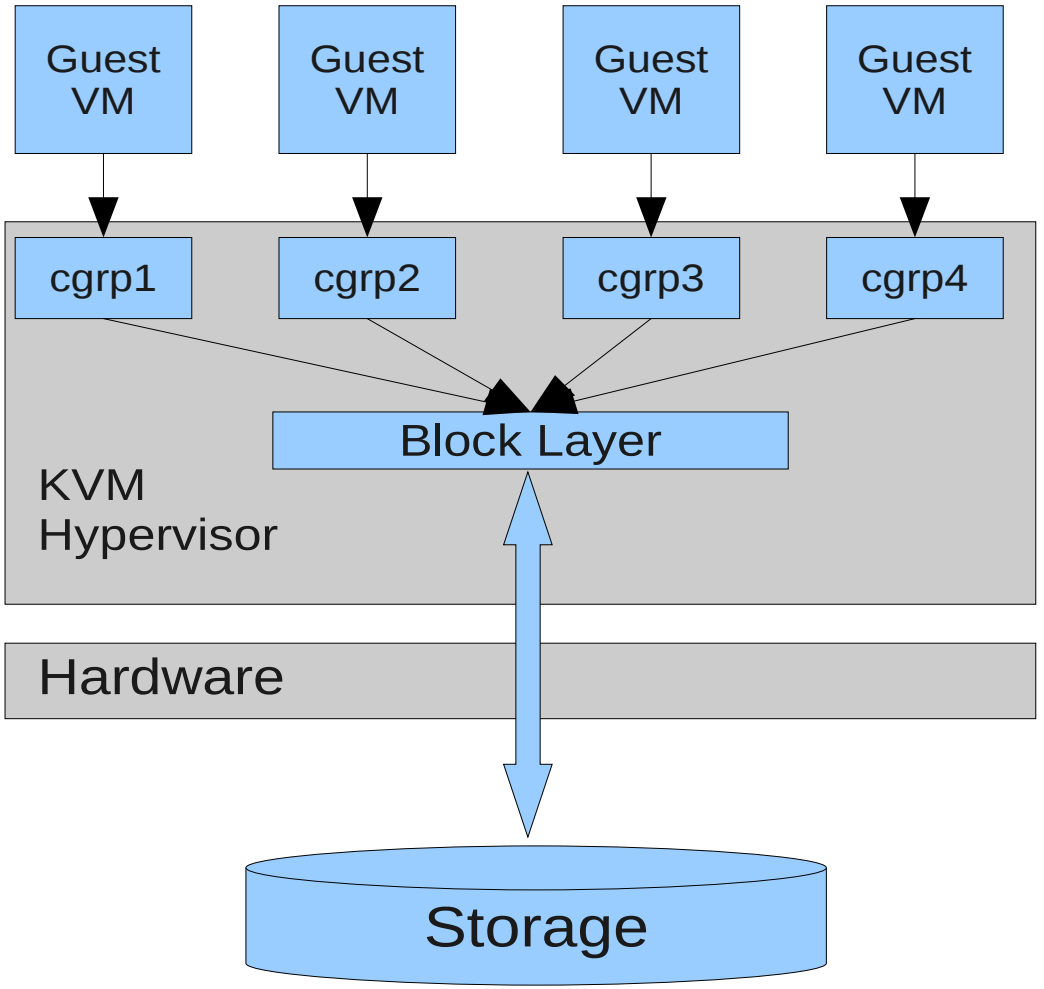
Other Reasons/Suggestions

- Check the I/O sizes against your RAID configuration
- Do other systems have access to the same storage?
- Have you identified any bottlenecks in the I/O path?
- Are you paying the NUMA penalty?
- Maybe it's time for a storage upgrade?

SSD used for DB logs



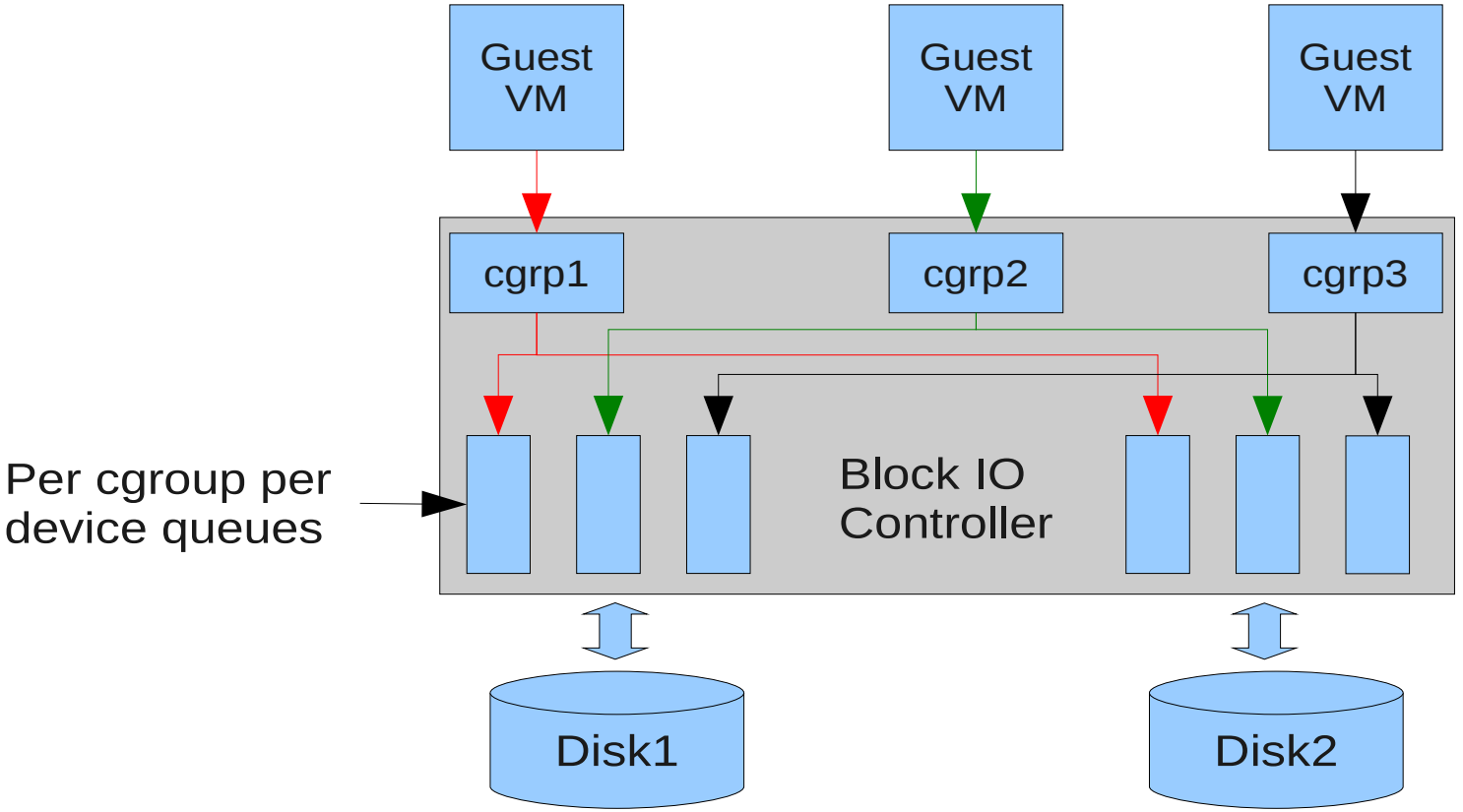
IO Cgroups Overview



Block IO Controller Policies

- IO Throttling
- Proportional weight based disk time division policy

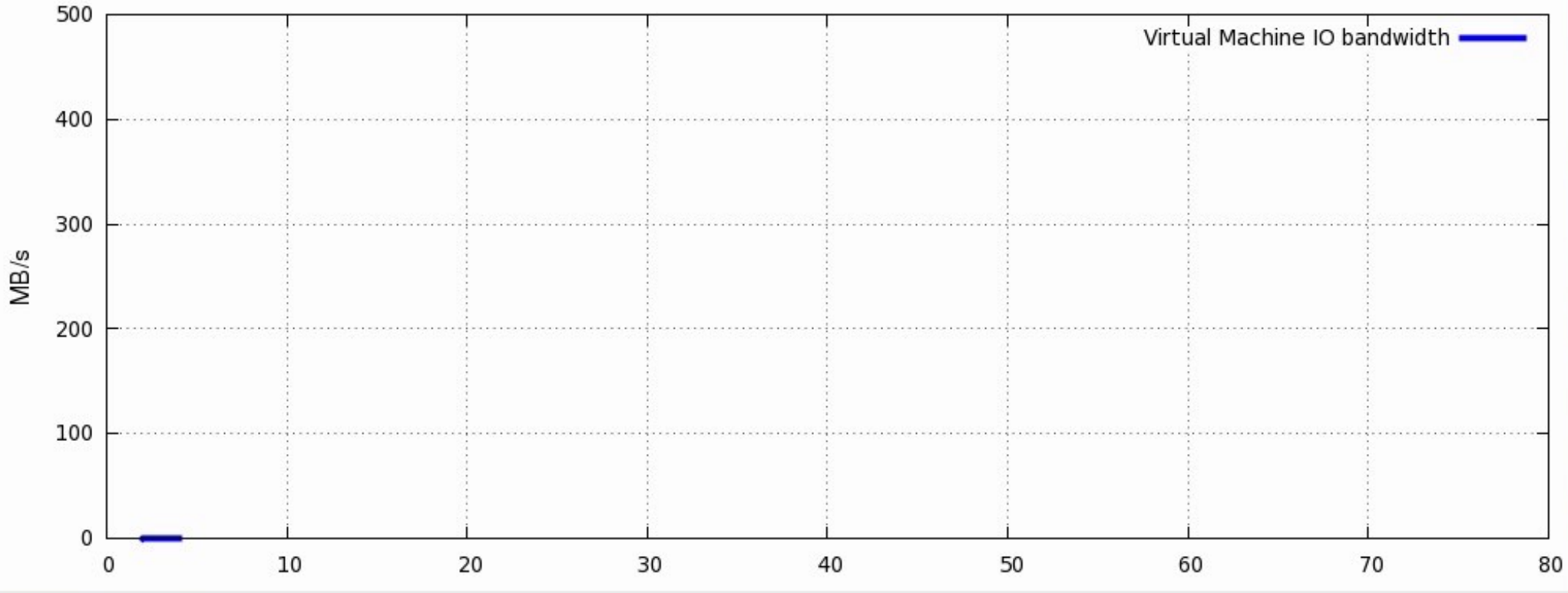
IO Throttling Policy



IO Throttling Interface

- Cgroup virtual file system interface
  - Mount -t cgroup -o blkio none /cgroup/blkio
- Bandwidth and IO per second Rules
- READ/WRITE rules
  - blkio.throttle.read_bps_device
  - blkio.throttle.write_bps_device
  - blkio.throttle.read_iops_device
  - blkio.throttle.write_iops_device

Gnuplot (window id : 0)



-6.71756, -50.8157

vgoyal@machine:~/demo-scripts/summit-2011/vivek-summit-presentation-experiments/plots/single-machine

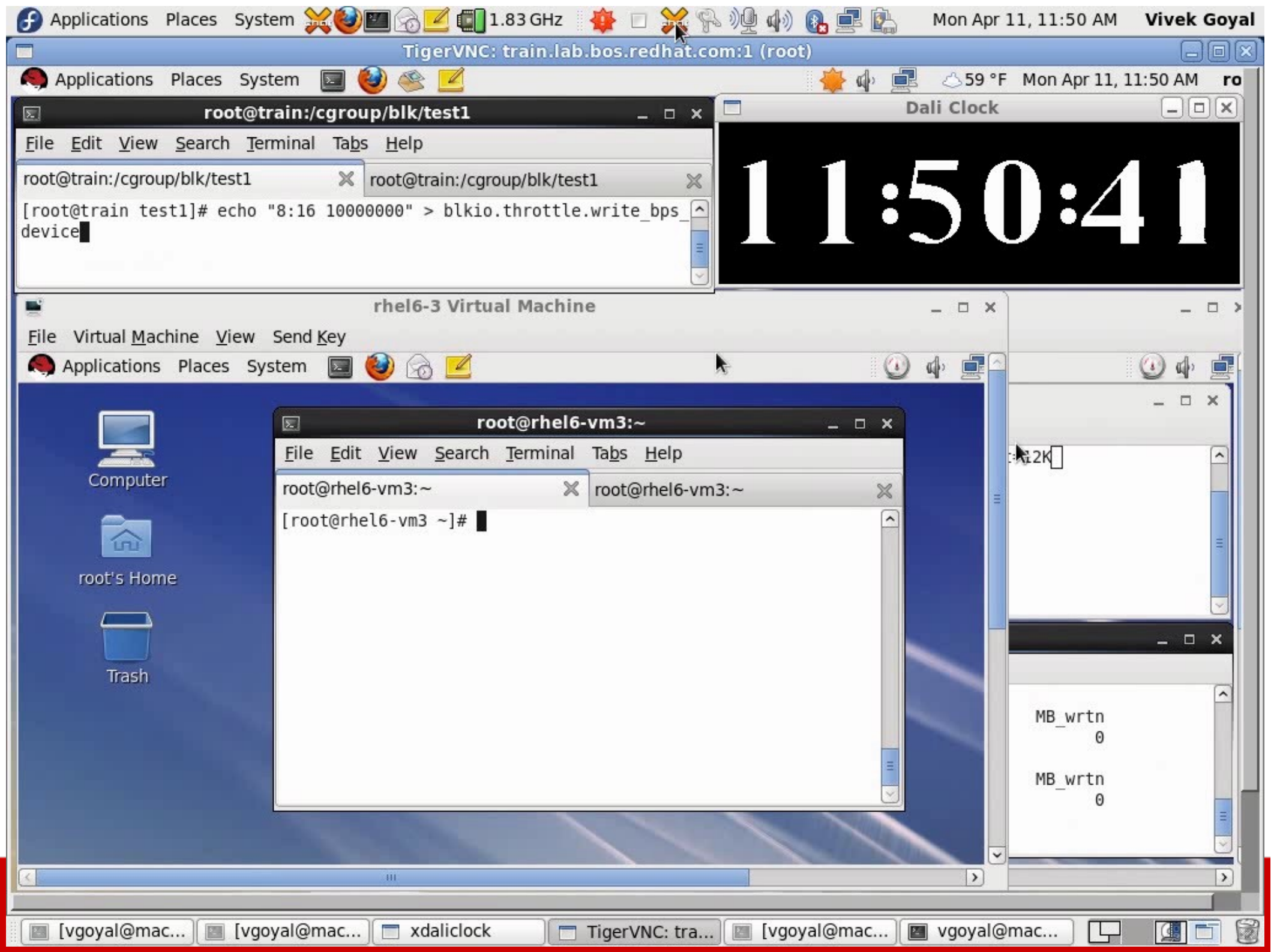
File Edit View Search Terminal Tabs Help

vgoyal@machine:~/demo-scrip... | root@train:~ | root@rhel6-vm4:~ | root@train:/cgroup/blk/test1

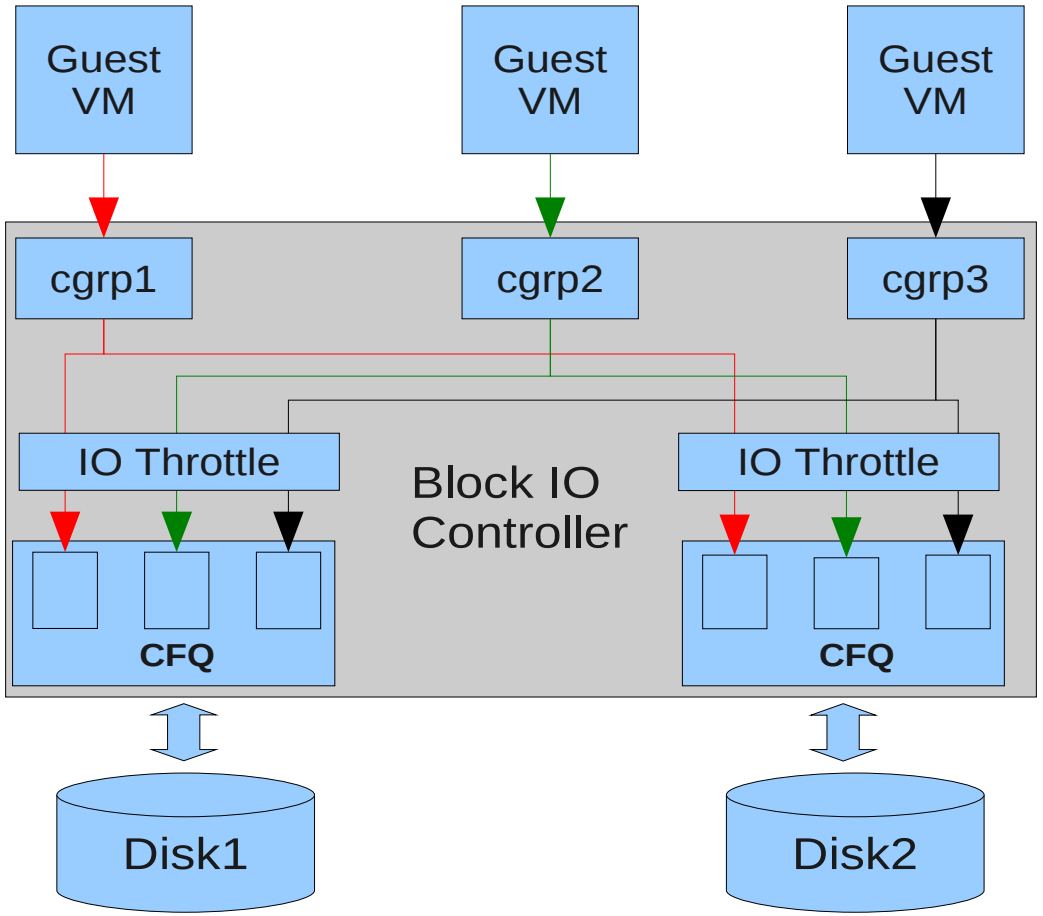
```
[vgoyal@machine single-machine]$ ./plot-data-iostat-single-machine.sh
```

Why Throttle

- Differentiated Quality of Service
- Resource Isolation

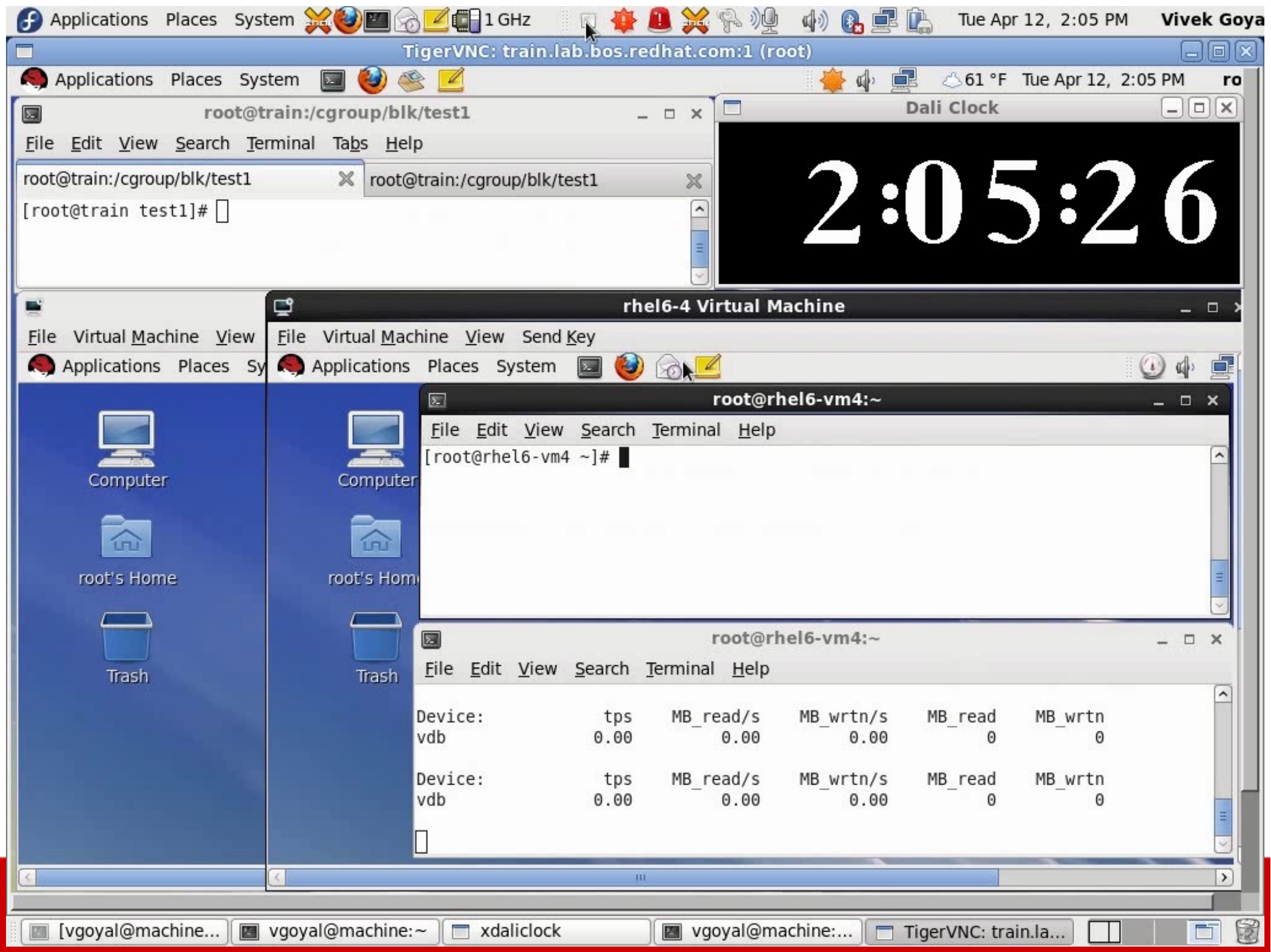

Applications Places System 1.83 GHz Mon Apr 11, 11:50 AM Vivek Goyal
TigerVNC: train.lab.bos.redhat.com:1 (root)
Applications Places System 59 °F Mon Apr 11, 11:50 AM ro
root@train:/cgroup/blk/test1
File Edit View Search Terminal Tabs Help
root@train:/cgroup/blk/test1
root@train:/cgroup/blk/test1
[root@train test1]# echo "8:16 10000000" > blkio.throttle.write_bps_
device
Dali Clock
11:50:41
rhel6-3 Virtual Machine
File Virtual Machine View Send Key
Applications Places System
Computer
root's Home
Trash
root@rhel6-vm3:~
File Edit View Search Terminal Tabs Help
root@rhel6-vm3:~
root@rhel6-vm3:~
[root@rhel6-vm3 ~]#
MB_wrtn
0
MB_wrtn
0
[vgoyal@mac...
[vgoyal@mac...
xdaliclock
TigerVNC: tra...
[vgoyal@mac...
vgoyal@mac...

Proportional disk IO



Proportional IO Interface

- Weight based proportional disk time division
  - blkio.weight
- Global as well as per device weights
  - blkio.weight_device
- Weight range 100 - 1000


Applications Places System 1 GHz Tue Apr 12, 2:05 PM Vivek Goya
TigerVNC: train.lab.bos.redhat.com:1 (root)
Applications Places System 61 °F Tue Apr 12, 2:05 PM ro
root@train:/cgroup/blk/test1
File Edit View Search Terminal Tabs Help
root@train:/cgroup/blk/test1
root@train:/cgroup/blk/test1
[root@train test1]#
Dali Clock
2:05:26
File Virtual Machine View
Applications Places Sy
Computer
root's Home
Trash
rhel6-4 Virtual Machine
File Virtual Machine View Send Key
Applications Places System
Computer
root's Hom
Trash
root@rhel6-vm4:~
File Edit View Search Terminal Help
[root@rhel6-vm4 ~]#
root@rhel6-vm4:~
File Edit View Search Terminal Help
Device: tps MB_read/s MB_wrtn/s MB_read MB_wrtn
vdb 0.00 0.00 0.00 0 0
Device: tps MB_read/s MB_wrtn/s MB_read MB_wrtn
vdb 0.00 0.00 0.00 0 0
[vgoyal@machine...
vgoyal@machine:~
xdaliclock
vgoyal@machine:...
TigerVNC: train.la...

Tips

- IO Throttling
  - Filesystem ordered mode issue on host
  - Useful in Cluster Configurations
- Proportional IO
  - Most effective on single spindle disks
  - Use group_isolation = 1

TODO

- IO Throttling
  - Buffered WRITE control
  - Global Limits
- Proportional IO
  - Buffered WRITE control

More Information

- git://git.kernel.dk/blktrace.git
- git://git.kernel.dk/fio.git
- http://oss.oracle.com/~mason/seekwatcher/

LIKE US ON FACEBOOK

www.facebook.com/redhatinc

FOLLOW US ON TWITTER

www.twitter.com/redhatsummit

TWEET ABOUT IT

#redhat

READ THE BLOG

summitblog.redhat.com

GIVE US FEEDBACK

www.redhat.com/summit/survey