Thomas-Krenn.AG®
The server experts

SSD & HDD Performance Testing with TKperf

Georg Schönberger <gschoenberger@thomas-krenn.com>

LinuxTag 2013

The Who

Agenda

1) The Why

2) The How

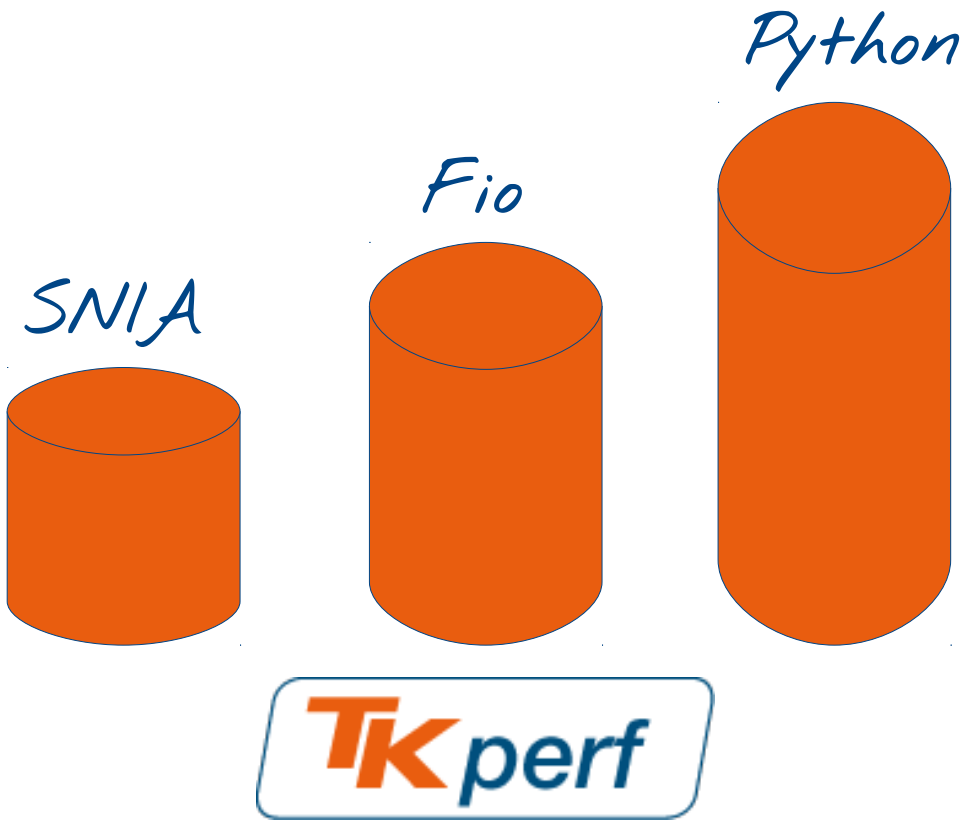
3) IOPS

4) Throughput

5) Write Saturation

6) Latency

7) Résumé



- Solid State Storage Performance Test Specification[1]

1 http://www.snia.org/tech_activities/standards/curr_standards/pts

The Why

1) The Why

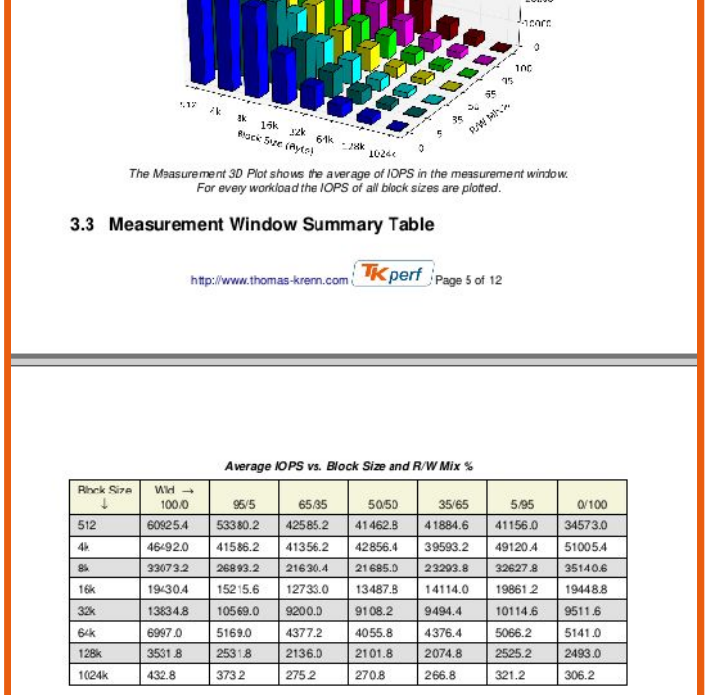
- Which devices are suited for enterprise
  - What is the difference to a consumer device?
- SNIA
  - SSDs are different
- Fio
  - Flexible, actively developed
- Python
  - Matplotlib
- Open Source
  - Get Feedback, improve tests

1) The Why

Thomas-Krenn.AG®
The server experts

- Automatism
  - Fio calls are logged
  - Write results to XML
  - Generate rst report

```
sudo tkperf ssd intelDCS3700 /dev/sdb -nj 2 -iod 16 -rfb
```



The Measurement Plot shows the average of IOPS in the measurement window. For every workload the IOPS of all block sizes are plotted.

IOPS 3D Measurement Plot

The Measurement 3D Plot shows the average of IOPS in the measurement window. For every workload the IOPS of all block sizes are plotted.

3.3 Measurement Window Summary Table

http://www.thomas-krenn.com TKperf Page 5 of 12

Average IOPS vs. Block Size and R/W Mix %

Block Size ↓	Wkl → 100/0	95/5	65/35	50/50	35/65	5/95	0/100
512	60925.4	53380.2	42585.2	41462.8	41884.6	41156.0	34573.0
4k	46492.0	41586.2	41356.2	42856.4	39593.2	49120.4	51005.4
8k	33073.2	26893.2	21630.4	21685.0	23293.8	32627.8	35140.6
16k	19430.4	15215.6	12733.0	13487.8	14114.0	19861.2	19448.8
32k	13834.8	10569.0	9200.0	9108.2	9494.4	10114.6	9511.6
64k	6997.0	5169.0	4377.2	4055.8	4376.4	5066.2	5141.0
128k	3531.8	2531.8	2136.0	2101.8	2074.8	2525.2	2493.0
1024k	432.8	373.2	275.2	270.8	266.8	321.2	306.2


Handy Tools

2) The How

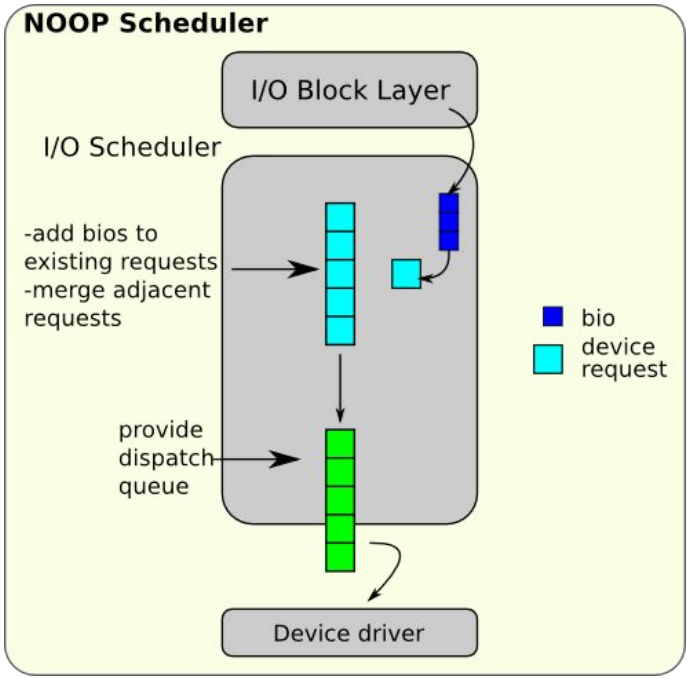
- Information gathering and Secure Erase
  - hdparm, lsblk, lsscsi
- Direct IO
- Number of jobs, IO depth
  - Reach the maximum of a device

```
# fio --rw=write --name=test --size=20M
[...]
Run status group 0 (all jobs):
  WRITE: io=20480KB, aggrb=930909KB/s
```

```
# fio --rw=write --name=test --size=20M --direct=1
[...]
Run status group 0 (all jobs):
  WRITE: io=20480KB, aggrb=28563KB/s
```

```
# hdparm -I /dev/sdb|grep -e NCQ -e depth
    Queue depth: 32
       *    Native Command Queueing (NCQ)
```

- Asynchronous IO
  - Outstanding IOs in flight
  - Requires direct IO
- Difference application ↔ device levels[1]
  - Block sizes are split
  - IO scheduler

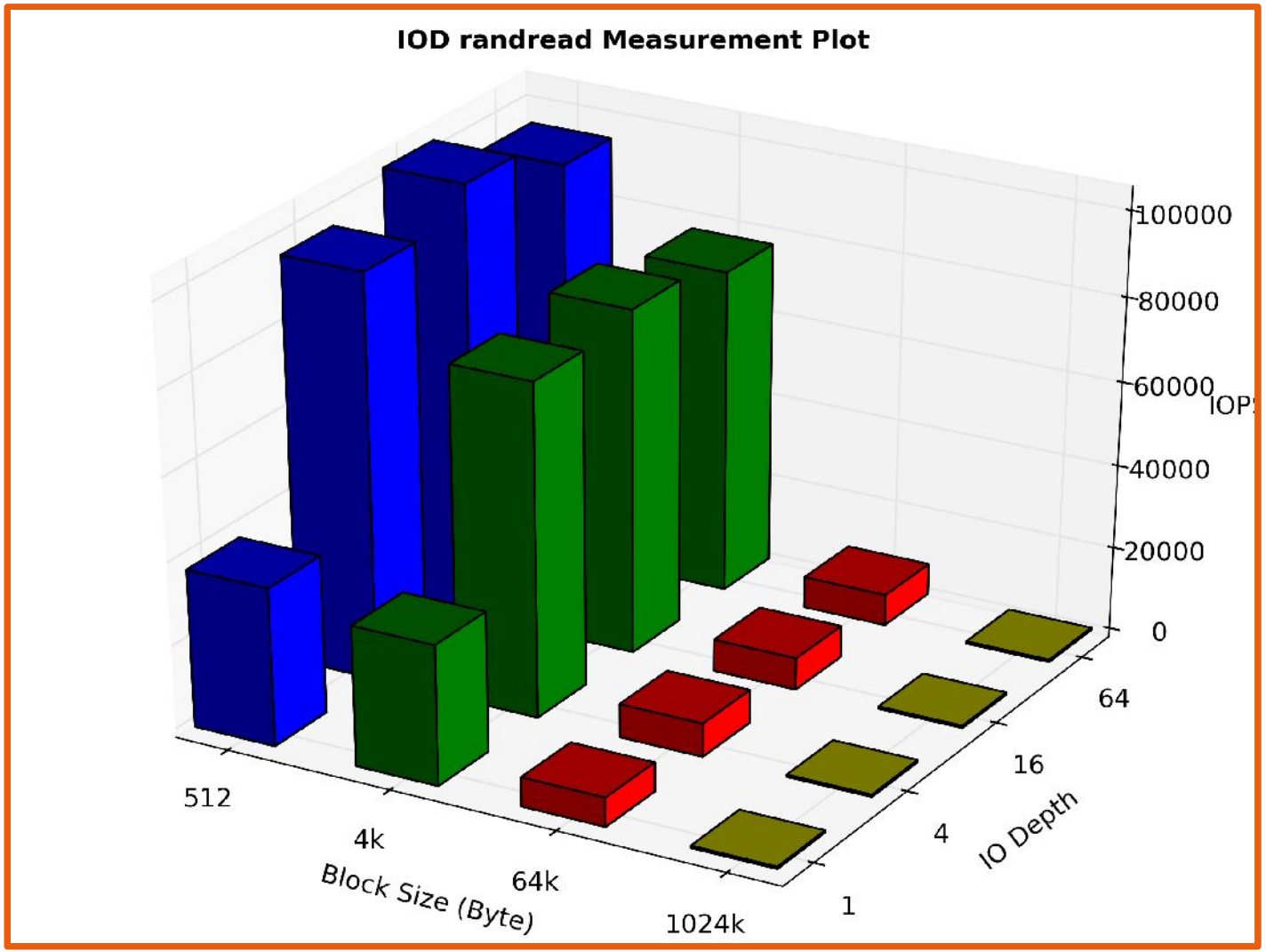


1 http://www.spinics.net/lists/fio/msg01526.html

IO Depth

Thomas-Krenn.AG®
The server experts

- It matters!



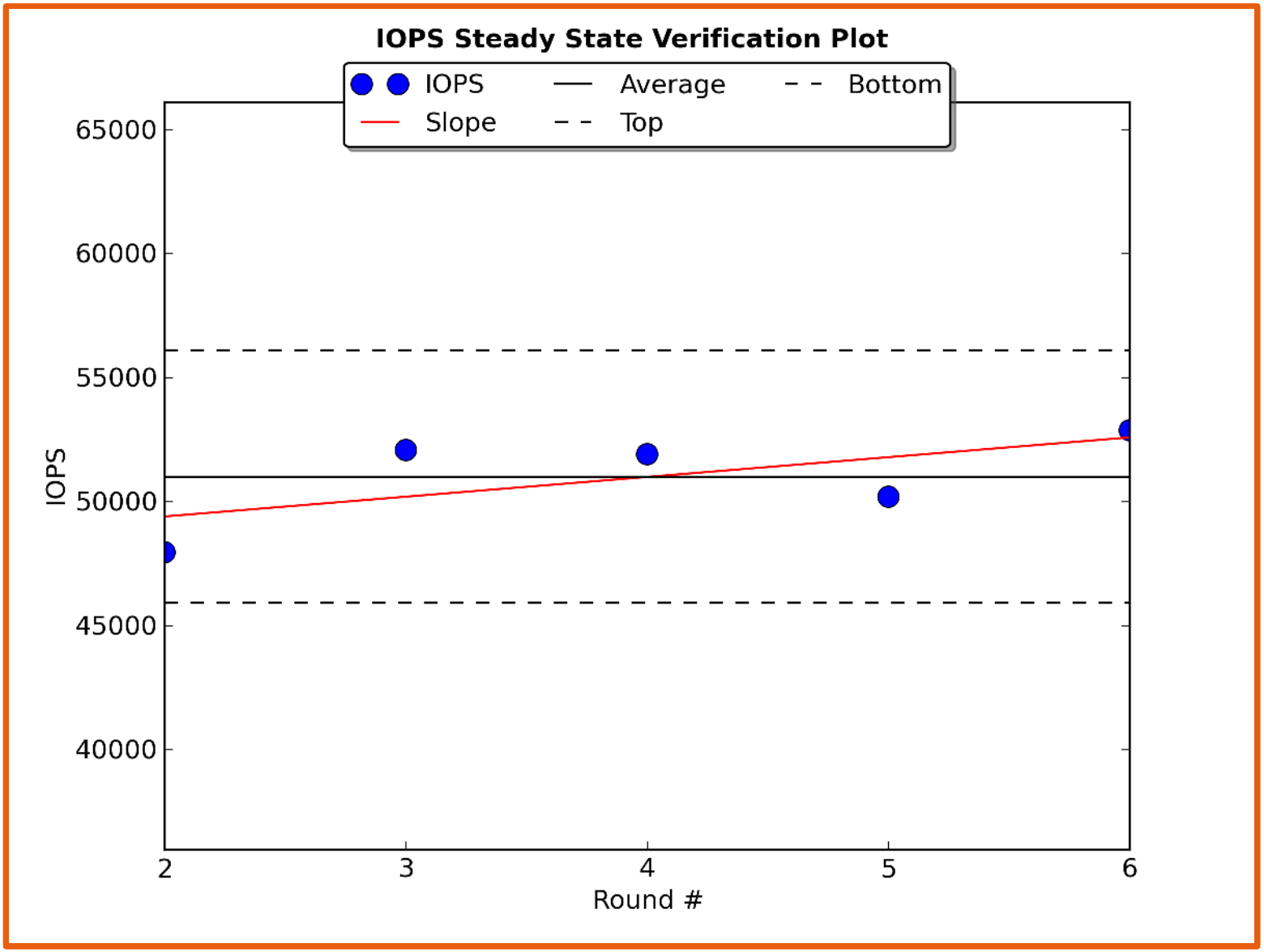
SNIA

Tests

- Synthetic
  - Test a specific scenario
  - Testing an individual component
  - Corner case behavior
- Get a first impression, compareable
- Not an application based test
  - blkreplay

New Words...

- Secure Erase
- Workload Independent Preconditioning
  - 2x device 128KB sequential write
- Workloads (IO access)
  - Random, sequential
  - Mixed (95/5 R/W)
- Steady State
  - Dependent Variable
    - States when device is stable
  - Performance values of last 5 rounds


IOPS Steady State Verification Plot
IOPS
Average
Bottom
Slope
Top
65000
60000
55000
50000
45000
40000
IOPS
2
3
4
5
6
Round #

Test Overview

Thomas-Krenn.AG®
The server experts

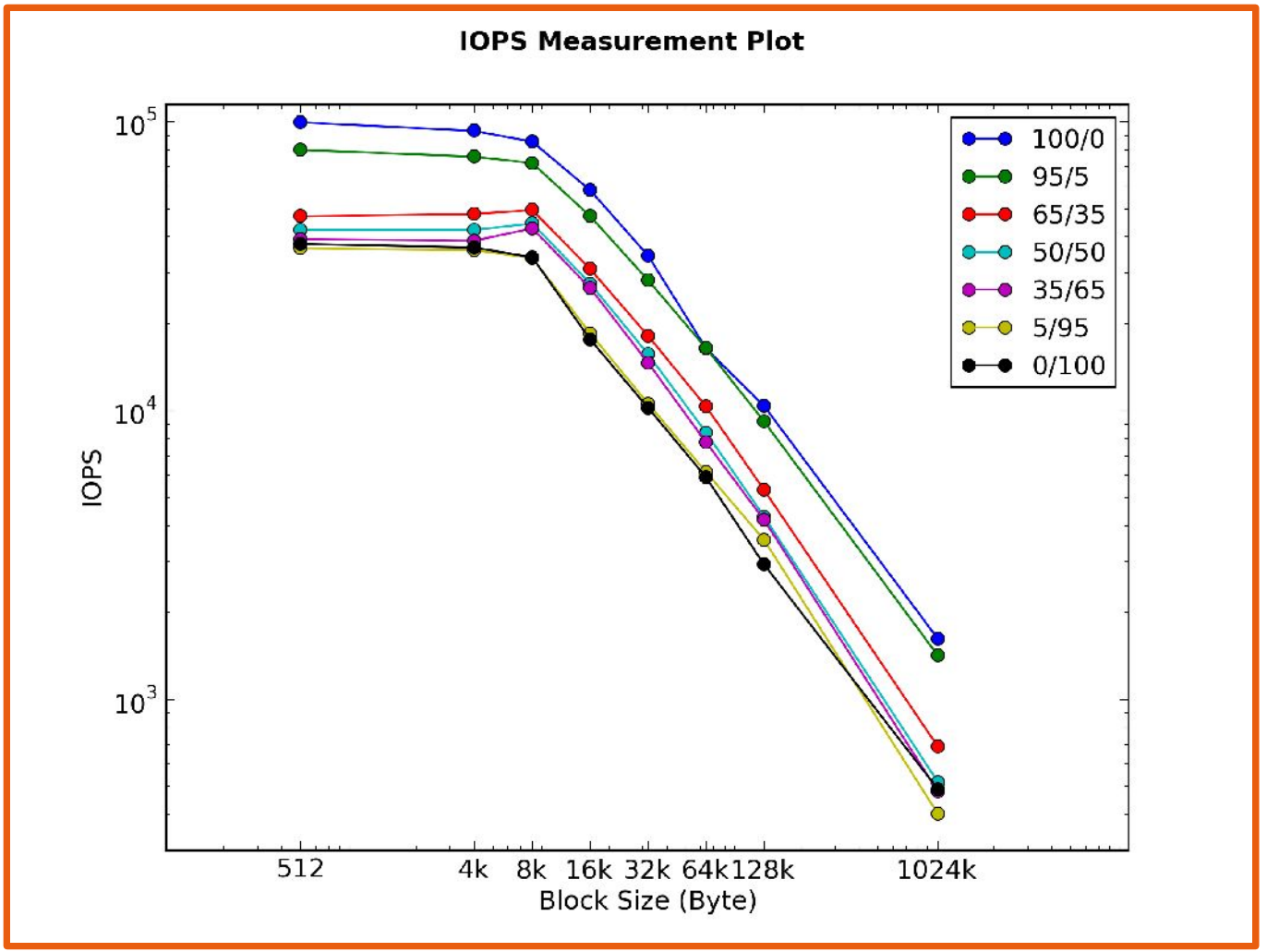
IOPS	TP	Write Sat.	Latency
Random Access	**Sequential Access**	**Random Access**	**Random Access**
R/W 100/0, 95/5, 65/35, 50/50, 36/65, 5/95, 0/10	**R/W** 100/0, 0/100	**R/W** 100% writes	**R/W** 100/0, 65/35, 0/100
BS 1M, 128K, 64K, 32K, 16K, 8K, 4K, 0.5K	**BS** 1M, 64K, 8K, 4K, 0.5K	**BS** 4K	**BS** 8K, 4K, 0.5K
DV 4K random writes	**DV** 1M seq. writes	-	**DV** 4K random write mean latency


IOPS
-> A random Workload

3) IOPS

- IO operations per second
- THE number for SSDs
- Dependent variable
  - 4KB random writes

```
Make Secure Erase
Workload Ind. Preconditioning
While not Steady State
  For workloads [100, 95, 65, 50, 35, 5, 0]
    For block sizes ['1024k', '128k', '64k', '32k', '16k', '8k', '4k', '512']
      Random Workload for 1 Minute
```



Average IOPS vs. Block Size and R/W Mix %

Block Size ↓	Wld. → 100/0	95/5	65/35	50/50	35/65	5/95	0/100
512	99793.0	80136.0	47181.6	42332.6	39282.0	36557.6	37873.8
4k	93192.0	75921.6	48053.0	42298.8	38834.0	36022.8	36788.2
8k	85564.0	72108.0	49605.0	44574.2	42869.2	33858.0	34035.0
16k	58048.8	47198.8	30972.6	27568.8	26603.8	18538.0	17653.4
32k	34408.0	28315.4	18163.4	15687.8	14661.6	10583.0	10239.2
64k	16424.4	16467.8	10337.2	8401.0	7772.2	6131.2	5879.4
128k	10392.4	9185.6	5322.8	4302.8	4191.6	3575.0	2936.2
1024k	1623.2	1420.8	686.8	518.2	480.2	401.8	489.2

Device: STEC s1120 PCIe Accelerator

Excursus: PCIe Cards

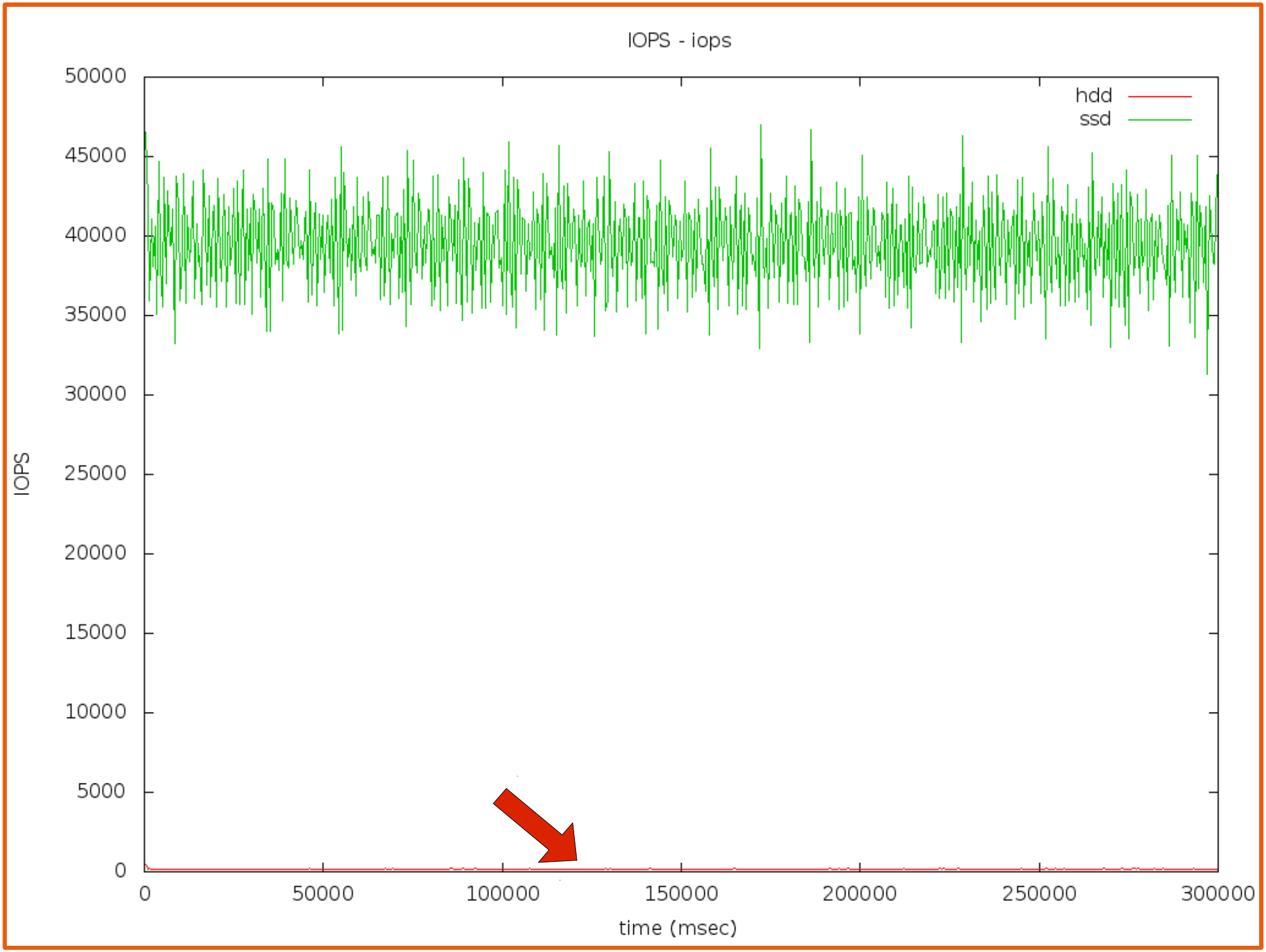
- Secure Erase must be done individually
  - Patch src/perfTest/DeviceTest.py

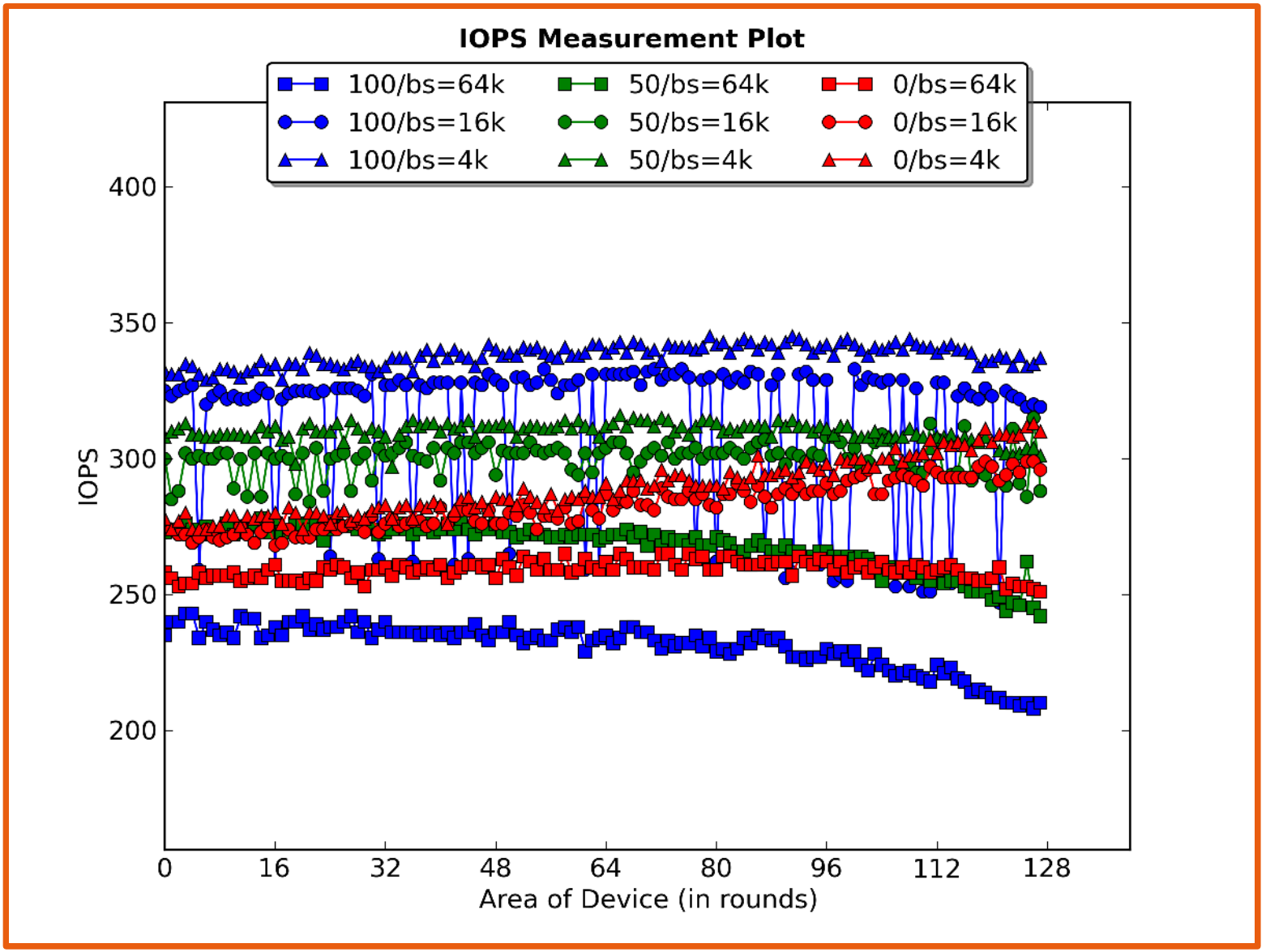
```
# sdmcmd64 sanitize target=gen4pcie:Drive0 sanitizetype=erase
```

```
# isdct -device 0 -drive 0 -erase -force
```

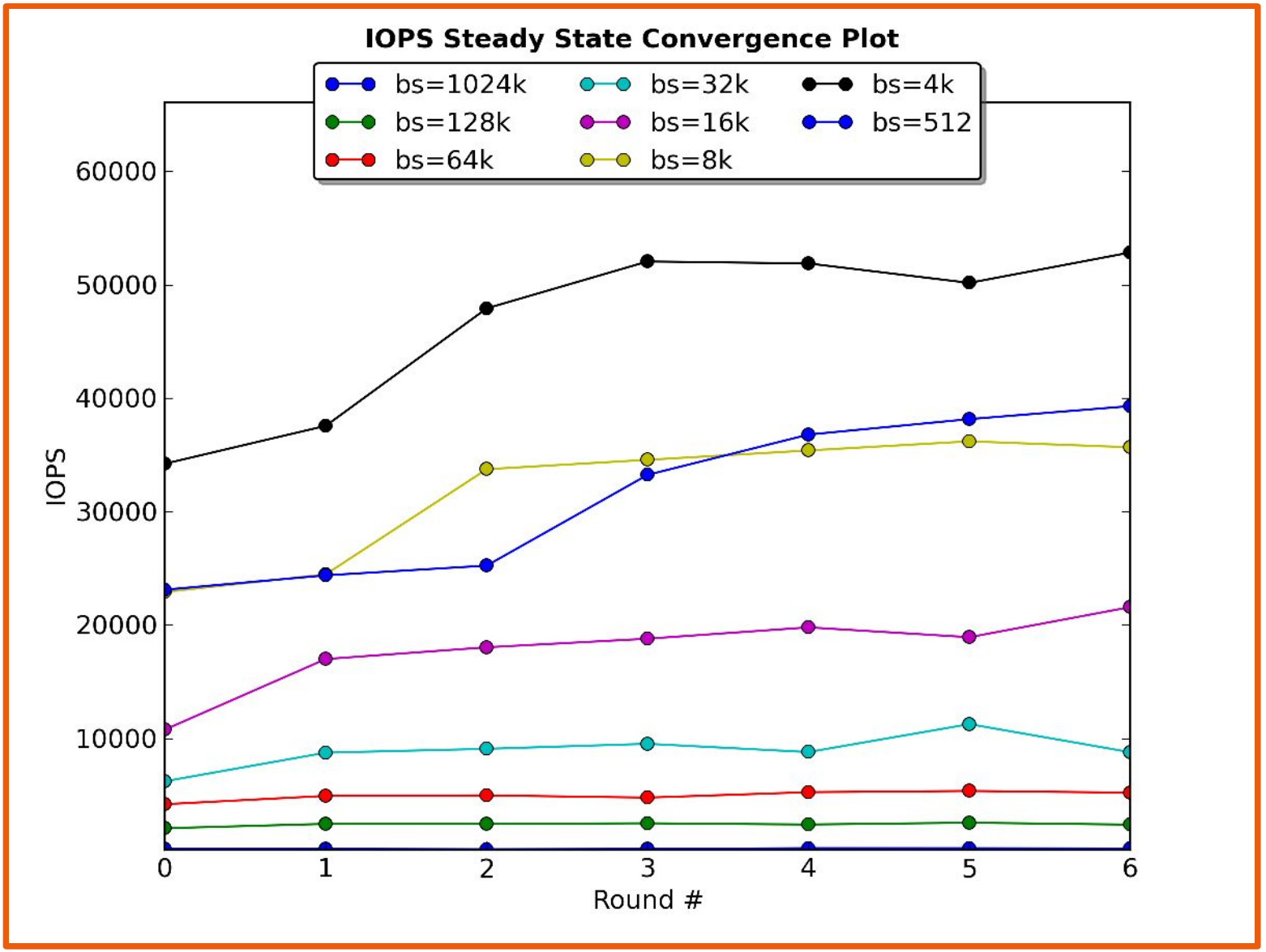
- hdparm doesn't work
  - Use a manually created description file

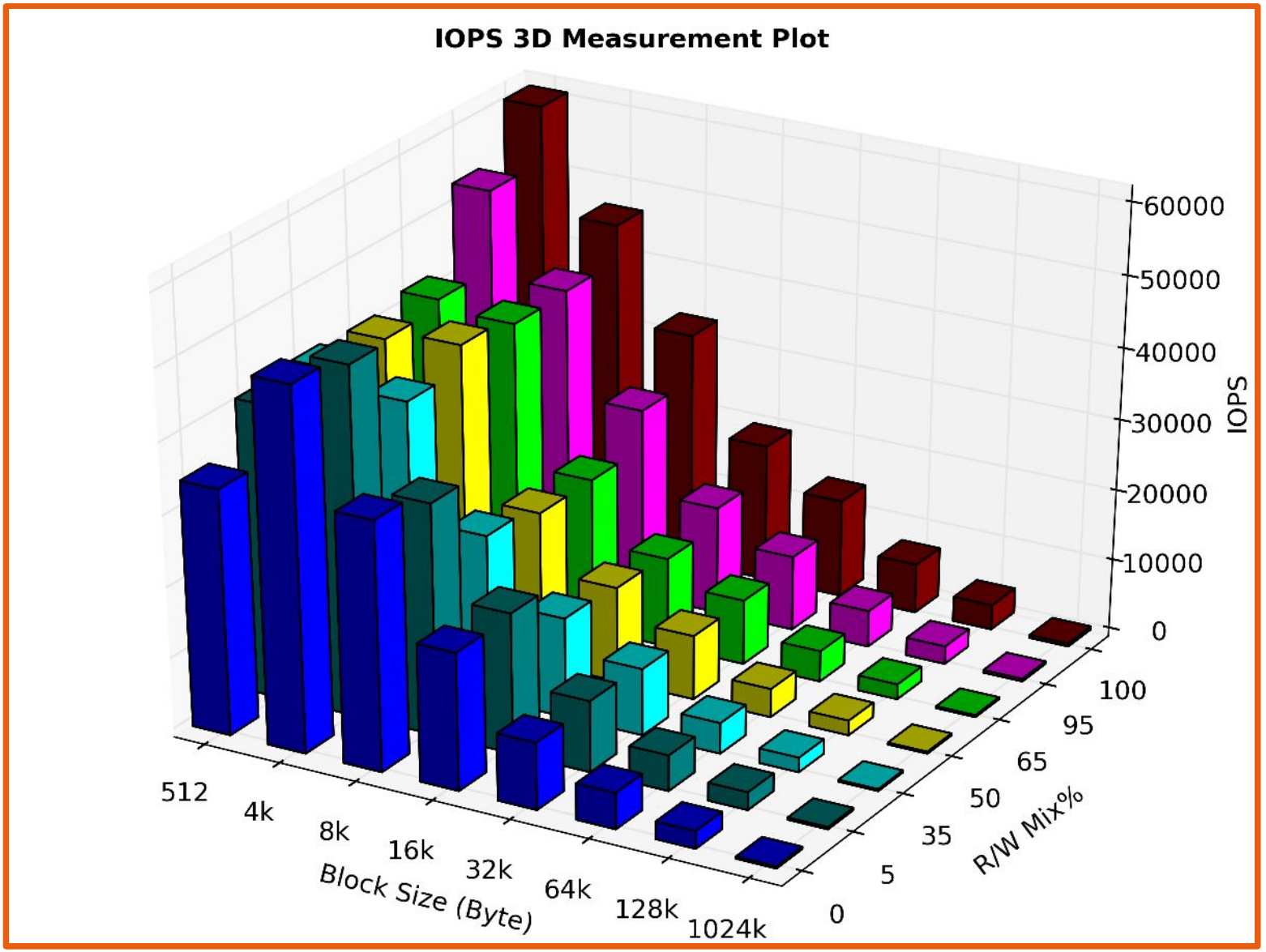
```
# sdmcmd64 GetInfo target=gen4pcie:Drive0 > S1120.dsc
# tkperf ssd S1120 /dev/skd1 -nj 2 -iod 16 -rfb -dsc S1120.dsc
```


IOPS - iops
hdd
ssd
IOPS
time (msec)
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
50000
100000
150000
200000
250000
300000



Device: HDD 4TB Western Digital WDC WD4000FYYZ-01UL1B0


IOPS Steady State Convergence Plot
bs=1024k
bs=32k
bs=4k
bs=128k
bs=16k
bs=512
bs=64k
bs=8k
60000
50000
40000
30000
20000
10000
IOPS
0
1
2
3
4
5
6
Round #



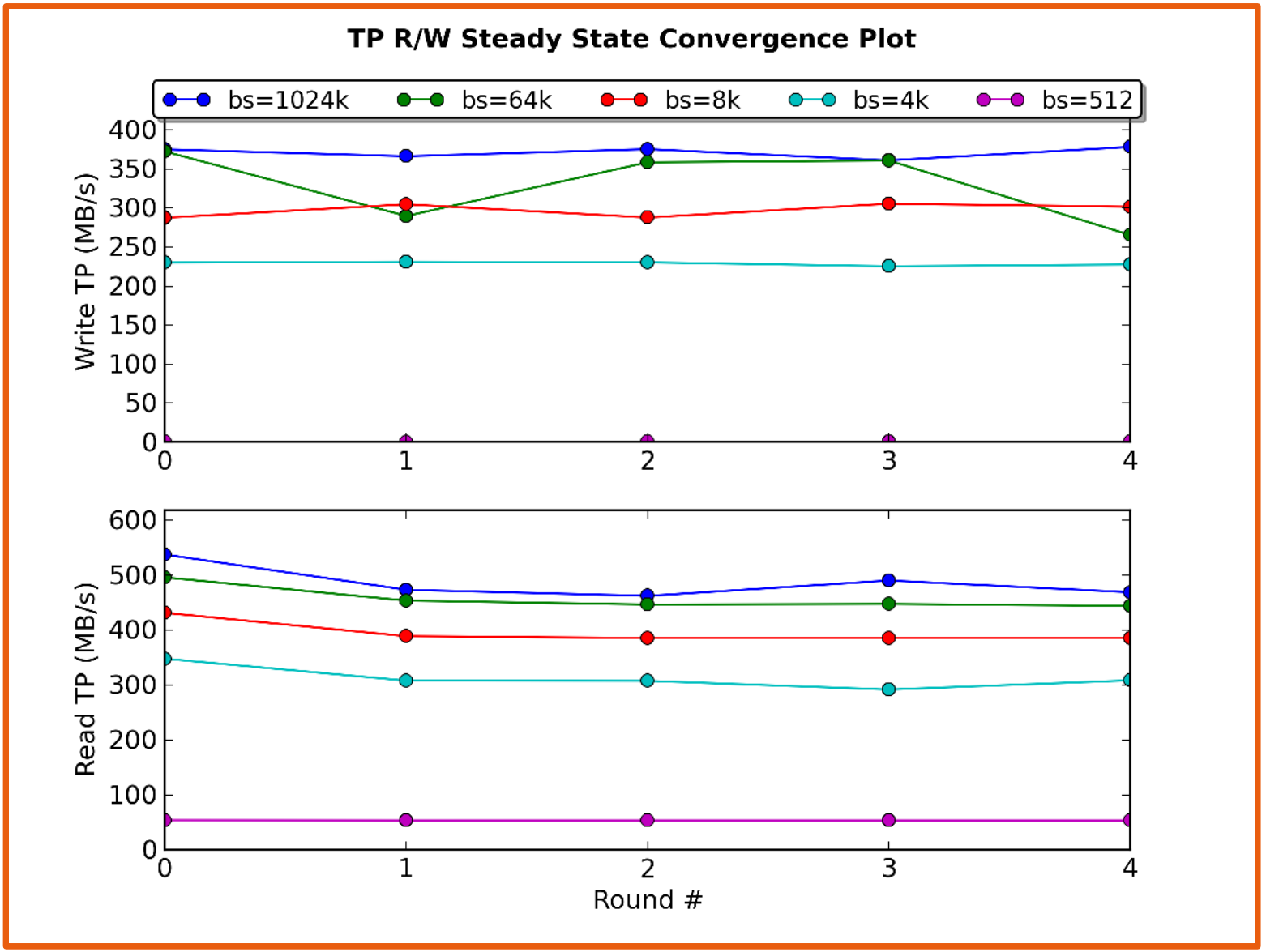
Device: Intel DC S3700 Series SSDs

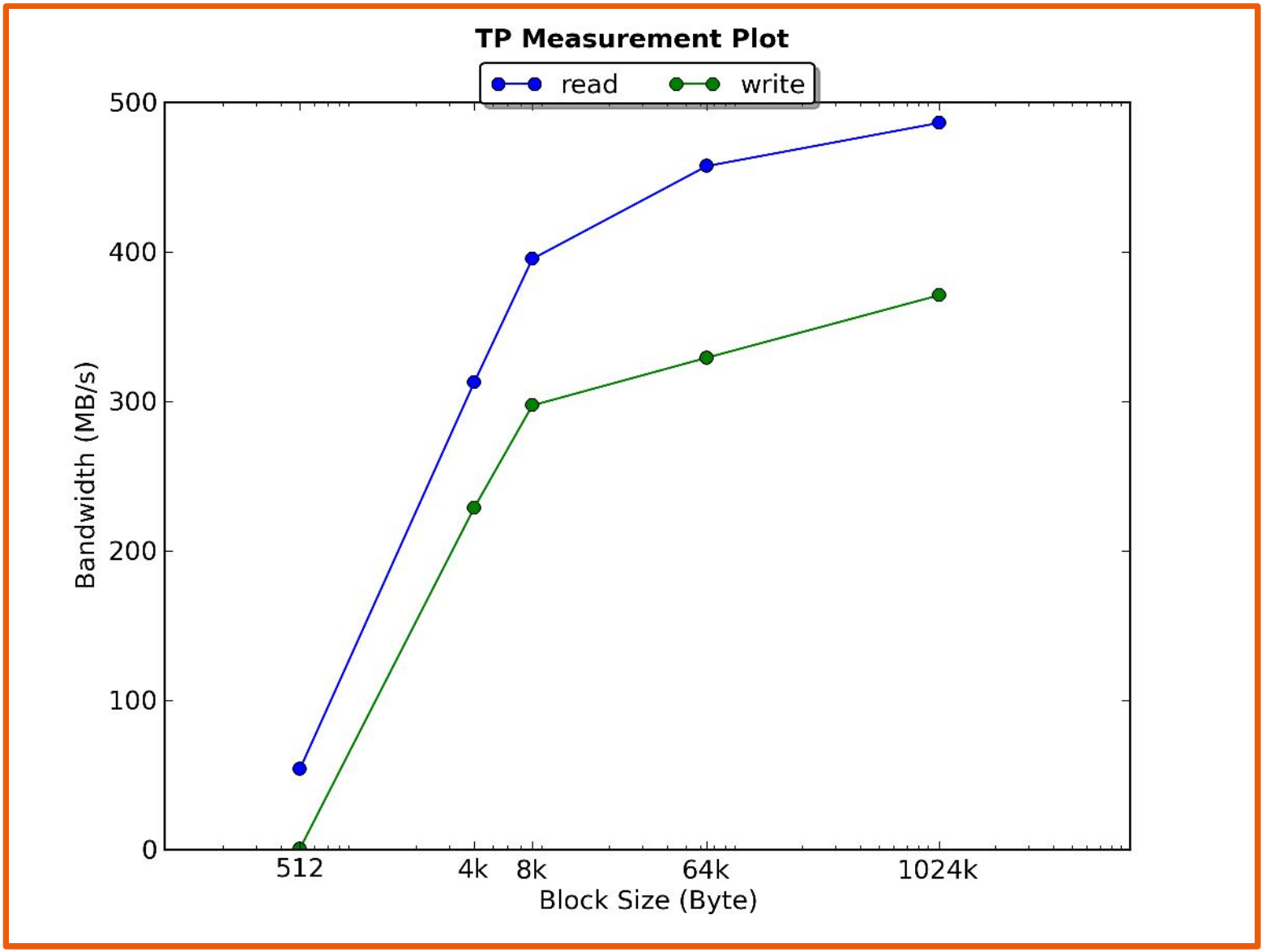

MBIs
-> A sequential Workload

4) Throughput

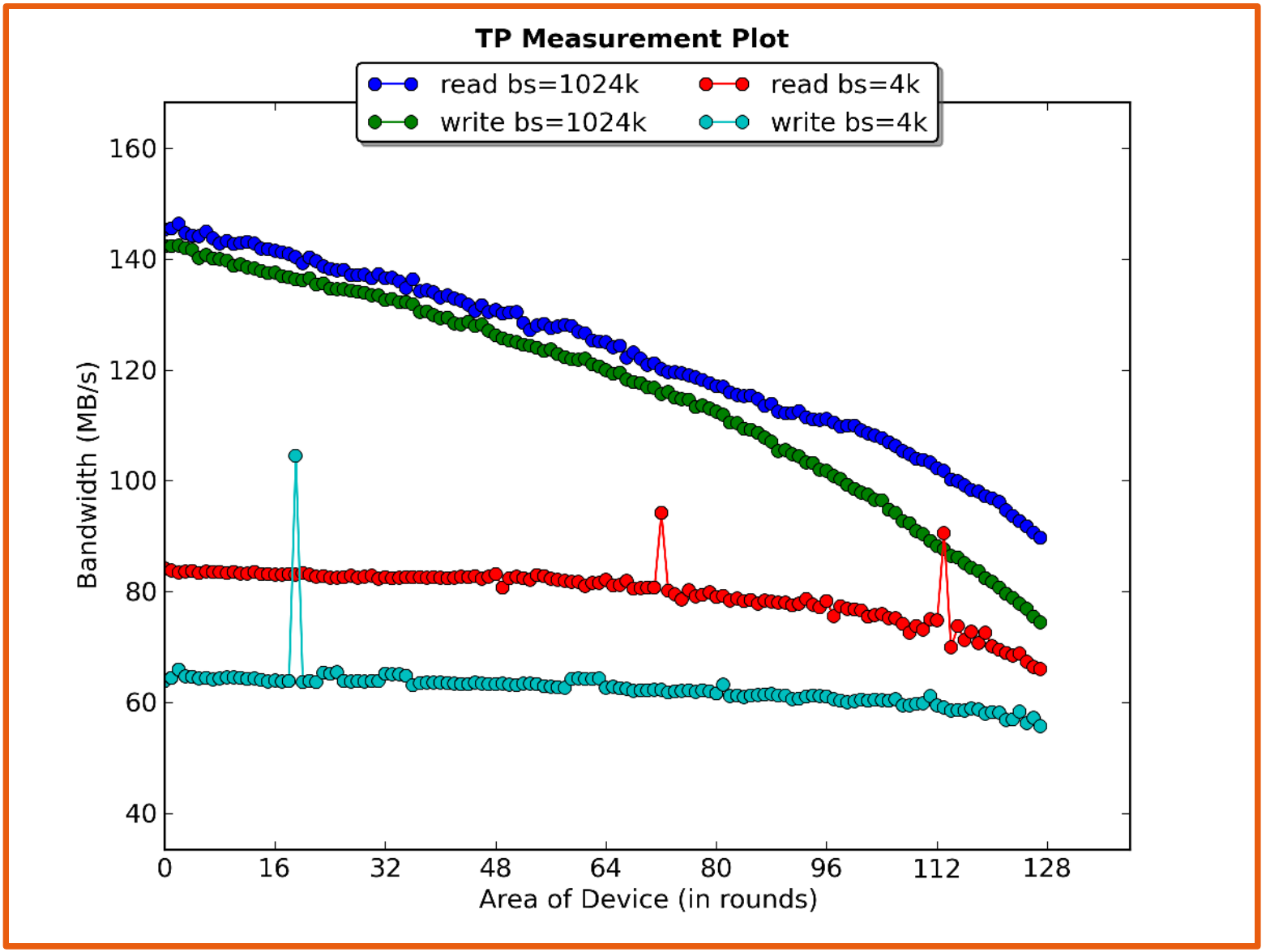
- MB per second
- Throughput, Streaming IO
- Interesting block sizes
  - 1MB, 512KB, 256KB
  - Smaller block sizes are more common for IOPS

```
For block sizes ['1024k', '64k', '8k', '4k', '512']
  Make Secure Erase
  While not Steady State
    Sequential read for 1 Minute
    Sequential write for 1 Minute
```


TP R/W Steady State Convergence Plot
bs=1024k
bs=64k
bs=8k
bs=4k
bs=512
Write TP (MB/s)
400
350
300
250
200
150
100
50
0
0
1
2
3
4
Read TP (MB/s)
600
500
400
300
200
100
0
0
1
2
3
4
Round #



Device: Intel DC S3700 Series SSDs

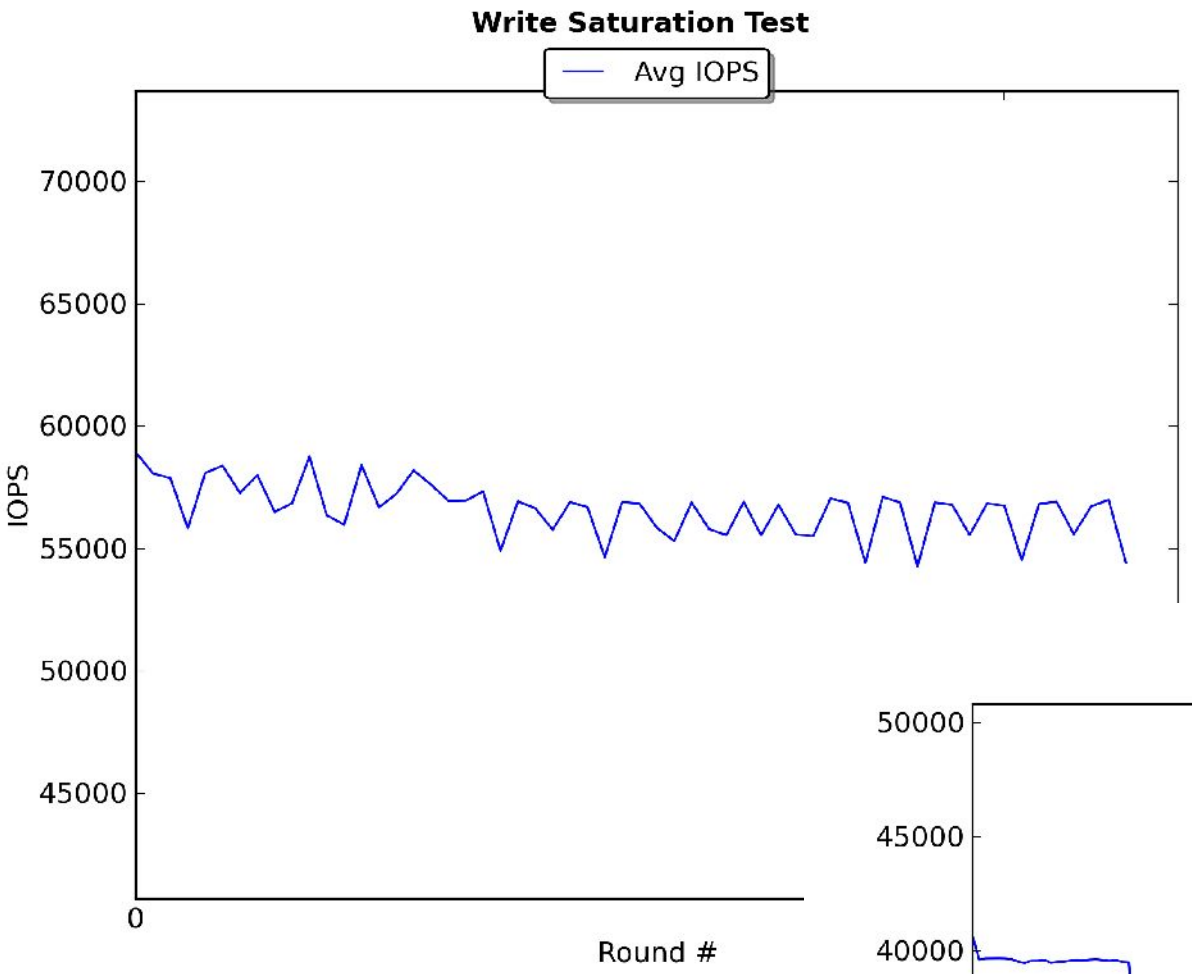


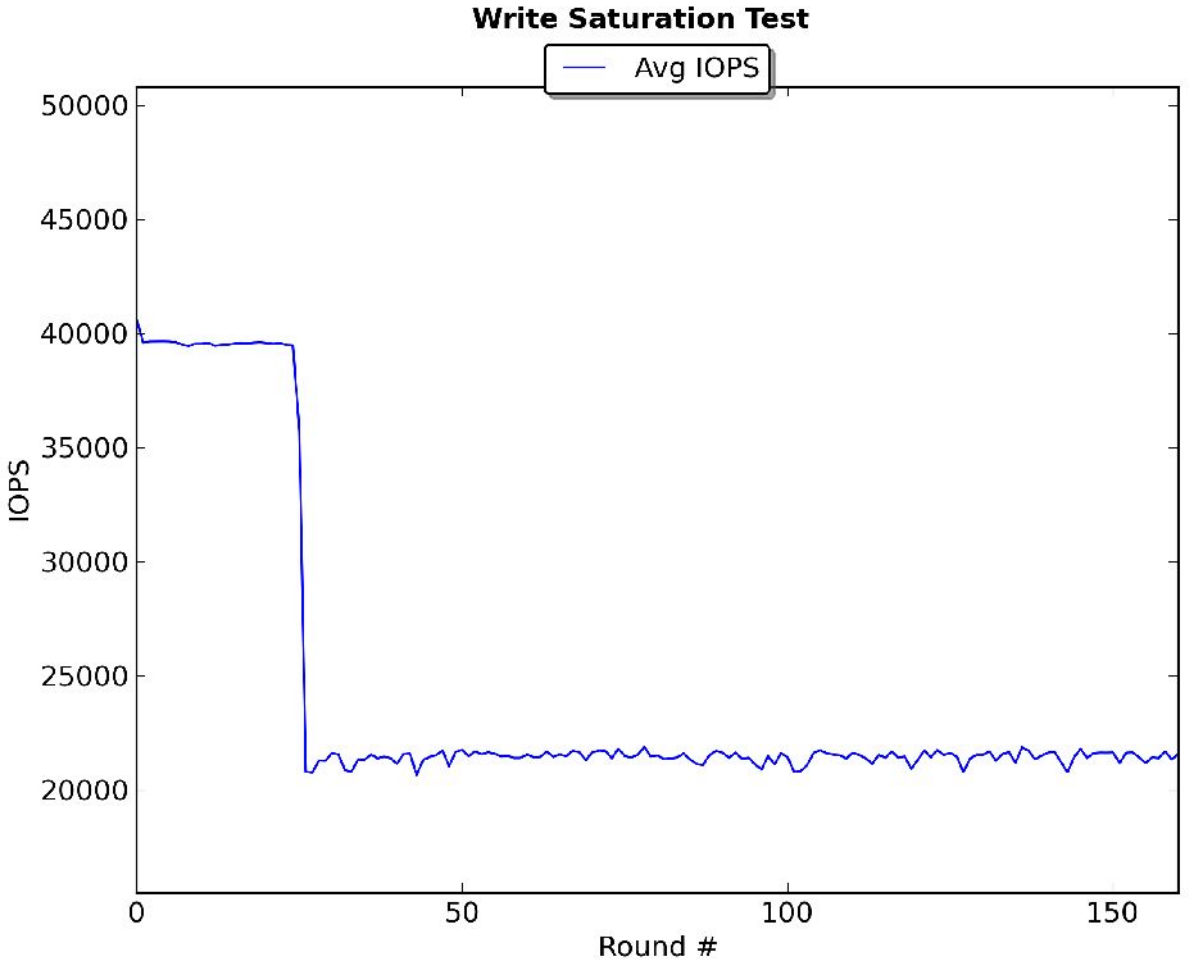
Device: HDD 4TB Western Digital WDC WD4000FYYZ-01UL1B0


Saturation

5) Write Saturation

- Continuous writes to the device
- Write enough data
- Show how stable the device is





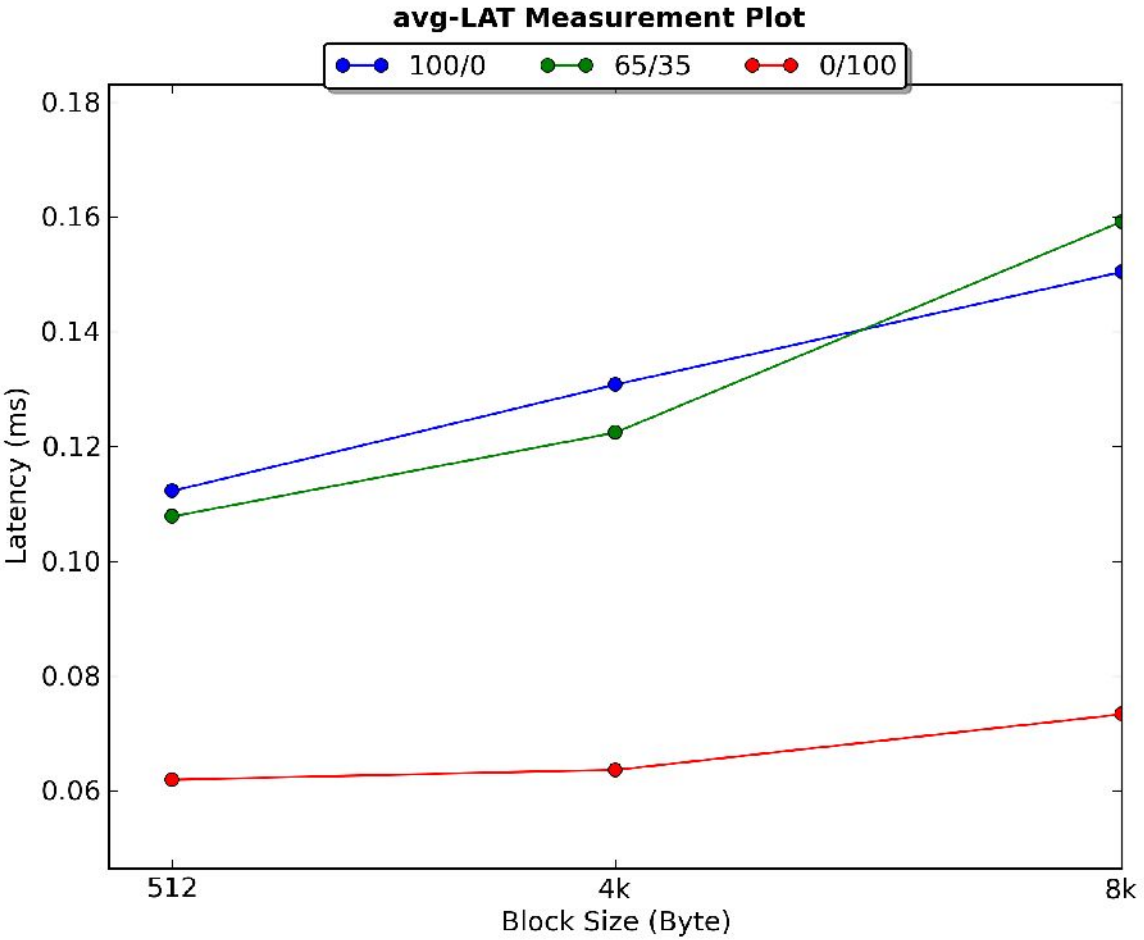
Devices

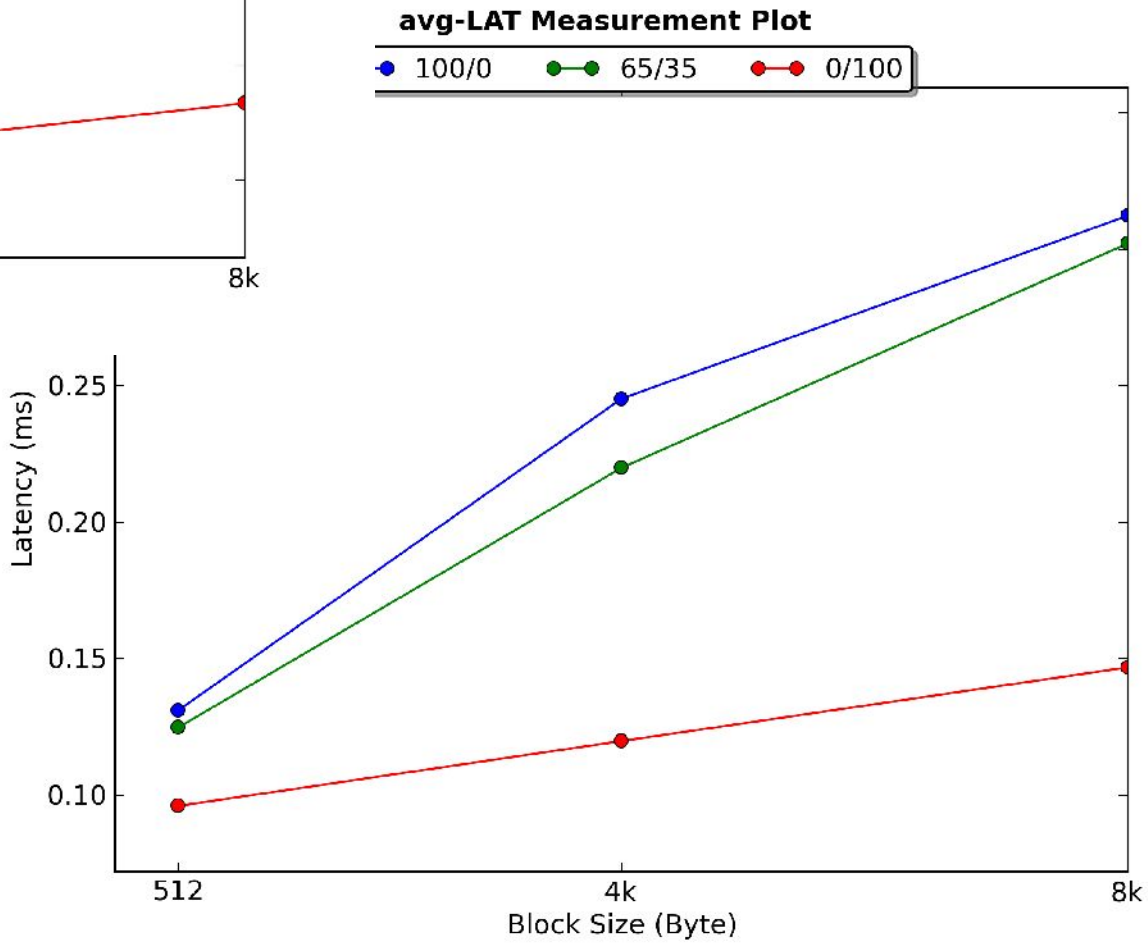
Left – Intel DC S3700 Series SSDs,

Right – Intel SSD 520

6) Latency

- Jobs and threads
  - Set to 1!
- DV
  - Average Latency 4KB writes

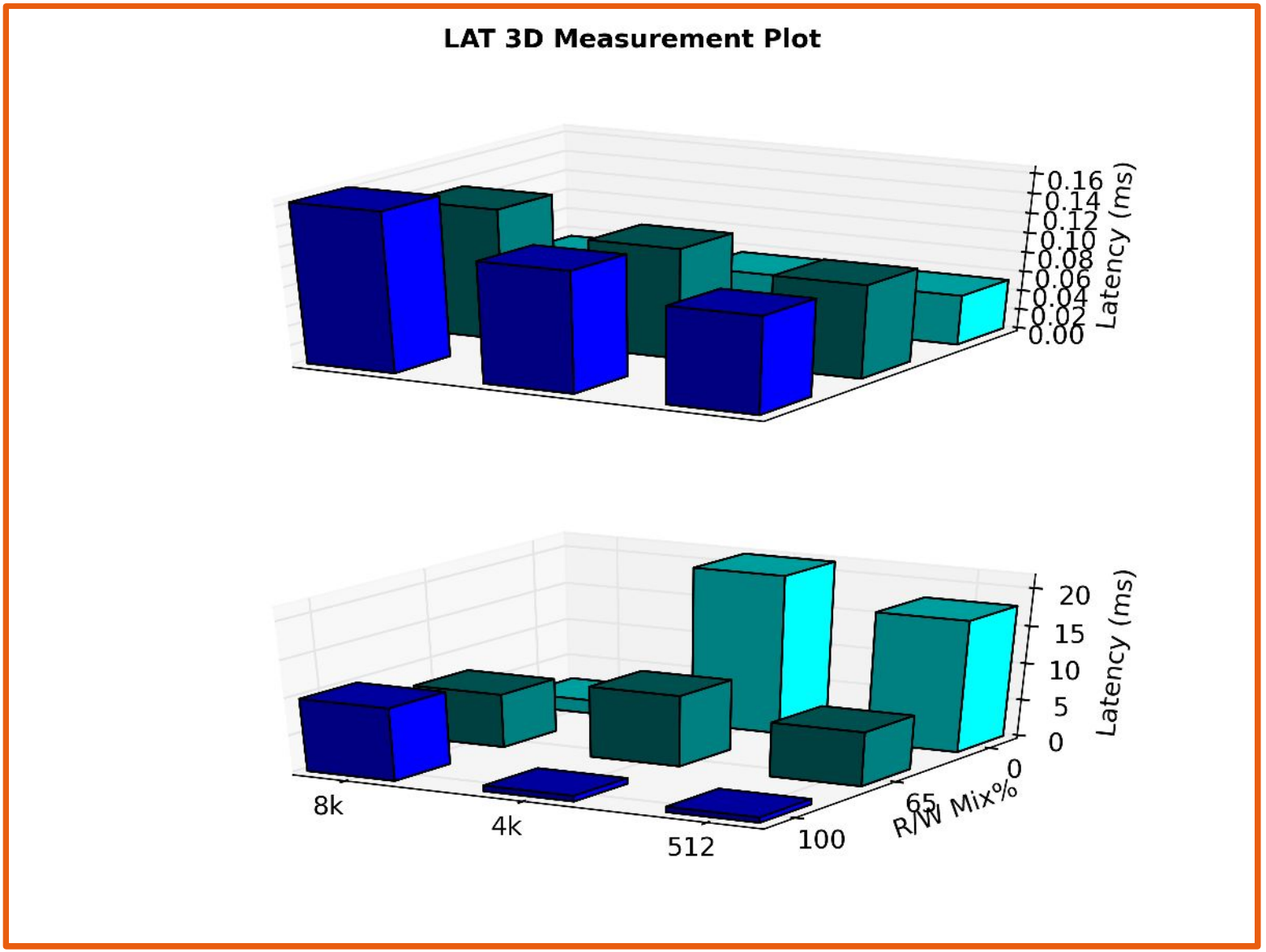




Devices

Left – Intel DC S3700 Series SSDs,

Right – Intel SSD 320



Device: STEC s1120 PCIe Accelerator

… because as we know, there are known knowns; ....
We also know there are known unknowns; …
But there are also unknown unknowns -- the ones we don't know we don't know.

(Donald H. Rumsfeld)

http://de.wikipedia.org/wiki/There_are_known_knowns

The Where

- http://www.thomas-krenn.com/de/wiki/Kategorie:T
- http://git.thomas-krenn.com/TKperf_v1.git
- http://www.snia.org/sites/default/files/SSS_PTS_En
- http://git.kernel.dk/?p=fio.git;a=summary

- Compression
  - Sandforce Controller!
  - Fio: refill_buffers

6) Resumé

Thomas-Krenn.AG®
The server experts

1. Stick tricky to find a correct setup
2. No perfect device
3. RAID/Caches

References

- Presentations
  - The Why and How of SSD Performance Benchmarking
- Pictures
  - https://commons.wikimedia.org/wiki/File:04KJER0243.jpg?uselang=en-gb
  - https://commons.wikimedia.org/wiki/File:Soca.jpg?uselang=en-gb
  - Augustinushaus Würfel
  - https://commons.wikimedia.org/wiki/File:Cgs_fat.JPG?uselang=en-gb
  - https://commons.wikimedia.org/wiki/File:Rusty_tools.JPG

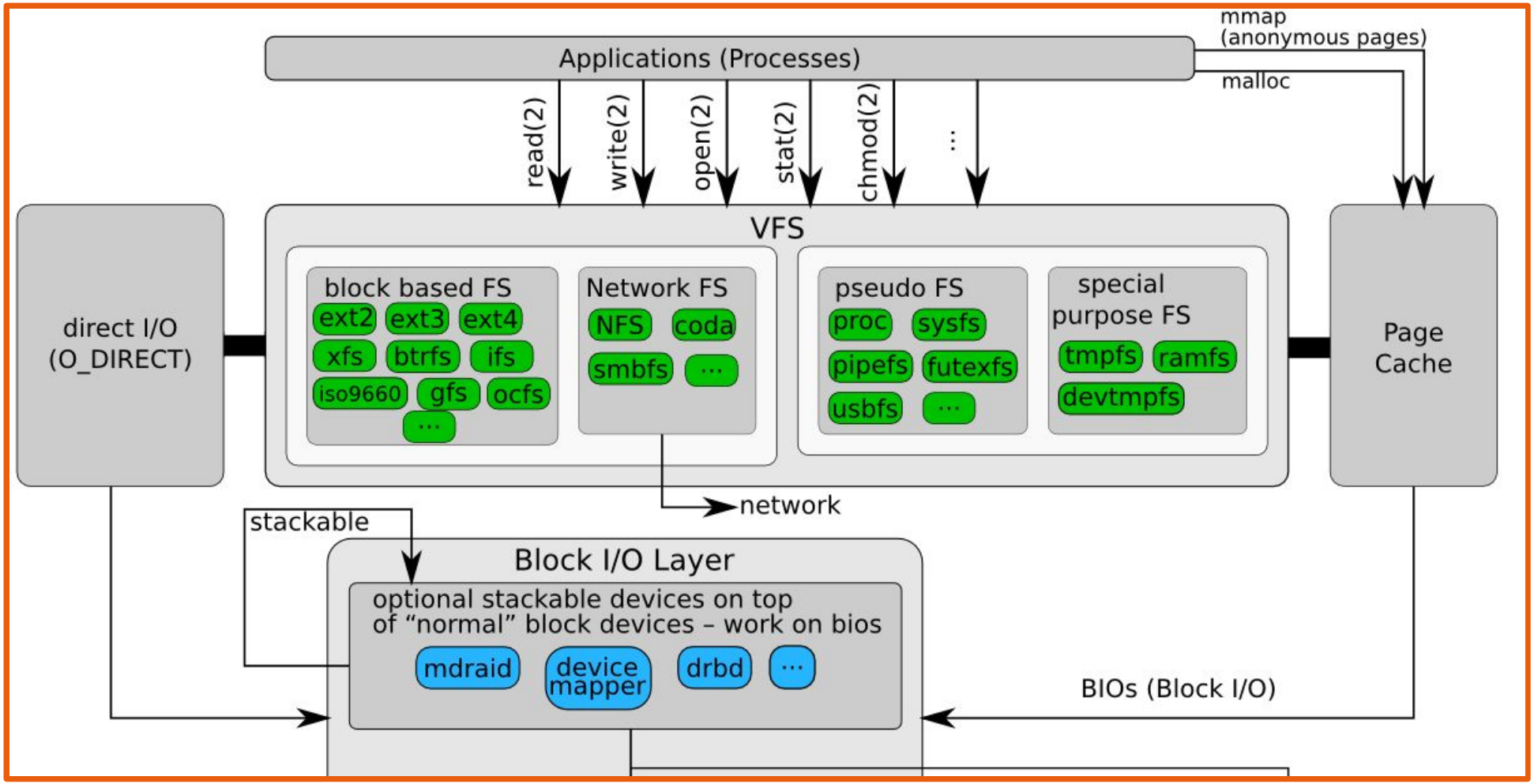
Backup Slides

- Synchronous IO-Engine[1]
  - Sync Engine → iodepth = 1
  - Submission = Completion
  - Application level: IO is done when system call returns
    - Read: IO is done by device
    - Write: Page Cache
  - Device level
    - Means not O_SYNC → can reside in drive cache

Women_Synchronized_10_metre_platform.jpg

1 Cf. http://www.spinics.net/lists/fio/msg00825.html

Page Cache


mmap
(anonymous pages)
Applications (Processes)
malloc
read(2)
write(2)
open(2)
stat(2)
chmod(2)
...
VFS
direct I/O
(O_DIRECT)
block based FS
ext2
ext3
ext4
xfs
btrfs
ifs
iso9660
gfs
ocfs
...
Network FS
NFS
coda
smbfs
...
pseudo FS
proc
sysfs
pipefs
futexfs
usbfs
...
special
purpose FS
tmpfs
ramfs
devtmpfs
Page
Cache
network
stackable
Block I/O Layer
optional stackable devices on top
of "normal" block devices – work on bios
mdraid
device
mapper
drbd
...
BIOs (Block I/O)

Steady State


```
stdyState = True
maxY = max(ys)
minY = min(ys)
avg = sum(ys)/len(ys)#calc average of values
#allow max excursion of 20% of average
avgRange = avg * 0.20
if (maxY - minY) > avgRange:
    stdyState = False

#do linear regression to calculate slope of linear best fit
y = np.array(ys)
x = np.array(xs)
A = np.vstack([x, np.ones(len(x))]).T
#calculate k*x+d
k, d = np.linalg.lstsq(A, y)[0]

#as we have a measurement window of 4, we calculate
#the slope excursion in  the window
slopeExc = k * self.testMesWindow
if slopeExc < 0:
    slopeExc *= -1
maxSlopeExc = avg * 0.10 #allowed are 10% of avg
if slopeExc > maxSlopeExc:
    stdyState = False
```

<iops>
<fioversion>"fio 2.0.7\n"</fioversion>
<numjobs>2</numjobs>
<iodepth>16</iodepth>
<roundmat>[[[126, 1087, 2157, 4235, 8052, 12196, 19534, 34917], [131,
976, 1916, 3717, 7101, 10179, 15669, 28266], [88, 600, 1153, 2230, 3870,
7975, 8766, 17704], [74, 544, 1013, 2137, 4117, 6277, 7410, 10754], [26,
261, 531, 999, 2524, 2999, 5101, 10291], [41, 415, 436, 865, 1390, 2626,
4758, 10799], [70, 635, 560, 666, 2185, 3231, 4625, 14559]],
[...]
 [[127, 1094, 2162, 4207, 7325, 10454, 13402, 20640], [130, 1032, 1920,
3130, 4813, 5161, 7352, 16104], [45, 565, 1265, 2485, 4833, 5933, 8703,
18716], [92, 521, 1050, 2017, 3446, 4774, 7641, 16926], [75, 505, 984,
2289, 3992, 6880, 8240, 17015], [94, 776, 1457, 3026, 5400, 9200, 11315,
15993], [102, 936, 1846, 3576, 6533, 10590, 13927, 17903]]]
</roundmat>
<stdyrounds>[7, 8, 9, 10, 11]</stdyrounds>
<stdyvalues>[13705, 14639, 14747, 14678, 13927]</stdyvalues>
<stdyslope>[48.300000000001042,13904.499999999989]</stdyslope>
<stdyavg>14339.2</stdyavg>
<reachstdystate>true</reachstdystate>
<rndnr>11</rndnr>
</iops>