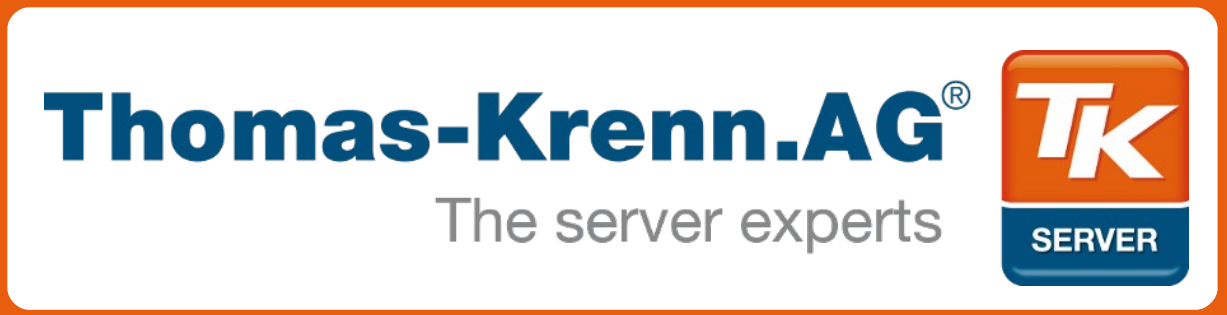

Thomas-Krenn.AG®
The server experts
TK
SERVER

Thomas-Krenn.AG®
The server experts

IO-Performance

Was MB/s, IOPS & Co wirklich bedeuten

TK Roadshow 2012

Agenda

1) Definitionen
2) Page Cache
3) IO-Tiefe
4) SNIA-Tests
5) MB/s & IOPS (seq. und zufällige Workloads)
6) Tipps & Resümee

… 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

1) Definitionen

- Transferrate / Datendurchsatz
  - MB/s
  - Vergleich: Personen/h auf einer Strecke

- Anzahl I/O Operationen pro Sekunde
  - IOPS
  - Vergleich: Anzahl mögl. individueller Fahrten

- Dazu kommt: Latenz!
  - Queue Depth
  - Vergleich: ab wie vielen Fahrzeugen fährt die Fähre los?

1) Definitionen

- Synchrone IO-Engine[1]
  - Sync Engine → iodepth = 1
  - Submission = Completion
  - Applikations-Ebene: IO ist fertig, wenn System Call retourniert
    - Read: IO von Device beendet
    - Write: Page Cache
  - Device-Ebene
    - Heißt nicht O_SYNC
    - Höhere iodepth bei Device möglich

Women_Synchronized_10_metre_platform.jpg

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

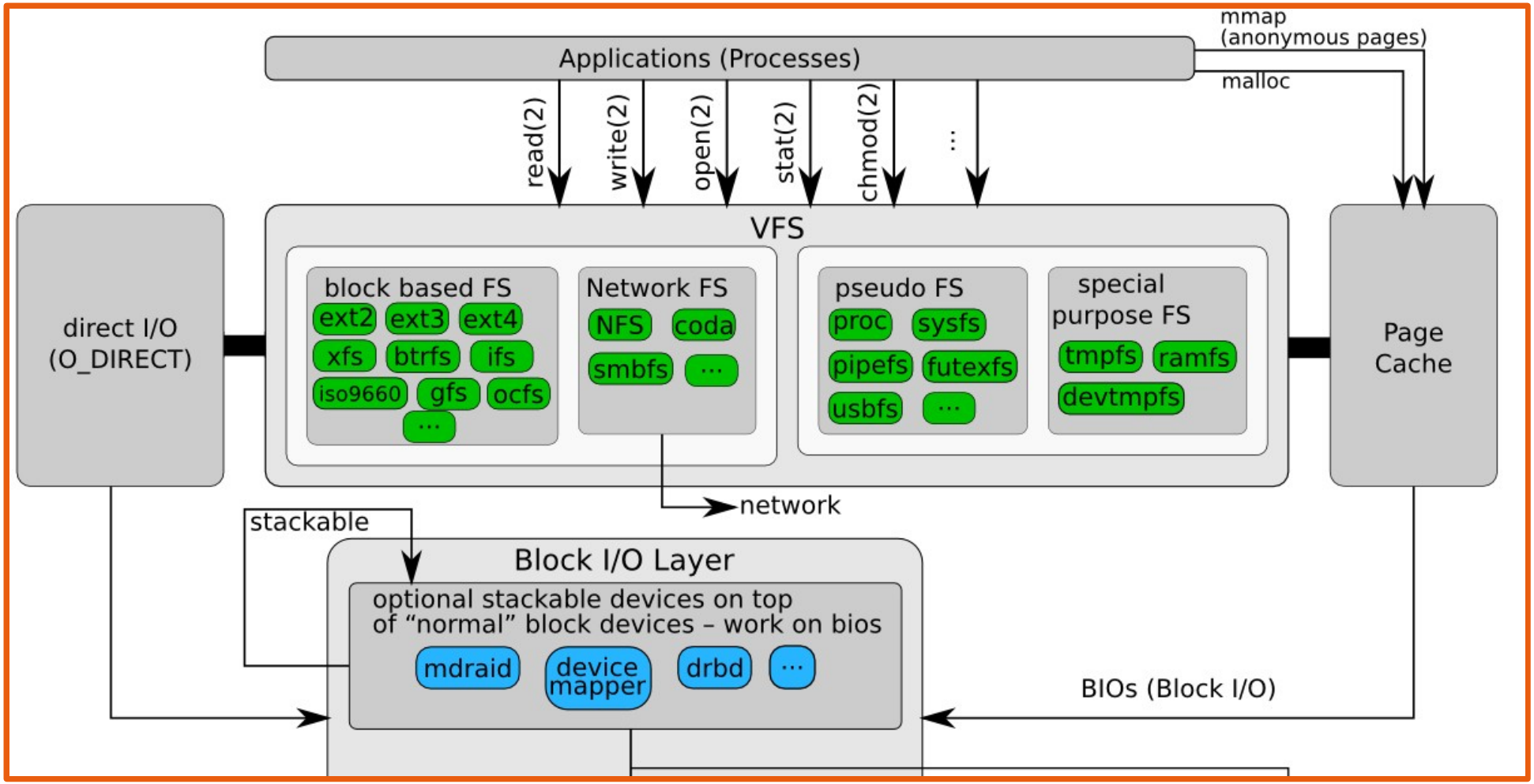
1) Definitionen

- Asynchrone IO-Engine
  - Absetzen von mehreren Requests
  - Warte/Prüfe auf Completion
  - Setzt „direct“ voraus (ohne Page Cache)
- Fio
  - Zumeist „libaio“ → „iodepth“ ausstehende IOs
  - Unterschied zwischen Applikations- und Device-Ebene[1]
    - Blockgrößen-Segmentierung
    - Scheduler

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

To cache or not to cache

2) 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)

2) Page Cache

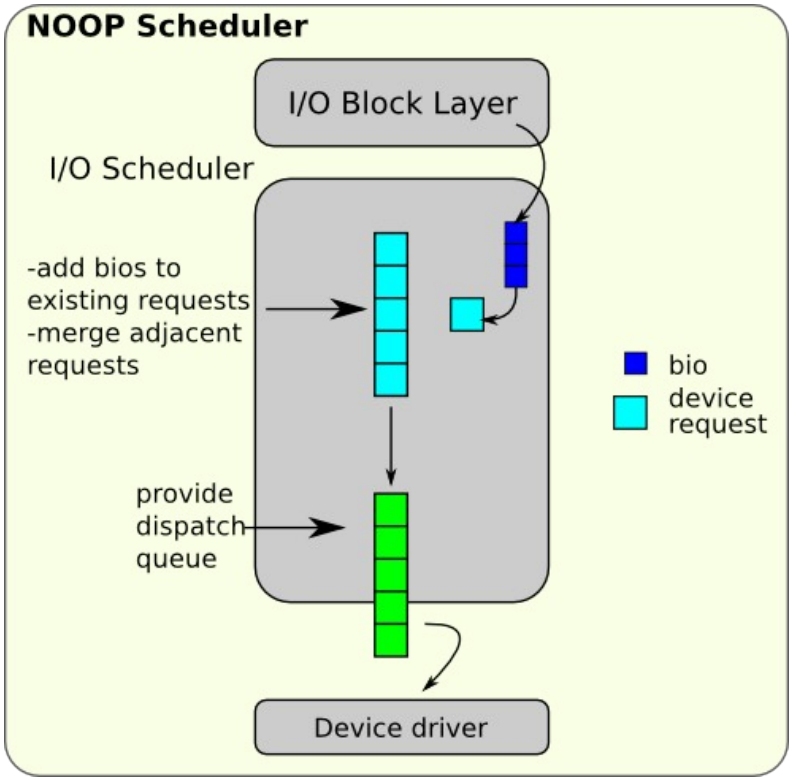
- Verwendung des Page Caches
  - Für Applikations-Performance interessant
  - Nur synchron möglich
  - Aus Applikations-Sicht iodepth = 1

```
$ 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
```

2) No Page Cache

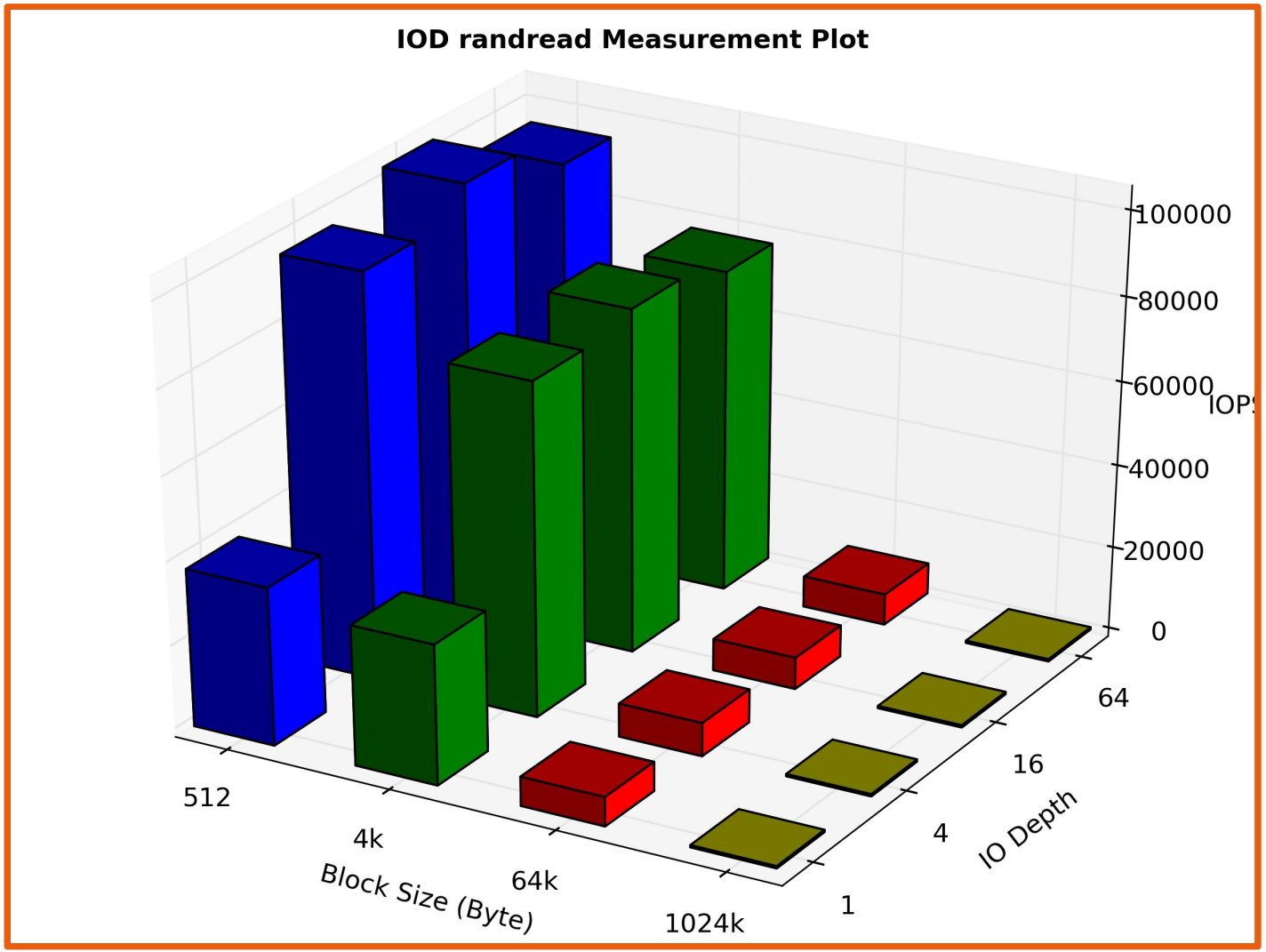
- Fio: direct = 1 umgeht Page Cache
  - Linux normalerweise O_DIRECT
- Für asynchrone Zugriffe (libaio) direct Vorraussetzung
  - Vgl. IO-Tiefe!
- Testet Performance des Devices „direkt“
  - IO-Scheduler




10-Tiefe

3) IO-Tiefe

- It matters!



3) IO-Tiefe

- Auslasten der Device Tiefe
  - NCQ
- SSDs
  - Paralleles bearbeiten von Requests

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

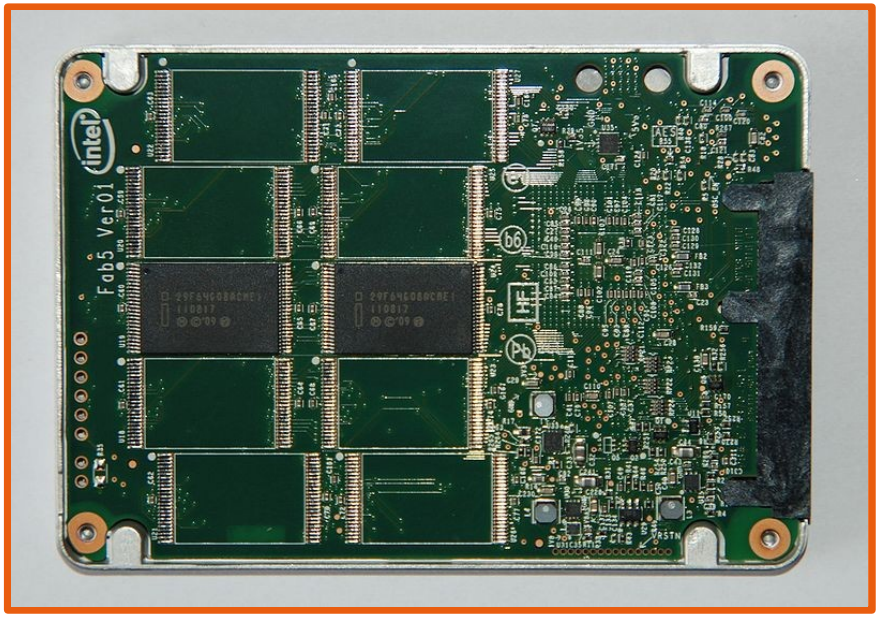
- Unterschiede Applikation ↔ Device
- Latenz wird größer!

Exkurs: SNIA Test Spezifikation

4) SNIA Tests

Thomas-Krenn.AG®
The server experts

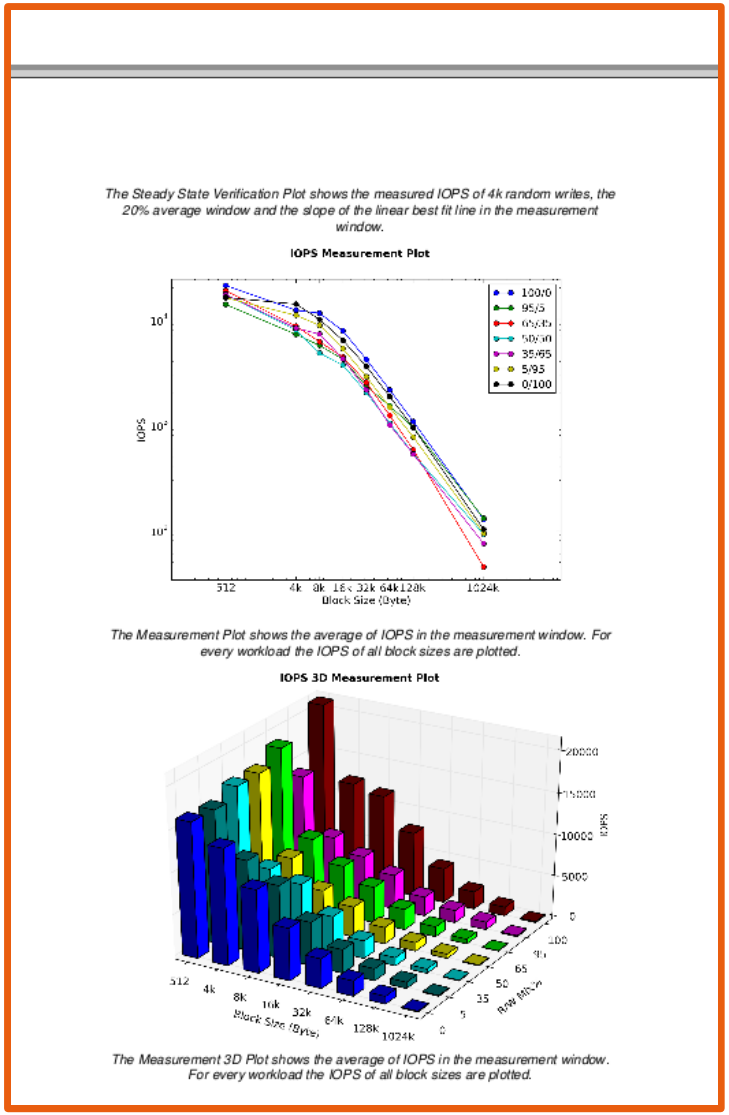
- Solid State Storage (SSS) Performance Test Specification (PTS)[1] → Enterprise-Bereich
- Speziell für SSDs
- Fokussiert auf „Stable State“
- Prekonditionierung
- Anforderungen an Dokumentation



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

4) SNIA Tests

- Applikation in Python
  - Log-Datei mit detaillierten Informationen
  - Graphen via pyplot
  - Xml-Datei mit Ergebnissen der Tests
  - Rst-Ausgabe für Pdf-Report
  - Open Source =)



The Steady State Verification Plot shows the measured IOPS of 4k random writes, the 20% average window and the slope of the linear best fit line in the measurement window.

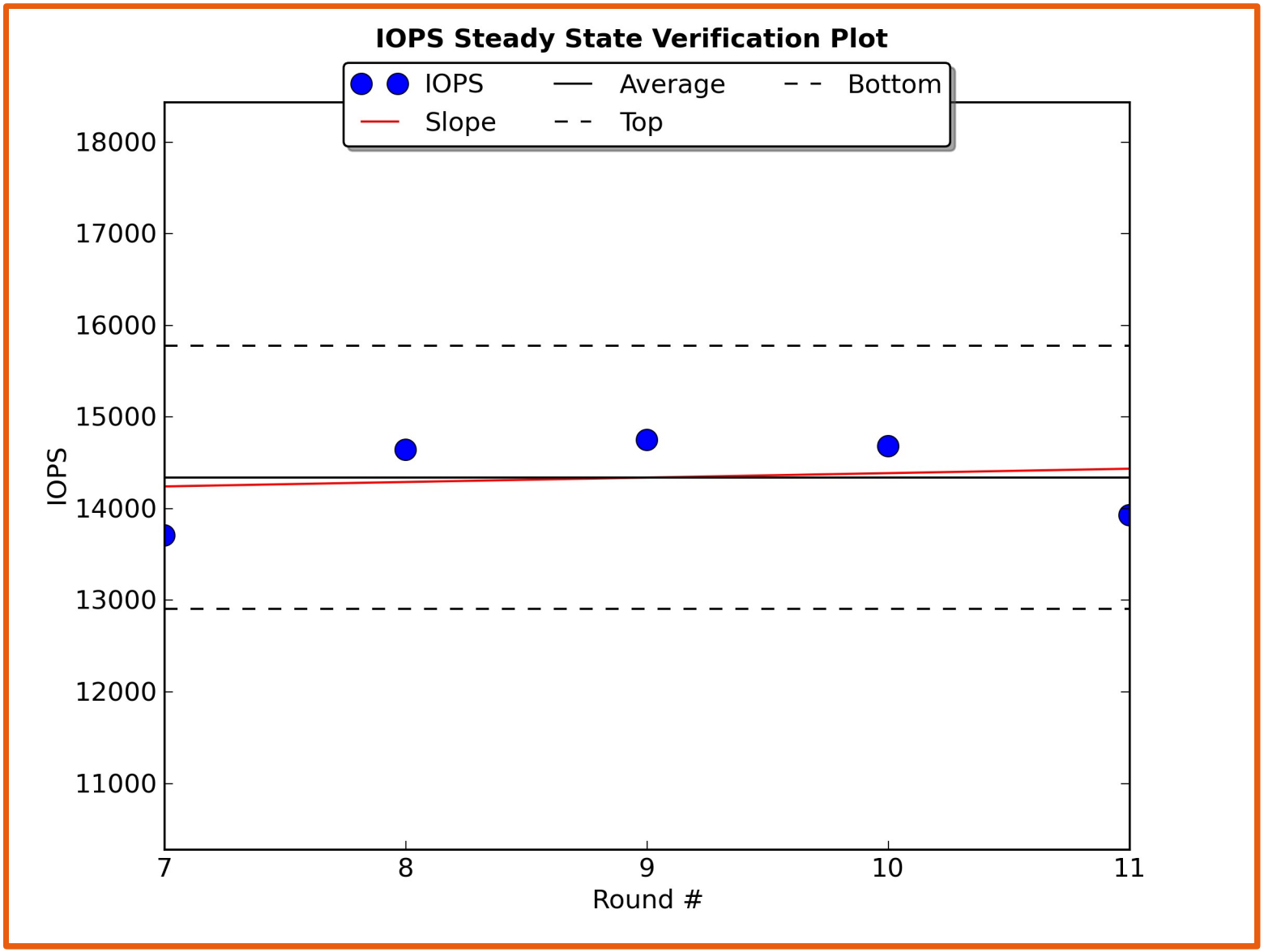
IOPS Measurement Plot

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.

<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>

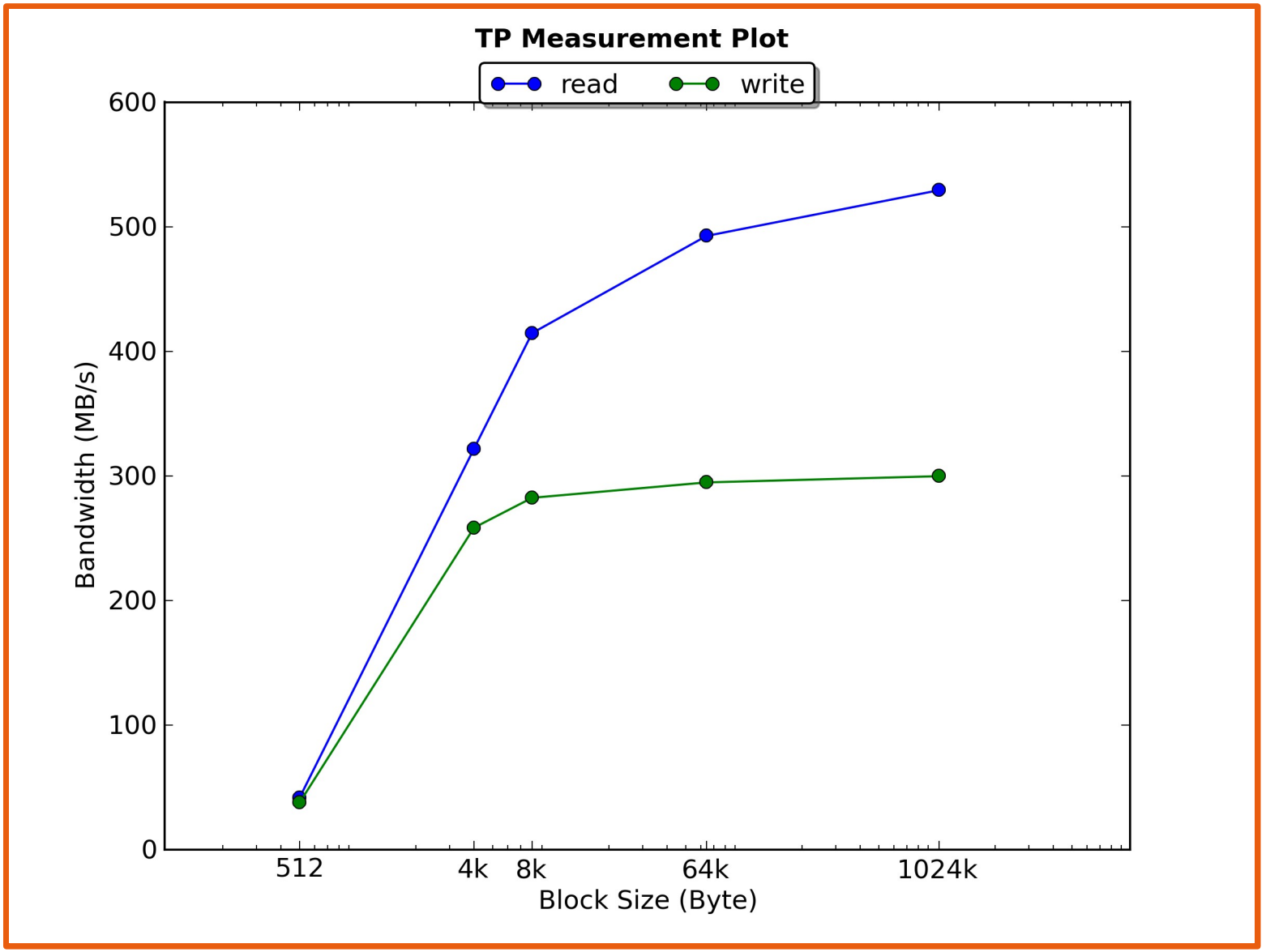

IOPS Steady State Verification Plot
IOPS
Average
Bottom
Slope
Top
18000
17000
16000
15000
14000
13000
12000
11000
IOPS
7
8
9
10
11
Round #

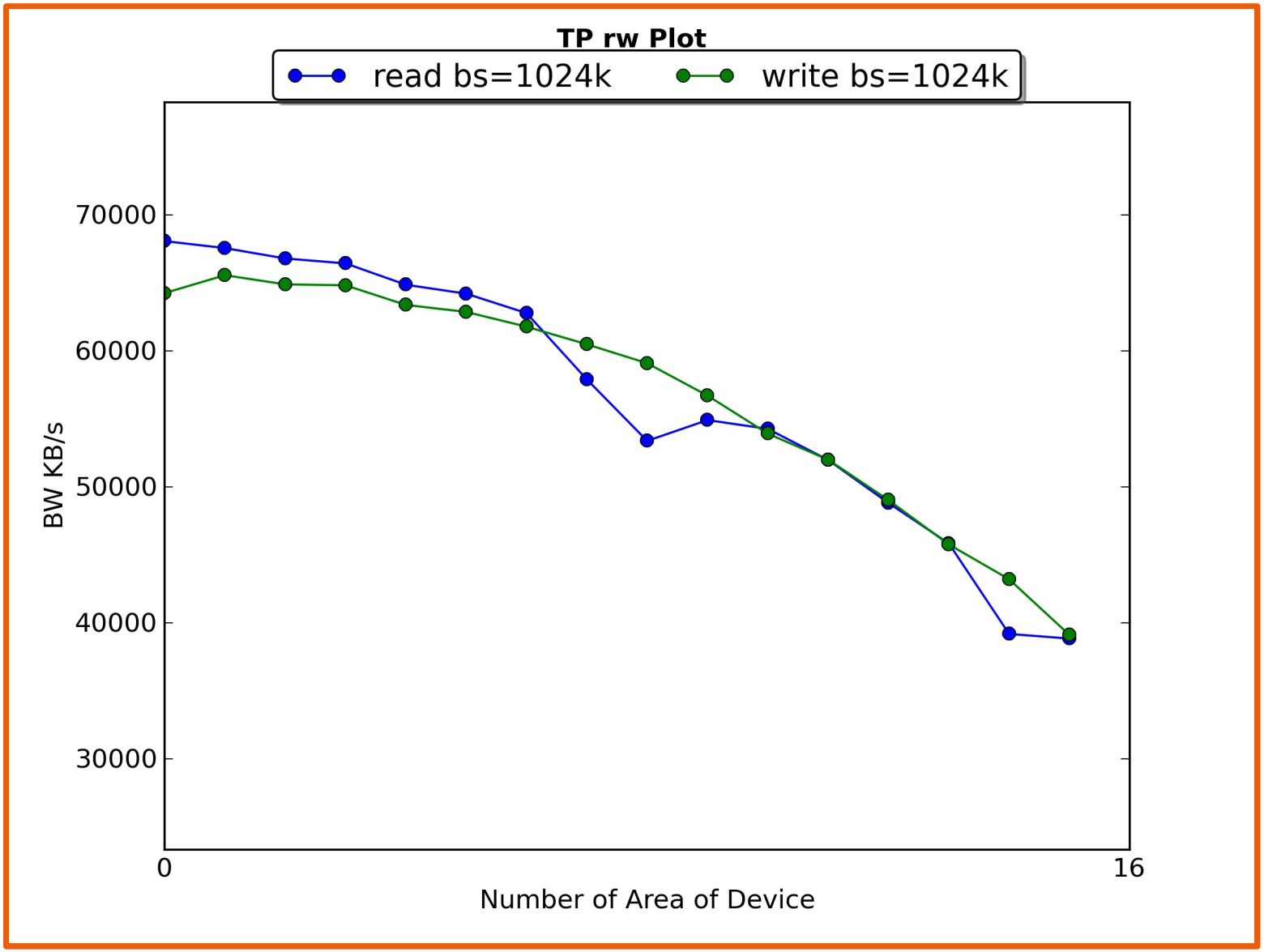

MBIs
-> Sequentielle Workloads

5) Sequentiell

- Messgröße: Mbyte/s
- Throughput, Durchsatz, Streaming IO
- Block-Größen
  - 1MB, 512KB, 256KB
  - Für sehr kleine Blockgrößen Durchsatz nicht interessant → IOPS
- Disk Write Cache?!

```
# /usr/local/bin/fio --rw=read --name=wd --bs=1024k --direct=1 --filename=/dev/sde
--offset=0 --runtime=300
wd: (g=0): rw=read, bs=1M-1M/1M-1M, ioengine=sync, iodepth=1
fio-2.0.9
Starting 1 process
Jobs: 1 (f=1): [R] [100.0% done] [106.0M/0K /s] [106 /0  iops] [eta 00m:00s]
wd: (groupid=0, jobs=1): err= 0: pid=2004: Wed Oct  3 08:19:36 2012
  read : io=31885MB, bw=108834KB/s, iops=106 , runt=300001msec
    clat (usec): min=4069 , max=42710 , avg=9404.33, stdev=384.97
     lat (usec): min=4070 , max=42711 , avg=9404.66, stdev=384.97
    clat percentiles (usec):
     |  1.00th=[ 8256],  5.00th=[ 9408], 10.00th=[ 9408], 20.00th=[ 9408],
     | 30.00th=[ 9408], 40.00th=[ 9408], 50.00th=[ 9408], 60.00th=[ 9408],
     | 70.00th=[ 9408], 80.00th=[ 9408], 90.00th=[ 9408], 95.00th=[ 9408],
     | 99.00th=[10560], 99.50th=[10560], 99.90th=[12096], 99.95th=[16768],
     | 99.99th=[25984]
    bw (KB/s)  : min=101386, max=109714, per=100.00%, avg=108974.61, stdev=574.84
    lat (msec) : 10=98.87%, 20=1.12%, 50=0.01%
  cpu          : usr=0.07%, sys=1.04%, ctx=32496, majf=0, minf=288
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued    : total=r=31885/w=0/d=0, short=r=0/w=0/d=0

Run status group 0 (all jobs):
   READ: io=31885MB, aggrb=108833KB/s, minb=108833KB/s, maxb=108833KB/s,
mint=300001msec, maxt=300001msec

Disk stats (read/write):
  sde: ios=63746/0, merge=0/0, ticks=442316/0, in_queue=442152, util=98.86%
```


TP Measurement Plot
read
write
600
500
400
300
200
100
0
Bandwidth (MB/s)
512
4k
8k
64k
1024k
Block Size (Byte)

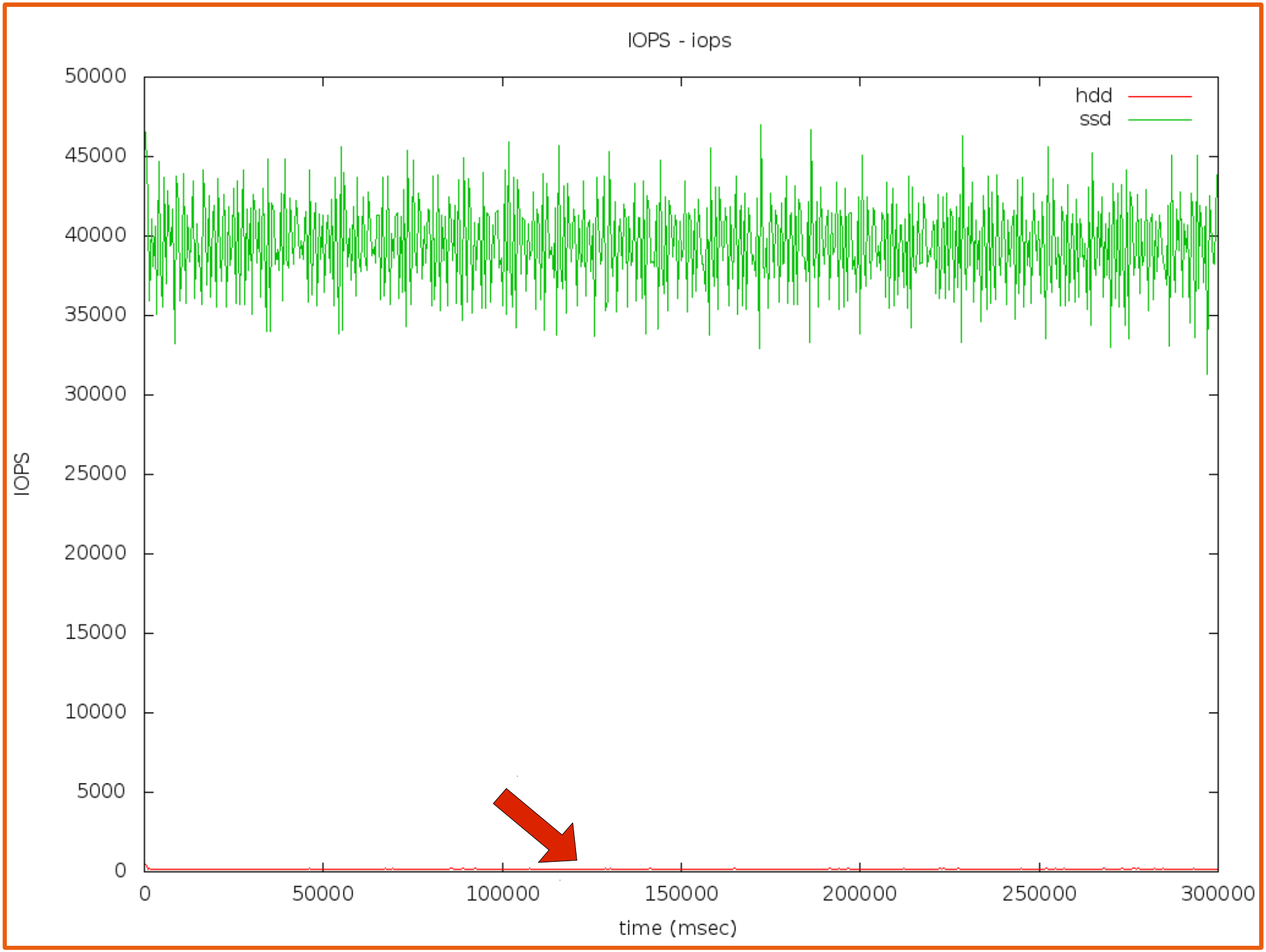

TP rw Plot
read bs=1024k
write bs=1024k
70000
60000
50000
40000
30000
BW KB/s
0
16
Number of Area of Device

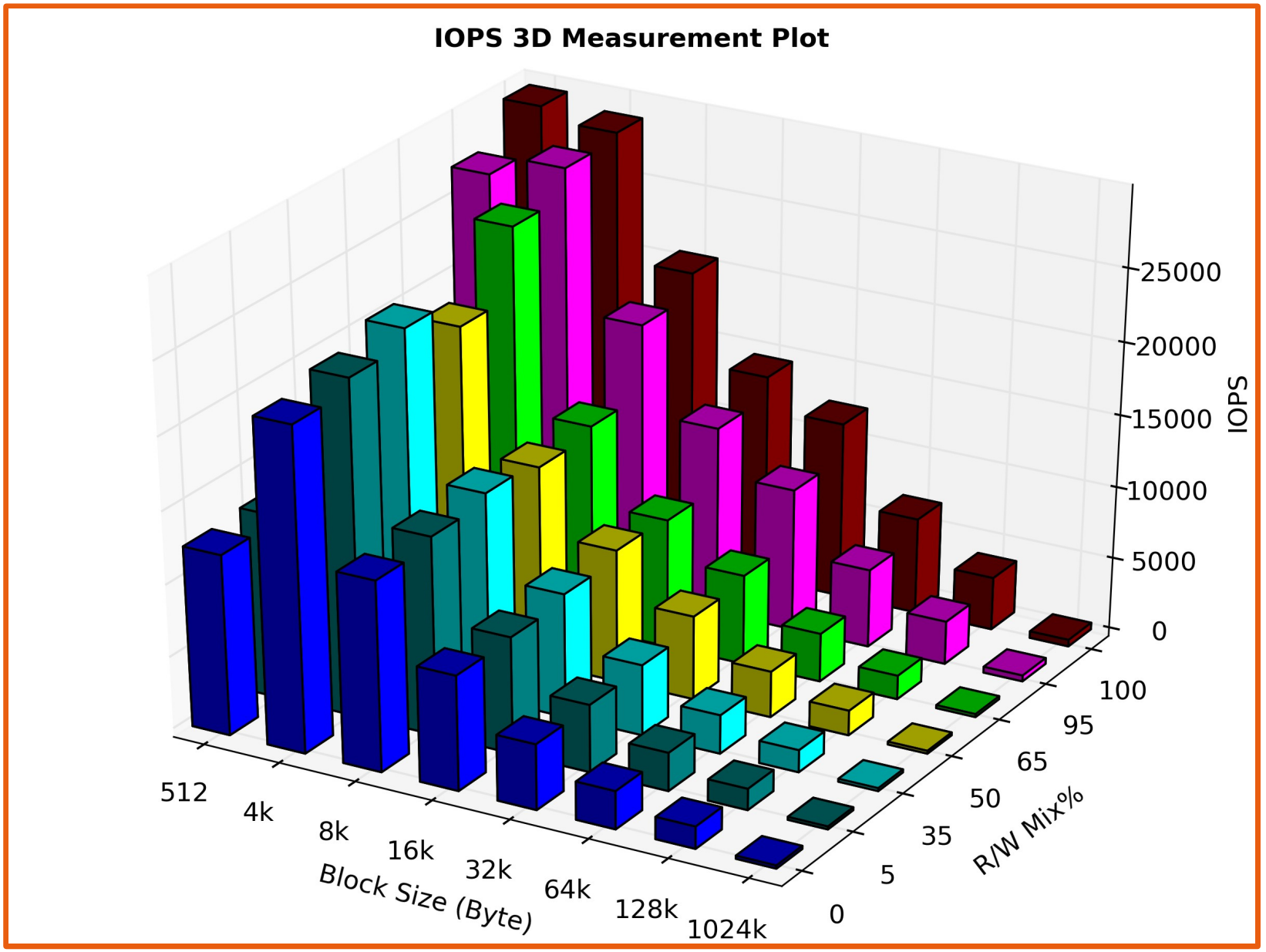

IOPS
-> Zufällige Workloads

5) Zufällig

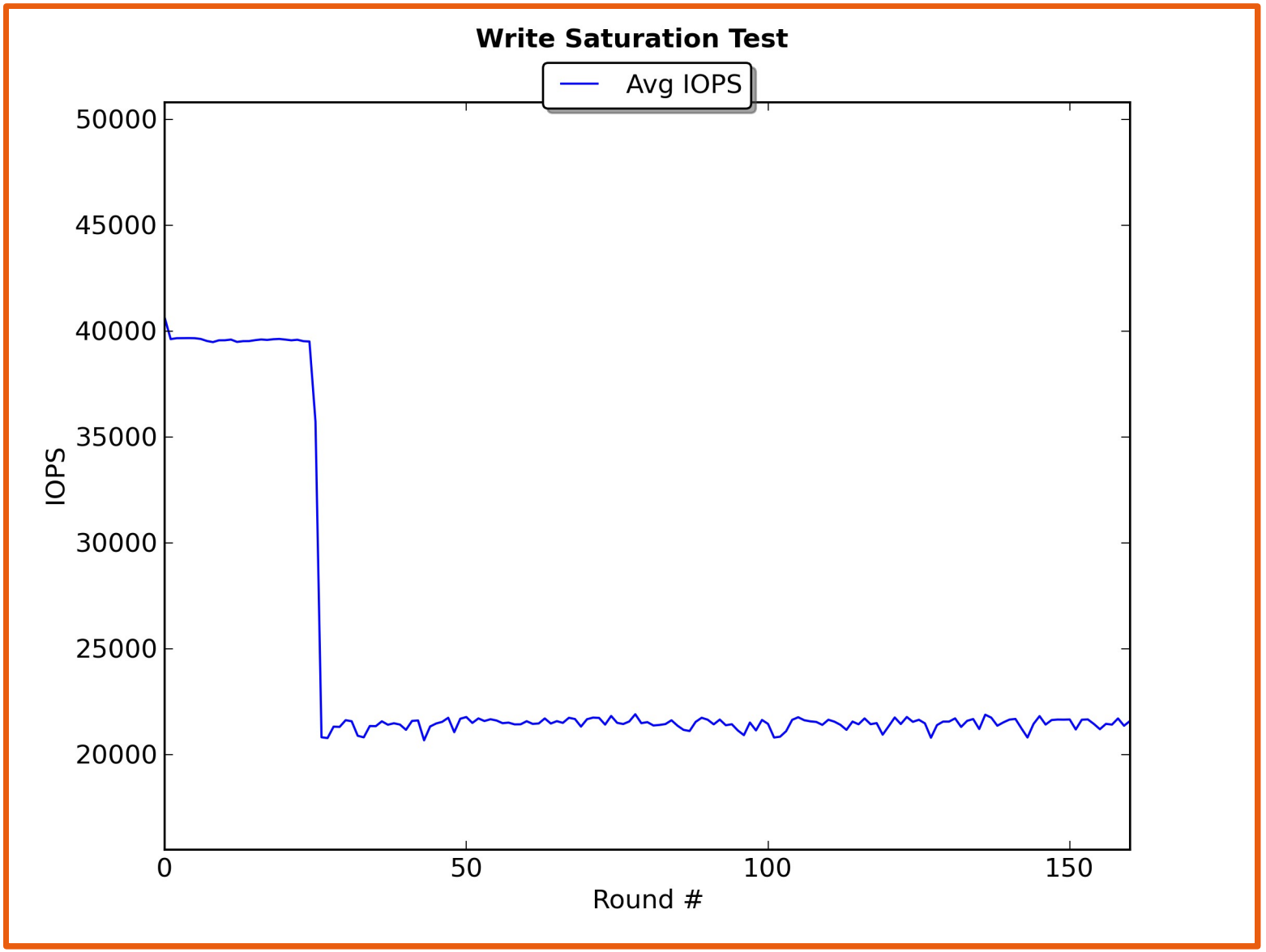
- Messgröße: IOPS
- Random Access
- Block Größen
  - Zumeist 4KB
- Große Vorteile auf Seiten der SSD
  - HDD mechanische Nachteile
  - Seek Zeiten

```
# /usr/local/bin/fio --rw=randwrite --name=ssd --bs=4k --direct=1 --filename=/dev/sdb
--runtime=300 --ioengine=libaio --iodepth=16 --write_iops_log=ssd --refill_buffers
ssd: (g=0): rw=randwrite, bs=4K-4K/4K-4K, ioengine=libaio, iodepth=16
fio-2.0.9
Starting 1 process
Jobs: 1 (f=1): [w] [100.0% done] [0K/149.8M /s] [0 /38.4K iops] [eta 00m:00s]
ssd: (groupid=0, jobs=1): err= 0: pid=4225: Thu Oct  4 15:56:34 2012
  write: io=46256MB, bw=157886KB/s, iops=39471 , runt=300001msec
[...]

Disk stats (read/write):
  sdb: ios=83/11833852, merge=0/0, ticks=8/4270928, in_queue=4267280, util=100.00%

# /usr/local/bin/fio --rw=randwrite --name=hdd --bs=4k --direct=1 --filename=/dev/sde
--runtime=300 --ioengine=libaio --iodepth=16 --write_iops_log=hdd --refill_buffers
hdd: (g=0): rw=randwrite, bs=4K-4K/4K-4K, ioengine=libaio, iodepth=16
fio-2.0.9
Starting 1 process
Jobs: 1 (f=1): [w] [100.0% done] [0K/752K /s] [0 /188  iops] [eta 00m:00s]
hdd: (groupid=0, jobs=1): err= 0: pid=4245: Thu Oct  4 16:14:08 2012
  write: io=215404KB, bw=735047 B/s, iops=179 , runt=300081msec
[...]

Disk stats (read/write):
  sde: ios=83/53816, merge=0/0, ticks=100/4794164, in_queue=4794824, util=100.00%
```


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


IOPS 3D Measurement Plot
25000
20000
15000
10000
5000
0
IOPS
512
4k
8k
16k
32k
64k
128k
1024k
Block Size (Byte)
0
5
35
50
65
95
100
R/W Mix%


Sättigung


Write Saturation Test
Avg IOPS
50000
45000
40000
35000
30000
25000
20000
IOPS
0
50
100
150
Round #

Genauigkeit ist noch lange nicht die Wahrheit.

(Henri Matisse)

6) Tipps

- Testszenarien überlegen
- Ausgangszustand festlegen
- Device direkt testen
- Aus Applikationssicht mit Page Cache
- Kompression beachten
  - Sandforce Controller!
  - Fio: refill_buffers
- Ergebnisse untersuchen, vergleichen, analysieren

6) Resümee

Thomas-Krenn.AG®
The server experts

1. Tests erfordern Engagement
2. Keine „Pauschal-Ergebnisse“
3. TK Performance-Tool kommt

- Bilder
  - 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