Benchmarks 29 May 2013 (c) napp-it.org

Hardware: SM X9 SRL-F, Xeon E5-2620 @ 2.00GHz, 65 GB RAM, 6 x IBM 1015 IT (Chenbro 50bay)
OS: napp-it appliance v. 0.9c1, OmniOS stable (May 2013)

Disks:
5 Seagate SAS ST3146855SS, 146 GB, 15k/rpm,
1 Intel 320, 300 GB SSD (MLC),
1 ATP SATA II SSD 16 GB (SLC)
1 Winkom ML-X8480, 480 GB MLC
1 ZeusRAM 8GB SAS (DRAM)

Intension of these benchmarks:
- verify some basic dependencies
- only a overview, no interest in absolute values
- quick tests with small files, larger files are more accurat but not too different

What I read from the benchmarks
Test 1: Sequential performance vs number of vdevs/disks via dd
- Sequential values scales with number of vdevs/disks (about 100-130 MB/s per disk)
- even a single disk is fast enough for 1 GB network
- a fast SSD is as good or better than 4 enterprise 15k rpm SAS disks

OPS/s (fileserver benchmark)
- OPS/s scales with number of vdevs
- a fast SSD is as good or better than 4 enterprise 15k rpm SAS disks

OPS/s (webserver benchmark)
- similar values with number of disks or SSD

Test 2: iSCSI vs SMB (sync disabled)
- iSCSI is similar to SMB regarding writes
- iSCSI is more than twice as fast compared to SMB regarding reads (needs some more tests)
- a fast SSD is as good or better than 4 enterprise 15k rpm SAS disks

Test 3: Async vs Sync Write
To check if a SSD is a good ZIL, set sync to always, create a volumebased iSCSI Target, run a Crystalmarbench and check 4k QD32 value
(With 32 concurrent small writes, this is the benchvalues that is of interest for a ZIL)
- Sync write perfomance is only 10-20% of async without dedicated ZIL !!!
- A ZIL build from a 3 years old enterprise class SLC SSD is mostly slower than without ZIL
  (this pool is build from fast disks, but a dedicated ZIL needs to be really fast or its useless)
- A Intel 320 SSD (quite often used because of the included supercap) is a quite good ZIL,
  You get up to 60% of the async values (at least with a larger 320, i used a 300 GB SSD)
- Only a DRAM based ZeusRAM is capable to deliver similar values like async write
- Some SSDs like newest SLC ones or a Intel S3700 are very good and much cheaper

Filebench: Randomwrite
Sync write values are quite bad, even with a ZeusRAM.
I suppose this is due the small 8 GB ZeusRAM (a ZIL needs to hold about 10s of writes, not ideal for a local benchmark)
but a single 8 GB ZeusRAM should be ok for a single 10 GbE link (about 1 GB/s x 10s = less than 10 GB needed Zilsize).

Test 4: Async vs Sync on a SSD only pool
- sync write performance is up to 40% of the async performance
- a slow SSD as extra ZIL, even a SLC one is a very bad idea (although may increase durability of MLC SSD's)
- Even with a SSD only pool, a ZeusRAM is a good idea. (Up to 70% or asny values and increase durability of MLC SSD's)
- ZFS seems quite well when a Pool is nearly full (at least with benchmarks from small files. Performance with large files
like ESXi VM's is a different thing from my experience, so try to stay below 70% fillrate)

The benchmarks

Test1: Use the Seagate in a Raid-0, test performance vs number of vdevs, sync: default (=disabled)
Remote tests are done from Windos via 10 GbE either via CIFS or iSCSI

Filebench, all Seagate SAS Disks in Raid-0, i do not check absolute values but differences plus dd write with 128GB, 2 MB blocks, writeonly,
NAS-Tester http://www.808.dk/?code-csharp-nas-performance. Because of the large RAM-Cache, i check mainly write values, readvalues are mostly similar without cache.

Stage 1.1: (fileserver.f, 30s), Raid-0 (one basic 15k disk disk per vdev)

Disks	OPS	OPS/s	RW		Latency		dd write	NAS tester write 400 MB (Windows SMB)
1	104987 ops	3499.449 ops/s	(318/636 r/w)	83.4 mb/s	1634us cpu/op	49.4ms latency	111 MB/s	143 MB/s
2	399095 ops	3302.761 ops/s	(1209/2419 r/w)	319.9mb/s,	428us cpu/op	13.0ms latency	229 MB/s	108 MB/s
3	233414 ops,	7779.562 ops/s	(707/1415 r/w)	185.9mb/s	1123us cpu/op	22.8ms latency	378 MB/s	117 MB/s
4	397243 ops	13238.229 ops/s	(1203/2407 r/w)	318.9mb/s,	542us cpu/op	13.1ms latency	475 MB/s	176 MB/s

Stage 1.2: (webserver.f, 30s), Raid-0 (one basic 15k disk per vdev)

Disks	OPS	OPS/s	RW		Latency	
1	13605195 ops	453490.7 ops/s	(146287/14631 r/w)	2405.3mb/s	56us cpu/op	0.2ms latency
2	13658179 ops	455255.654 ops/s	(146856/14688 r/w)	2414.6mb/s,	56us cpu/op	0.2ms latency
3	13595568 ops,	453166.862 ops/s	(146182/14620 r/w)	2404.3mb/s,	56us cpu/op,	0.3ms latency
4	13553535 ops	451769.074 ops/s	(145731/14575 r/w)	2396.3mb/s,	56us cpu/op,	0.2ms latency

Stage 2.1: Compare to a single SSD (480 GB), (fileserver.f)

Disks	OPS	OPS/s	RW		Latency		dd write	NAS tester write 400 MB (Windows SMB)
1	633773 ops,	21123.501 ops/s,	(1920/3841 r/w),	509.5mb/s,	428us cpu/op,	8.1ms latency	470 MB/s	141 MB/s

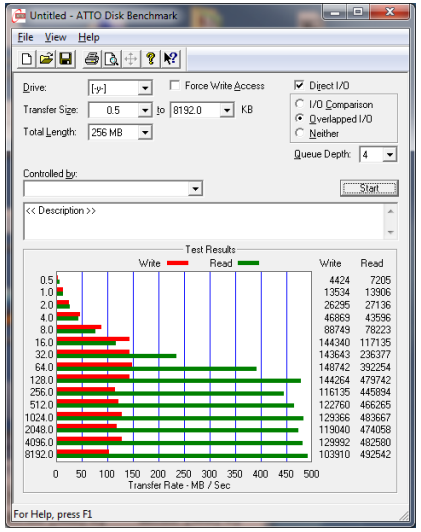
Stage 2.2: Compare to a single SSD (480 GB), (webserver.f)

1	13649111 ops,	454954.630 ops/s	(146759/14678 r/w),	2413.5mb/s,	56us cpu/op,	0.3ms latency

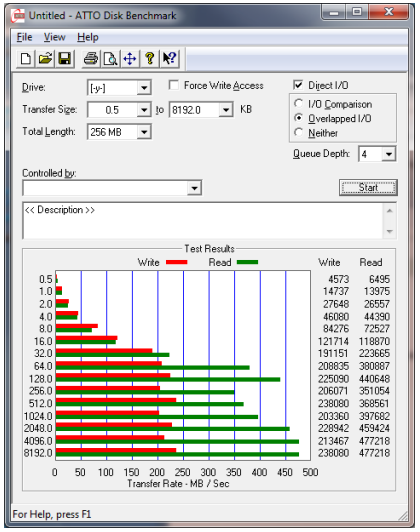
Test 2. iSCSI vs SMB, disks vs SSD, sync disabled, volume based LU

iSCSI Benchmark: Windows 7-64, 8GB RAM, 10 GbE via iSCSI Target (volumebased, 50 GB, 64k blocksize, thin prof., writeback cache enabled, NTFS formatted

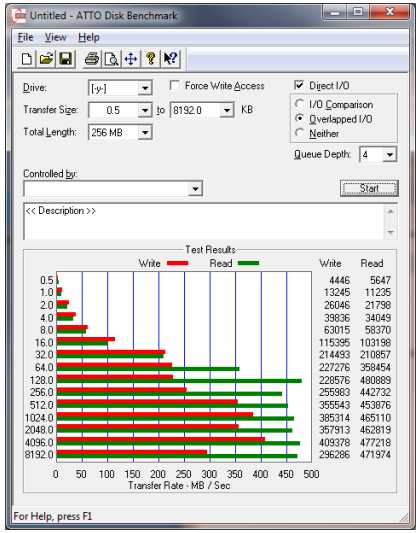
Pool from single Seagate disk via iSCSI



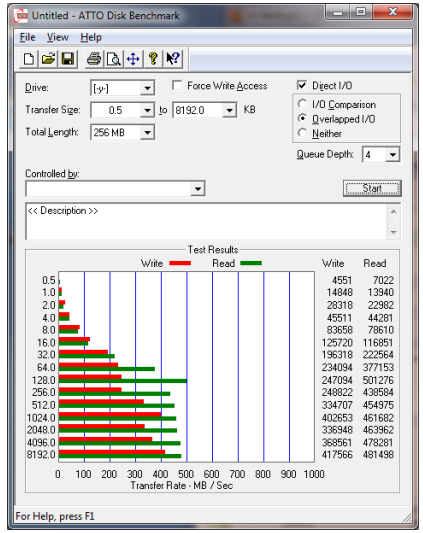
Pool from 2 disks, 2 vdev=Raid-0



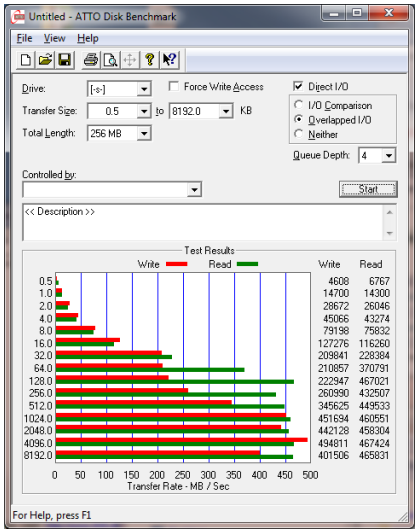
Pool from 3 disks, 3 vdevs in Raid 0



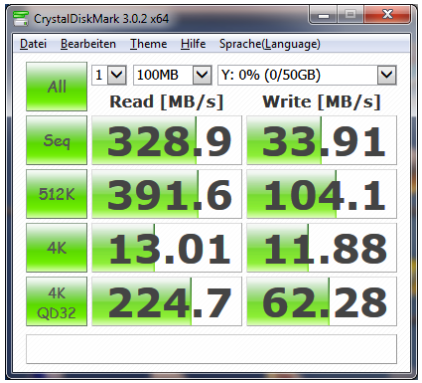
Pool from 4 disks, 4 vdevs in Raid 0

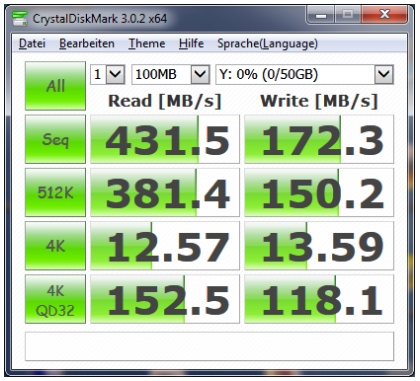


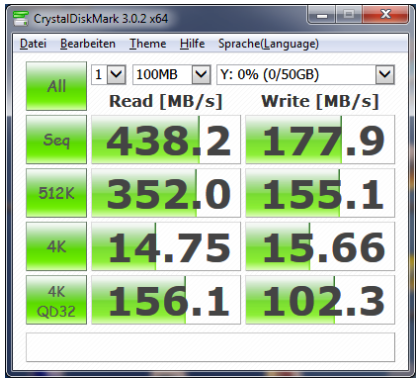
Pool from Single 480 GB SSD

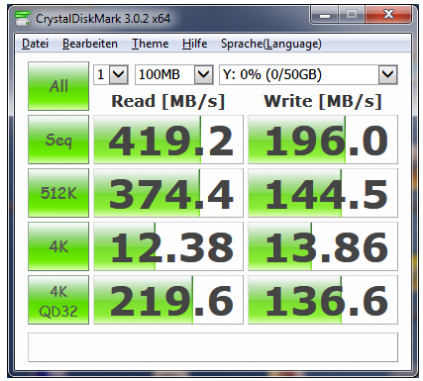


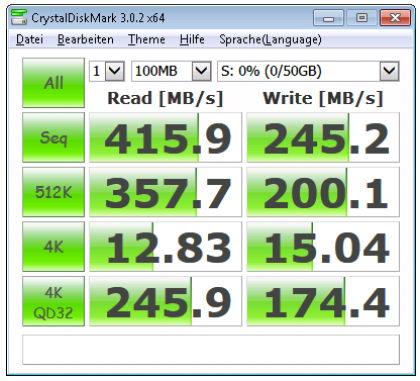
Drive Y: iSCSI 50 GB



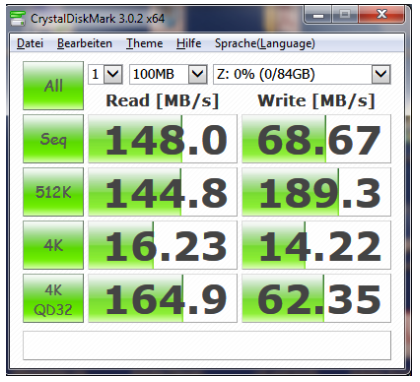


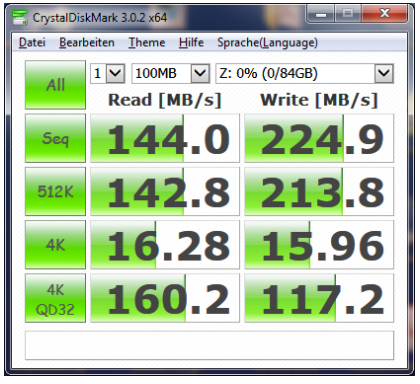


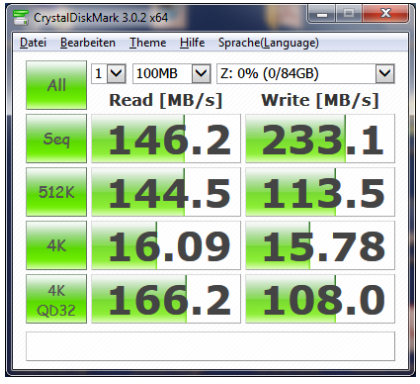


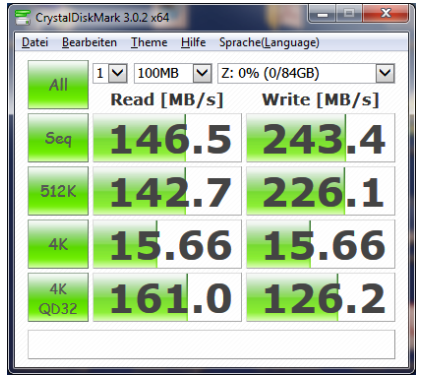


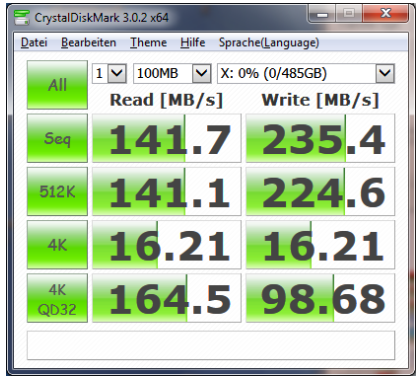
Drive Z: same Pool via SMB





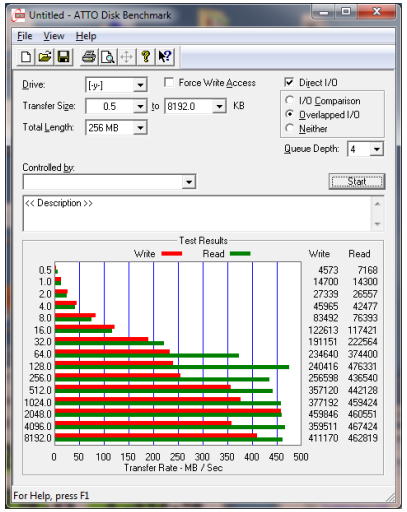




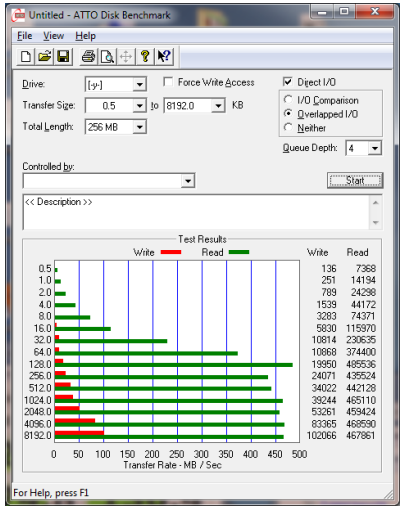


Test 3. Async vs sync write depending on ZIL, Pool build from 5 x vdevs, each from a basic Seagate 15k/m disks (Raid-0)

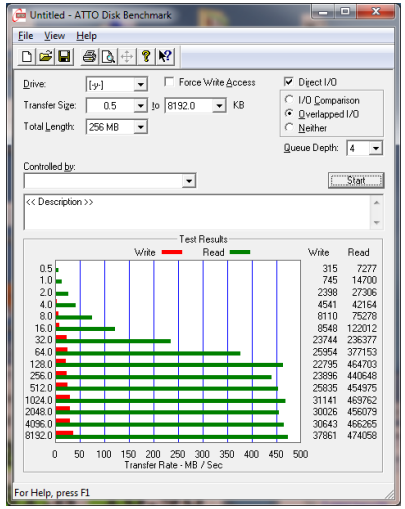
sync=disabled



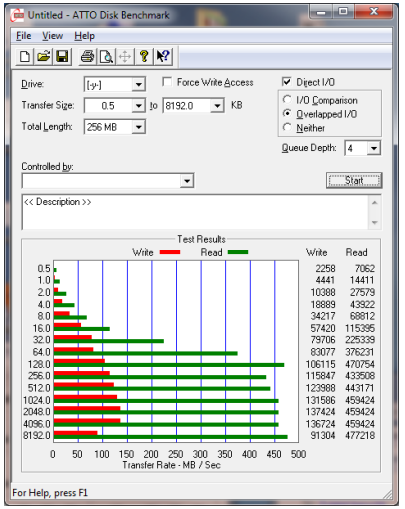
sync, no ZIL



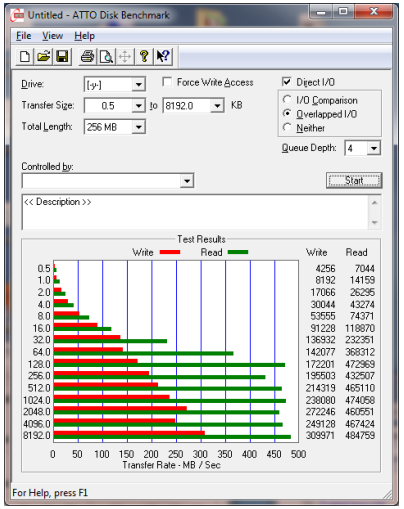
sync, Adata 16GB SLC



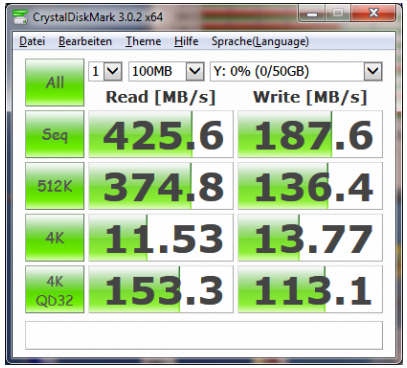
sync, Intel 320-300GB MLC

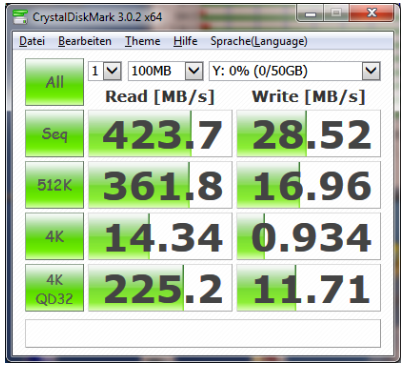


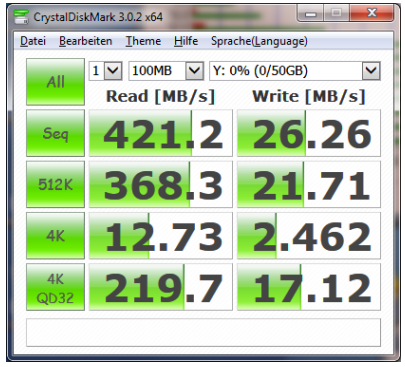
sync, ZeusRAM, DRAM 8 GB

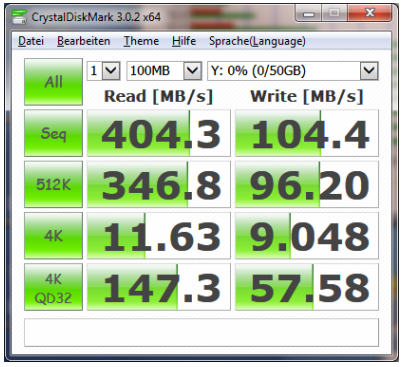


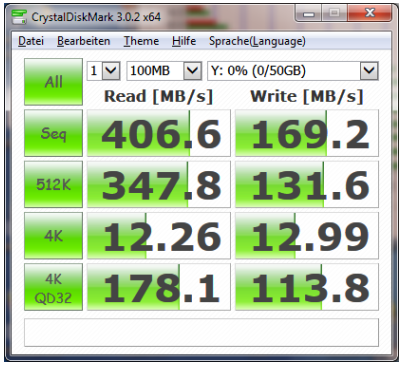
Drive Y: iSCSI 50 GB



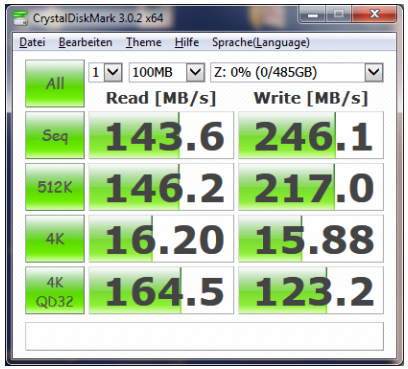


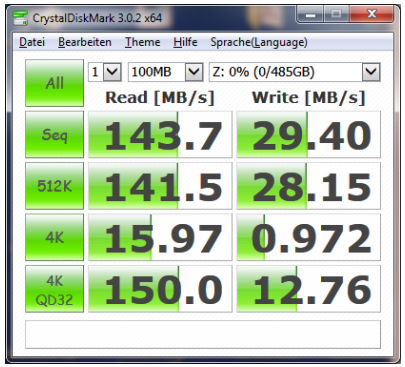


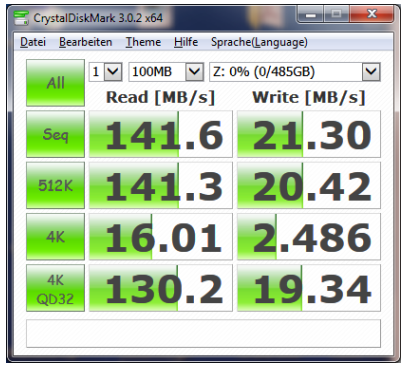


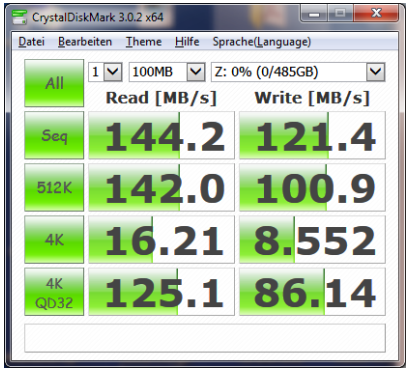


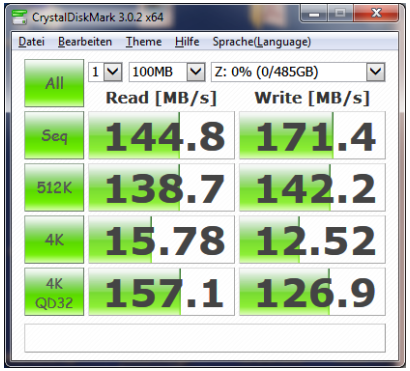
Drive Z: same Pool via SMB











Filebench randomwrite.f, 30s
44393.296 ops/s, 346.8mb/s,

Filebench randomwrite.f 30s
8808.833 ops/s, 68.8mb/s

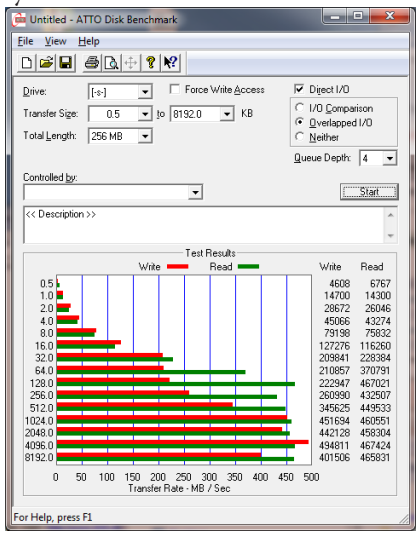
Filebench randomwrite.f 30s
12240.467 ops/s, 95.6mb/s

Filebench randomwrite.f 30s
2283.002 ops/s, 17.8mb/s

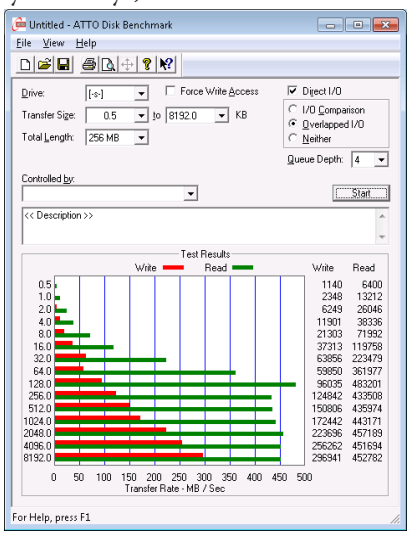
Filebench randomwrite.f 30s
4068.654 ops/s, 31.8mb/s

Test 4. Async vs sync write depending on ZIL on a SSD Pool, Pool build from 1 x vdev from a basic Winkom SSD 480 GB, important is the 4k QD32 value

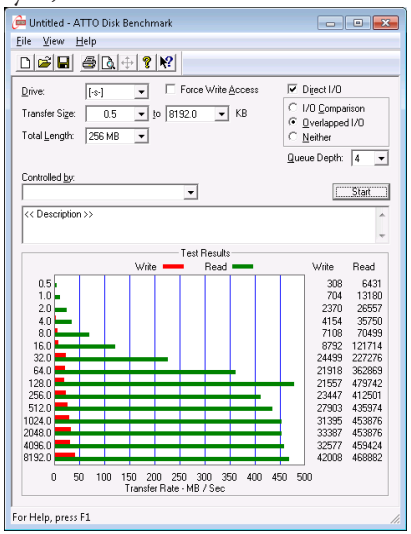
sync=disabled



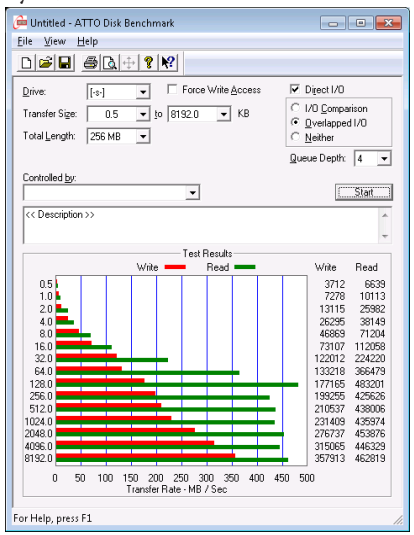
sync=always, no ZIL



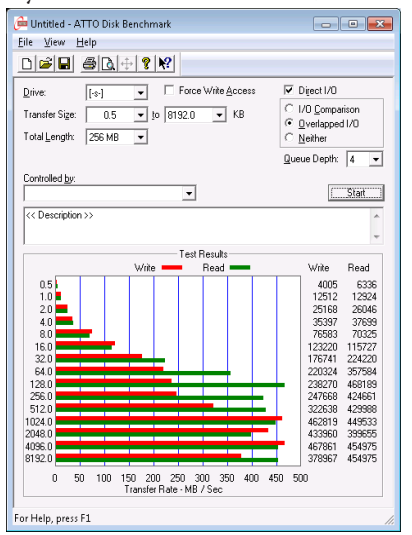
sync, ATP SSD 16 GB SLC ZIL



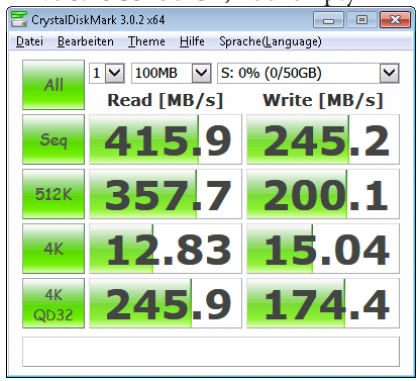
sync, ZeusRAM Dram ZIL

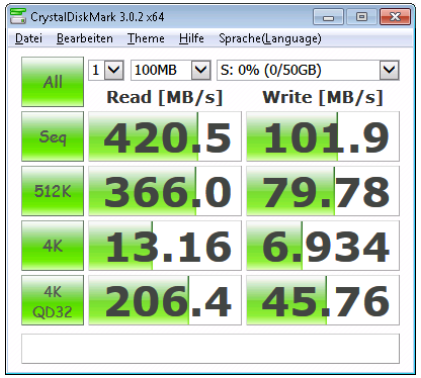


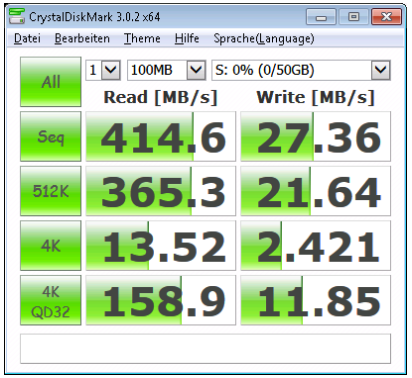
sync=disabled, Pool 95%full, iSCSI

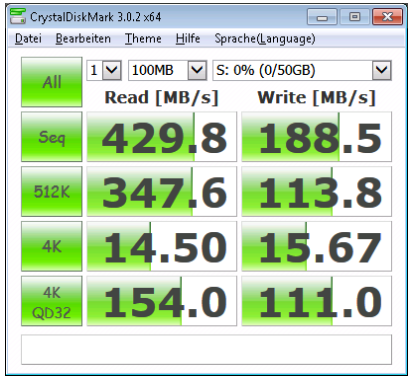


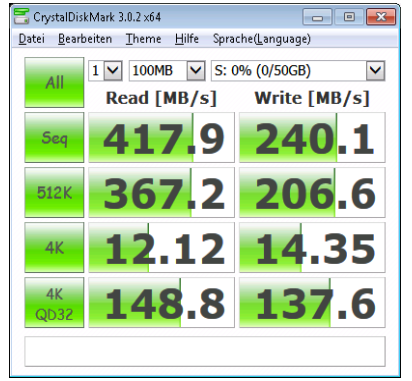
Drive S: iSCSI 50 GB, Pool empty



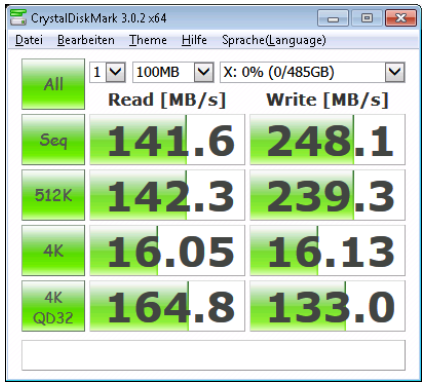


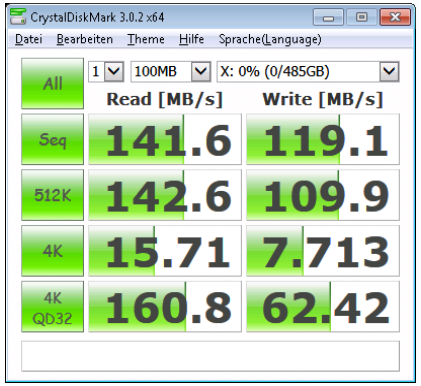


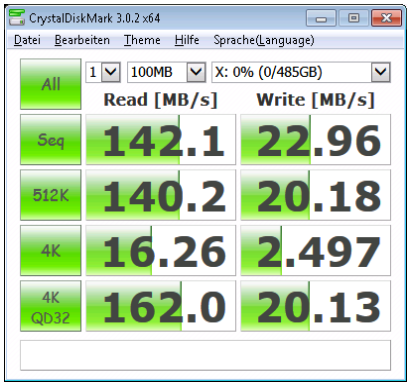


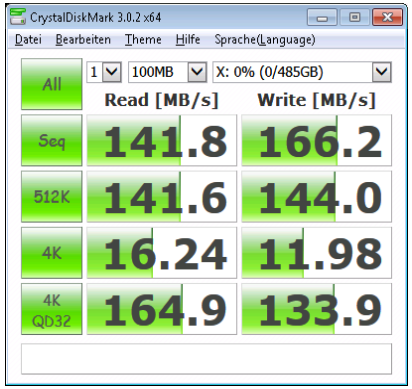


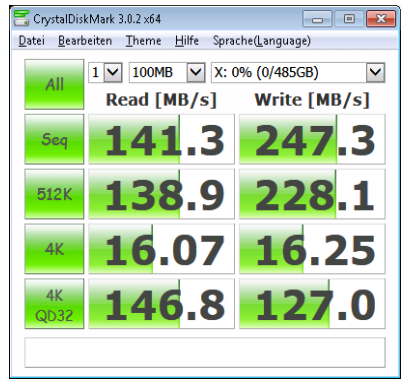
Drive X: same Pool via SMB





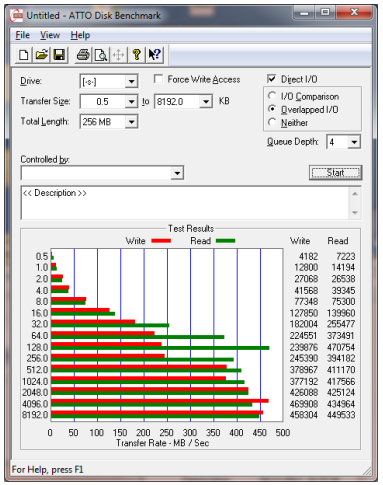


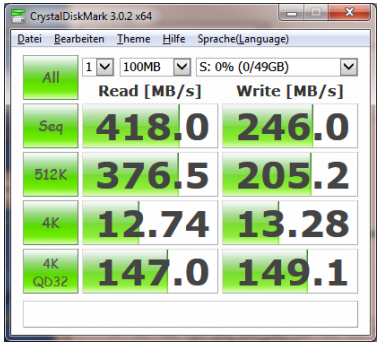


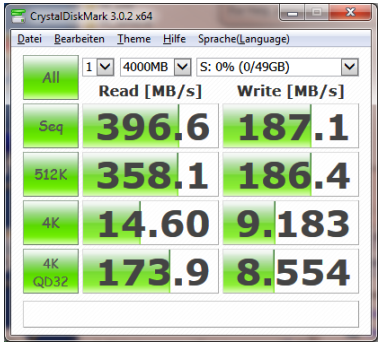


Test 5: special configurations

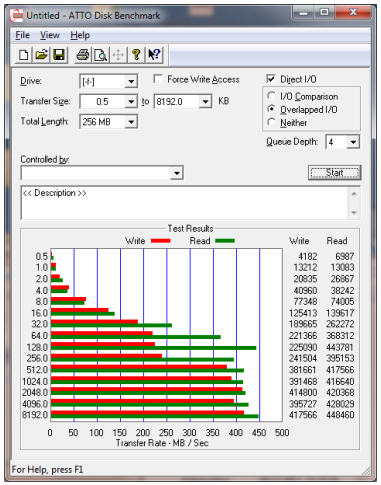
sync=off, iSCSI, volume LU, SSD

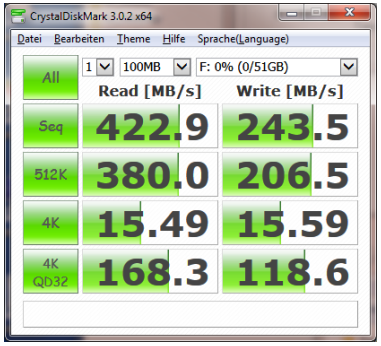


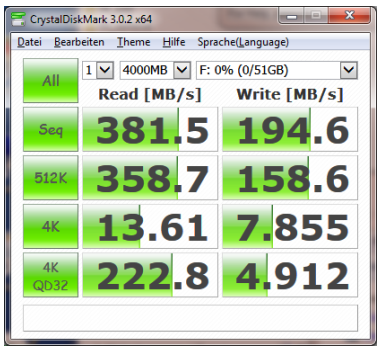




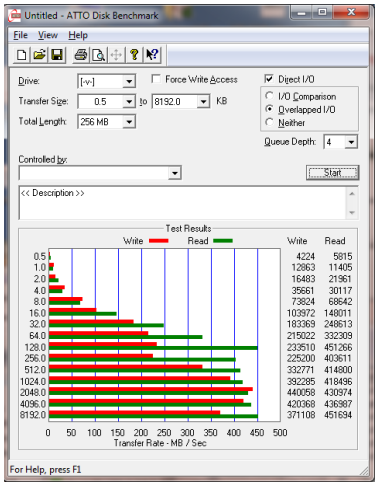
sync=off, iSCSI, file LU, SSD

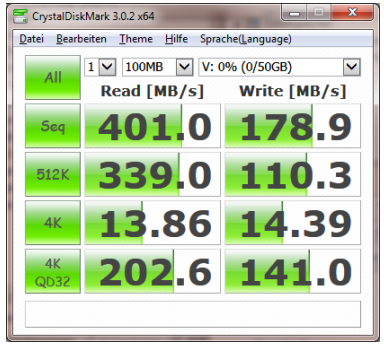






4 x vdevs, each from a basic disk



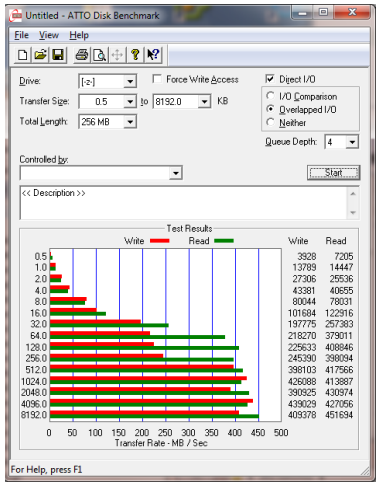


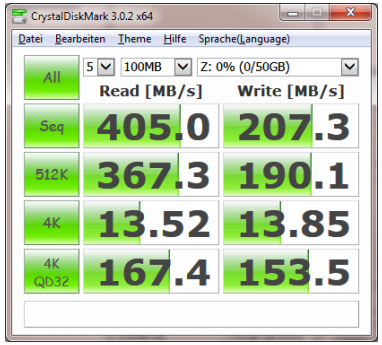
Filebench fileserver.f
13594.182 ops/s, (1236/2472 r/w),
327.4mb/s, 393us cpu/op, 12.8ms latency

Filebench randomrw.f
88637.352 ops/s, (86004/2634 r/w),
692.5mb/s, 13us cpu/op, 0.0ms latency

Filebench webserver.f
458002.397 ops/s, (147742/14777 r/w),
2430.2mb/s, 55us cpu/op, 0.3ms latency

1 x vdev Z1 from 4 datadisks (4+1)



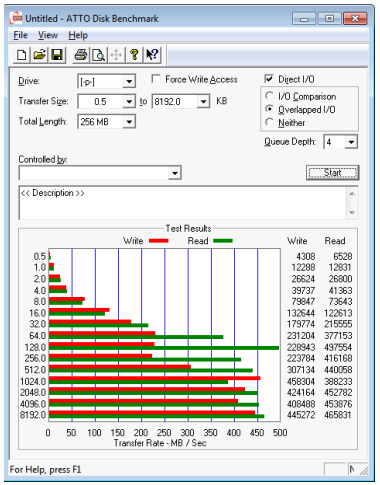


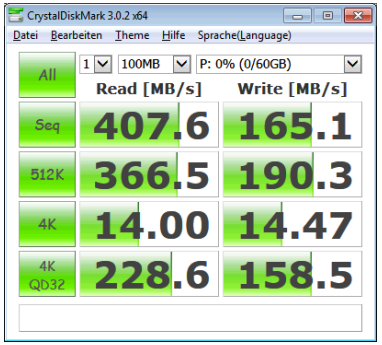
Filebench fileserver.f
9352.514 ops/s, (850/1701 r/w), 224.4mb/s,
474us cpu/op, 18.9ms latency

Filebench randomrw.f
86419.294 ops/s, (83691/2728 r/w),
675.1mb/s, 17us cpu/op, 0.0ms latency

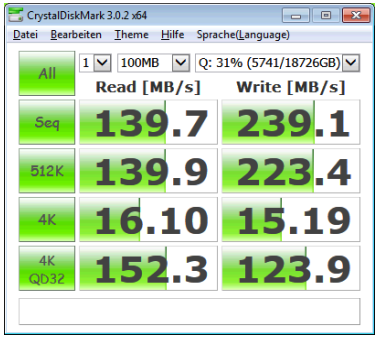
Filebench webserver.f
456351.152 ops/s, (147209/14723 r/w),
2420.4mb/s, 55us cpu/op, 0.3ms latency

4 x Z2, each 7 disks RE4 5400rpm





iSCSI



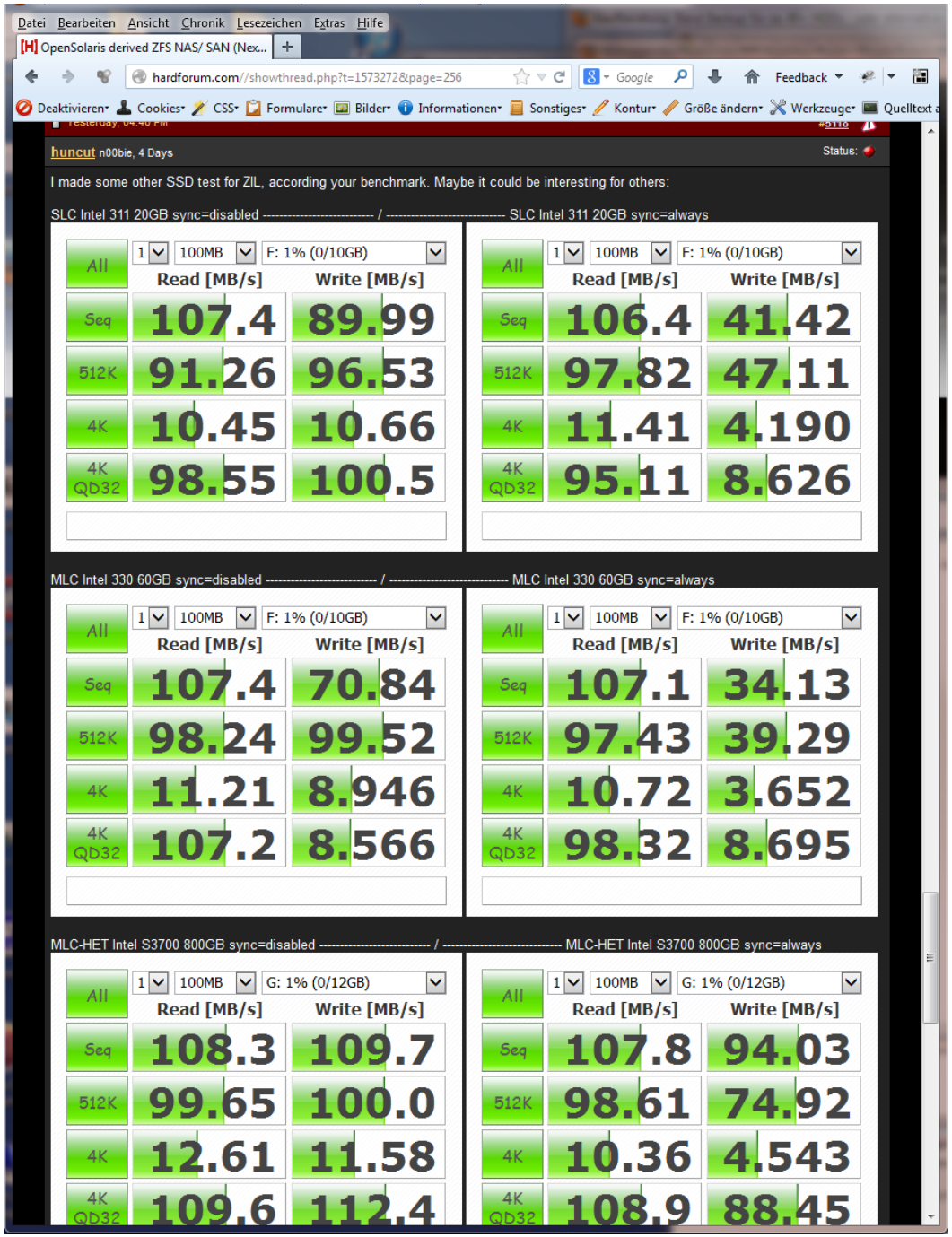
SMB

Question: Volume or Filebased Logical Units?
Volumbased LUs are minimal faster, but not as easy to handle compared to filebased LUs regarding copy/move/backup/restore from snap.

More vdevs or Raid-Z? (same amount of datadisks/poolsize)
If you look at sequential performance, they are similar, Z1 even slightly faster. If you look at the fileserver-filebench, the multi-vdev option is up to 50% faster on latency, r/w and cpu/op than the Raid-Z1.

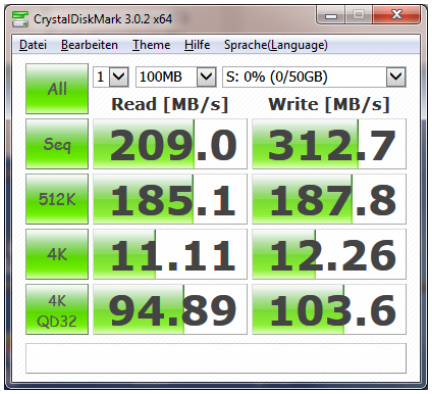
Backup pool (green WD disks RE4)
dd: 1800 MB/s write, 4000 MB/s read
fileserver.f
29950.846 ops/s, (2723/5446 r/w),
726.0mb/s, 604us cpu/op, 5.1ms lat

More Benchmarks (sync vs async Performance - Is this a good Zil?
Look mostly at 4k QD32 with sync=always and 32 concurrent small 4k writes

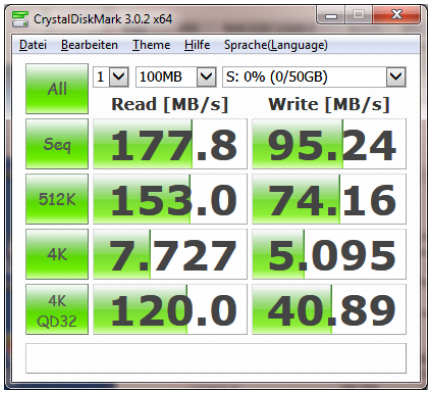


1GB network,

Winkom SSD 120 GB (SF1222, Intel SLC Nand, high IOPS)

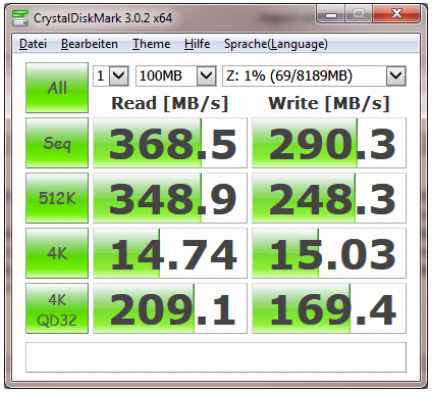


10 Gbe iSCSI, sync=disabled

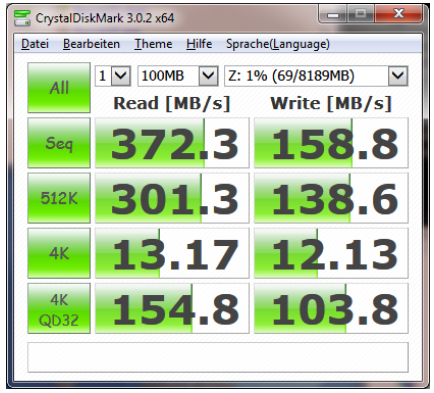


10 GbE iSCSI, sync=always

ZeusRAM (8 GB DRAM based)



10 Gbe iSCSI, sync=disabled



10 GbE iSCSI, sync=always, best of all 4k QD32