Benchmarks October 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 accurate 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 volume-based iSCSI Target, run a Crystalmarbench and check 4k values

- Sync write performance 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 to 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 async 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

Test 1: 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)

<table>
<thead>
<tr>
<th>Disks</th>
<th>OPS</th>
<th>OPS/s</th>
<th>RW</th>
<th>Latency</th>
<th>dd write</th>
<th>NAS tester write 400 MB (Windows SMB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>104987 ops</td>
<td>3499.449 ops/s</td>
<td>(318/636 r/w)</td>
<td>83.4 mb/s</td>
<td>1634us cpu/op</td>
<td>49.4ms latency</td>
</tr>
<tr>
<td>2</td>
<td>399095 ops</td>
<td>3302.761 ops/s</td>
<td>(1209/2419 r/w)</td>
<td>319.9mb/s</td>
<td>428us cpu/op</td>
<td>13.0ms latency</td>
</tr>
<tr>
<td>3</td>
<td>233414 ops</td>
<td>7779.562 ops/s</td>
<td>(707/1415 r/w)</td>
<td>185.9mb/s</td>
<td>1123us cpu/op</td>
<td>22.8ms latency</td>
</tr>
<tr>
<td>4</td>
<td>397243 ops</td>
<td>13238.229 ops/s</td>
<td>(1203/2407 r/w)</td>
<td>318.9mb/s</td>
<td>542us cpu/op</td>
<td>13.1ms latency</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Disks</th>
<th>OPS</th>
<th>OPS/s</th>
<th>RW</th>
<th>Latency</th>
<th>dd write</th>
<th>NAS tester write 400 MB (Windows SMB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13605195 ops</td>
<td>453490.7 ops/s</td>
<td>(146287/14631 r/w)</td>
<td>2405.3mb/s</td>
<td>56us cpu/op</td>
<td>0.2ms latency</td>
</tr>
<tr>
<td>2</td>
<td>13658179 ops</td>
<td>455255.654 ops/s</td>
<td>(146856/14688 r/w)</td>
<td>2414.6mb/s,</td>
<td>56us cpu/op</td>
<td>0.2ms latency</td>
</tr>
<tr>
<td>3</td>
<td>13595568 ops</td>
<td>453166.862 ops/s</td>
<td>(146182/14620 r/w)</td>
<td>2404.3mb/s,</td>
<td>56us cpu/op,</td>
<td>0.3ms latency</td>
</tr>
<tr>
<td>4</td>
<td>13553335 ops</td>
<td>451769.074 ops/s</td>
<td>(145731/14575 r/w)</td>
<td>2396.3mb/s,</td>
<td>56us cpu/op,</td>
<td>0.2ms latency</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Disks</th>
<th>OPS</th>
<th>OPS/s</th>
<th>RW</th>
<th>Latency</th>
<th>dd write</th>
<th>NAS tester write 400 MB (Windows SMB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>633773 ops</td>
<td>21123.501 ops/s</td>
<td>(1920/3841 r/w),</td>
<td>509.5mb/s,</td>
<td>428us cpu/op,</td>
<td>8.1ms latency</td>
</tr>
<tr>
<td>2</td>
<td>13649111 ops</td>
<td>454954.630 ops/s</td>
<td>(146759/14678 r/w),</td>
<td>2413.5mb/s,</td>
<td>56us cpu/op,</td>
<td>0.3ms latency</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Disks</th>
<th>OPS</th>
<th>OPS/s</th>
<th>RW</th>
<th>Latency</th>
<th>dd write</th>
<th>NAS tester write 400 MB (Windows SMB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13649111 ops</td>
<td>454954.630 ops/s</td>
<td>(146759/14678 r/w),</td>
<td>2413.5mb/s,</td>
<td>56us cpu/op,</td>
<td>0.3ms latency</td>
</tr>
</tbody>
</table>
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 (volume based, 50 GB, 64k blocksize, thin prof., writeback cache enabled, NTFS formatted)

- Pool from single Seagate disk via iSCSI
- Pool from 2 disks, 2 vdevs in 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)

<table>
<thead>
<tr>
<th>Drive</th>
<th>Description</th>
<th>Filebench randomwrite.f 30s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drive Y: iSCSI 50 GB</td>
<td>sync, no ZIL</td>
<td>44393.296 ops/s, 346.8mb/s</td>
</tr>
<tr>
<td>Drive Z: same Pool via SMB</td>
<td>sync, Adata 16GB SLC</td>
<td>8808.833 ops/s, 68.8mb/s</td>
</tr>
<tr>
<td></td>
<td>sync, Intel 320-300GB MLC</td>
<td>12240.467 ops/s, 95.6mb/s</td>
</tr>
<tr>
<td></td>
<td>sync, ZeusRAM, DRAM 8 GB</td>
<td>2283.002 ops/s, 17.8mb/s</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4068.654 ops/s, 31.8mb/s</td>
</tr>
</tbody>
</table>
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 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
1 x vdev Z1 from 4 datadisks (4+1)
4 x Z2, each 7 disks RE4 5400rpm

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?
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), 173.9 8.554
396.6 187.1
358.1 186.4
14.60 9.183
13.61 7.855
222.8 4.912

Filebench fileserver.f
13594.182 ops/s, (1236/2472 r/w), 327.4mb/s, 1470 149.1
376.5 205.2
12.74 13.28
147.0 149.1

Filebench randomrw.f
88657.352 ops/s, (86004/2634 r/w), 692.5mb/s, 13us cpu/op, 0.0ms latency
380.0 206.5
15.49 15.59
168.3 118.6

Filebench webserver.f
458002.397 ops/s, (147742/14777 r/w), 2430.2mb/s, 55us cpu/op, 0.3ms latency
422.9 243.5
339.0 110.3
14.38 14.39
202.6 141.0

Filebench fileserver.f
9352.514 ops/s, (850/1701 r/w), 224.4mb/s, 474us cpu/op, 18.9ms latency
401.0 178.9
339.0 110.3
13.86 14.39
202.6 141.0

Filebench randomrw.f
86419.294 ops/s, (83691/2728 r/w), 675.1mb/s, 17us cpu/op, 0.0ms latency
405.0 207.3
367.3 190.1
13.52 13.85
167.4 153.5

Filebench webserver.f
456351.152 ops/s, (147209/14723 r/w), 2420.4mb/s, 55us cpu/op, 0.3ms latency
407.6 165.1
366.5 190.3
14.00 14.47
128.6 158.5

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 Benchmarks (sync vs async Performance - Is this a good Zil?
Look mostly at 4k with sync=always

<table>
<thead>
<tr>
<th></th>
<th>Read [MB/s]</th>
<th>Write [MB/s]</th>
<th>Seq</th>
<th>512K</th>
<th>4K</th>
<th>4K QD32</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC Intel 311 20GB sync=disabled</td>
<td>107.4</td>
<td>89.99</td>
<td>91.26</td>
<td>96.53</td>
<td>10.45</td>
<td>98.55</td>
</tr>
<tr>
<td>SLC Intel 311 20GB sync=always</td>
<td>106.4</td>
<td>41.42</td>
<td>97.82</td>
<td>47.11</td>
<td>11.41</td>
<td>95.11</td>
</tr>
<tr>
<td>MLC Intel 320 80GB sync=disabled</td>
<td>107.4</td>
<td>70.84</td>
<td>98.24</td>
<td>99.52</td>
<td>11.21</td>
<td>98.32</td>
</tr>
<tr>
<td>MLC Intel 320 80GB sync=always</td>
<td>107.1</td>
<td>34.13</td>
<td>97.43</td>
<td>39.29</td>
<td>10.72</td>
<td>98.32</td>
</tr>
<tr>
<td>MLC HET Intel S770 100GB</td>
<td>108.3</td>
<td>109.7</td>
<td>99.65</td>
<td>100.0</td>
<td>12.61</td>
<td>109.6</td>
</tr>
<tr>
<td>MLC HET Intel S770 100GB</td>
<td>107.8</td>
<td>94.03</td>
<td>98.61</td>
<td>74.92</td>
<td>10.36</td>
<td>108.9</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Read [MB/s]</th>
<th>Write [MB/s]</th>
<th>Seq</th>
<th>512K</th>
<th>4K</th>
<th>4K QD32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>209.0</td>
<td>312.7</td>
<td>185.1</td>
<td>187.8</td>
<td>11.11</td>
<td>12.26</td>
</tr>
<tr>
<td></td>
<td>94.89</td>
<td>103.6</td>
<td>153.0</td>
<td>74.16</td>
<td>7.727</td>
<td>5.095</td>
</tr>
<tr>
<td></td>
<td>120.0</td>
<td>40.89</td>
<td>301.3</td>
<td>138.6</td>
<td>13.17</td>
<td>12.13</td>
</tr>
</tbody>
</table>

ZeusRAM (8 GB DRAM based)

<table>
<thead>
<tr>
<th></th>
<th>Read [MB/s]</th>
<th>Write [MB/s]</th>
<th>Seq</th>
<th>512K</th>
<th>4K</th>
<th>4K QD32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>368.5</td>
<td>290.3</td>
<td>348.9</td>
<td>248.3</td>
<td>14.74</td>
<td>15.03</td>
</tr>
<tr>
<td></td>
<td>209.1</td>
<td>169.4</td>
<td>154.8</td>
<td>103.8</td>
<td>15.12</td>
<td>12.13</td>
</tr>
</tbody>
</table>

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

1 GB network,
More Benchmarks (SSD only pools), 15 X Sandisk Extreme2-480 GB, Benchmarks done via volume-based iSCSI via 10 GbE

- sync disabled
- one vdev Raid-Z2 (15 SSD)
- sync=always, no ZIL
- sync=always, 120 GB WinKom SLC SSD (ZIL)
- sync=always, ZeusRAM (8GB DRAM ZIL)

Result for SSD only pool: No need for mirrors, Raid-Z vdevs are ok, a dedicated very fast ZIL is recommended.
some user benchmarks

Intel S3700-100 GB (the cheapest 3700), with a comparison sync vs nonsync on FreeNAS and OmniOS, see

FreeNAS 9.1 sync=disabled FreeNAS 9.1 sync=always OmniOS, sync=disabled OmniOS, sync=always

especially with small writes on iSCSI, OmniOS and Comstar seems dramatically better