A Comparison of ZFS and ESS Managed Large SSD Arrays

October 27, 2013 – Doug Dumitru – CTO, EasyCo LLC

This paper discusses the performance, functionality, and durability differences between using ZFS and ESS to manage a large array of SSDs storage disks.

Benchmark Hardware:

Single socket Xeon E-1650 CPU (3.33 GHz six core)Supermicro X9SRL-F Motherboard
64 GB of DDR3-1666 registered ECC RAM
3 LSI 8-port SAS controllers (2 x 9207 + 1 x 9211)
16 x Samsung 830 128GB SSD

ZFS Configuration:

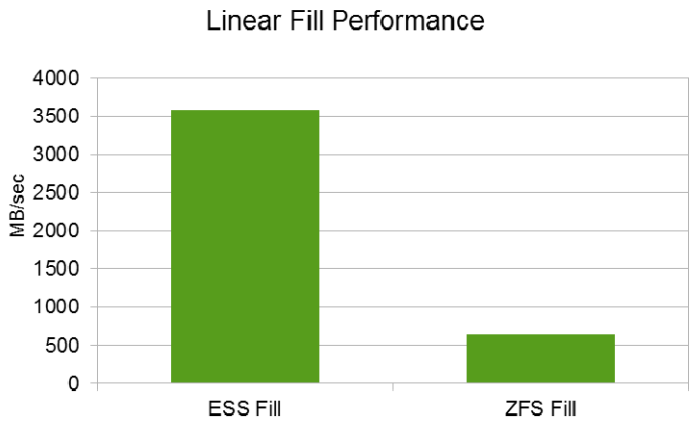
ZFS on Linux v 0.6.2 (module) on Ubuntu 12.04
24 drives raid-0
300GB "zvol"

ESS Configuration:

ESS v 4 w/ de-dupe, LBA in VM, PBA in dedicated RAM
Linux 'md' raid-0 24-drive array
Linux 'lvm' 300GB logical volume

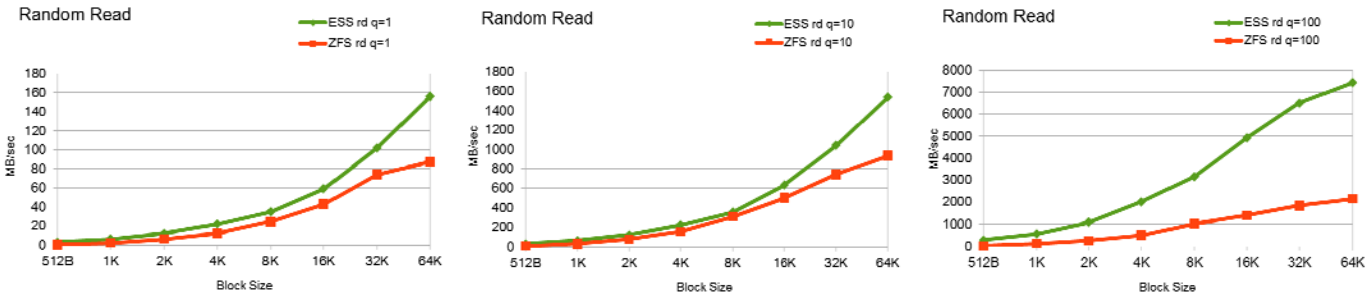
Results:

Linear Fill Performance



As you can see, the ESS volume is much faster.

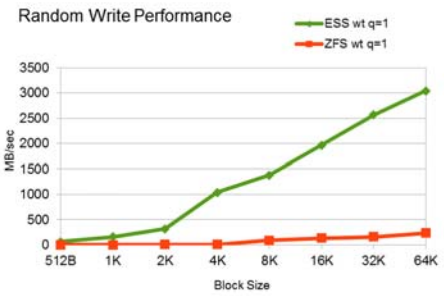
Random Read Performance

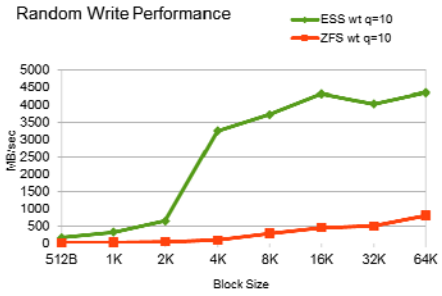


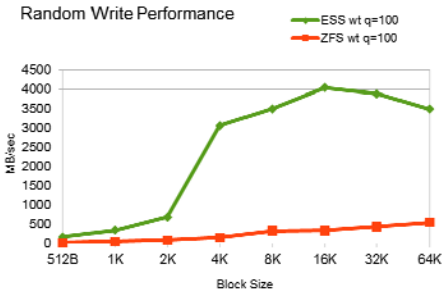
a

At low queue depths, the ZFS volume is only slightly slower than the ESS volume. At higher queue depths, ESS pulls ahead.

Random Write Performance:







Again, for random writes, ESS is much faster than ZFS, often more than an order of magnitude.

Analyzing ZFS Flash Wear:

This is a much harder task, as actually measuring wear at the flash chip level is difficult. ESS has proven wear amplification reduction in the field in the range of 1.3:1 to 1.5:1. For ZFS, we ran random write tests at q=10 and looked at the underlying IO to the individual SSDs.

	4K Random Writes				64K Random Writes			
Raid Level	Usable BW	Disk Reads	Disk Writes	Wear Amp	Usable BW	Disk Reads	Disk Writes	Wear Amp
0	69	187	365	5.3:1	602	40	723	1.2:1
Z1	67	191	768	11.5:1	661	46	1627	2.5:1
Z2	65	188	1058	16.3:1	594	42	2195	3.7:1
Z3	55	161	1268	23.0:1	523	41	2574	4.9:1

The “Disk Reads” and “Disk Writes” numbers are derived by adding up the bandwidths reported by ‘iostat’ over 30 seconds while the workload was active. These numbers can be hard to repeat, as the “state” of the volume impacts performance. These tests were run after a fill test followed by 512 byte to 64K byte random writes at q=1, q=10, and q=100 (the same test that produced the read/write graphs above).

Some of these numbers are very concerning. Realize that these are writes to the drives at the drive interface level. Any wear amplification caused by the drive FTL multiplies this numbers, so effective wear amplifications are easily above 40X for some configurations.

ZFS is also reported to be sensitive to “full volumes”. SSDs and ESS are also subject to this behavior, althouth the behavior with ESS (and even SSDs) is much easier to understand. Regardless, with a full volumes these wear amplification numbers could really skyrocket.

Conclusion:

Read Performance:

ZFS has a significant performance overhead compared to bare SSDs or SSDs managed through ESS. This is probably a byproduct of the SHA-256 error checking algorithms. This test system runs SHA-256 at 211MB/sec (according to openssl speed …) on a single core

Write Performance:

ZFS falls even further behind the linear write engine of ESS, and in many cases is well behind the bare array.

ZFS SSD Wear:

This is probably a “show stopper” for many applications. With a > 20X wear amplification before the SSDs FTL is taken into account, enterprise media is a must, and may not be enough. ESS with consumer and even “three bit per cell” media will outlast ZFS with enterprise media for some workloads. Raid-z# just does not seem to be designed for flash.

Workload Limits:

One common workload for ESS is a centralized host for virtual servers, including VDI. These tests were all run with ESS “de-duplication” active, although the test data contained 100% unique blocks so no actual de-duplication was occurring. ZFS did not have either de-dupe or compression turned on as this would have impacted performance to an even greater extent.

Many SSD vendors actually target 100+ IOPS per VDI client. This insured no boot storms and “SSD laptop level” performance for most installations. ZFS could not crack 40K 4K random writes implying an upper limit of 400 VDI clients. ESS was well over 500K implying 5000 VDI clients on the same SSD array.

Limits of this Analysis

This analysis was exclusively run with zvol’s on ZFS on Linux. ZFS from Oracle (Sun) might be a different animal. Also, local file-system performance might easily mask some of the zvol limitations.