Enterprise Storage Stack for CMR and SMR HDDs

Yes, hard disks can write quickly too

Optimizing Hard Disks

ESS for HDDs is a customized version of ESS designed specifically to optimize the behavior of hard disks. ESS-HDD is fully usable on CMR hard disks, but additionally follows all of the rules for HA-SMR and HM-SMR shingled hard disks. When ESS-HDD is in use CMR and SMR disks exhibit identical performance.

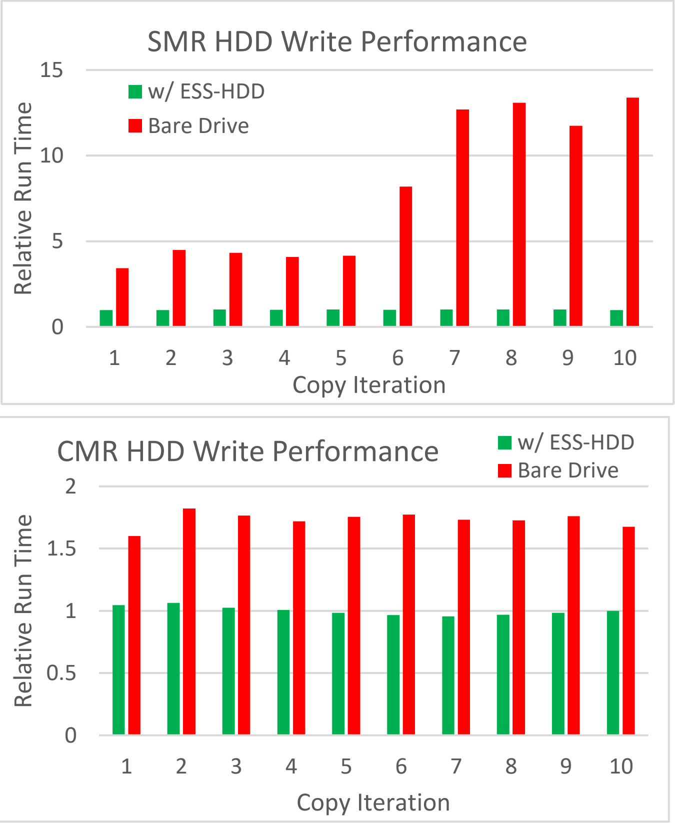
The Challenges of HDDs

The size of the block translation map:

The first challenge for a block translation layer for hard disks is actually economic. When you are working with Flash drives and arrays, the cost differential between the Flash media and DRAM is comparatively low (currently about 40:1). When you look at this with HDDs, it is much larger (in the neighborhood of 300:1). Because DRAM is so much more expensive in "capacity" terms, a simple, single layer map is not practical. ESS-HDD does a two-layer map with the second layer either on Flash or on a small segment of the HDD. This "map layer" device is also ESS mapped and it's blocks use application specific compression. The result is that the map layer has very little impact on read latency and almost no impact on write latency.

Random reads are much more expensive

Hard disks have significant differences compared to SSDs. Not only is the latency a lot higher, the available queue depth is much lower. In some ways, an SSD is an array of smaller parts. To mitigate this, ESS-HDD uses an aggressive, learning algorithm to force read aheads on both data blocks and map blocks.



Benchmark run-time. Smaller numbers are faster.

Data compression is worth the overhead

With very fast Flash arrays, the data speeds are so high that block-level compression is not always practical. With hard disks, the data rates are lower and the benefits of compression are greater. This is true for both data blocks and map blocks. The compression algorithm used for data blocks is fungible. We test with a variant of LZ4. The compression for map blocks is an algorithm that is specific to the map encoding and often exceeds 95% compression ratios.

Mount times are a big issue

With ESS, the mount operations needs to get summary data from the end of each zone. With hard disks, this is impractical. To mitigate this, the summary data is stored near the front of the drive in the CMR region of an SMR disk. This region is tightly packed so that mounts can do linear reads to slurp the entire drive status at linear read speeds.

Finally, our mapping layer allows mounts to be incremental only requiring a scan of the last few zones. This keeps mounts at around 5 seconds, even for the largest drives.

ESS-HDD on a single HA-SMR Hard Disk

We have run performance tests on an 8TB Seagate 5900 RPM SATA drive with an HA-SMR controller. When used with stock Linux file systems, data fill rates are 4X to 12X faster than using the drive's built in controller. Read performance scanning the recorded file system are also faster by about 2X. Most notably, ESS-HDD has consistent write performance which the SMR controller would frequently need to stop to do large amounts of garbage collection.

When tested on a 7200 RPM CMR disk, ESS-HDD improved writes by about 1.7X and reads by about 40%.

The test data set consisted of making copies of a Linux root file system. About 50K files averaging 45K bytes. Most files are not compressible. Both EXT4 and XFS were tested.

ESS-HDD and Arrays

HDDs are often deployed in very large arrays. With ESS-HDD mapping blocks in front of the array, arrays can scale perfectly. Because of the extra "zone operations" required for SMR disks, the array must be customized for SMR drives, but the performance advantages for ESS remain.

Arrays with 4+ parity drives and 200+ drives total are supported with linear scaling as drives are added. Drives can be added and removed while online. With some limitations, drive sizes can be mixed while still using all of a drive's capacity and performance. If adequate interface bandwidth is available, write bandwidths well above 10 GB/sec are achievable.

This compares with conventional arrays where only mirrors scale with drive count. A conventional 16 drive array might be configured RAID-60 as two RAID-6 stripes. This configuration would be expected to write at about 200 MB/sec. The ESS-HDD array would be a simple parity +2 array with all drives active and reach nearly 2 GB/sec.

ESS-HDD arrays are also dynamic in their handling of drive failures. If a drive fails, it immediately is no longer a part of the current "write set", so new data is never degraded. Existing data parity is recovered with extreme levels of parallelism, 100% linear IO, and only recovering the actual active dataset shortening recovery times.

ESS-HDD in the Controller

The ESS-HDD logic is currently implemented in software and talks to a "drive" via the standard disk interface. The design and resources used by ESS-HDD are such that a controller implementation is not only possible, but practical with current controller limitations. For example, a 20+ TB SMR disk can be completely mapped in under 512 MB of DRAM. It would be up to a drive OEM to decide whether to include a small Flash module which would further accelerate performance. The nature "map volume" updates is such that Flash endurance is not an issue.

WildFire Storage

https://wildfire-storage.com	+1 (610) 237-2000
sales@wildfire-storage.com	+1 (888) 473-7866