Enterprise Storage Stack
Concepts for Storage OEMs

When ESS does not add value:

ESS does not always make sense. If your storage usage is already:

- Using the lowest cost media.
- Using the lowest cost "redundancy" (ie, parity RAID instead of mirrors).
- Does not have compressible data.
- … and performance is not an issue.
- … and media wear is not an issue.

So, there are use cases where ESS does not gain you anything. Now that that is out of the way …

ESS for Block Storage

ESS can enhance block storage in many scenarios. The main feature of ESS is that writes are linearized. Block storage is nearly always more efficient and effective with linear write IO than with random write IO. In particular:

- Linear writes are faster at the device.
  - HDDs are much faster with linear IO
  - SSDs are faster with linear IO
- Linear writes wear media less:
  - Write amplification on SSDs is reduced (and usually reaches 1:1)
- Linear writes eliminate performance overhead associated with parity RAID
  - read/modify/write operations can be avoided completely as whole stripes are updated as single operations.
- Linear write restricted media is now an option:
  - ZNS SSDs
  - HM-SMR and HA-SMR HDDs
- Linear writes are move efficient, everywhere on the stack:
  - Bus overhead is lower.
  - Fabric / network overhead is lower.
- ESS can optionally compress blocks:
  - If your data rates and data content can benefit from this.

The benefits of linear writes cannot be over-stated. Some items like migrating away from traditional raid can results in a massive bandwidth, IOPS, and system overhead benefits. For example, 1M 4K random writes into a 14+2 array (16 drive RAID-6) imposed obnoxious overhead. Remember every 4K random write will require 13 random reads, and then 3 random writes. SSDs are "OK" with random reads. HDDs despise random reads. This is the "layout" of RAID. Better raid "parity comp" like NVMe off-loads don't help. The IO operations are still there.

ESS linearizes the writes, so the overall system overhead drops dramatically.

	ESS	Traditional Array	ESS Benefit
Member Drive read OPs	None	13 M	Infinite
Member Drive read GBs	None	52 GB	Infinite
Member Drive write OPs	75 K	3 M	40 X
Member Drive write GBs	4.5 GB	12 GB	2.6 X
Total system OPs (interrupts)	75 K	16 M	213 X
Total system IO GBs	4.5 GB	64 GB	14 x
Drive Wear (after write amp)	~ 10 GB	~ 60 GB	6 X

Some vendors advertise "all flash" random write performance of several hundred thousand IOPS. Some of the new "all NVMe" solutions with low drive count can get above 1 million IOPS. ESS has been over 2.5 million IOPS since the days of SATA SSDs. Remember that with ESS it is raw drive bandwidth and raw data bandwidth and not "randomness". The randomness of the write stream is eliminated.

ESS Block Deployment

ESS as a block device can be deployed on:

- SSDs
  - Conventional SSDs
    - Read-intensive SSDs are preferred as they have identical performance and wear compared to write intensive models
    - SSDs with “emulated block sizes” are 100% acceptable:
      - for example, some QLC drives have terrible 4K random write amp because their FTL maps 16K blocks. ESS Linear writes will do high performance 4K random writes to these drive without penalty.
  - Zoned SSDs
- HDDs
  - Conventional CMR HDDs
  - Host Managed and Host Aware SMR HDDs
    - SMR and CMR HDDs perform identically on both reads and writes.
- Single Drives
- Parity Protected Arrays
  - Unlike conventional arrays, as the drive count grows, the array gets faster. The upper limit is > 64 drives, depending on the system architecture.

ESS’s block device is a stock block device and can be used anywhere a /dev/sda or /dev/md0 style device can be used:

- Partition with FDISK
- Apportion with LVM
- Host a file system
- Host a direct block application
- Export over a fabric
  - iSCSI.
  - NVMe-OF.
  - Fiber Channel
  - InfiniBand SRP

ESS for Object Storage

A new file system (working name WFFS), optimized for object storage application, and based on ESS, has been developed. This file system is not "production ready" and WildFire Storage is looking for OEM to partner with to bring this solution to market.

Object Storage Workloads:

File systems can be presented with several styles of workloads.

- Object Storage:
  This is where files are created and accessed all at once, as opposed to incrementally.
- Append Storage:
  Mostly log files that are slowly built over time.
- Database Storage:
  Files that are written to and read 'in place".

WFFS is not intended for use with logs and databases. This does not preclude logs and database using other file systems on the same host, perhaps on top of an ESS block solution.

How Does WFFS Differ from Traditional File System

WFFS is built on top of an estended version of ESS. This allows for file system structures that look more like a database that a file system.

- All writes are linear (this is still ESS)
- All allocations are variable sized.
  - Every file system element uses only the number of bytes needed. There is no padding to 4K or some other allocation unit size.
  - Large files are stored in 16 MB "extents":
    - This eliminates fragmentation.
    - This keeps the overhead of extent tables low.
- All directory lookups are "single IO"
  - Regardless of directory file count
- All updates are inherently atomic:
  - This eliminates the overhead of journals, etc.

WFFS is currently envisioned to operate on a "single disk". This matches the operating mode of many object server applications (like MinIO). WFFS on an integrated array is "under design review". We are convinced that it will not only work but run "like a bat out of hell", but array features like on-line expansion are still being tweaked.

With a single drive, the current limits are:

- up to 512 TB
- > 1B files per directory
- > 16 TB file size limit

WFFS can be deployed on:

- A single HDD
    - Either CMR or SMR
- A single HDD paired with 3% to 10% of SSD
- A single SSD
    - Either conventional or ZNS

The Hybrid environment is the ideal target for traditional “mass object storage” applications, including public hosting. This gives you the cost structure of SMR HDDs with the performance of SSDs for small objects. In general, performance can be expected to be > 10X for small objects and > 3X for large objects.

The pure SSD environment is ideal for either very small objects or cases where extreme bandwidth is needed.

Working with WildFire

WildFire works with OEMs to customize our solutions for your system design, expectation, and needs. We don't want to re-invent what you have already done.

We also understand that you need "your solution" to be "yours".

- You can private label:
  - ESS and WildFire do not need to be disclosed.
- You can get ESS source for business continuity.
- You can do as much or as little in-house support as you need.

EasyCo LLC dba WildFire-Storage

Doug Dumitru – CTO
doug@wildfire-storage.com
+1 (610) 237-2000 x43
+1 (888) 473-7866 x43
+1 (949) 291-0184 (cell)
https://wildfire-storage.com