HPSS Community Tech-Topic	
Topic: ZFS ZVOL Settings on Commodity HDD Hardware For Use By HPSS Disk Movers	
Contributing Site: LLNL Livermore Computing HPC Archive	**Date:** November, 2021

Background: LLNL has deployed disk movers with ZFS-managed JBODs (just a bunch of disks) in lieu of disk array controllers as both a cost savings measure and an acknowledgment as to how far ZFS has come. The deployment class is called Adaptable Storage Platform (ASP) and is intended to serve a multitude of use cases in the HPC center. The ASP Disk Building Block (DBB) configuration utilizes two 4U drawers of 90 drives each frontended by a pair of hosts in 2U that have access to all 180 drives for a footprint of 10U. Figure 1 shows 4 DBBs in a single 42U rack for a total rack scalable unit size of 11.5PB raw. ZFS presents dRAID-protected ZVOL block devices to HPSS for use as HPSS disk devices. The following assessment illustrates how LLNL tuned for maximum ZFS ZVOL device performance on a single host of a DBB (i.e., half a DBB worth of hardware).

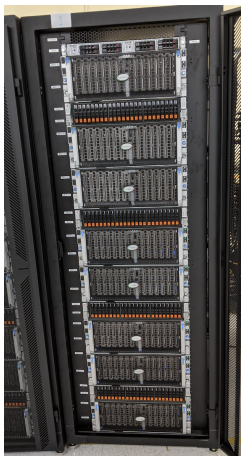

Figure 1

ZFS dRAID was utilized for all testing. For more information on dRAID, see here[1] and here[2]. All testing done against a ZFS pool with a configuration of 2 parity, 8 data, 2 hot-spares, and 90 children. In ZFS vdev nomenclature, this would be a `draid2:8d:2s:90c` zpool.

When deploying new hardware, it can be difficult to decide how to first approach disk tuning given multiple "knobs" in the full software stack. To achieve optimal performance for an application like HPSS, first test and tune outside of the application on the specific operating system which will be used for production deployments.

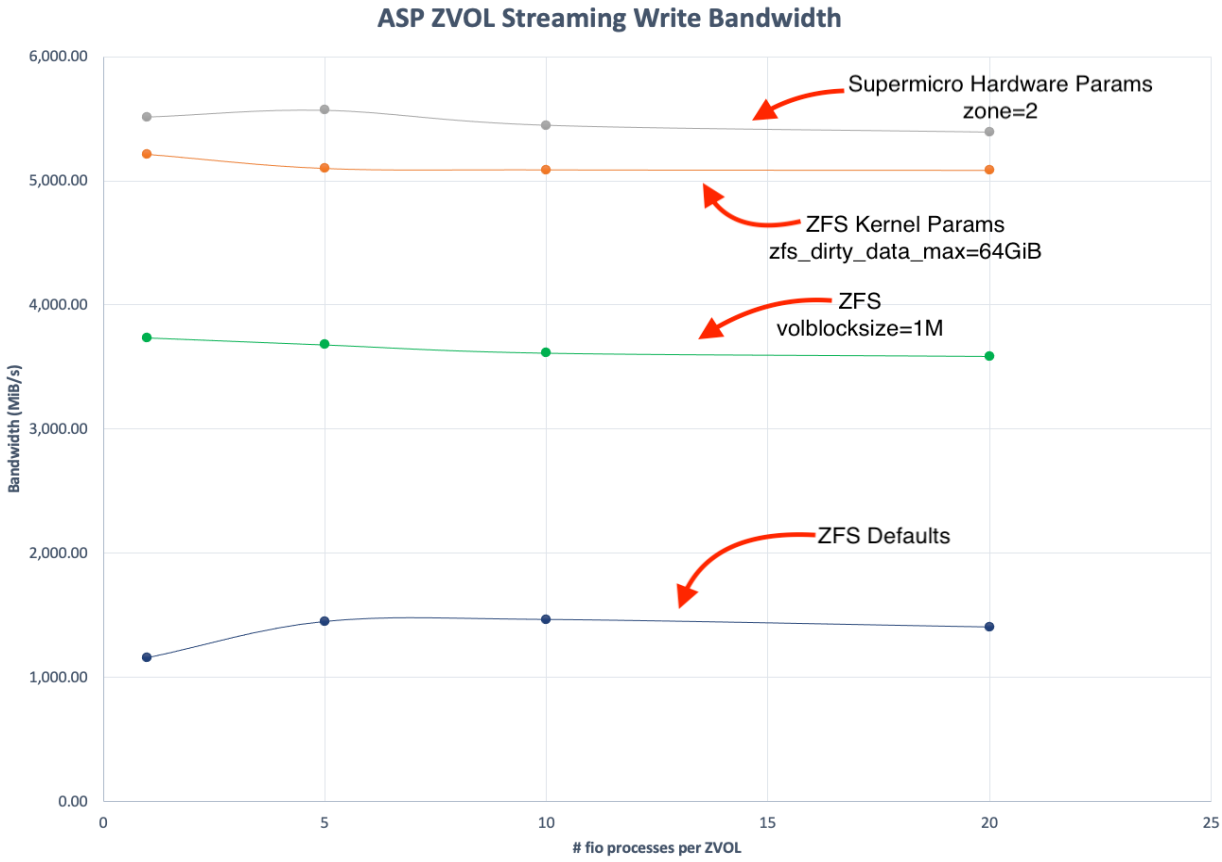

Figure 2

Release Number LLNL-PRES-841477

Figure 2 illustrates `fio` streaming write performance directly against the ZFS ZVOL.

Using the "out of the box" ZFS ZVOL default settings, streaming write bandwidth is poor. Modifying one parameter at a time shows sequential performance increases up to around 5.5 GB/s.

Note when running tests with Supermicro *zone* set to 2, the intent was that all 90 disks in the zpool would be located in a single enclosure. However, due to an error in the /etc/zfs/vdev_id.conf file, this might not have been the case during actual testing. Thus, the pool may have spanned enclosures.

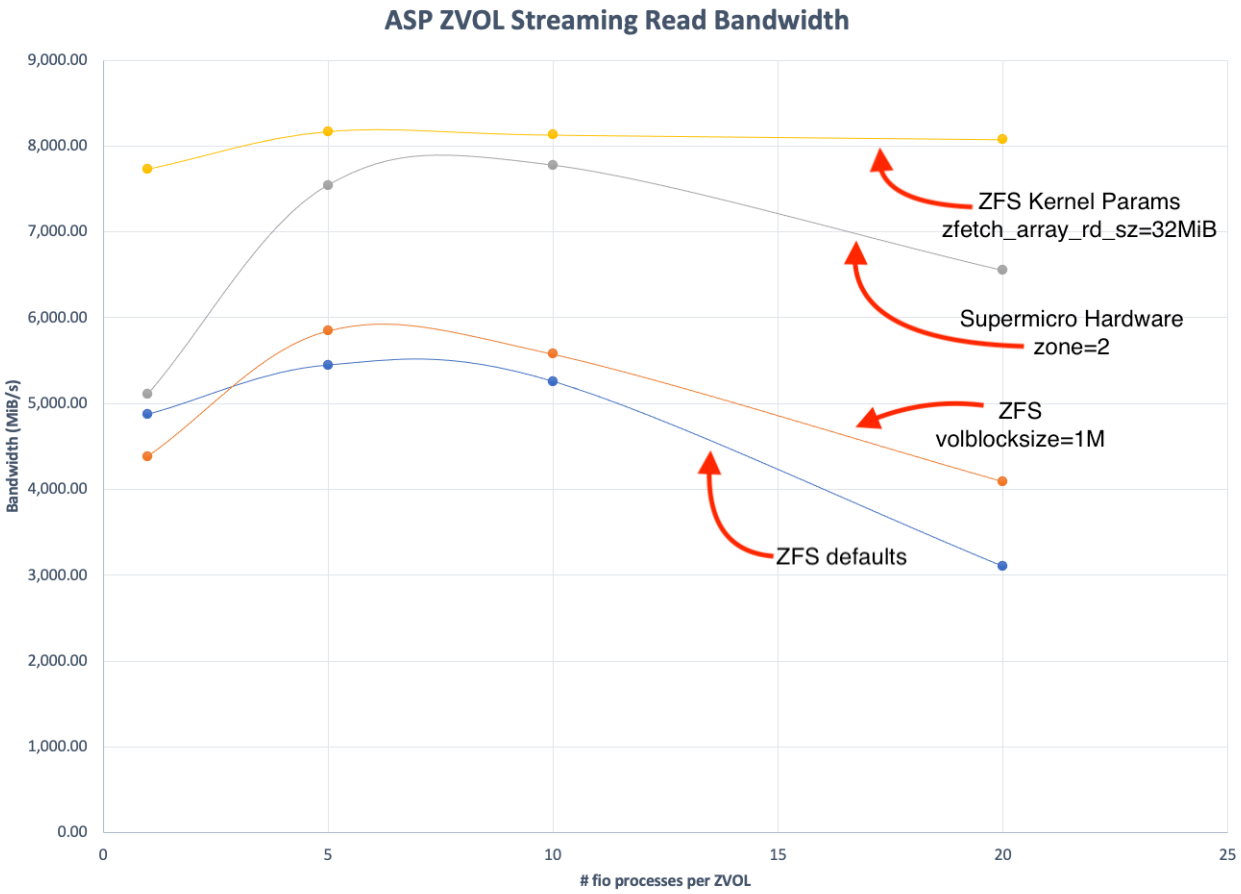

Figure 3

Figure 3 shows `fio` streaming read performance directly against the ZFS ZVOL, again without HPSS involved.

It's important to properly tune ZFS readahead kernel module parameters for the application. The parameters in question are *zfetch_array_rd_sz, zfetch_max_distance,* and *zfetch_max_*streams.

Ultimately, LLNL's ongoing tests led us to deploy to production with these settings:

```
options zfs zfetch_array_rd_sz=67108864
options zfs zfetch_max_distance=67108864
options zfs zfetch_max_streams=500
```

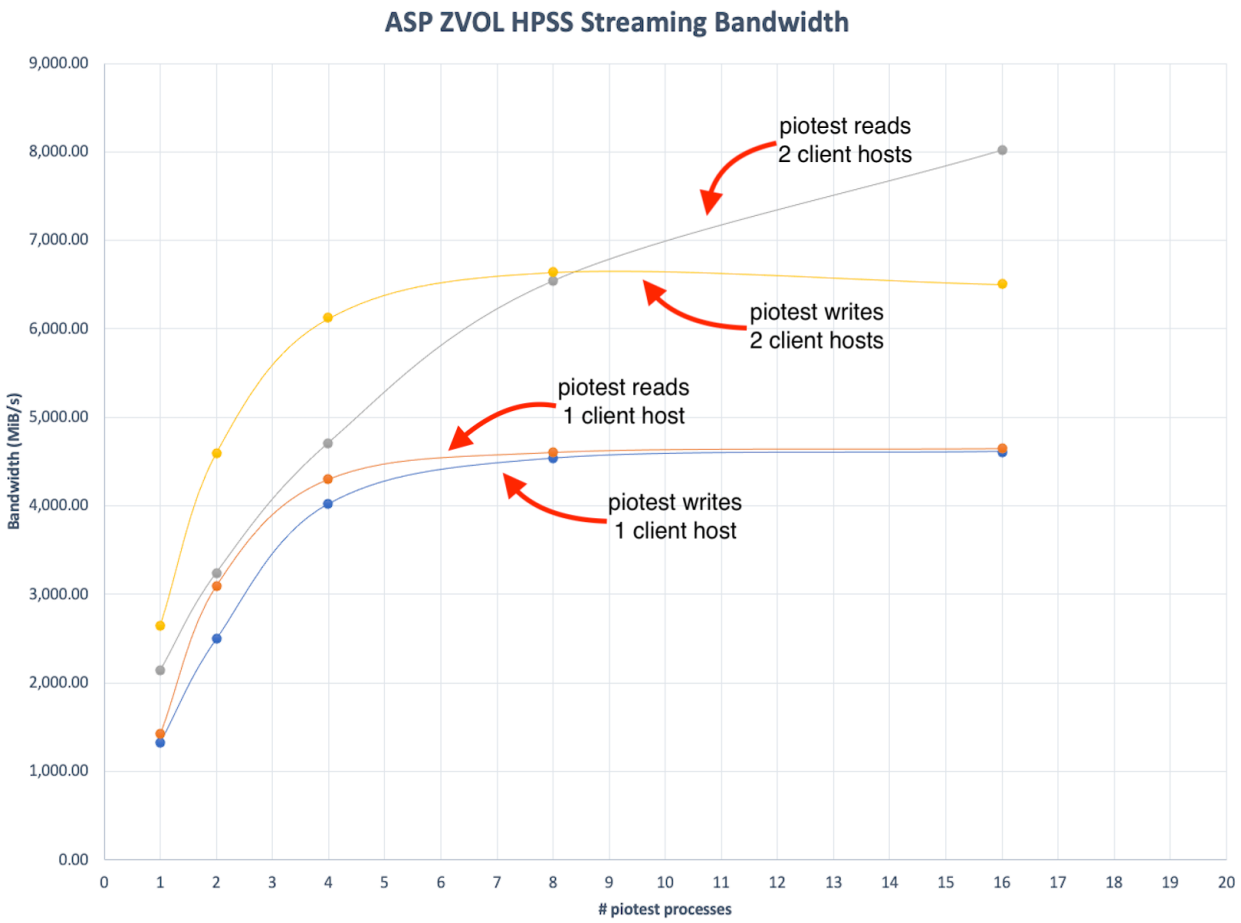

Figure 4

Streaming write and read tests were conducted on HPSS disk devices built on ZFS ZVOLs with the settings from above. Note that test results from the `piotest` HPSS performance benchmark utility match the maximum performance from the `fio` tests against the ZFS ZVOLs.

Figure 4 shows `piotest` read/write streaming performance results. A single client host comes close to saturating its 40GbE link. And the addition of multiple client hosts shows nearly the same performance with HPSS as demonstrated with `fio`. Single I/O stream (process) performance needs improvement. HPSS CR617 is intended to address this issue.

[1] https://docs.google.com/presentation/d/1uo0nBfY84HIhEqGWEx-Tbm8fPbJKtIP3ICo4toOPcJo/edit#slide=id.p1
[2] https://openzfs.github.io/openzfs-docs/Basic%20Concepts/dRAID%20Howto.html

Summary:
We have shown how spending some time tuning before application integration can increase HPSS disk mover performance. With ZFS ZVOLs we have shown that using the default tuning parameters results in poor performance. Large gains in performance can be had by setting ZFS volume blocksize, maximizing the hardware drive pool layout, setting ZFS readahead options for improved read performance, and increasing ZFS's ability to absorb more workload variation before throttling to improve write performance.

For more info: email Herb Wartens (wartens2@llnl.gov) or Todd Heer (heer2@llnl.gov)

Release Number LLNL-PRES-841477

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Release Number LLNL-PRES-841477