Project Report (June, 2017)

-- George (Zhi Qiao)

In June, I have worked on the ZFS recovery experiment, and the HPDC 2017 poster and presentation. For ZFS recovery, I have complete experiment on HDD and SSD for RAID 5 and RAID 6 setting, each with different percentage of disk utilization rate. Below will discuss the methodology and experiment result.

Methodology

1. Fill the pool with random number

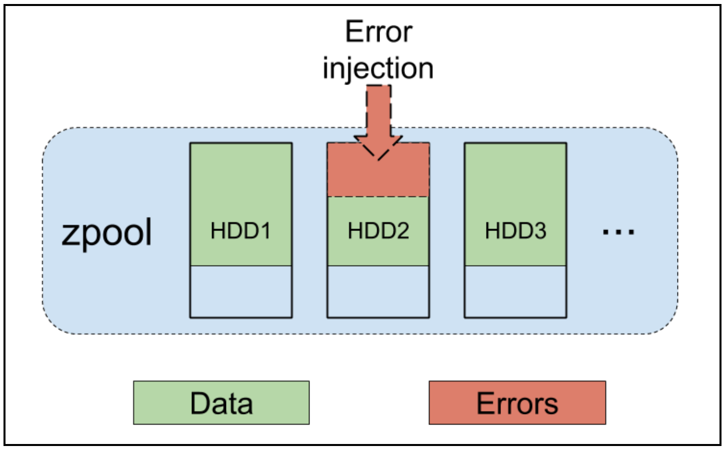
First, we fill the zpool with random values to certain percentage, such as 25%, 50%, and 75% of the usage. Random data is generated via openssl PRNG engine. We use “pv” utility to fill the disk and inject errors as it's faster than “dd” command.

2. Inject errors to drive directly

We directly inject zeros to disk using “*pv* < */dev/zero* > *dev/sdx*” to flood certain disk.

3. Notify ZFS to recover/resilver

We signal the ZFS to check data corruption using “zpool scrub”. After ZFS scan the drive, it will start to resilver the lost data. If metadata also corrupted, we need to replace the drive. In this case, “zpool replace -f pool sdx” will reformat the disk and replace the “faulty” drive.

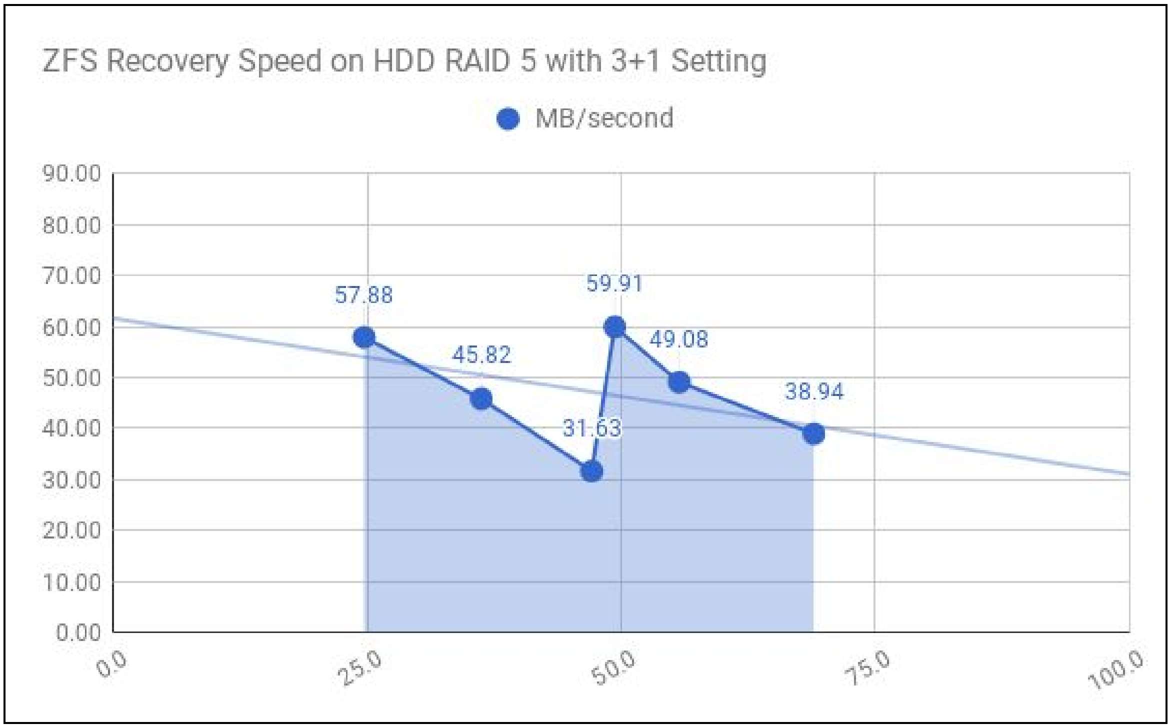


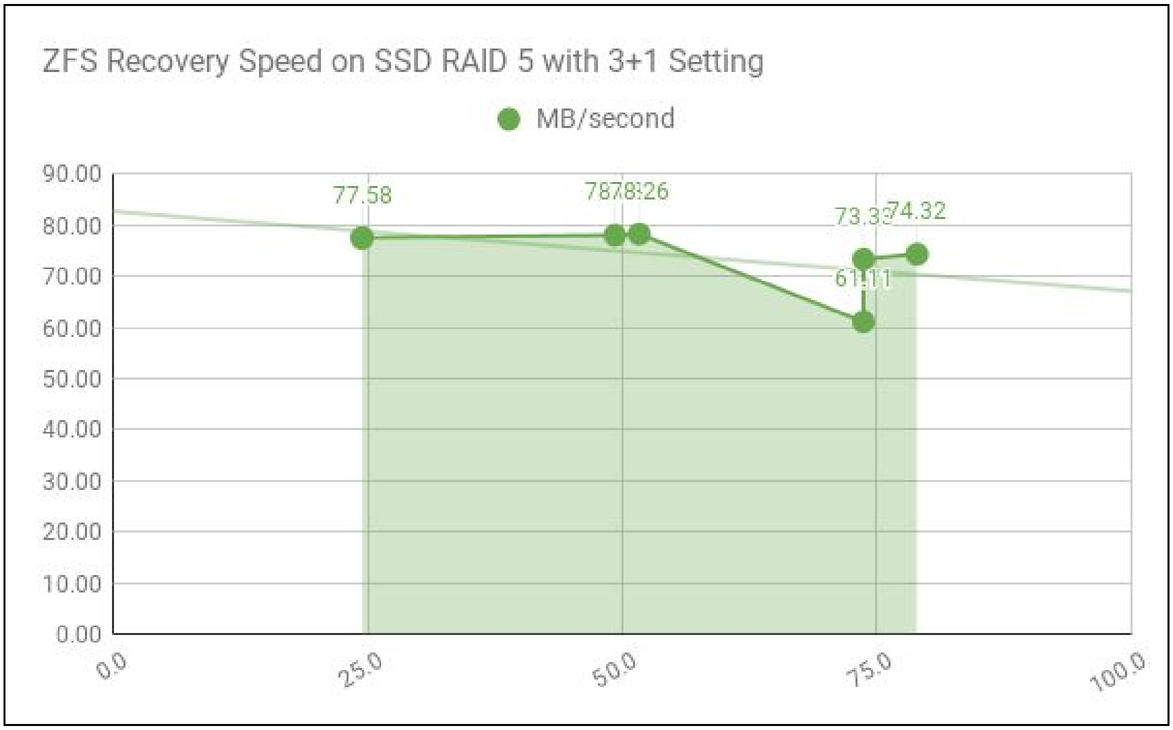
4. Log recovery time and repeat the experiment

Repeat the experiment on HDD and SSD server, with different pool utilizations, and different RAID setup. All experiment are performed twice to see the variance.

Experiment Result

After 2 round of experiment, we have the following result. X-axial shows the disk utilization rate in percentage. Y-axial is the recovery speed in MB/second.





From the experiment, we can conclude following observations.
1) SSD pool recovery speed are more stable than HDD, at 75MB/s. disk usage % do not have significant impact on performance. RAID5 vs RAID6 result are very close. We ran the experiment twice, the result between each run are very close.
2) HDD pool recovery speed varies from 32MB/s to 60MB/s. Since the result from two round, with same RAID setting and same disk usage, yields significantly different speed, we cannot tell if disk usage or RAID setting have any impact on recovery speed.
3) Both SSD and HDD recovery speed is much lower than IO benchmark result from bonnie++. I think recovery computation is the major cause of such slow down.

We were trying to find the speed of clone one drive to another, so we can have a comparison of pro-active disk replacement vs. ZFS recovering. We tried to use "*dd if=/dev/sdb of=/dev/sde*" to clone disk sdb to sde. But this didn't work well. After disk clone, ZFS will not accept new drive to replace the old disk in zpool. Moreover, the avg. speed of "*dd*" only at around 30MB/s, which is a little bit slower than the worst ZFS recovery speed.

There are still more to test for ZFS recovery speed, such as the impact of larger RAM in system, since ZFS is RAM hunger. So far we used only 8GB of RAM in each server, this can add up to 32GB to see if it generates different result. And we can also use SSD as the cache for HDD pool to see if there are any performance improvement.