Storage

This page covers RAID configurations, performance implications and availability, I/O behavior for sealed and unsealed sectors, andf read/write performance considerations.

RAID configurations

Storage systems use RAID for protection against data corruption and data loss. Since cost is an important aspect for storage providers, and you are dealing with cold storage mostly, you will be opting for SATA disks is RAID configurations that favor capacity (and read performance). This leads to RAID5, RAID6, RAIDZ and RAIDZ2. Double parity configurations like RAID6 and RAIDZ2 are preferred.

Thewidth of a volume is defined by how many spindles (SATA disks) there are in a RAID group. Typical configurations range between 10+2 and 13+2 disks in a group (in a VDEV in the case of ZFS).

RAID implications

Although RAIDZ2 provides high fault tolerance, configuring wide VDEVs also has an impact on performance and availability. ZFS performs an automatic healing task calledscrubbing which performs a checksum validation over the data and recovers from data corruption. This task is I/O intensive and might get in the way of other tasks that should get priority, likestorage proving of sealed sectors.

Another implication of large RAID sets that gets aggravated with very large capacity per disk is the time it takes to rebuild. Rebuilding is the I/O intensive process that takes place when a disk in a RAID group is replaced (typically after a disk failed). If you choose to configure very wide VDEVs while using very large spindles (20TB+) you might experience very long rebuild times which again get in the way of high priority tasks like storage proving.

It is possible though to configure wider VDEVs (RAID groups) for the unsealed sectors. Physically separating sealed and unsealed copies has other advantages, which are explained inCustom Storage Layout .

I/O Behavior

Storage providers keep copies of sealed sectors and unsealed sectors (for fast retrieval) on their storage systems. However the I/O behavior on sealed sectors is very different from the I/O behavior on unsealed sectors. Whenstorage proving happens only a very small portion of the data is read by WindowPoSt. A large storage provider will have many sectors in multiple partitions for which WindowPoSt requires fast access to the disks. This is unusual I/O behavior for any storage system.

The unsealed copies are used for fast retrieval of the data towards the customer. Large datasets in chunks of 32 GiB (or 64 GiB depending on the configured sector size) are read.

In order to avoid different tasks competing for read I/O on disk it is recommended to create separate disk pools with their own VDEVs (when using ZFS) for sealed and unsealed copies.

Write performance

Write access towards the storage also requires your attention. Depending how your storage array is connected (SAS or Ethernet) you will have different transfer speeds towards the sealed storage path. At a sealing capacity of 6 TiB/day you will effectively be writing 12 TiB/day towards the storage (6 TiB sealed, 6 TiB unsealed copies). Both your storage layout and your network need to be able to handle this.

If this 12 TiB were equally spread across the 24 hrs of a day, this would already require 1.14 Gbps.

12 TiB1024 / 24 hr / 3600 sec 8 = 1.14 Gbps

The sealing pipeline produces 32 GiB sectors (64 GiB depending on your configured sector size) which are written to the storage. If you configuredbatching of the commit messages (to reduce total gas fees) then you will write multiple sectors towards disk at once.

A minimum network bandwidth of 10 Gbps is recommended and write cache at the storage layer will be beneficial too.

Read performance

Read performance is optimal when choosing for RAIDZ2 VDEVs of 10 to 15 disks. RAID-sets using parity like RAIDZ and RAIDZ2 will employ all spindles for read operations. This means read throughput is a lot better compared to reading from a single or a few spindles.

There are 2 types of read operations that are important in the context of Filecoin:

- random read I/O:

- When storage proving happens, a small portion of a sector is read for proving.
- sequential read I/O:
- When retrievals happens, entire sectors are read from disk and streamed towards the customer via Boost.
-

Previous Infrastructure Next Network

Last updated7 months ago