Reference architectures

This page contains some reference architectures that storage providers can use to build out their infrastructure.

1 PiB raw architecture

?

The following reference architecture is designed for 1 PiB of raw sectors or raw data to be stored. Let’s discuss the various design choices of this architecture.

Virtual machines

- 32 CPU Cores
- 512 GB RAM
- 8x 2 TB SSD storage
- 2x 10 GbE ethernet NICs
-

Lotus daemon and Boost run as Virtual Machines in this architecture. The advantages of virtualization are well-known, including easy reconfiguration of parameters (CPU, memory, disk) and portability. The daemon is not a very intensive process by itself, but must be available at all times. We recommend having a second daemon running as another VM or on backup infrastructure to which you can fail over.

Boost is a resource-intensive process, especially when deals are being ingested over the internet. It also feeds data payload of the deals into the Lotus miner.

We recommend 12-16 cores per VM and 128 GiB of memory. Lotus daemon and Boost need to run on fast storage (SSD or faster). The capacity requirements of Boost depend on the size of deals you are accepting as a storage provider. Its capacity must be sufficient to be landing space for deals until the data can be processed by your sealing cluster in the backend.

Both Lotus daemon and Boost require public internet connectivity. In the case of Boost you also need to consider bandwidth. Depending on the deal size you are accepting, you might require 1 Gbps or 10 Gbps internet bandwidth.

Lotus miner

- 16 CPU Cores
- 256 GB RAM
- 2x 1TB SSD storage
- 2x 10 GbE ethernet NICs
-

Lotus miner becomes a less intensive process with dedicated PoST workers separated from it (as in this design). If you use a dedicated storage server or NAS system as the storage target for your sealed and unsealed sectors, Lotus miner eventually could also become a VM. This requires additional CPU and memory on the hypervisor host.

We opted for a standalone Lotus miner in this design and gave it 256 GiB of memory. This is because we operate ZFS at the storage layer, which requires a lot of memory for caching. Lotus miner has enough with 128 GiB of memory when you opt for a dedicated storage server or NAS system for your storage.

SATA Storage

In this architecture we have attached storage shelves to the Lotus miner with 2.4 PiB of usable capacity. This is the capacity after the creation of a RAIDZ2 file system (double parity). We recommend vdevs of 12 disks wide. In RAIDZ2 this results in 10 data disks and 2 parity disks. Storage systems also don’t behave well at 100% used capacity, so we designed for 20% extra capacity.

PoST workers

- 16 CPU Cores
- 128 GB RAM
- 2x 1TB SSD storage
- 1x GPU 10+ GB memory, 3500+ CUDA cores
- 2x 10 GbE ethernet NICs
-

We have split off the Winning PoST and Window PoST tasks from the Lotus miner. Using dedicated systems for those processes increase the likelihood of winning block rewards and reduces the likelihood of missing a proving deadline. For redundancy, you can run a standby WindowPoSt worker on the WinningPoSt server and vice versa.

PoST workers require 128 GiB of memory at the minimum and require a capable GPU with 10GB of memory and 3500 or more CUDA cores.

Sealing workers

The sealing workers require the most attention during the design of a solution. Their performance will define the sealing rate of your setup, and hence, how fast you can onboard client deals.

Keep in mind that usingSealing-as-a-Service reduces the requirements to have a fast performing sealing setup. In this design, however, we plan for an on-premise sealing setup of maximum 7 TiB/day. This theoretical sealing capacity is based on the entire sealing setup running at full speed for 24 hrs/day.

AP / PC1 worker

- 32 CPU Coreswith SHA-extensions
- 1 TB RAM
- 2x 1TB SSD OS storage
- 15+ TB U.3 NVMe sealing / scratch storage
- 2x 10 GbE (or faster) ethernet NICs
-

We put the AddPiece and PreCommit1 tasks together on a first worker. This makes sense because AddPiece prepares the scratch space that will be used by the PC1 tasks thereafter. The first critical hardware component for PC1 is the CPU. This must be a CPU with SHA-256 extensions. Most storage providers opt for AMD Epyc (Rome, Milan or Genova) processors, although Ice Lake and newer Intel Xeon processors also support these extensions.

To verify if your CPU has the necessary extensions, run:

```

Copy cat/proc/cpuinfo|grep--colorsha_ni

```

PC1 is a single-threaded process so we require enough CPU cores to run multiple PC1 tasks in parallel. This reference architecture has 32 cores in a PC1, which would allow for ~30 parallel PC1 processes.

For this, we also need 1TB of memory in the PC1 server.

Every PC1 processes requires approximately 450 GiB of sealing scratch space. This scratch space is vital to the performance of the entire sealing setup. It requires U.2 or U.3 NVMe media. For 30 parallel PC1 processes we then need ~15 TiB of scratch space. RAID protection on this volume is not mandatory, however losing 30 sectors during sealing and having to start over does have an impact on your sealing rate.

PC2 / C1 / C2 workers

- 32 CPU Cores
- 512 GB RAM
- 2x 1TB SSD
- 1x GPU 10+ GB memory, 3500+ CUDA cores
- 2x 10 GbE (or faster)
-

The next step in the sealing pipeline is PreCommit2 (PC2). You could decide to keep it together with PC1, but given the size of our setup (1 PiB) and the likely requirement to scale beyond that later, we split off PC2 in this architecture.

We plan for twice the amount of PC2 workers compared to PC1, as explained undersealing rate . Apart from the memory requirements this process specifically requires a capable GPU with preferably 24GB of memory and 6000 or more CUDA cores.

The scratch space contents from PC1 is copied over to the PC2 worker. This PC2 worker also requires fast NVMe scratch space. Since we plan for 2 PC2 workers against 1 PC1 worker, the capacity of the scratch space per PC2 worker is half of the total scratch space capacity of the PC1 worker, 8 TiB in our case.

C1 doesn’t require much attention for our architecture. C2 however requires a capable GPU again.

Solo storage providing

Please take a look at the presentation Benjamin Hoejsbo fromPIKNIK gave, in which solo storage provider setups are examined. The presentation is from 2022, but the content is still relevant as of March 2023.

?

We are working to improve this section. If you would like to share your mining setup, please create an issue in the Filecoin documentation GitHub repository !

Previous Backup and disaster recovery Next Skills

Last updated6 months ago