Contribution ID: **223** Type: **Presentation**

Seeking Cost-Optimal Infrastructure Size for Distributed File Systems: A Ceph Case Study

Wednesday 19 March 2025 16:15 (15 minutes)

Here, we present a preliminary study to evaluate how hardware configuration choices can affect the performance of a distributed file system.

While it is straightforward to size hardware for intensive computational tasks to achieve a given performance target, the complexity of the I/O hardware and software stack makes it challenging to predict - and even assess [1] - file system performance based solely on hardware specifications.

However, being able to predict overall performance and associated hardware cost is of utmost importance in many cases, like, for instance, large scale sharing platforms (i.e., NextCloud and others) based on distributed storage solutions as their backbone and HPC facilities with I/O intensive workload (i.e., large scale ML training).

Our experiments are conducted using the Ceph file system; the test infrastructure comprises 10 storage nodes, each of them is equipped with 192 GB of RAM, 2 processors with 16 cores each, while the storage comprises 12x22TB HDD for data and 2x14TB NVMe for metadata.

We explore three different HW parameters: number of CPU cores, amount of memory, and disk speed to see how the performances are affected in terms of bandwidth and IOPS of sequential and random read/write operations.

As a benchmarking tool, we use FIO [2], which is run multiple times, adjusting the number of available CPU cores and the memory capacity by means of the Linux hotplug interface while controlling disk I/O speeds through Linux cgroups.

Our preliminary results reveal that for some workloads, increasing hardware resources does not yield proportional performance gains. Surprisingly, sequential I/O operations are not significantly influenced by additional CPU cores or memory, indicating that the lower tiers of the hierarchical storage model do not benefit enough to justify noteworthy resource increases.

In contrast, IOPS-intensive tasks benefit from increased resources (CPUs and RAM).
Finally, performance consistently improves as disk speeds increase, but only up to a certain point; this suggests that the maximum potential performance of the disks is never fully realized in practice.

We are currently performing additional tests and examining different network configurations, including speed, link layer, and LAG settings, to achieve a comprehensive hardware analysis.

[1] Vasily Tarasov, Saumitra Bhanage, Erez Zadok, and Margo Seltzer.
“Benchmarking File System Benchmarking: It IS Rocket Science.” Proceedings of the 13th Workshop on Hot Topics in Operating Systems
(HotOS XIII), May 2011. Napa, CA: USENIX Association.

[2] Jens Axboe. Flexible I/O Tester (fio). Version 1.2.0, 2022.

Authors: PASIANOTTO, ISAC; TOSATO, NICCOLÒ; Dr LOT, Ruggero (Area Science Park)

Co-author: Prof. COZZINI, Stefano (Area Science Park)

Presenter: TOSATO, NICCOLÒ

Session Classification: Storage Technology

Track Classification: Main sessions: Scalable Storage Backends and Integration with Data Processing