opendrives

Don't be late to the game.

Latency May Be Your Biggest Storage Bottleneck

Too often, performance issues within the storage infrastructure are pinned on lack of sufficient bandwidth. In this sense, bandwidth refers to the overall ability (or inability) of the storage solution to process data efficiently and effectively. The truth of the matter is that, while bandwidth is an important factor in the overall equation of an optimally functioning storage solution, a bigger culprit may actually be latency. At OpenDrives, we understand how latency affects the overall storage topology and workflow, and we've built our storage solutions so that latency is minimized as much as possible.

When users start to experience performance issues when accessing data within the enterprise storage solution, they often have a canned reaction: the storage system just doesn't have the bandwidth to handle all the requests it receives from multiple or many concurrent users. While this is definitely something that can and does happen, it's not the sole culprit when reading from or writing to storage. Very often, available system bandwidth and other solution resources are quite sufficient. The issue in many cases actually stems from inherent latency within the storage solution itself.

While ***bandwidth*** is an important factor in the overall equation of an optimally functioning storage solution, a bigger culprit may actually be ***latency***.

Latency and Its Effects

Latency is loosely correlated with the response time it takes for a storage solution to complete a read or write operation for a client workstation. It is actually a fairly complex notion, though, encompassing many factors, including data transport distance along with any electrical and/or mechanical processes that the storage solution must carry out in order to fulfill the overall command. Latency is simply the lag time experienced between submitting the request and the storage solution completing it. What the end user experiences is always a primary indicator of adequately performing storage.

Latency may not be a critical factor to many end users. In many industries latency may not even something perceptible or detectable by end users. To generic enterprise users, the critical factors are storage capacity, accessibility, security, or extensibility. These are users whose needs are not for real-time responsiveness, but rather for dependability and integrity of the data stored within the solution. In some industries, the focal goal is near-instantaneous responsiveness from the storage solution. These industries are the ones that combine very large data which is hyper dynamic and used or consumed simultaneously by multiple or many users. These are environments in which the data itself changes rapidly and is extracted from and written back into storage many times by numerous concurrent users and their workstations and client applications. Any measurable latency in these situations has serious negative business outcomes.

For the media workflows, latency equates to an experiential deficiency. Post-production has many concurrent users all accessing and working with project data collaboratively.

The Latency **Bottleneck**

The most common perception of latency is that it is a result of the distance that data must travel across a transport medium. Of course, this is a partially true statement. By the definition above, latency is a measurement of the time between issuing the read/write command and the fulfillment of it by the storage solution. If long data transport is a part of this interaction—for example, packets of data traversing LAN and WAN transport networks to some geographically distant storage solution resource—then certainly distance is part of the latency problem. An increased number of hops, slower transport media, and overall volume of traffic only worsen the latency situation. Even at the speed of light, packets take a certain amount of time to go from origination to destination.

Focusing only on the distance factor is a bit facile, though. The electrical and mechanical components involved in the read/write operations introduce latency characteristics of their own. Physical hard drives, take time to spin platters and access data from or write data to the physical storage medium.

Logical components such as software interfaces and operating systems also add to overall latency. And unlike transport variabilities, these latency-inducing factors are entirely under the control of storage solutions vendors. Finding the right combination of technologies to reduce latency at the component level has a noticeable effect on the end user experience.

Attacking Latency Through RAM

OpenDrives has worked closely with our customers in the media and entertainment industry to provide storage solutions that greatly reduce latency and improve responsive times and the overall user experience. As mentioned earlier, this is an absolute necessity in post-production workflows, where many users are accessing project files simultaneously, and smooth high-quality playback is key to the overall workflow. Because of our deep institutional knowledge of the industry, we've implemented architectural features to overcome the latency factor. As an overarching guideline, OpenDrives storage solutions reflect a "memory first" approach. What we mean by this is implementing control logic to ensure the use of high-speed memory over slower read/write operations performed by mechanical drives. OpenDrives privileges memory caching and high-speed secondary caching to avoid latency-inducing mechanical operations. For example, when data is about to be written to storage, smart logic with the OpenDrives operating system ensures that the data lands first in memory prior to being committed to disk. Data remains in memory for a certain period of time in case the data is rapidly accessed again. In that case, it is read from memory, not disk. This vastly reduces latency within the overall transaction.

This same approach is used for read operations as well. When data is read from either an HDD or SSD disk within an OpenDrives solution, our software puts that data into memory at the moment it's read, and it stays there. That way, if the application requests that data again, it's read directly from the much faster memory cache. To reduce latency even more, OpenDrives uses predictive algorithms to "pre-fetch" data. When a file sequence is being read, our software preloads the next file into the faster access tier so that it's ready even before the requesting application requests it. The operating system is constantly evaluating the files in memory, and as the files become less active over time, they eventually are retired from cache to make room for new pre-fetched data. All of this activity is embedded at the operating system level, is completely transparent to the user (except for the snappiness of the system due to reduced latency), and does not depend on a specific type of drive (flash or HDD).

You should not confuse latency with the bandwidth of a storage solution. Latency is the lag-time between read/write command issued and the response of that command, created either by distance or by physical operations at the storage component level. Latency certainly is affected by other performance factors such as bandwidth and throughput, but ultimately certain aspects of latency can be controlled within the storage solution itself. OpenDrives greatly reduces latency through a memory-first approach to data storage, leveraging the speed and efficiency of RAM memory to cache data and the intelligence of the operating system to control how that RAM memory is used. To learn more, visit us at www.opendrives.com.

OpenDrives 2018 All Rights Reserved.