Why Low Latency Matters

Enterprise Applications and Dynamic Business Workloads Demand Faster and Faster Response Times

Prepared by: George Crump, Lead Analyst
Prepared: February 2016

Applications are driving the enterprise, whether it is a relatively simple application used by millions of customers or a complex, scalable database that drives an organization's back end. These applications, and the users that count on them, expect rapid response times. In a world that demands 'instant gratification,' forcing a customer, prospect or employee to wait for a response is the kiss of death.

For most data centers the number one cause of these "waits" is the data storage infrastructure, and improving storage performance is a top priority for many CIOs. The problem is that the storage industry often misleads IT professionals as to where they should direct their attention when trying to eliminate wait time. There is an overwhelming emphasis on IOPS (Input/Output Operations Per Second) and the impact of latency is conspicuously ignored.

The reality is that three factors intertwine to impact overall storage performance: IOPS, latency, and bandwidth. Again, the industry often focuses on IOPS, and to a lesser extent bandwidth, but the most important of the three factors is latency, especially as the modern data center moves to integrate flash storage into the storage infrastructure. Flash storage raises the IOPS potential and eliminates latency at the storage media level, but it also exposes latency elsewhere in the architecture. The latency of the rest of the storage infrastructure becomes the key differentiator for enterprises examining storage systems that will deliver a consistently responsive application experience.

The Storage Performance Ecosystem

Storage systems have four basic components that create an ecosystem. First is the media on which users store and access data. The second component is the storage software. It controls how data is written to the media as well as providing advanced features like data protection, snapshots, and replication. This software component should also dispatch and schedule I/O traffic to back-end media. The third component is the CPU processing that drives the storage software. Finally, there is the storage network. It transfers data back and forth to the application tier.

IOPS essentially impacts the performance of just one of these four components; the media. The media and network capabilities impact bandwidth. Latency is the time it takes for a read or write to traverse all four components - it measures the full cycle impacting response times from application input to final output. Only latency impacts all four components, factoring in the efficiency of the storage software and its ability to use the available CPU power to process I/O.

There are two steps to reading and writing data; the time it takes to get the data on or off of the media and the time it takes that data to traverse the storage system itself. In the hard disk drive (HDD) based storage array, the time it takes to rotate a hard disk platter into place is an order of magnitude greater than the time it takes for the data to traverse the rest of the storage system.

The problem is data centers are moving into the flash era. Their active data is now typically being written to and read from a flash storage area. The use of flash for storage means the time it takes to get data onto and off the storage media is now measured in microseconds instead of milliseconds, exposing for the first time the inefficiency in the rest of the storage ecosystem. Eliminating this latency, or at least reducing it, from the rest of the storage ecosystem is critical for vendors to achieve optimal performance at minimal cost.

While NAND vendors continue to innovate and will continue to decrease latency within the flash module itself, most latency reductions will come from the rest of the storage ecosystem. But this does not mean necessarily that hardware needs to be custom designed for flash. Certainly a fast internal and external network and powerful processors help, but it is the storage software that plays the critical role in reducing latency.

Historic Latency

Prior to the introduction of flash, all storage systems were hard disk drive-based. Latency was determined by how quickly the platter could rotate data into place so it could be read by the hard disks head. The faster the drive could rotate, quantified by its revolutions per minute (RPM), the lower the latency. But even the fastest drive measured response time in multiples of milliseconds. To help with response time, multiple hard disks (dozens to hundreds) were stripped together into a single volume. The data was split across dozens of hard drives, so rotation only impacted performance once. While wide-striped volumes helped, the storage media was still the primary source of latency. These latency problems meant that storage software paid little attention to CPU efficiency because it was always waiting for media to rotate into place.

Another historic source of latency was the storage network. The storage network typically operated on 4Gbps fibre channel (FC) or 1Gbps Ethernet. Again, in the hard disk drive era, the latency of network transfers was minor compared to the latency of the HDD.

Latency Today

Most data centers have moved, or are moving, to flash based storage for their active data. Flash storage responds instantly and has almost no latency. Most IT planners assume that replacing hard disk drives with flash drives solves their latency problem, and on a per-drive basis that is true. The problem is the differences between hard disk and flash drives discussed earlier. The introduction of flash shifts latency away from the storage media and onto the storage controller. The storage controller is responsible for getting data onto and off the storage media. In the HDD era this controller was typically a single core CPU which ran the storage software. Because of the latency inherent to hard disk drives, a single threaded CPU was more than enough to service I/O demands since most of the time it was waiting for the hard disk to position itself into place.

In the modern era, the storage media is now flash. It can respond instantly to I/O requests, a group of flash drives in a storage system could easily overwhelm a single core CPU. Fortunately, processing technology continues to advance at a pace faster than storage media, and today's modern CPUs have dozens of cores. A modern storage controller now has enough cores to keep pace with flash media. The problem is the software that runs on that controller is not able to fully exploit multi-core processors. Storage software that is single-thread or limited in its threading artificially drives up the price of storage because it requires more expensive, faster processors instead of taking advantage of all available cores in less expensive but slower (per core) processors.

In the flash era the media now finds itself waiting on the storage controller to either send or receive data, but the cores on that controller are typically sitting idle or under-utilized because the storage software is not taking full advantage of parallel processing.

Networking with the advent of 8 Gbps FC or 10 Gbps Ethernet and next generation 16 Gbps FC and 25 Gbps Ethernet bandwidths has advanced in the modern data center and is also better prepared to support the low latency of flash storage.

What's Left? Eliminating Latency from the Storage Ecosystem

Given the realities of very low latency storage media as well as low latency storage networking, the largest contributor to latency is now found in the storage controller. Again, the problem with the storage controller is not the lack of processing power. Multi-core processors are up to the challenge. The challenge is in the storage software. Most storage software is focused on providing external features that administrators see and appreciate like LUN/Volume management, snapshots, replication, deduplication and compression. Storage software also needs to focus on features that the storage administrator does not "see" like I/O distribution and scheduling. This means taking advantage of multiple cores in parallel as we discussed in our article "Software Defined Storage meets Parallel I/O".

Without this parallel I/O capability, the storage software becomes the choke point between the now fast storage network and the even faster flash media. The I/O is serialized and therefore becomes a bottleneck. The result is most of the gains in latency reduction are lost almost entirely as data is typically funneled to just one of the available cores for I/O distribution and servicing. It is like an 8-lane superhighway that is only using one lane with a toll booth; traffic or in this case I/O, gets backed up. Parallel I/O is like a 8 lane superhighway with 'EZ pass' on all the lanes. This avoids the toll booth wait time and opens up the other cores (all the "lanes" in this analogy) for I/O distribution so that data can continue to flow back and forth between the application and the storage media.

The Impact of Low Latency Storage I/O

The effect of low latency storage I/O is more data flows through the same hardware infrastructure in the same amount of time as legacy storage systems. The traditional

three-tier infrastructure of servers, network, and compute benefits by having storage systems that directly respond and service existing I/O requests faster and thus have the capability of supporting significantly more applications and workloads on the same platforms.

The efficiency of a low-latent parallel architecture is potentially more critical in hyper-converged architectures, which are a "shared-everything" infrastructure. If the storage software is more efficient in its use of computing resources, that means that it returns more available processing power to the other processes on which it runs. Again, the result is a hyper-converged architecture that can support more virtual machines while providing a more predictable pattern of performance. In other words, it can 'do far more work' and thus maximizes productivity.

As an example, DataCore recently ran an independently audited storage industry benchmark which reported the fastest response times ever recorded, with results anywhere from 3X to 10X faster in overall performance on the same storage hardware. The 3X lower latency times on less expensive hardware, surpassed all previously published results, including those from all-flash arrays. Since DataCore is software that harnesses available multi-cores to do parallel I/O, it can better leverage existing storage hardware and infrastructure investments, putting an end to the "throw hardware at it" problem-solving approach that is expensive and inefficient.