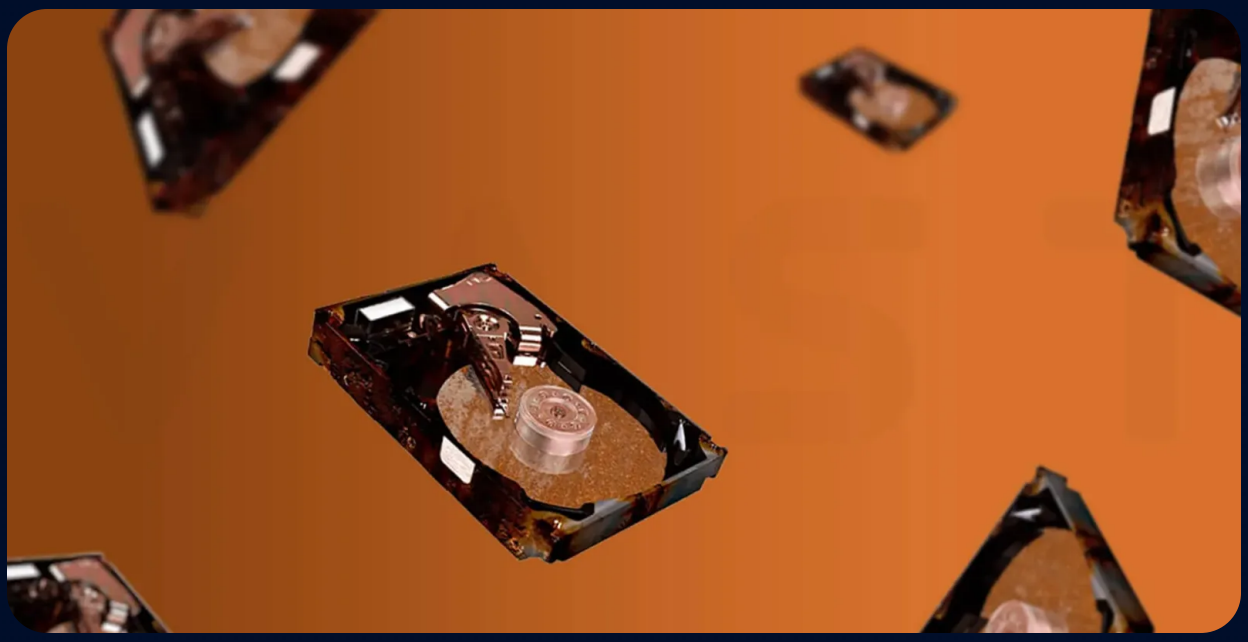
Perspectives Jan 26, 2021

The Diminishing Performance of Disk Based Storage



Posted by
Howard Marks

to remain relevant is to also get less expensive over time. Bigger hard drives have always been the key to low storage costs but bigger hard drives lead to slower disk arrays.

As disk drives get bigger, the number of drives needed to build a storage system of any given size decreases and fewer drives simply deliver less performance. The table below shows the number of disk drives, and the aggregate performance those drives can provide, in a 1 PB array based on 225 MB/s of sustained bandwidth and 150 IOPS per drive.

Drive Size	1	2	4	8	12	16	20
Drives per PB	1000	500	250	125	84	63	50
Drives w/25% overhead	1250	625	313	157	105	79	63
Total IOPS	187500	93750	46950	23550	15750	11850	94
Total B/W GB/s	281	141	70	35	24	18	14
IOPS/TB useable	187.5	93.75	46.95	23.55	15.75	11.85	9.

IOPS per TB. At that level, even applications we think of as sequential, like video servers, start to stress the hard drives.

A video-on-demand system doesn't serve one user a single sequential video stream; it serves 1,000s of users 1,000s of streams and the hard drives have to move their heads back and forth between Frozen, The Mandalorian and the thousands of other videos customers are watching.

Don't forget the slot costs

It's important to remember that it's been a long time since the vast majority of the cost of a disk array was the disk drives. Today, you'll pay more for the JBODs, servers, software and support than for spinning disk drives, what we call the slot costs.

To build a storage system, you'll need hardware and software. You can buy a 60 drive SAS JBOD for about $6,000 or $100/slot but that JBOD needs to be connected to an existing storage system or server. At the low end, a SuperMicro 1U server holding 12 3.5" disks, like you might use for an object store, sells for $5,600 or $466/slot, their 4U 90 HDD behemoth with redundant servers brings that down to $180/slot. Enterprise storage hardware usually ends up costing $500-$1,000 a slot.

The per slot cost of software covers a much bigger territory from essentially free open source software to the software licenses for enterprise storage which can be several times the cost of the hardware. When we priced out a leading scale-out file system, the software and feature licenses were three times the cost of the hardware reducing the cost of disk drives to less than a quarter of storage system costs.

biggest drives.

Drive Size (TB)	4	6	8	12	14	16
NewEgg Price 11/20	$72.00	$120.00	$153.00	$220.00	$290.00	$370
$/TB	$18.00	$20.00	$19.13	$18.33	$20.71	$23.1
$/TB w/$200 Slot Cost	$68.00	$53.33	$44.13	$35.00	$35.00	$35.6
$/TB w/$500 Slot Cost	$143.00	$103.33	$81.63	$60.00	$56.43	$54.1
$/TB w/$800 Slot Cost	$218.00	$153.33	$119.13	$85.00	$77.86	$73.1

Hard Drive Storage Costs

Ceph at ~$200/slot, using 4TB hard drives will cost you twice as much as using 12 TB or bigger drives. Move to the 1U server with 12 HDDs at $500/slot and a system with 18TB drives costs 1/3rd of what the same system with 4TB drives would.

The $200-$800 per slot costs included in the table, are low, even for a roll-your-own solution covering basically just the server the drives will plug into. Put commercial storage software like VSAN (~$11,000/node) on that server and your cost per slot goes up another $180 (with 60 drives per server) to $900 with 12 HDDs/server.

Flash Caches Won't Save You

We also didn't include in our discussion of slot costs; the cost of any SSDs for cache, or metadata store, and who runs a pure HDD storage system today?

You would think that with big hard drives delivering so few IOPS/TB caching the hot data and metadata is a good idea...and it is, but a cache can't completely make up for a lack of IOPS in it's backing store, especially when facing demand peaks. The system may be able to satisfy 90% of I/Os from a cache of a few SSDs under so called normal conditions even then users will report inconsistent performance as the latency of each cache miss is the 5-10ms it takes to read from disk 20-40 times the 500µs SSD read latency.

The problem comes on the weekend our video service releases the hotly expected new movie, thousands of new viewers arrive, and the number of cache misses exceeds the small number of IOPS a back end of 20 or 40 TB hard drives can handle. Now requests back up in queues and latency climbs dramatically from 5-20ms to hundreds of ms, this

more latency.

You would think that you could just make the cache bigger, to keep the number of cache misses down, but that ignores the fact that cache effectiveness falls off dramatically with cache size. Put a small cache on a system and the very hottest data on that system, the half a percent that gets 50% or more of the I/Os doing metadata, or database index lookups.

As you increase the size of the cache it has to hold more data, but that new data is always less active, and soon you've hit the land of diminishing returns where you need 10X as much cache to get any significant increase in the cache hit rate. As the cache gets bigger all that flash adds to the system cost and soon you end up with a hybrid system that costs so much you could buy an all-flash VAST system instead.

Why Hard Drives Don't Get Faster

While advancements like NVMe and PCI 4.0 have made each generation of SSD faster, as well as bigger than its predecessor, hard drives are limited by their physics, so hard drives have only gotten a bit faster since the first "nearline" 7200 RPM disks arrived about 15 years ago.

Over those years drive vendors have boosted capacity in three basic ways:

- **Packing more tracks on each platter.**

  Higher track density doesn't have a direct impact on performance though packing tracks tighter may increase seek times as positioning the heads on narrower tracks could take longer.The ultimate expression of increased track density are SMR (shingled magnetic

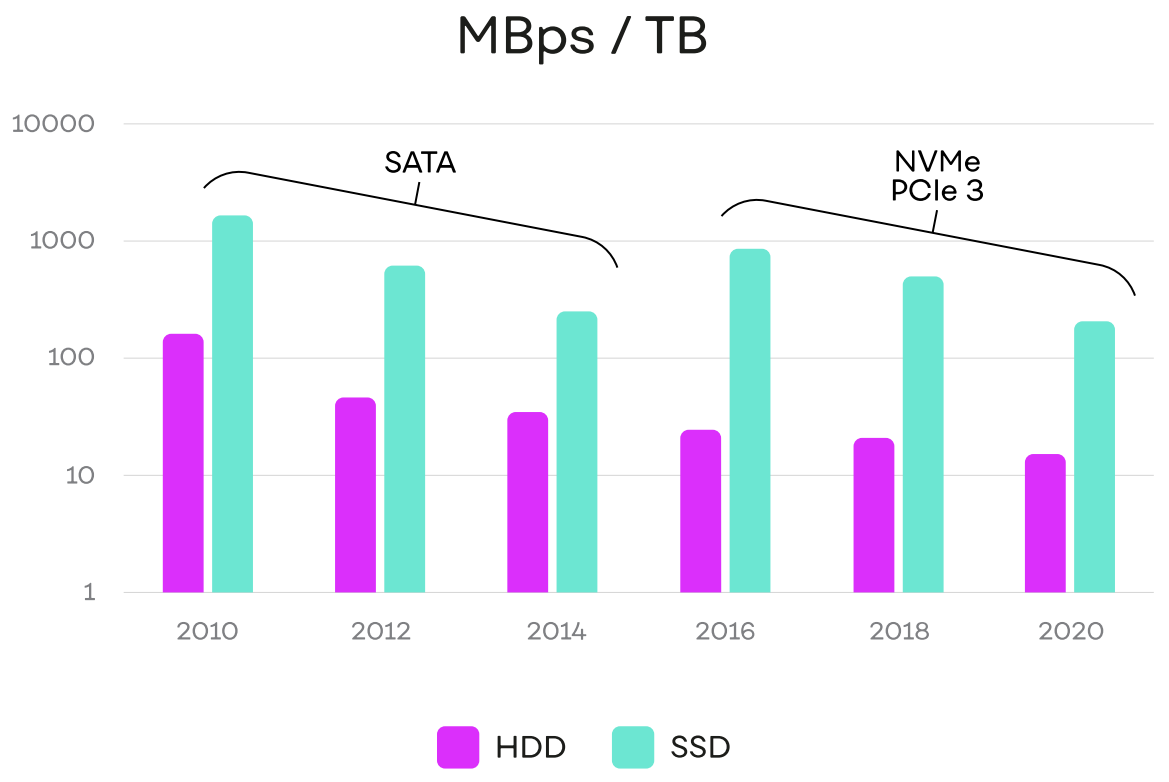
- **Packing more disks in each drive.**

  Thinner disks, thinner heads and the better thermal connectivity of Helium let vendors pack more disks in the same space which increases capacity but since a disk drive can only read from one side of one platter at a time increasing the number of platters in the drive has no impact on performance.

- **Packing more bits on each track.**

  Since disks spin at a constant rate, packing the bits closer together does boost drive bandwidth as more bits will pass under the heads per second.

The net result is that while the first generation of what we then called nearline HDDs, Seagate's 500GB NL35 of 2005 could deliver 65 MB/s of sustained throughput. Today's drives deliver 215-249 MB/s an improvement of about 3.5 times over 15 years.



to move the heads, which takes an average of 5ms, then it as to wait for the sector(s) it wants to read to rotate under the heads, which takes an average of 4.16ms for a 7200 RPM drive.

Since hard drives can only perform one I/O at a time if it takes 12ms to perform an I/O a disk drive can only perform 83.3 I/Os per second. In the real world I/O is never strictly random or sequential, short seeks can take as little as 0.2ms, and drives can process some I/Os from cache which allows vendors to claim up to 200 IOPS. We use 150 IOPS in our math as a generous compromise.

Can Hard Drives Get Faster?

The obvious way to make disk drives faster is simply to spin the platters faster, reducing rotational latency, we had 15,000 RPM disks in the past, why can't we bring them back? The main reason is power. Spinning a disk twice as fast takes roughly four times as much power, which is why the 15K RPM disks of yesteryear used small 2" diameter platters, which limited their capacity.

Disk drive vendors are talking about returning to another old idea, using more than one positioner. The tight spacing of tracks on todays disks requires that the track following servo be embedded between sectors on each track so the drive can properly align heads over the track. Since switching to another head causes the drive to adjust the head position over the track on the new surface a disk drive can only read or write from one head at a time, even on drives with 8 or more platters.

Seagate's multi-positioner drive, like the IBM 3780 of old, uses two positioners, with each serving half the platers in the drive allowing it to read and/or write from two heads at a time. A 20 TB dual positioner drive looks to a storage system as if it were two 10 TB drives doubling

than 1/1000th what an equal sized QLC SSD can deliver.

Conclusion

Hard drive storage cost is a function of the size of the hard drives used to build that storage system. While big hard drives reduce a system's cost per PB they also reduce the resulting systems performance. As hard drives pass 20TB, the I/O density they deliver falls below 10 IOPS/TB which is too little for all but the coldest data.

Flash caches and distributed RAID that worked with 1 and 2 TB HDDs just can't provide consistent performance with many fewer 12 and 16 TB HDDs. When users are forced to use 4 TB hard drives to meet performance requirements, their costs skyrocket with systems, space and power costs that are several times what an all-flash system would need.

The hard drive era is coming to an end. The VAST Data Platform delivers far greater performance, and doesn't cost any more than the complex tiered solution you'd need to make big hard drives work.

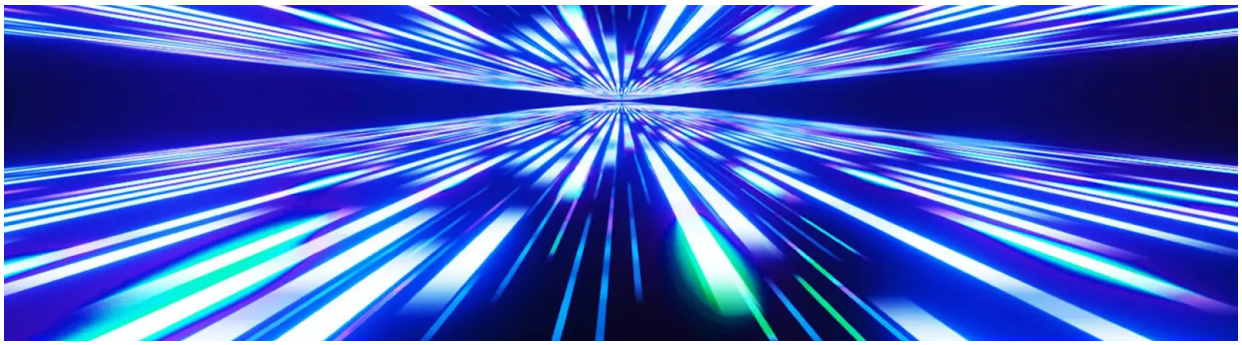
More from this topic

Shared Everything Storage Breaks the Tradeoffs of Shared Nothing Clusters

While it should be obvious, we sometimes forget that storage system architectures are defined by technologies that were available when that system was designed.

The End of the Shared-Nothing Era

The world's first Disaggregated and Shared-Everything (DASE) storage system this is the start of a generational shift in storage solutions.



To Infinity and Beyond AI: GPUDirect Storage is Happening

In support of NVIDIA's launch of GPUDirect Storage 1.0, we wanted to discuss how GDS makes I/O better with the VAST Data Platform and dispel some misconceptions

Explore the VAST Possibilities

Talk to a Solution Engineer

Get in Touch

Email Us

Contact hello@vastdata.com for a 24-hour response.

Chat With Us

Start a live conversation with a VAST expert now.

Call Us

Speak with a team member today at 212-658-1753.

Platform

VAST Data Platform
VAST DataStore
VAST DataSpace
VAST DataBase
VAST DataEngine
Gemini: Consumption Model
DASE Architecture
Supported Platforms
Scale-Out Solutions

Company

About
Customers
News
Partners
Careers
Contact

Resources

Resource Library
Blog
Events
VAST Data Platform White Paper
Documentation
Knowledge Base