Western Digital®

WHITE PAPER OCTOBER 2018

Choosing the Right High-Capacity Hard Drives for Apache Hadoop® Clusters

The storage world has changed dramatically since the early days of Hadoop® and HDFS. The first Apache Hadoop clusters were rolled out at a time when hard drives of 500GB were common, but hard drives today can store 28 times that amount of data with the same power and space requirements. While hard drive performance has improved, it has not grown by the same factor.

Big Data architects must reconcile this disconnect between capacity growth and performance growth. They must balance the number and types of hard drives deployed to match current and future workloads, while keeping an eye on both acquisition and operational costs.

This whitepaper will help architects design Hadoop storage systems and demystify the process of choosing the right size and quantity of high-capacity hard drives. Factors such as bottleneck analysis, I/O performance estimation, and acquisition vs. operational costs will be modelled for several types of clusters.

Modern Hard Drives are Bigger, Faster, and Smarter

Current hard drives range in capacity from 2TB all the way to 14TB. The platters inside of these drives can spin in either an environment of air or helium. Air is used for smaller capacity points (i.e. fewer platters), and results in lower initial acquisition cost. Helium enables higher density (i.e. more platters), reduces power and results in the highest capacity point for the best TCO.

SATA and SAS are the primary interfaces with most Hadoop clusters using SATA drives because many of the SAS enterprise-class features are unimportant for Hadoop and you can often eliminate the cost of a SAS Host Bus Adapter (HBA) by connecting the SATA HDDs directly to the SATA connectors on the motherboard. The SATA interface typically consumes less power than SAS, which means lower operating costs for clusters.

The hard drives examined in this paper are the latest enterprise models from Western Digital (Figure 1). Hard drive designers have taken pains to improve performance given the physical limitations inherent in spinning media. Two powerful features that improve density, performance, and power needs are HelioSeal and our media cache architecture.

HelioSeal–Lowering Power

Traditionally, hard drives have been filled with regular air, which allows for a simpler case design and acceptable density and performance. Air allows the drive heads to float just above the individual hard drive platters, but it is also relatively viscous and induces both friction on the rotating disks and flutter in the drive heads due to its variable densities. Overcoming this friction consumes additional power, and the resulting air turbulence limits the maximum density of bits on the platters, limiting drive capacities.

Western Digital pioneered HelioSeal, the technology that replaces the air in hard drives with helium. HelioSeal allowed Western Digital to increase the MTBF rating to 2.5M hours and still offer a 5-year limited warranty. The Ultrastar DC HC530 drive is the fifth generation of this proven technology. The helium reduces friction and requires less power to spin the hard drive platters. Not only is power consumption reduced, but less heat is generated. This improves overall reliability, and reduces cooling cost. Reducing turbulence allows increases in both platter count and the maximum areal density per platter. HelioSeal enables the most power efficient and highest capacity HDDs available.

Media Cache Architecture–Increasing Write IOPS

Hard drive random-write performance is limited because you need to physically move the drive head to the track being written, which can take several milliseconds. Western Digital's media cache architecture is a disk-based caching technology that provides large non-volatile cache areas on the disk. This architecture can improve the random small-block write performance of many Hadoop workloads by providing up to three times the random write IOPS. Unlike a volatile and unsafe RAM-based cache, the media cache architecture is completely persistent and improves reliability and data integrity during unexpected power losses. No Hadoop software changes are required to take advantage of this optimization, which is present on all 512e (512-byte emulated) and 4Kn (4K native) versions of the drives in this paper.

	Ultrastar® DC HC310 (7K6)	Ultrastar DC HC320	Ultrastar DC HC530
Technology	Air-based	Air-based	Helium-based
Capacities	4TB, 6TB	8TB	14TB
Interface	12Gb/s SAS, 6Gb/s SATA	12Gb/s SAS, 6Gb/s SATA	12Gb/s SAS, 6Gb/s SATA
Platters	4	5	8
HelioSeal® Technology	No	No	Yes
Operational Power (SATA)	7.0W	8.8W	6.0 W
Power Efficiency Index (Operating)	1.16 W/TB	1.1 W/TB	0.43 W/TB

Figure 1. Latest models of Western Digital enterprise-class hard drives.

Workloads Matter for Hard Drive Selection

In order to select the best hard drive configuration, you must understand your current and future workloads. The workload that you run on a Hadoop cluster defines its bottlenecks. All systems have bottlenecks, even perfectly balanced ones (which, in fact, have bottlenecks at all stages). These bottlenecks can be due to the CPU, the number of nodes, the data center networking architecture, the rack configurations, the memory, or the storage medium. Every workload will have a unique bottleneck, not necessarily storage, so recognizing these bottlenecks allows clusters to be optimized to reduce the bottlenecks.

While there are thousands of different Hadoop workloads, they often fall into the following categories:

- Compute or Node Limited Workloads
- I/O Bandwidth Limited Workloads
- Ingest Constrained Workloads
- Random, Small-Block Constrained Workloads

Picking the right hard drive for these different workloads requires the following tasks:

- Identifying the correct total capacity per server
- Identifying the network, CPU, and storage requirements
- Using fewer larger-capacity drives or using more lower-capacity drives

Cluster Configuration Assumptions

Every Hadoop cluster is unique, but they all follow the same general architectures imposed by physical and logical realities. For illustration purposes in Figure 2 we assume a multi-rack configuration of servers, along with an intra-rack switch capable of line-speed switching at 40Gbps and a top of rack switch for cross-rack communication capable of 100Gbps. Assuming 2U Hadoop nodes and 10% overhead for power distribution and other miscellaneous needs, this configuration allows for 18 Hadoop nodes per rack. Each of these Hadoop nodes will have a single 40Gbps connection to the intra-rack switch and will be able to support eight 3.5-inch hard drives over a SATA or SAS backplane.

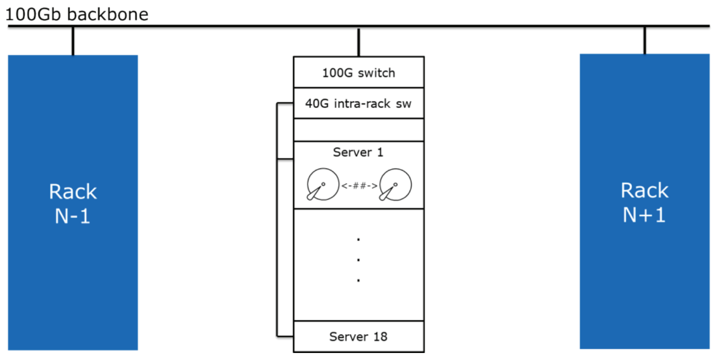

Figure 2. Hadoop Cluster Architecture

Compute-Limited Workloads

Many non-trivial Hadoop tasks are not limited at all by HDFS performance but by the number of compute nodes and their processing speeds. Highly compute-intensive workloads such as natural language processing, feature extraction, complex data mining operations, and clustering or classification tasks all generally work on relatively small amounts of data for relatively long periods of time.

The way to verify these workloads and prove that they are not dependent on I/O speeds is to use your Hadoop distribution's cluster monitoring tool (such as Apache Ambari™ or commercial alternatives). Over the lifetime of any given task, if HDFS traffic between nodes is low while CPU usage is nearly 100%, that's a very good indicator that you have a CPU-limited task.

For CPU-limited workloads the selections of hard drive type and count are not critical. To speed up such workloads you can either invest in higher-speed processors and memory (i.e., scale-up the nodes) or invest in additional processing nodes of the same general configuration (i.e., scale-out the nodes). When scaling-up individual compute nodes it may make operational sense to consider 14TB Ultrastar DC HC530 high-capacity drives to help combat the increase in power usage caused by those faster CPUs and memory. Conversely, when scaling-out the cluster the lower initial cost of air-based hard drives may help to keep initial acquisition costs in check when smaller amounts of total storage are required.

Compute-Limited Workloads–Key Takeaways

- Many Hadoop installations are CPU limited, not I/O bound
- Hard drive selection can focus on operational issues, not performance
- Consider 14TB Ultrastar DC HC530 for scaled-up Hadoop nodes
- Consider 4TB/6TB Ultrastar DC HC310 (7K6) or 8TB Ultrastar DC HC320 for scaled-out Hadoop nodes

I/O-Bandwidth Constrained Workloads

One of the first uses of Hadoop was to scan large quantities of data using multiple nodes, and such tasks are still present in many applications such as databases or data transformation. These workloads can become limited by the available hard drive bandwidth in individual Hadoop nodes. Compounding the problem is Hadoop's replication algorithm, which triples the required cluster write bandwidth due to three-way replication. In clusters where the HDFS replication factor is set higher, this problem only compounds.

For bandwidth-constrained jobs it is first important to determine if the I/O is local to the rack or is going cross-rack and is constrained by the top-of-rack networking and not by storage at all. Poor job placement may cause individual Hadoop MapReduce jobs to be located on servers that don't actually contain the data that they need to process. This situation would cause data requests to be serviced over the global rack-to-rack interconnect, which can quickly become a bottleneck (as seen in the Ingest Constrained Workload section). In this case it is important to adjust the job placement before trying to optimize storage configurations.

After determining that rack-local I/O bandwidth is really the bottleneck, a simple calculation of available theoretical bandwidth per server and per CPU core can help in selecting the proper hard drive configuration. To simplify for illustrative purposes, we assume that all servers are performing the same task at the same time and that the intra-rack networking bandwidth is capable of supporting local HDFS traffic.

First we need to determine the *average* sequential bandwidth of a given drive, not the datasheet maximum. Note that hard drives do not have a constant sequential bandwidth. Because hard drive tracks are concentric rings, and hard drive spindles spin at a constant RPM, the inner tracks read out less data than the outer tracks per revolution, which results in differing bandwidths. We normally take the average of the inner track bandwidth and outer track bandwidth:

Average Hard Drive Bandwidth =
(Outer Bandwidth + Inner Bandwidth) / 2

The total server available sequential bandwidth is simply the average hard drive bandwidth multiplied by the number of drives per server, and dividing that by the number of available CPU cores can explain how much is available per running MapReduce job (assuming no core over committal):

Per-Server Bandwidth =
Average Hard Drive Bandwidth * Drives per Server

Per-Core Bandwidth =
Per-Server Bandwidth / CPU Cores

It's obvious that the more drives available, the higher the bandwidth available. It's also clear that, all things being equal, as the number of CPU cores increases the amount of data they can get to process decreases. These two observations lead us to a very simple rule for these types of workloads: put as many hard drives in each server as physically possible and adjust their capacity as desired. In these cases, the lower capacity of Ultrastar DC HC310 (7K6) drives can help keep total storage capacity reasonable.

I/O-Bandwidth Constrained Workloads–Key Takeaways

- Verify that the I/O is rack- and node-local, not cross-rack
- Determine average per-hard-drive bandwidth, per-server bandwidth, per-core bandwidth
- Try and fill all server hard drive bays, potentially with Ultrastar DC HC310 (7K6) drives for moderate capacities

Ingest-Constrained Workloads

Loading data from an external system into a Hadoop HDFS file system can often require a significant amount of time. During this time the Hadoop cluster's HDFS file system may be taxed, resulting in the lower performance of other jobs during this load phase.

Determining where the bottleneck is for these kinds of jobs is normally very simple: it requires no more than examining the flow of data into the top-of-rack switch and comparing it to the aggregate throughput of the intra-rack switch and hard drives.

In the example cluster configuration, ingest is constrained by the top-of-rack networking. Even at 100Gbit/s, only a maximum of 12 GByte/s may be fed into the entire rack. Spreading that 12GB/s over 18 servers requires that each server have only slightly more than 650MB/s of hard drive bandwidth which can be handled by 4 HDDs if we assume an average sequential throughput of about 180MB/s per HDD (using the average hard drive bandwidth formula given in the prior section, and assuming outer bandwidth is around 240MB/s and, conservatively, that the inner bandwidth is half that at 120MB/s). In this case any configuration of four or more hard drives will suffice to match the performance needs, and only capacity and operational needs will dictate the selection. For the same capacity, you can choose four 14TB HDDs or seven 8TB HDDs.

An even more interesting configuration derives from examining the reason why this large amount of data needs to be ingested in the first place: namely, a lack of HDFS space to contain it economically. If the data were already present on the HDFS file system and were able to be stored as economically as if it were on archive storage, the load stage could be skipped and compute could begin directly. A storage-only disaggregated subset of nodes, filled with 14TB Ultrastar DC HC530 drives, may be able to provide for direct access to your data without requiring archive and avoiding ingest delays altogether.

Ingest-Constrained Workloads–Key Takeaways

- Top-of-rack networking is often the bottleneck
- Any reasonable configuration of four or more hard drives should suffice
- Choose hard drives for operational or capacity reasons
- Consider expanding HDFS space using 14TB Ultrastar DC HC530 drives for a storage-heavy subset of nodes

Random-I/O Bound Workloads

Hard drives today are many times larger than in years past, but they are only somewhat faster due to the limitations of being a mechanical device. Since the IOPS available per drive have remained almost constant while the capacity has grown, the ratio of IOPS per terabyte (IOPS/TB) of these drives has decreased. For heavy random I/O-bound Hadoop jobs, such as building indexes on large datasets, this reduction in IOPS/TB can be a real issue, but fortunately there are ways to ameliorate it.

Much random I/O is caused by temporary and immediate files in MapReduce. For hard-drive-based Hadoop nodes, when more hard drives are present the aggregate random I/O performance supporting these jobs is higher. Media cache technology can also have a dramatic impact here, as the random I/O in MapReduce jobs is often highly write dependent. As a side benefit, by reducing the number of seeks needed to write data to different locations on a hard drive, media cache can increase the random read performance in a mixed workload by being able to dedicate a larger portion of the hard drives' potential head seeks to reads.

Another option for handling workloads with high IOPS, allowing reduction in the total number of hard drives and increasing performance, is to side-step the random I/O issue entirely by deploying local SATA SSDs for the MapReduce temporary files. Adding a single or RAID-mirrored pair of enterprise-class SSDs may augment random performance by 10x to 100x, while allowing the hard drives to be fully dedicated to delivering large block performance. Note, however, that for such workloads only an enterprise-class SSD with a moderate-to-high endurance should be used. Alternatively, any of the open source caching tools such as block cache (bcache) or logical volume cache (lvmcache) may be used to transparently cache accesses to the hard drive array using the SSD.

Random-I/O-Bound Workload—Key Takeaways

- Keep total number of drives per server as high as possible
- Use 512e/4Kn media cache-enabled drives to accelerate write performance
- Consider an enterprise-class, medium- or high-endurance SSD for temporary storage or caching

Choosing Between Helium and Air Hard Drives

In many cases it's not clear from performance modeling whether to deploy the largest HelioSeal hard drive or smaller air-based drives. Assuming that capacity needs can be met with either configuration, the choice can come down to initial and ongoing costs to deploy these drives.

For large Hadoop installations, operational expenses can match or exceed acquisition expenses. Power and cooling are a function of network, server, and storage power. While any individual hard drive's power is not large compared to that of a modern server, it is important to multiply that power by the 8 or 12 drives per server to see true per-server usage. Small changes in the power consumption of drives can therefore have large impacts on this total power. In addition, higher-density drives require fewer drives to match any particular capacity needs.

For certain workloads, the extra capacity per drive of the 14TB Ultrastar DC HC530 can provide operational savings by requiring under half of the disk drives for capacity as the 6TB Ultrastar DC HC310 (7K6). This reduction of drives does impact, of course, the total I/O capacity of each Hadoop node. Therefore, it is imperative to ensure that this reduction will not impact workload performance.

Let's examine the potential savings of a 42TB/node compute-limited Hadoop cluster using either 14TB or 6TB Ultrastar drives.

	Ultrastar DC HC310 – 6TB	Ultrastar DC HC530 – 14TB
Drives Required per Server	42TB / 6 TB = 7 drives	42TB / 14TB = 3 drives
Operational Power per Drive	7.0W	6.0 W
Total Operational Power per Server	7 * 7.0W = 49W	3 * 6.0 W = 18.0 W

This simple analysis shows a difference of a total of 31W per server. Over a five-year lifetime this power savings comes to:

31W * 24 hours * 365 days * 5 years = 1358 kWh savings

Data center power costs vary, of course, but assuming $0.11/kWh and cooling costs equal to power, the savings in operational expenses for the Ultrastar DC HC530 over the Ultrastar DC HC310 come to:

1358 kWh * (0.11 + 0.11) = ~$299 total operational savings per server

Spreading that operational savings over each Ultrastar 14TB drive shows that the individual savings per drive and per terabyte of storage are:

Savings per drive = ~$299 / 3 drives = ~$100 / drive savings

Savings per terabyte = ~$100 / 14 Terabytes = ~$7.14 / terabyte savings

Given this math, it makes economic sense to use the highest capacity Ultrastar 14TB drives even if they have an acquisition cost per terabyte of up to $7.14 above the acquisition cost per terabyte of the Ultrastar 6TB drives.

If the total desired capacity per server is a more modest 24TB, the same sort of calculation can be performed using the Ultrastar DC HC320:

	Ultrastar DC HC310 (7K6) – 4TB	Ultrastar DC HC320 – 8TB
Drives Required per Server	24TB / 4TB = 6 drives	24TB / 8TB = 3 drives
Operational Power per Drive	7.0W	8.8W
Total Operational Power per Server	6 * 7.0W = 42W	3 * 8.8W = 26.4W

This analysis shows a difference of 15.6 watts per server. Over a five-year lifetime this difference in power comes to:

15.6W * 24 hours * 365 days * 5 years = ~680kWh savings

Assuming the same $0.11/kWh power and cooling costs, the savings in operational expenses for the Ultrastar DC HC320 over the Ultrastar DC HC310 (7K6) come to:

~680 kWh * (0.11 + 0.11) = ~$150 total operational savings per server

Spreading that operational savings over each Ultrastar DC HC320 drive shows that the individual savings per drive and per terabyte of storage are:

Savings per drive = $150 / 3 drives = $50 / drive savings

Savings per terabyte = $50 / 8 Terabytes = $6.25 / terabyte savings

Given this math, it makes economic sense to use the highest capacity Ultrastar DC HC320 drives even if these drives have an acquisition cost per terabyte of up to $6.25 above the acquisition cost per terabyte of the Ultrastar DC HC310 (7K6) drives.

Choosing Between Helium and Air HDDs—Key Takeaways

- Is the cluster CPU or ingest constrained?
- Total hard-drive power usage can be a significant part of OPEX
- 14TB Ultrastar DC HC530 or 8TB Ultrastar DC HC320 drives may offer significant savings over other drives, even with an initial cost premium

Additional Considerations

This paper has focused on the bottom-line effects of modern, high-capacity hard drives in Hadoop clusters. These drives also bring with them potential top-line advantages that can be quantified only by your own application and organization.

Deeper Data Can Provide Deeper Insight

In many Hadoop applications such as advertisement placement optimization, the amount of time-based data that the Hadoop cluster reviews before producing its results can impact its accuracy and final results. Allowing for larger datasets to be stored in the same physical space may open up deeper insights into existing data.

More HDFS Space Can Mean Less Administration Overhead

It's a fact of life that no matter how much space you provide to users, they will find a way to fill it up. To preserve space, administrators need to define policies and scripts to move valuable source data to archive and delete unneeded files. On space-constrained file systems, this migration of data back and forth between archives and the HDFS file system can become a serious bottleneck for running jobs. By increasing space available, fewer data migrations may be necessary.

Conclusion

Hard drives today are much larger and smarter than ever before. High-capacity enterprise hard drives, like the 4TB and 6TB Ultrastar DC HC310, 8TB Ultrastar DC HC320, and 14TB Ultrastar DC HC530 can be ideal for delivering the optimal Hadoop experience. Their increased capacity, coupled with performance optimizations such as media cache and power optimizations such as HelioSeal, can have a significant impact upon your Hadoop infrastructure's costs and performance.

Selecting the hard drive for your Hadoop cluster depends upon the cluster's expected workloads. For clusters that are compute or ingest limited, selection of hard drives can focus on capacity and costs; whereas for clusters that are bandwidth limited, the total number of drives per server is an important factor. For those clusters that are running into random I/O bottlenecks, it may make sense to augment your chosen hard drives with a flash-based SSD to obtain the random performance of SSDs with the capacity of hard drives. Be sure to calculate your operational costs to determine if it makes sense to use fewer higher-capacity drives or additional lower-capacity drives. Finally, consider that using a larger total capacity might have non-obvious gains in data quality and administration overhead.

Western Digital.

5601 Great Oaks Parkway
San Jose, CA 95119, USA
US (Toll-Free): 800.801.4618
International: 408.717.6000

www.westerndigital.com