WHITE PAPER

# Deploying Flash-Accelerated Hadoop with InfiniFlash by SanDisk®

a Western Digital brand

Western Digital Technologies, Inc.
951 SanDisk Drive, Milpitas, CA 95035

www.SanDisk.com

## Table of Contents

## Introduction

Deriving value and key insights from growing amounts of structured and unstructured data is now critical across myriad diverse organizations. At the same time, cloud applications in particular are driving unprecedented data growth, and a need to apply new techniques and technology to achieve effective large-scale and hyperscale big data infrastructure. While Hadoop deployments have been very successful to date, actually delivering Hadoop at scale based on existing deployment models brings a host of challenges that can impact both performance and the critical ability to scale.

While these challenges are significant, Hadoop technology itself is undergoing rapid and dramatic change to address them. With Hadoop 2.x, the platform has transitioned to be considerably more flexible, allowing innovation from across the Hadoop community. Network technology too is evolving quickly, delivering ever-increasing bandwidth and fundamentally redefining how Hadoop can be deployed. Building on these trends, flash technology now represents a new frontier in accelerating big data storage at scale—either by augmenting traditional spinning media, or by replacing it entirely.

Ideal for hyperscale Hadoop deployments, the InfiniFlash™ system delivers both high performance and capacity in an all-flash storage array. The system provides a choice of both direct-attached and object-based storage models in a high-density fault tolerant architecture that demonstrates dramatically improved reliability and reduced service requirements over spinning media. A building-block approach to infrastructure means that even very large hyperscale deployments can be scaled out predictably. Beyond providing new levels of capacity and performance, InfiniFlash provides total cost of ownership (TCO) that is highly comparable to Hadoop deployments based on traditional hard disk drives (HDDs).

When compared to typical HDDs deployed in Hadoop infrastructure, the InfiniFlash system provides:

- Up to five times the density, minimizing the physical rack space required.
- Up to 50 times the performance, allowing answers to be generated more quickly.[1]
- Up to five times the power efficiency, reducing both power and cooling costs.
- Up to four times greater reliability, ensuring lower downtime and lower maintenance costs.

## Delivering Hadoop at Hyperscale

Hadoop and its original DataNode model was designed to provide both data storage and analytics capabilities. While this model fit well with the technology and expectations of the time, it is beginning to strain with regular computational

[1] As compared to 200 IOPS HDD.

performance gains and the growth of storage capacity without an accompanying growth in storage performance. Hyperscale deployments in particular need the flexibility to take advantage of new technologies without arbitrary limitations that restrict either scale or performance.

## Large-Scale Hadoop Challenges

Hadoop was originally designed with the fundamental assumption of data distribution across compute and storage, with data placement as close to compute as possible (striving for co-locality of compute and data). It was also designed with the notion that all jobs were created equal with respect to compute and data storage needs. While these assumptions may have been accurate for initial Hadoop deployments, the model can start to fail with cluster growth and asymmetric growth in processor performance relative to storage performance.

### *Data Distribution Across Compute and Storage*

Designed with local storage on each DataNode, the original Hadoop model works well when there is an approximate 1:1 core-to-spindle ratio. This ratio was a fair assumption several years ago when the number of processor cores roughly approximated the number of drives available in an enterprise rack-mount server. Since that time, however, available processor core counts and speeds have continued to increase thanks to Moore's Law, while the number of physical spindles in a given number of rack units cannot increase significantly. Growing the number of cores without appropriate throughput or IO Operations Per Second (IOPS) starts to slow performance as Hadoop jobs start to contend for I/O resources.

The Hadoop Distributed File System (HDFS) is intended to provide streaming data access, ideally maintaining the maximum I/O rate of the drive. Unfortunately, HDDs have an upper limit of aggregate throughput regardless of queue depth, and latency and performance both suffer as queue depth increases. Moreover, typical Hadoop practices can interfere with streaming data access. Standard Hadoop deployments partition the HDDs to at least two partitions, with one EXT3/4 partition providing space for HDFS, and a second partition used for logs, temp files, and transient MapReduce files. The result is that although most of the I/O is comprised of large sequential reads, streaming is often violated by temp file changes, MapReduce shuffle data, and by data writes due to rebalancing or extract transform load (ETL) operations.

In most deployments, HDFS writes to full and very few deletes ever happen. Data scientists are notorious for wanting to keep all their results data because they never know when they will need them again. If the cluster is constrained with search or other specific tasks generating new data, older unused data can result in necessary periodic purging. These purges add to the list of streaming data access violations. At scale, Hadoop operations architects often complain about "long tail jobs" or ETL batch failures due to network bandwidth issues, or cluster slowdowns caused by these issues.

### *Data Locality to Compute*

In a perfect world, Hadoop clusters are sized and data is evenly distributed across those nodes. Jobs would then be scheduled and sized appropriately to evenly distribute the MapReduce processes, thus ensuring a balanced job execution. In reality, unless it is brand new, most Hadoop deployments rapidly develop data distribution issues.

- **Cluster growth**. It is a simple fact that most Hadoop clusters grow over time. Unless the cluster was deployed fully populated at maximum size, some level of cluster growth over time has likely occurred. In practice, new data lands on the servers where there is available space, and old data doesn't tend to move unless a node or storage has failed or an administrator forces a move.  After two to three years, a given cluster might have tripled in size, with data locality at only 30%-60% on average. This non-locality of data to compute results in much heavier reliance on the network to access data to send to MapReduce containers to process the job.

- **Result or data pruning**. As mentioned earlier, results deletion is seldom maintained in typical Hadoop clusters. If data pruning does occur, then the potential for data fragmentation at the distribution level can occur. This issue is more related to data locality than to performance impacts on HDDs due to fragmentation.

### *All Jobs Are Not Created Equal*

Hadoop performance can be very deterministic for batch-oriented jobs such as search, where a new dataset is typically spread across a fixed number of servers for processing. However, ad-hoc usage of a growing cluster by independent data scientists can create other performance issues. Data scientists typically want to assign the most available resources to process their data, but they typically do so without any insight into the data locality issues listed above. As a result, they may well schedule 1500 MapReducers against a dataset that is distributed across only 500 DataNodes due to space constraints in a growing cluster. The result is that 1000 of the MapReducers will have to stream data across the network from the 500 DataNodes, potentially impacting performance.

Of course, sharing a cluster in such a situation can be problematic for performance. Beyond the performance impact from the network access, running jobs can be impacted by streaming data access requests for MapReducers running on other physical servers. This problem is exacerbated by the fact that traditional Hadoop clusters grow with both compute and storage simultaneously, even though jobs are typically bound by either compute or storage resources.

## Hadoop Technology at an Inflection Point

Most of the issues above stem from the traditional Hadoop deployment model, with a network of servers with locally attached storage. This model was chosen for Hadoop both to provide data locality, and because of practical limitations in network throughput. Fortunately, a number of technology advancements and trends are contributing to an inflection point that makes alternative Hadoop deployment models possible. In fact, these trends and innovations fundamentally change the ways that Hadoop can be deployed.

- **Hadoop 2.x enhancements**. Hadoop development is highly dynamic, and Hadoop 2.x brings significant enhancements that make Hadoop considerably more flexible. The YARN (yet another resource negotiator) mechanism is at the architectural center of Hadoop 2.x, allowing an entirely new approach to analytics. Support for Amazon's Simple Storage Service (S3) means that distributed object storage platforms such as Ceph can be used directly as a storage platform for Hadoop clusters.

- **Rapid network bandwidth improvements**. Traditionally, improvements in network bandwidth have moved very slowly, over a multi-year cadence. For example, it took years for 100 Mb Ethernet to be replaced by Gigabit Ethernet. Now networking advancements are happening much more quickly—on a roughly 18-month cadence—with 10 Gb Ethernet, 40 Gb Ethernet, and 100 Gb Ethernet all in the deployment plans of many organizations.

- **Affordable flash acceleration for Hadoop infrastructure**. Flash technology is evolving beyond merely providing high performance. Combined with Hadoop 2.x advancements and improvements in network bandwidth, all-flash arrays such as the InfiniFlash system now represent a viable alternative to spinning media. Not only can all-flash arrays deliver superior performance, but new deployment models can rival the total cost of ownership of HDD-based solutions. Flash can also deliver substantial benefits in terms of density and maintenance costs.

## InfiniFlash™ System

As a manufacturer of flash memory, SanDisk is in a unique position to deliver end-to-end flash-accelerated solutions with full integration, performance, and reliability. SanDisk can innovate at all layers of the hardware and software stack, from unique flash-based building blocks through software, shared storage systems, and application solutions.

### InfiniFlash™ Technology

Illustrated in Figure 1, InfiniFlash is a high-density "just a bunch of flash" (JBOF) all-flash storage system architected to deliver massive scale and quality of service (QoS). As a scale-out, high density, all-flash system, InfiniFlash is designed for high capacity, high density and maximum efficiency, making it ideal for Hadoop deployments at scale. Unlike SSD-based devices, InfiniFlash is based on a new flash form factor, resulting in a platform that can deliver compelling TCO with space and energy savings in a fault-tolerant architecture with low servicing requirements. Characteristics that make InfiniFlash ideal for hyperscale Hadoop deployments include:

- High capacity with 512TB[2] hosted in only three rack units (3U) with up to sixty-four 8TB InfiniFlash cards.

- Scalable performance over HDD-based approaches with throughput of 7GB/s and 780K IOPS.[3]

- Operational efficiency and resiliency with hot-swappable everything, low power consumption (480W average), and high reliability, with a mean time between failure (MTBF) of 1.5+ million hours.

[2] 1TB = 1,000,000,000,000 bytes. Actual user capacity less.

[3] Results based on internal testing. Results available upon request.

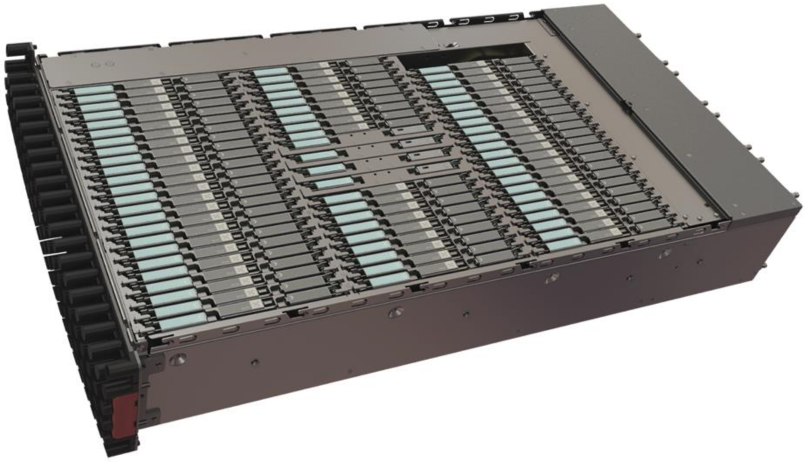

*Figure 1: InfiniFlash by SanDisk provides up to 512TB of high-performance flash storage in only 3U.*

**A Choice of InfiniFlash Models for Hadoop**

Those deploying Hadoop at scale need the flexibility to design innovative infrastructure that meets their unique needs. For some, traditional direct-attached models make the most sense. Others are beginning to explore data tiering, object storage, and other mechanisms. The InfiniFlash system is offered in a choice of models to help enable flexible deployments:

- **InfiniFlash IF100**. With eight SAS ports to support one to eight direct-attached servers, InfiniFlash IF100 provides competitive TCO with space and energy savings for Hadoop deployments. This option is a good choice for organizations wanting to deploy large-scale Hadoop with traditional HDFS.

- **InfiniFlash IF500**. Powered by the InfiniFlash Operating System (OS), InfiniFlash IF500 is a scale-out all-flash system that provides unified block, object, and file support with enterprise-class and web-scale features including snapshots, replication, and thin provisioning. This option is a good choice for organizations wanting to deploy large-scale Hadoop on object storage, with the benefits of disaggregated compute and storage. Organizations gain the flexibility of deploying the right amount of compute and storage for their needs, with the ability to change the mix to handle load spikes—all with compelling TCO.

### Software Stack

The InfiniFlash software stack (Figure 2) is designed to provide both performance and flexibility for a wide range of flash-accelerated applications. Software development kits (SDKs) offer both platform-focused control path services as well as application-specific data path optimization services to supported hyperscale applications. Ceph and OpenStack SWIFT based block and object services are offered as a middleware layer for major applications. Ceph in particular has been transformed by SanDisk for flash performance with over a 10-fold improvement for block reads and a five-fold improvement for object reads. Hadoop can interact with InfiniFlash either through traditional direct-attached models, or via Amazon S3 support provided through Hadoop 2.x.

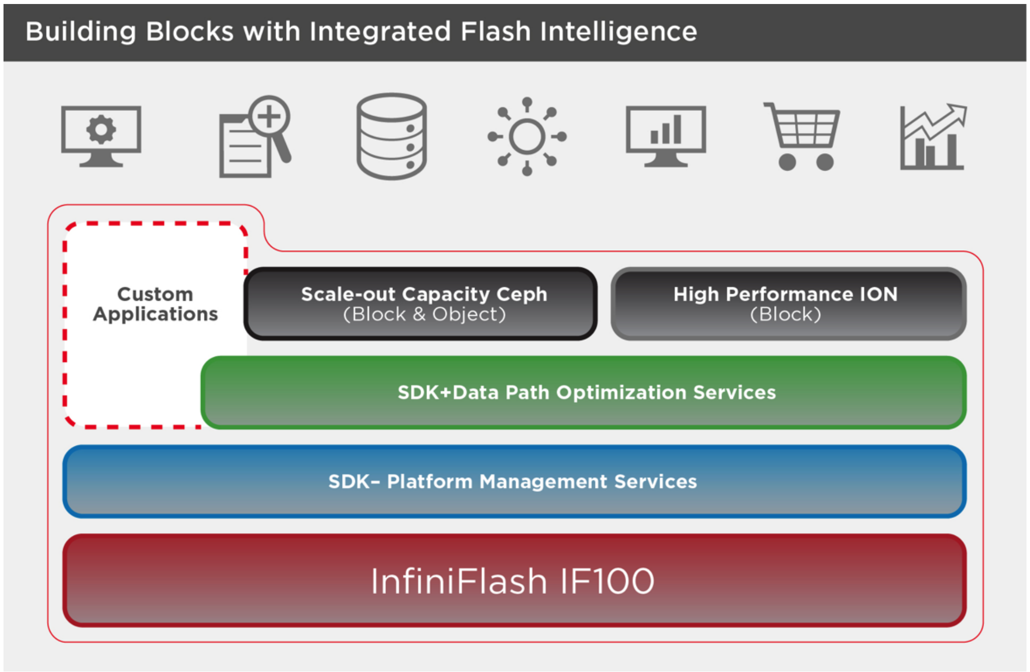
Building Blocks with Integrated Flash Intelligence
Custom Applications
Scale-out Capacity Ceph (Block & Object)
High Performance ION (Block)
SDK+Data Path Optimization Services
SDK- Platform Management Services
InfiniFlash IF100

*Figure 2. The SanDisk software stack provides maximum flexibility to enable InfiniFlash deployments.*

## Flash-Based Solution Architectures for Hadoop Deployments

The InfiniFlash system can be integrated with Hadoop 2.x in a number of different solution architectures, depending on the needs of the organization. The sections that follow detail several alternative approaches. All of these architectures are designed to be deployed as repeatable building blocks that can be scaled out to facilitate large Hadoop deployments.

### Traditional Symmetric HDFS with Data Tiering

Most traditional Hadoop deployments are deployed with relatively uniform storage on DataNodes, which can cause many of the issues described above. Just as data tiering has merit in the enterprise, some hyperscale organizations are evaluating tiered models for Hadoop (http://www.ebaytechblog.com/2015/01/12/hdfs-storage-efficiency-using-tiered-storage/). The flexibility of Hadoop 2.x and new archival models are demonstrating that data tiering can be accomplished by retaining only highly-active data on high-performance storage, while migrating less active data to more cost-efficient storage. InfiniFlash can work well in these more traditional environments, bringing performance, capacity, and throughput to accelerate Hadoop while low-activity data is migrated to a lower-performance HDD-based storage tier. Figure 3 illustrates a traditional three-replica Hadoop deployment with the compute tier for each replica provided by an InfiniFlash IF100 system directly attached to eight DataNodes. In this example, the HDD tier might be implemented cost effectively via a larger-capacity 60-disk server. In this illustration, 40% of capacity is provided by InfiniFlash with the remaining 60% provided by the HDD-based servers, though this could be easily adjusted as needed.

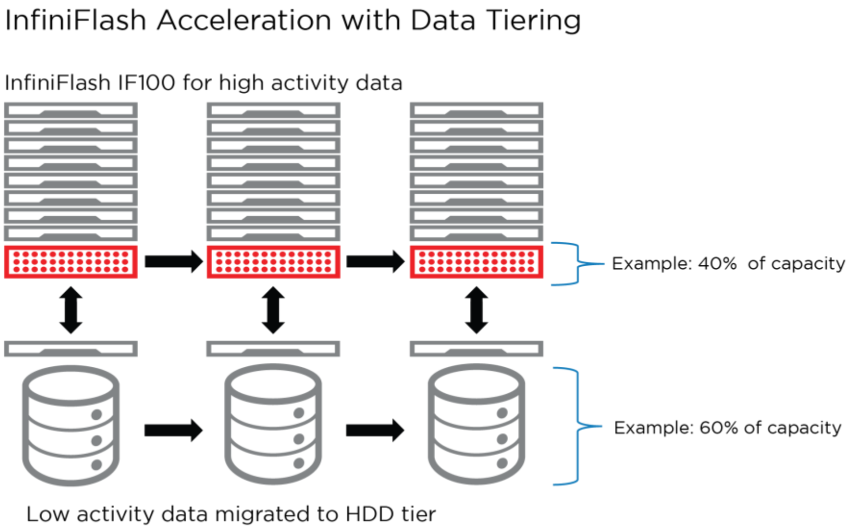
InfiniFlash Acceleration with Data Tiering
InfiniFlash IF100 for high activity data
Example: 40%  of capacity
Example: 60% of capacity
Low activity data migrated to HDD tier

*Figure 3. For traditional Hadoop models, InfiniFlash IF100 provides flash acceleration with low-activity data migrated to cost-effective HDD-based data tier.*

Deploying data tiering with InfiniFlash acceleration for high-activity data presents several distinct advantages:

- The solution architecture provides a performance increase with the least impact to the traditional model.
- The compute tier is able to take advantage of local flash in the MapReduce shuffle phase, keeping data streaming.
- With the HDD bottlenecks removed, organizations can now follow Intel’s cadence for processor core development, deploying higher-performance DataNodes as they become available.
- The ability to employ relatively low-cost, high-capacity storage for the HDD tier helps offset costs for providing flash acceleration at the compute tier.

### Hadoop on Object Store with InfiniFlash Primary

Building on the strengths of traditional models, InfiniFlash IF500 can be deployed with Hadoop as a flash-optimized object store. Figure 4 illustrates a model in which a disaggregated compute farm runs Hadoop compute jobs only, with all data access streamed over the network from a Ceph-based object store. In this solution architecture, two servers attach directly to the InfiniFlash IF500 system, with Ceph providing the data tiering capability to migrate replicas to attached HDD-based storage.

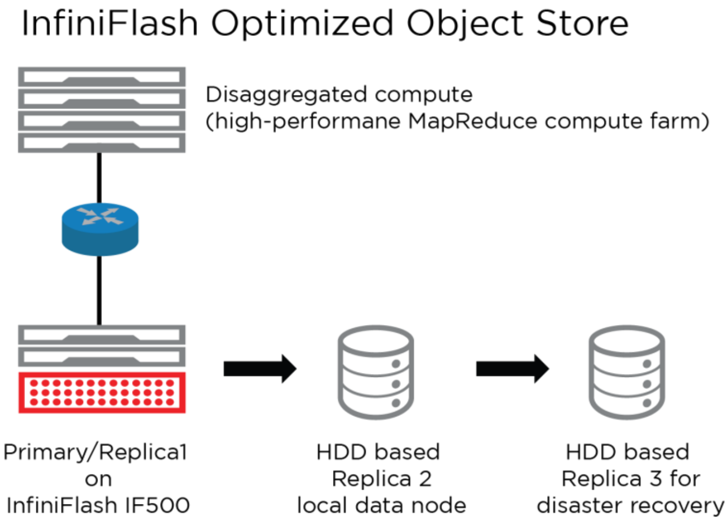

*Figure 4. A flash-optimized object store disaggregates Hadoop compute resources from storage for greater flexibility.*

Disaggregating Hadoop compute resources in this fashion brings distinct advantages that have simply not been possible before:

- Deployments can scale orthogonally, scaling compute and storage independently as dictated by need (e.g. easily adjusting to load spikes that occur during peak periods).
- Data integrity capabilities provided by the InfiniFlash OS and data tiering provided by Ceph offload those processes from Hadoop.
- With compute resources disaggregated, multiple Hadoop distributions can be easily intermixed, all without the need to reload data.
- Importantly, with this model, high-performance compute nodes can be dynamically repurposed as needed for other tasks, and don’t need to be constantly dedicated to Hadoop—in concert with virtualized infrastructure trends.

### All-Flash Object Store with Erasure Coding

Taking the object store model to the next level, InfiniFlash can also support high-performance Hadoop installations that eliminate HDDs entirely. In this solution architecture, all three replicas are maintained on InfiniFlash, with erasure coding providing low-overhead redundancy and data integrity. Figure 5 illustrates an all-flash object store with two servers attached to each InfiniFlash IF500 appliance.

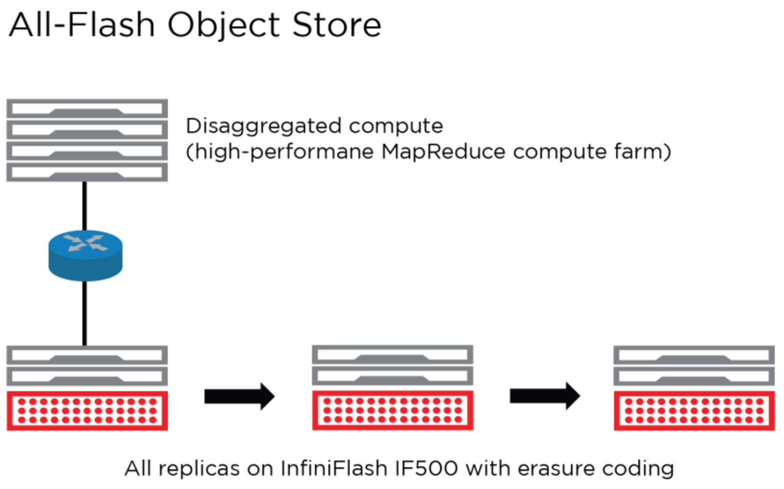
All-Flash Object Store
Disaggregated compute
(high-performane MapReduce compute farm)
All replicas on InfiniFlash IF500 with erasure coding

*Figure 5. An all-flash object store with erasure coding eliminates HDDs entirely from the Hadoop deployment.*

Deploying an all-flash object store presents several additional advantages:

- Using flash throughout the cluster provides more than 2.5 times the performance advantage over HDDs.[4]
- With dramatically fewer racks required for the same capacity, an all-flash architecture can be realized at a compelling TCO as compared to HDDs (as described in the following section).
- Eliminating HDDs and their high failure and replacement rates can dramatically lower operational costs.

[4] Test results available here: http://www.sandisk.com/assets/docs/increasing-hadoop-performance-with-sandisk-ssds-whitepaper.pdf

## Flash Performance with Compelling TCO

To evaluate the cost effectiveness of flash-accelerated Hadoop architecture, SanDisk has performed analysis that evaluates TCO across the three solution architectures as compared to traditional HDD-based three-replica Hadoop. SanDisk's price reduction roadmap allows for further InfiniFlash cost improvements not shown in this average model. In addition to performance gains, and the ability for flash to resolve many of the issues encountered with traditional Hadoop deployments, all three flash-accelerated architectures demonstrate compelling TCO when compared to traditional HDD-based Hadoop (Figure 6). Despite a slightly higher cost of acquisition, all three InfiniFlash architectures are much more cost effective when three-year TCO is taken into account.

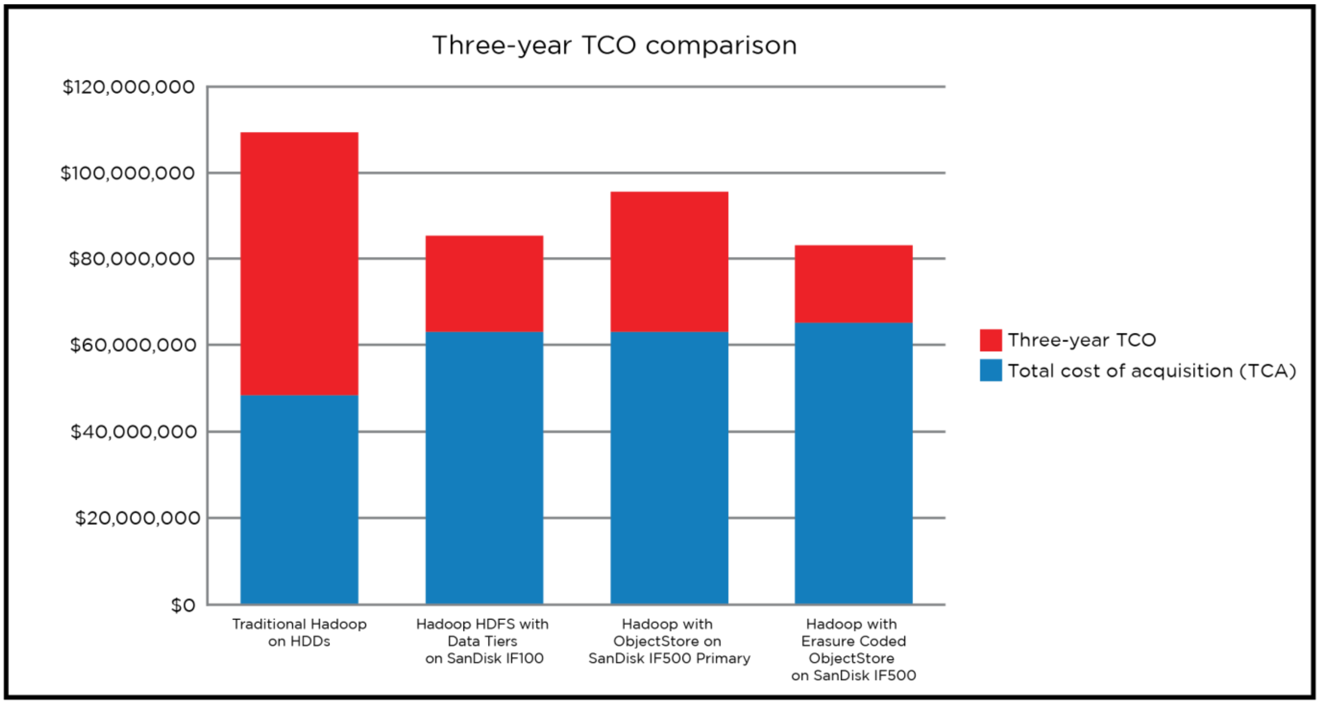

*Figure 6. Flash-accelerated Hadoop compares favorably with HDDs in a three-year TCO calculation for a 200PB Hadoop cluster.*

Importantly, performance benefits and operational and maintenance costs are not included in these models. To give some idea of the impact of flash-accelerated Hadoop, Figure 7 shows the dramatically reduced number of racks required to host a 200PB flash-accelerated Hadoop cluster using each of the three solution architectures as compared to standard HDD-based Hadoop. The density and low power profile of InfiniFlash can contribute greatly to savings in terms of real estate, power and cooling in the data center.

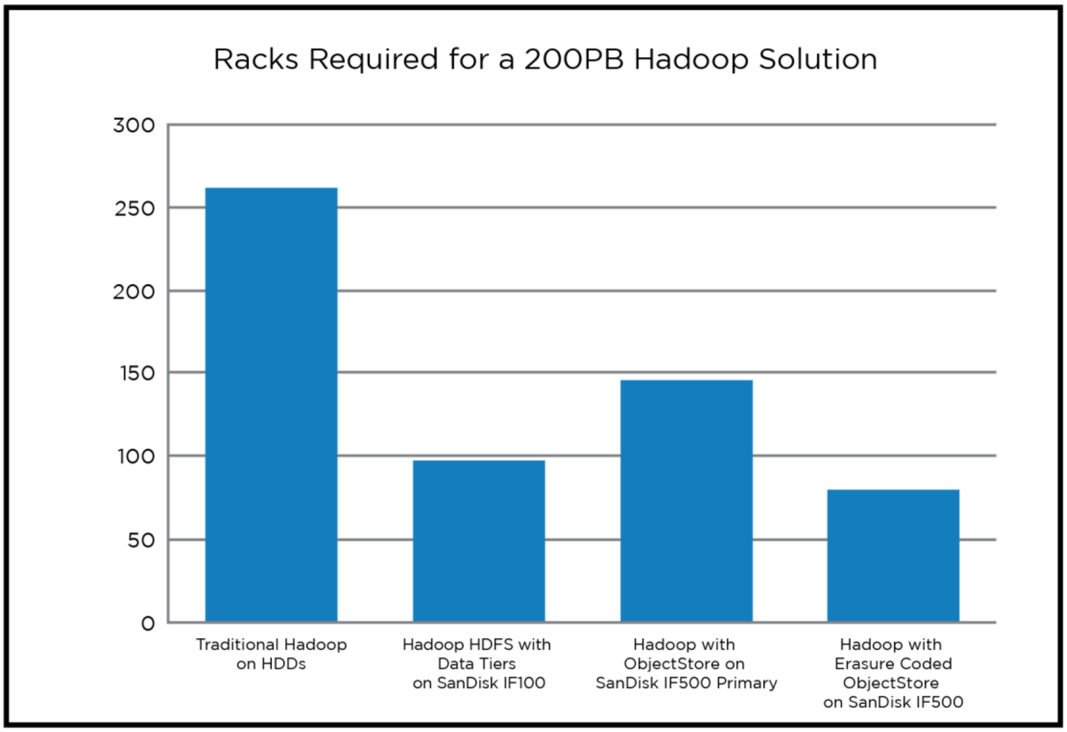

*Figure 7. Flash-accelerated Hadoop requires dramatically fewer racks, and provides commensurate savings in power and cooling.*

## Conclusion

With the combination of Hadoop 2.x, rapid improvements in network bandwidth, and flash technology, organizations deploying Hadoop at hyperscale can now begin to free themselves from HDD-based Hadoop scalability issues. InfiniFlash presents significant performance improvements over standard HDDs, and a choice of flash-accelerated solution architectures resolves many scale-related issues related to traditional HDD-based deployment models for Hadoop. Replacing spinning media with innovative solid-state InfiniFlash systems from SanDisk along with data tiering and object storage can literally redefine how Hadoop is deployed at scale—with TCO comparable to HDDs.