MACH.2™

TECHNOLOGY PAPER

SEAGATE

INTRODUCTION

Seagate, as a leader in storage technology, is releasing to market MACH.2™ multi-actuator technology hard disk drives (HDD). HDDs provide the storage industry with consistent capacity growth at a consistently declining cost, making them a dominant storage technology across all markets and applications. Enterprise 3.5-inch, 7200-RPM HDDs, commonly known as nearline drives, have doubled capacity in the last five years, enabling the lowest cost, while maintaining a consistent level of performance. These HDDs have met the growing exabyte demands of cloud infrastructures. As HDDs deliver more capacity at lower costs in the future, to extract maximum value from these storage devices, a consistent performance level must be maintained. MACH.2 multi-actuator technology is on an innovation vector that will maintain performance to match the capacity growth of these 3.5-inch high-capacity HDDs while meeting or exceeding total cost of ownership (TCO) goals.

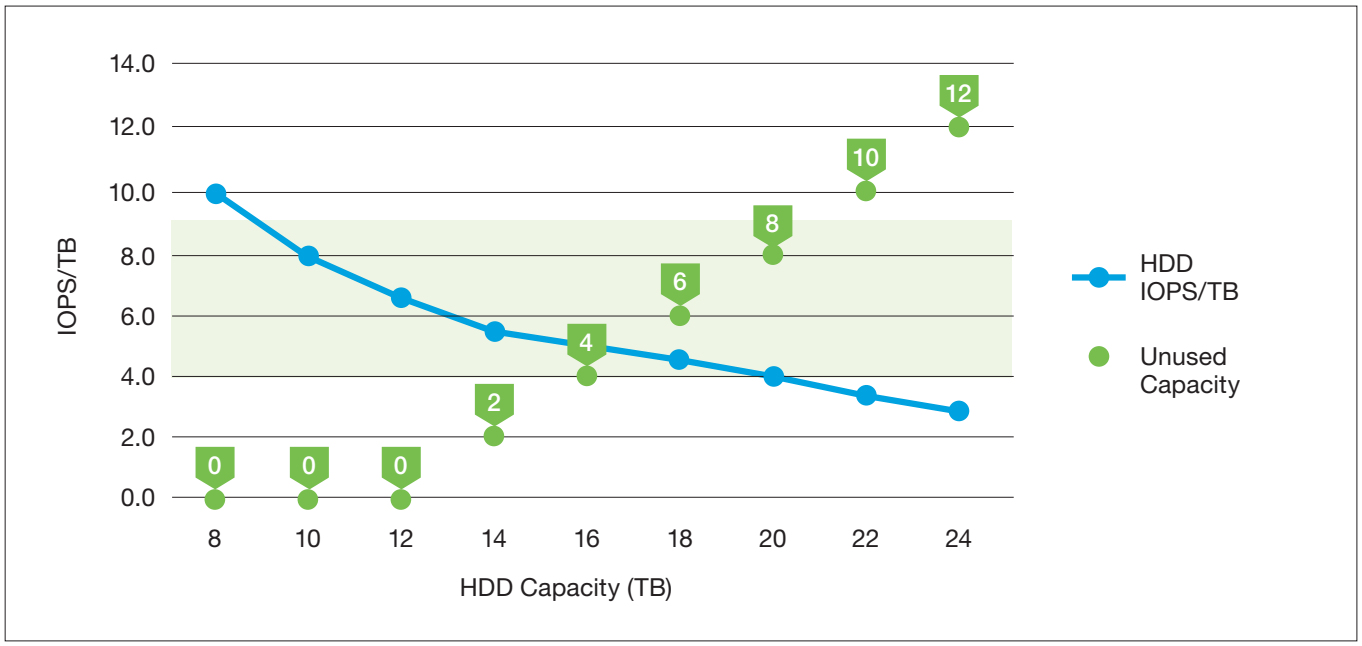
HDD PERFORMANCE

HDD technology involves moving parts, and performance is driven largely by RPM and the number of channels used to transfer I/Os (read and write commands). Historically, HDDs have gained performance by increasing RPMs, which led to 2.5-inch mission-critical 15K- and 10K-RPM drives. But these 2.5-inch drives cannot deliver higher capacities due to form factor constraints. Similarly, 3.5-inch HDDs deliver the highest capacity but cannot be spun at higher RPMs due to higher power consumption and not enough performance gains to justify higher power. Other challenges at higher RPMs include chassis-level vibration and operational shock performance degradation.

An alternate means of improving 3.5-inch HDD performance is via caching within the drive, which helps only at certain workload conditions and provides limited benefits. Queueing is another means of deriving more performance from a drive, but it comes at a latency penalty and having to architect applications to work at higher queue depths (QD). You may also increase random read/write performance by reducing the maximum usable capacity of a drive (also known as short-stroking) which limits the seek distance across the disk. This technique, though effective for gaining additional performance, is not cost-efficient due to the amount of capacity lost.

PROBLEM STATEMENT: IOPS/TB AND TCO

Cloud customers deploy large fleets of 3.5-inch HDDs in their data centers and manage their application storage nodes across multiple devices spread across multiple racks. In order to meet their Service Level Agreements (SLAs) these cloud customers need to achieve a certain level of performance, measured in IOPS (I/O per second) and command latency, across the deployed HDD storage capacity. This required performance varies by workload, but can be characterized as IOPS/terabyte (IOPS/TB) at a particular latency. For many cloud workloads, IOPS/TB is simply derived from the drive's random IOPS capability at a certain workload divided by the available capacity. As long as a particular HDD device meets or exceeds a threshold IOPS/TB as required by the customer, all of the capacity on the HDD device can be utilized. If the IOPS/TB on an HDD is below a threshold value as defined by a customer based on application workloads, then the customer cannot utilize all of the drive capacity with the targeted workload. This problem of not being able to utilize all of the capacity due to IOPS limitation at a targeted workload then impacts another critical metric of TCO: $/TB of utilizable capacity.



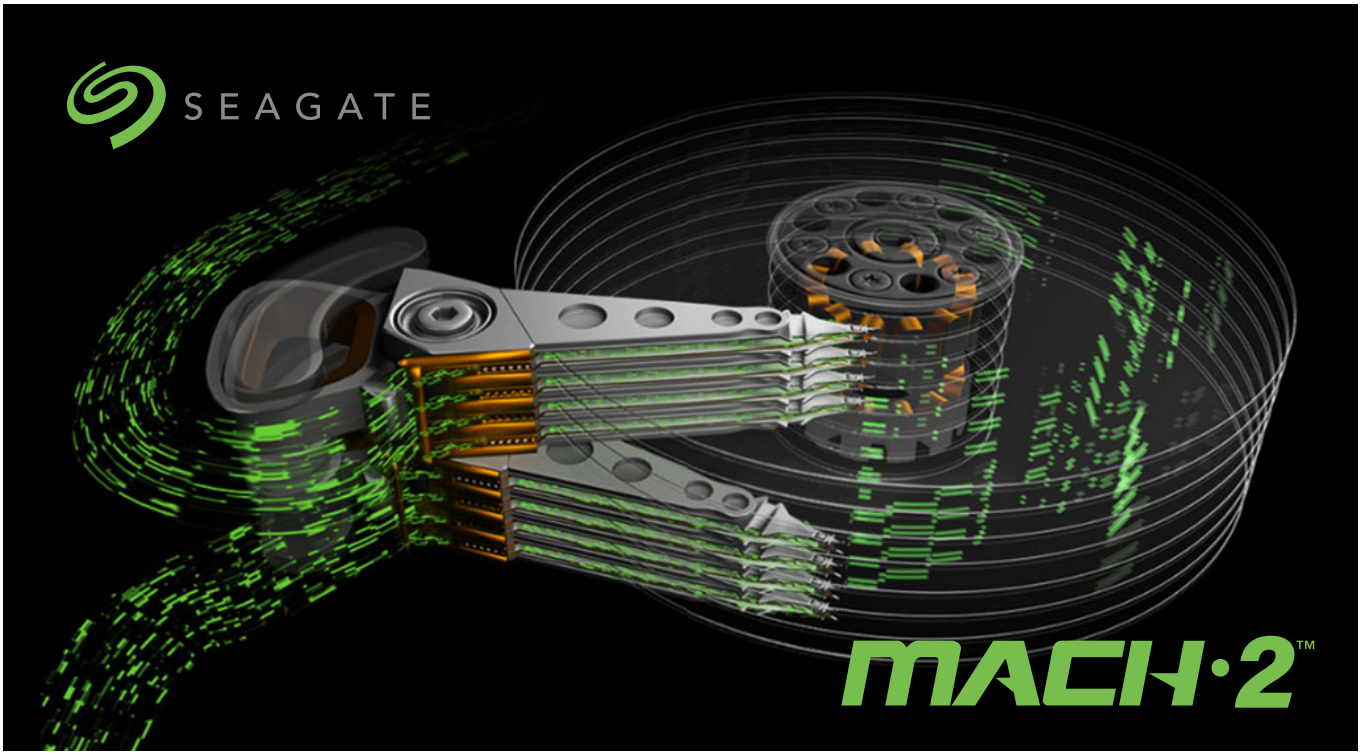
Above chart illustrates this problem.

The X-axis is HDD capacity per device for 3.5-inch nearline products. The Y-axis is IOPS/TB, which declines as capacities grow due to an effectively fixed servo-mechanical capability (thus a fixed IOPS capability). The blue region shows the expected range of IOPS/TB required to meet the performance needs at the application level across a data center fleet. For example, if a customer's threshold IOPS/TB is 7, then the customer can use all of the available capacity at 7 IOPS/TB, but beyond that point, the customer gets stranded at approximately 12TB. Any more available capacity on the HDD may be inefficiently utilized resulting in higher customer TCO.

Today's HDDs use a single actuator to transfer I/Os from the device to the host using a single read/write channel. This results in fixed performance irrespective of the capacity gain and number of heads/media per drive.

MACH.2 TECHNOLOGY

MACH.2 technology addresses the IOPS/TB challenge by using two actuators that can transfer I/Os independent of each other within a single HDD, creating parallelism that enables up to double the performance. Within a drive, the top half of the heads/media are addressed by one actuator, while the bottom half of the heads/media are addressed by a second actuator. Each actuator addresses one half the total capacity of the drive. The image below helps visualize the two actuators working simultaneously, yet independently, of each other. The table below helps explain the gain in performance using MACH.2 vs. a single-actuator drive (today's HDD).

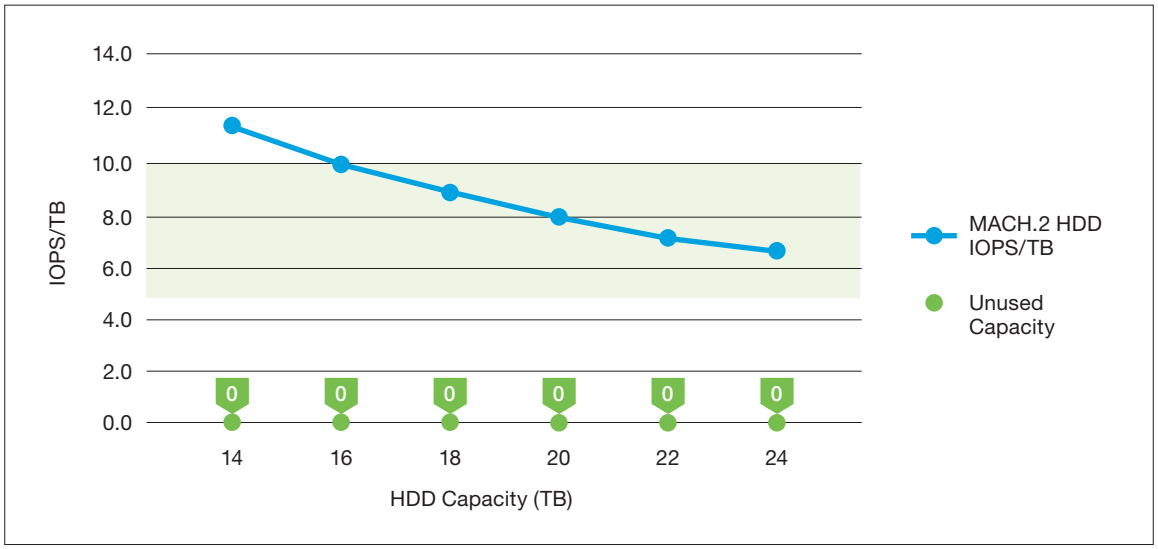


MACH.2 HDD	Number of Actuators	Capacity (TB)	Random IOPS	IOPS/TB	SDR (MB/s)
Actuator-Level Metrics	N/A	7	80	5.7	260
Drive-Level Metrics	2	14	160	11.4	520

MACH.2 HDDs delivers up to 2× performance that can be harnessed as random I/O performance and/or sequential data rate, depending on the needs of the application and workloads.

IOPS/TB AND TCO

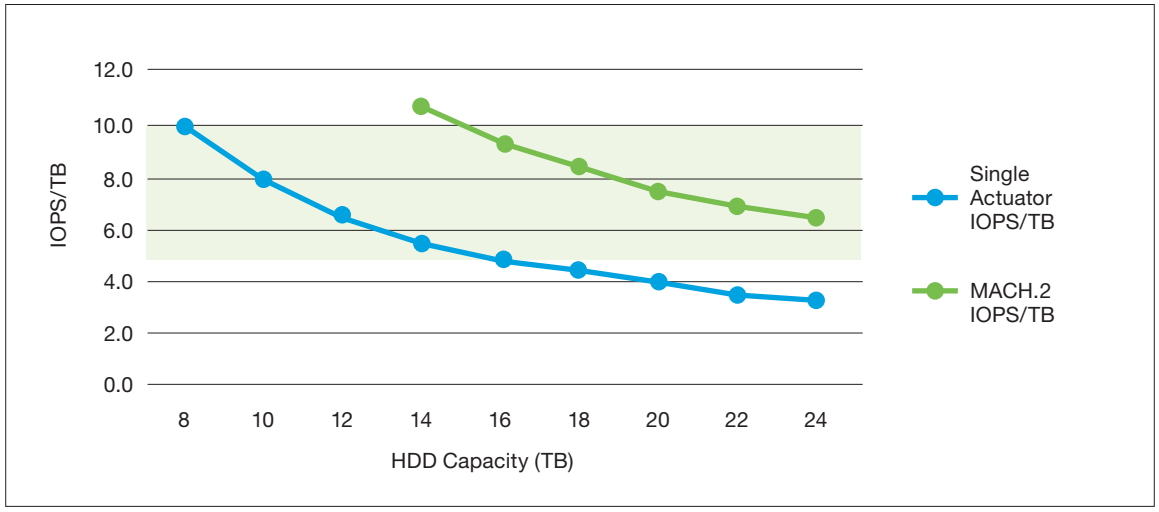
MACH.2 delivers up to double the performance, and the impact shows up on both the performance as well as a customer's ability to utilize all of the capacity available at the HDD level.



The chart above illustrates the same customer who was previously stranded at 12TB capacity utilization due to IOPS/TB constraints, but can now extend beyond 24TB of HDD device capacity and be able to use all the available capacity.

TRANSITION FROM SINGLE ACTUATOR TO MACH.2

As previously stated, each customer is different depending on their storage architectures and applications and hence has a different IOPS/TB requirement. Understanding that value determines the capacity point at which a customer would want to switch from single-actuator drives to MACH.2 drives. Seagate believes most data center customers service applications that need HDD device performance to range between 5 to 10 IOPS/TB. The chart below illustrates the value MACH.2 provides by enabling customers to maximize their capacity utilization at a threshold IOPS/TB, while enabling lowest TCO $/TB.

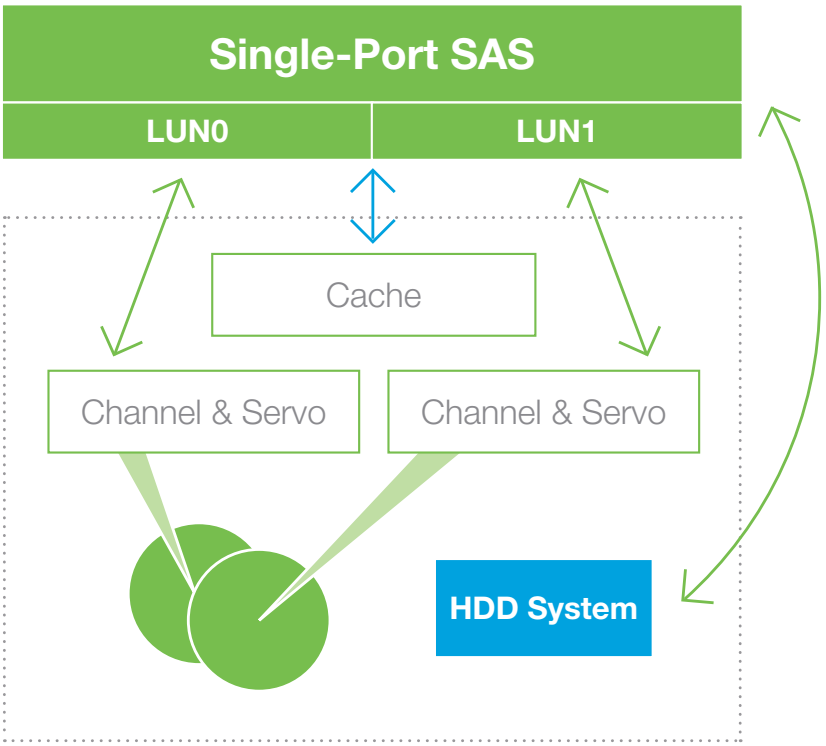


MACH.2 DESIGN

MACH.2 design is enabled by two actuators that transfer data via dedicated channels. There are multiple proprietary design innovations to make sure there is no coupling between the two actuators while they are servicing I/Os. The design is intended to maximize IOPS capability (i.e., doubling the performance compared to single-actuator design).

Choosing the right interface (SAS vs. SATA) for MACH.2 was an important design decision. MACH.2 technology can be implemented using both interfaces, but the decision to choose SAS was made based on performance characteristics and architecture simplicity.

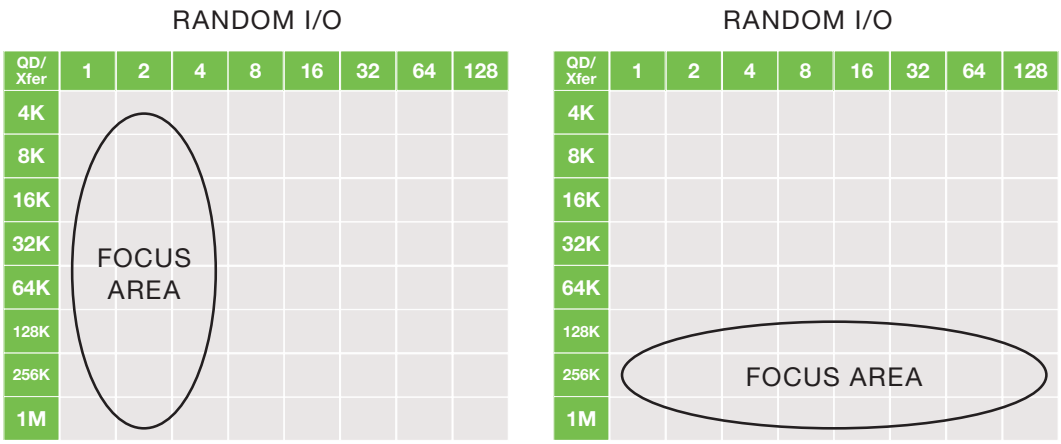
- Performance Characteristics: SAS 12Gb/s vs. SATA 6Gb/s. The first-generation MACH.2 product will come very close to saturating the SATA 6Gb/s interface (600MB/s throughput). As stated above, the drive will have a 520MB/s sustained data rate (SDR) and next-generation drives will perform even better. Hence, the design decision was made to enable ecosystem and storage architectures to readily deploy next-generation MACH.2 drives without changing interface protocols.
- Architecture Simplicity: The two actuators are addressed by the host via SAS logical unit number (LUN) protocol. The LUN protocol is already built into the SAS interface architecture and can be leveraged to address both actuators independently. LUN0 and LUN1 present the same worldwide name (WWN) to the host with a unique identifier to specify LUN0 vs. LUN1 and, when plugged into the host system, one MACH.2 drive shows up as two independent devices of equal capacity—e.g., one 14TB drive will show up as two 7TB drives. The image below shows an illustration of the LUN-based SAS architecture.



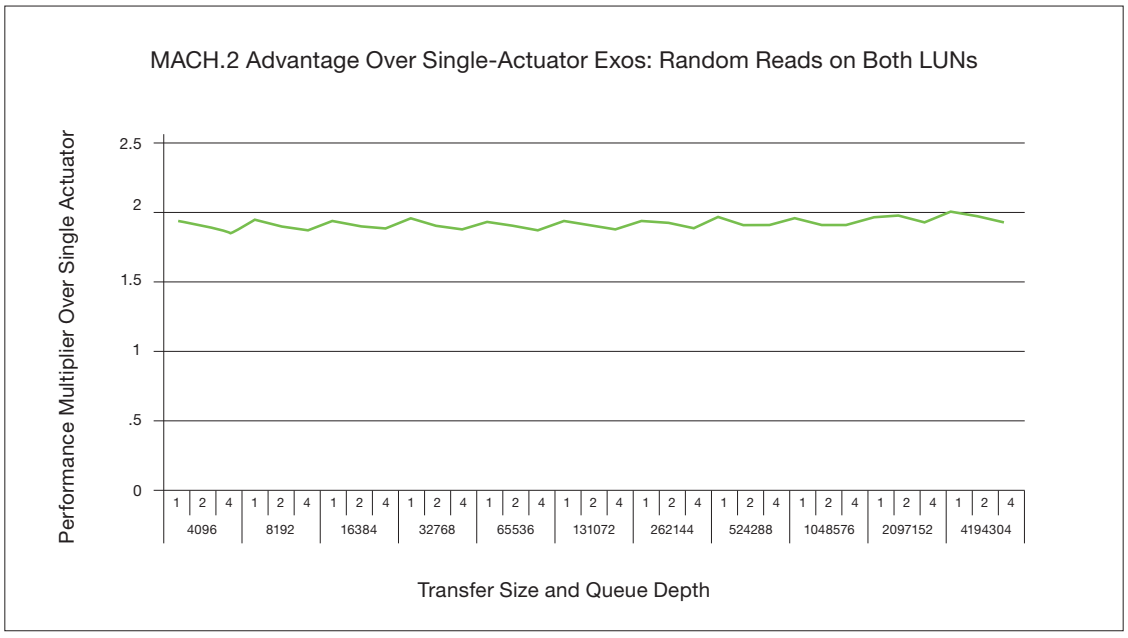
Seagate MACH.2 Conceptual Block Diagram

MACH.2 PERFORMANCE DATA— SEAGATE INTERNAL AND CUSTOMER TESTING

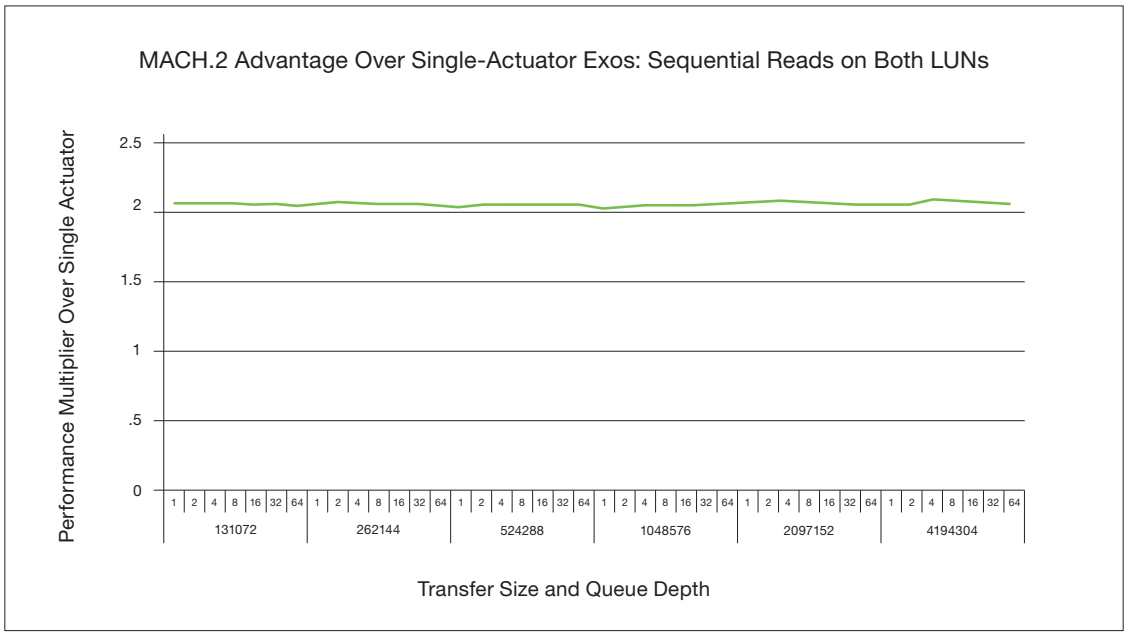
MACH.2 provides up to 2× performance for random IOPS and sequential data rate. In general, customer feedback has been to tune the following queue depths (QD) and transfer sizes (xfer) to maximize performance, since these are the most popular workloads for most applications.



Shown below are some real examples of IOMeter benchmarks in these areas of focus, comparing MACH.2 to a Seagate® Exos® drive (comparable single-actuator drive). The multiplier (benefit) of MACH.2 over single-actuator Exos is shown on the y-axis and QD and transfer size is shown on the x-axis. Note that the exact throughput advantage may vary depending on the model of HDD being compared.



For random reads, you can see that MACH.2 shows 1.85× to 2× the performance of a single-actuator Exos drive when running the same workload on both LUNs.



For sequential reads at higher transfer sizes (where both MACH.2 and single-actuator Exos are in full sequential streaming), you can see that MACH.2 shows 2× the throughput of a single-actuator drive when running the same workload on both LUNs.

POWER USAGE

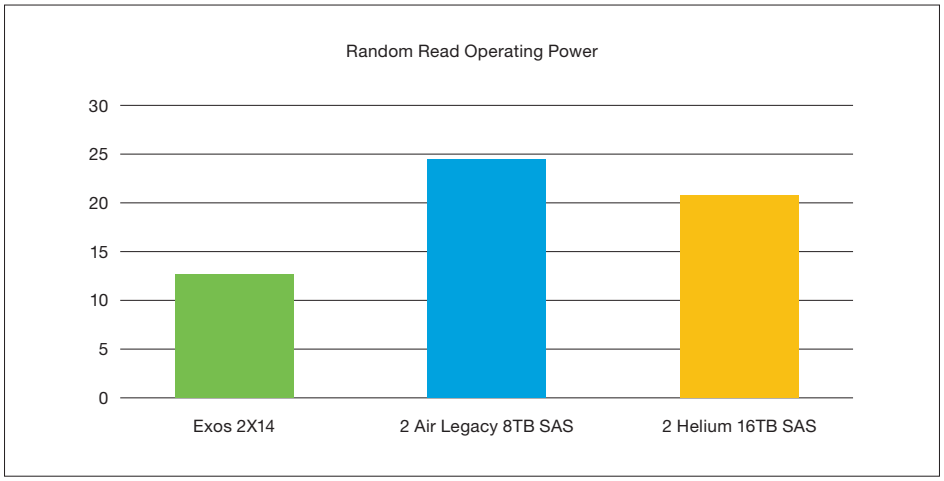
As expected, with more electronics and moving parts, MACH.2 HDD draws more power compared to a single-actuator drive. However, the design goal is to fit approximately within a 12W power envelope for random workloads that are most critical to a majority of cloud applications. The table below illustrates power comparisons between the current generation of Seagate Exos products.

	SINGLE ACTUATOR			MACH.2		
	Amps	Amps	Watts	Amps	Amps	Watts
Voltage	**5V**	**12V**		**5V**	**12V**	
Avg idle current DC	0.3	0.32	**5.3**	0.65	0.32	**7.1**
Peak operating current (Random Read 4K 16Q)						
Typical DC	0.5	0.64	**10.2**	0.85	0.67	**12.2**
Peak operating current (Random Write 4K 16Q)						
Typical DC	0.4	0.38	**6.6**	0.82	0.41	**9.0**
Peak operating current (Sequential Read 256K 16Q)						
Typical DC	0.88	0.31	**8.1**	1.31	0.57	**13.3**
Peak operating current (Sequential Write 256K 16Q)						
Typical DC	0.73	0.31	**7.4**	1.3	0.44	**11.6**

Looking at the above comparison, one will note that the highest power mode for dual actuator is sequential reads, primarily driven by the increase in 5V current over single actuator due to the additional electronics that are used when doing two sequential data streams.

Seagate has found that many existing multi-drive systems are in fact bandwidth limited by the SAS host topology and will not be able to support the full sequential bandwidth that multiple MACH.2 drives can output (520MB/s × number of drives). This means that for most customers, the highest power mode will actually be random reads, which has a modest power increase over single actuator in exchange for up to 2× the performance.

One can look at it a different way: although per-device power draw is more for a MACH.2 drive, such as Exos 2X14, at the TCO level it saves power since it provides a relief on the slot cost. Deploying one drive consumes lower power than deploying two drives to achieve the same throughput requirement. This is illustrated in the chart below.



MACH.2 COMMAND USAGE

SAS LUN behavior leads to ambiguity where some commands affect the individual LUN and others affect the device (both LUNs). It's important for users of the drive to note these differences to successfully deploy the drive. High priority commands (HPC) such as Read and Write are LUN-based. Low priority commands (LPC) are a mix of LUN and device-based. Some examples of the more impactful device-based commands are noted in the table below:

COMMAND	LUN/DEVICE	DETAILS
Test Unit Ready (0×00)	Device	Command will only report ready if both LUNs are ready
Format Unit (0×04)	Device	Format to either LUN initiates the data loss format of both LUNs
Start/Stop Unit (0×1B)	Device	Start/Stop Unit affects the single motor in the drive
Power Modes	Device	Idle A, B, C, and Standby modes are all device-based
Flush Cache	Device	Cache is shared, so this command will affect both LUNs
Sanitize	Device	Sanitize sent to either LUN sanitizes the entire device

You can query the device vs. LUN effect of each command on the drive by issuing the REPORT SUPPORTED OPERATION CODES command and noting the multiple logical units (MLU) field for each command.

Seagate is actively working on the inclusion of these nuances into a T10 proposal to standardize the usage of the MLU field on multi-actuator drives. You can access the T10 proposal here: http://www.t10.org/cgi-bin/ac.pl?t=d&f=18-102r1.pdf

MACH.2 FOR OTHER APPLICATIONS/INDUSTRIES

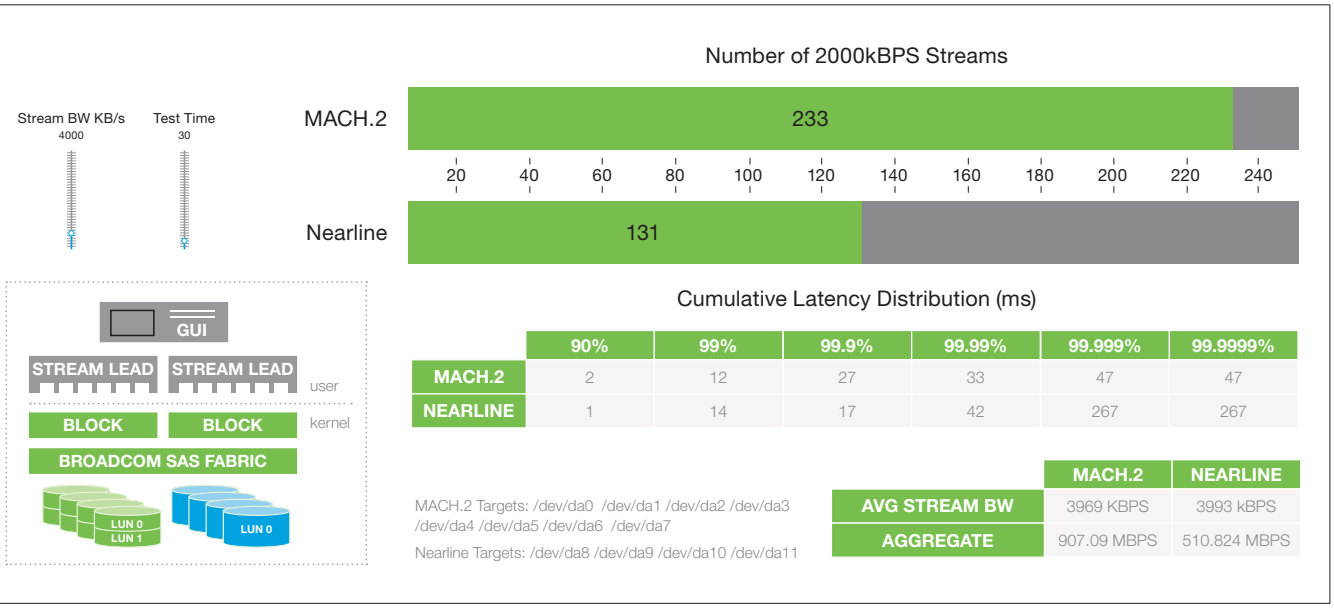
Although MACH.2 is designed to solve cloud-centric applications in data centers, its performance can be harvested for significant gains in other applications as well. Some notable applications are: content delivery networks (CDN, specifically online streaming services), redundant array of independent disks (RAID), Hadoop, backup/shuttle, surveillance, and scalable filesystems.

The following are examples of some Seagate demonstrated test results:

Online Streaming Services: MACH.2 performance gains helps customers who use HDDs for streaming digital content to improve their QoS (Quality of Service) to their existing customer base and/or expand the number of users they can service with the same deployed capacity. In this type of application, users will be streaming a variety of different video files at any moment in time. The commands issued by the streaming application/operating system (OS) are a series of sequential reads at various locations across the recording media. Because the file locations on media are not common, the HDD must seek to each file and read a portion of it before servicing the other streams; the resulting workload at a drive level is random reads of various transfer lengths. As shown in the MACH.2 performance data above, random reads consistently provide 1.8×+ the throughput of a standard single-actuator drive.

At the 2019 OCP Global Summit, Seagate demonstrated the benefit of a CDN configuration using a simple storage server using the FreeBSD OS. This demo created as many fixed-bitrate sequential non-overlapping read streams (using the BSD libc read system calls) as possible from an array of 4× 14TB MACH.2 drives and an array of 4× 14TB Exos NL SAS drives. Stream quantity was increased until the storage subsystem could no longer maintain the requested average bandwidth on all the streams. The max number of streams supported reflects the multi-stream read capability of the underlying storage.

The results of that demo can be seen below.



Number of 2000kBPS Streams

Cumulative Latency Distribution (ms)

	90%	99%	99.9%	99.99%	99.999%	99.9999%
MACH.2	2	12	27	33	47	47
NEARLINE	1	14	17	42	267	267

MACH.2 Targets: /dev/da0 /dev/da1 /dev/da2 /dev/da3 /dev/da4 /dev/da5 /dev/da6 /dev/da7

Nearline Targets: /dev/da8 /dev/da9 /dev/da10 /dev/da11

	MACH.2	NEARLINE
AVG STREAM BW	3969 KBPS	3993 kBPS
AGGREGATE	907.09 MBPS	510.824 MBPS

As you can see, MACH.2 was able to support 233 simultaneous streams compared to 131 streams on the single-actuator drives. That is a ~1.8× increase in throughput over the single-actuator solution, and, in addition, MACH.2 showed lower average and cumulative command latency distributions.

RAID: RAID is another application that shows the benefit of MACH.2 technology. RAID is a method of combining multiple storage devices into a virtual array for the purposes of data redundancy, performance improvement, or both. Because MACH.2 technology presents two logical units to a host, a RAID array can be created with as few as one dual-actuator HDD. That said, precautions must be made to not build redundancy within one MACH.2 drive without understanding the possibility of data loss.

One example of MACH.2 being used for performance gains is via a demo that Seagate, Microsoft, and Broadcom demonstrated at the 2019 OCP Global Summit. For this demo 11× MACH.2 drives and 11× single-actuator Exos NL SAS drives were loaded into a Microsoft J2010 storage system connected to a Microsoft C2010 server containing Broadcom 9400 host bus adapters (HBAs). Each group of drives had a dedicated SAS expander and HBA ports and they were combined into a simple striped Windows Storage Spaces volume (virtual disk) with no parity and no cache tier.

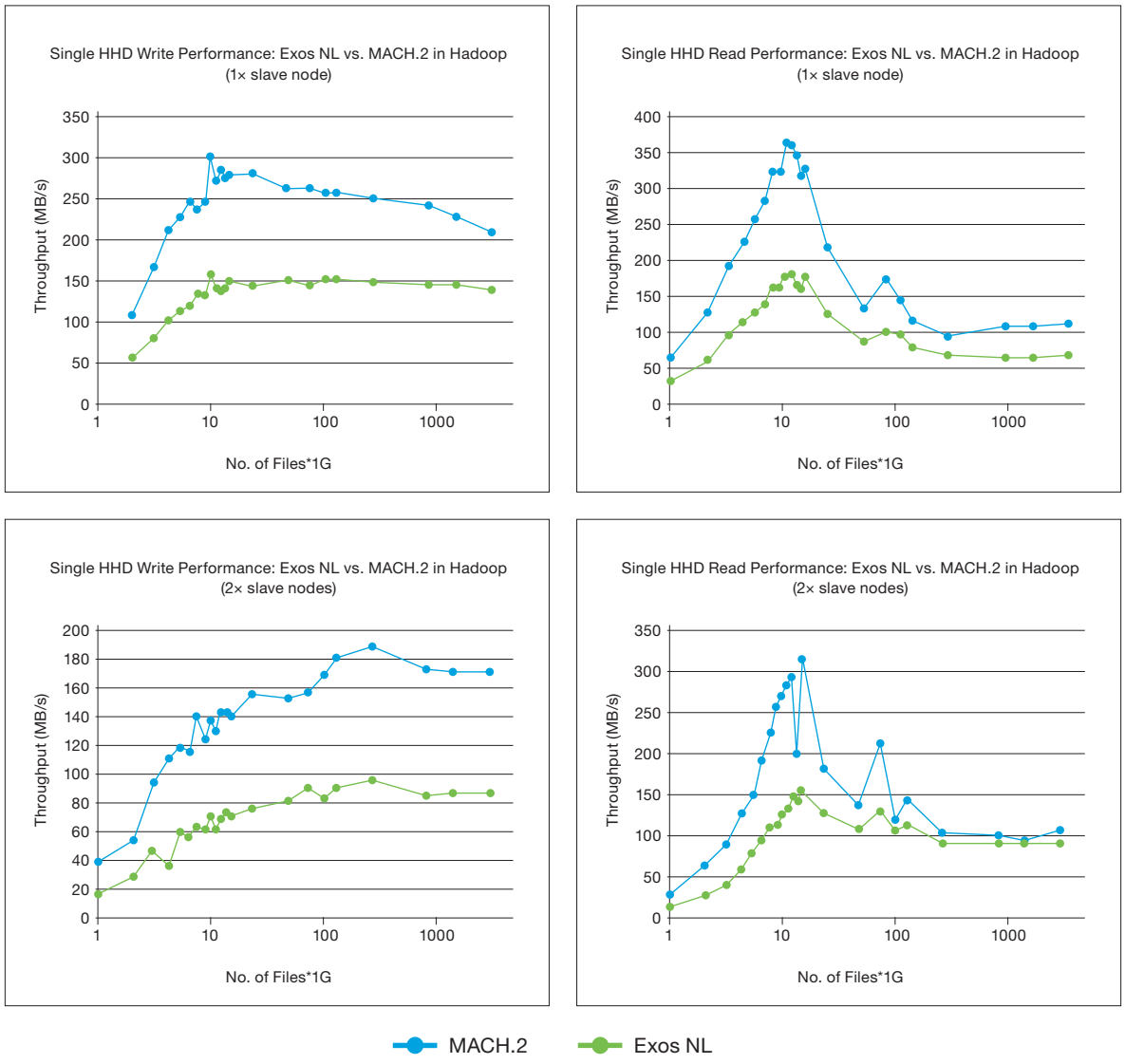
Both virtual disks were issued full-stroke 16K or 32K random reads while controlling queue depth to maintain a user-set target average latency (30ms in this demo), using Microsoft's Diskspd micro-benchmark. The QD applied to the virtual disks was increased until the average latency fell below the target. Average latency and cumulative latency distribution were graphically displayed in real time.

QD	IOPS	Average Latency (ms)	Throughout (MBPS)
35	**1517**	**23**	**47.43**
27	**1011**	**27**	**31.62**

You can see from the results above, the MACH.2 drives were able to produce 1517 IOPS at a QD of 35 with an average latency of 23ms, whereas the single-actuator drives produced 1011 IOPS at a QD of 27 with an average latency of 27ms. This demo showed a ~1.5× increase in throughput with lower average latency. Note that had the queue depth been driven a bit deeper on the MACH.2 RAID set (which would increase latency), you would see even more throughput and a greater advantage, but in this demo the main goal was to demonstrate MACH.2 performance gains while lowering command latencies.

Hadoop: Hadoop provides an efficient method for processing of data using the Hadoop distributed file system (HDFS) when combined with MapReduce (which provides the distributed processing for Hadoop). To demonstrate how MACH.2 can show a performance benefit using the HDFS, a small-scale demonstration was created by Seagate using three servers, forming a small Hadoop cluster. In this demo, one master node (server) is connected to either one or two slave nodes, each containing a MACH.2 or single-actuator Exos NL test drive.

In this demo we used the Hadoop TestDFSIO benchmark, which uses a MapReduce job to read and write files in parallel, to test the benefit of dual-actuator technology. In this test the file size to be read/written was 1GB and the test scaled the number of files from 1 to 1000+ while measuring the read and write throughput.



You can see from the above results that dual-actuator HDDs outperform a single-actuator drive in Hadoop MapReduce jobs for both writes (left graphs) and reads (right graphs) at a small scale and show the potential to outperform at larger scales. Future testing is planned to demonstrate results using larger quantities of MACH.2 drives and with more data nodes.

MACH.2 SUMMARY

MACH.2 technology provides a proven solution to accelerate data transfer while reducing overall command latency. As more and more applications encounter IOPS/TB constraints, MACH.2 technology will resolve performance constraints, reduce TCO, and help customers meet the growing demand for increased performance as defined in SLAs.

seagate.com