White Paper:

Understanding and Preparing for the Transition to 4K-Sector Disc Drive Designs

Summary

Across the storage industry, a move is underway from the traditional 512-byte sectors on hard disc drives to a new format whose sectors extend to 4096 bytes. The main reasons for this transition are the need for improved format efficiency and better error correction, which the longer sectors help to achieve. However, the 4096-byte sector formats also provide a number of new opportunities to enhance the performance and reliability of hard disc drives.

This paper highlights how Seagate is readying its enterprise-level disc drives to adopt the new format and assisting customers to accomplish a smooth transition with an emulation disc drive configuration that works within the 512-byte environment, yet delivers all the advantages of the new sector format. Seagate introduces new drive components, such as the non-volatile cache and media cache, to ensure the improved performance and data integrity of 4K-sector drives. Performance benchmarks illustrate the validity of the new drive designs. They generally perform significantly above what legacy disc drives with 512-byte sector can achieve. The paper showcases several test results.

Finally, the paper summarizes how the 4K-sector disc drives handle errors and failures, how they track and maintain the alignment of data and commands from the host with sectors on the drive media, and how they operate with standard commands.

To realize the best outcomes in terms of system performance and reliability as well as user experience, OEMS and systems builders might want to consider the implications of the 4K sector format and take advantage of the opportunities it presents. Systems that are 4K-aware in all their components will deliver the strongest results by using 4K-sector disc drives.

In this paper, best-practice suggestions for storage technologists and systems builders are prefaced by a green arrow: ➡.

Introducing Disc Drives with 4K Sectors

For decades, hard disc drives have written data to storage media in logical block sizes, known as sectors, of 512 bytes (often informally referred to as '5k'). With increasing drive capacities and larger data volumes to be stored on hard disc drives, this format is no longer efficient and practical. Seagate and the entire storage industry are transitioning to a new sector size of 4096 bytes, generally referenced as '4K'. Not only are 4K sectors more efficient, but the move to the new format also brings opportunities to increase the performance of hard drives. In its drive design, Seagate is taking full advantage of the efficiency and performance enhancements that become possible with 4K sectors.

This section of the white paper highlights some basic ideas related to the error correction, data flow, configurations, and design of Seagate hard disc drives with 4K physical sectors.

Better Efficiencies for Error Correction in High Areal Density Conditions

Recent years' dramatic improvements in hard disc drive technology make the industry's shift to 4K sectors a timely and valuable innovation. For one thing, Seagate and other manufacturers have advanced **areal density** to levels that were simply unimaginable at one time. However, with increasing areal density, even extremely small physical defects of the storage media have an ever-stronger impact on drive performance.

In 1999, when areal density amounted to 213,000 bits per inch (BPI), 30 bytes of error correction code (ECC) in a 512k data sector sufficed. More recently, with 1,100,000 BPI or higher, 90 bytes of ECC serve to protect the data. As 5k sectors require less and less physical space and the ECC has increased in size, the efficiency of the drive format declines and error correction becomes more and more challenging.

The move to 4K sectors corrects this unsustainable trend while enabling drives to deliver the capacity that the evolving industry and applications demand. Instead of 512 bytes, in the new format the ECC protects 4096 bytes, but the ECC itself only needs to increase by 50 bytes**. Format efficiency, which had gone down to 85 percent, receives a boost that elevates it to 96 percent.** That enhanced format efficiency amounts to an increase in the amount of capacity available for data storage. Reliability also improves, because the ECC can now correct larger error bursts—lowering the rate of unrecoverable-error incidents.

Native and Emulation Disc Drive Configurations

During the next few years, Seagate is phasing out all 512k drives and replacing them with drives that incorporate the new 4K sector format. Seagate will be offering 4K-native disc drives, meaning that the host transmits data and commands in 4K strings and the drives write them to storage media in 4K sectors. In addition, to make the transition easy and seamless for customers, Seagate also will ship 5k emulation-design drives, where the host communicates data in 512k segments and the drives manage them in such a way that they can write them to media as 4K sectors. In the emulation, the drives aggregate eight 5k data strings within a single 4K sector. Thus, three types of transfers can take place between the host and the drive:

- **5k native:** The host communicates in 5k format. The disc drive writes data to media in the same format.

- **Emulation:** The host communicates in 5k format. The disc drive manages the data and stores them in 4K sectors.
- **4K native:** The host transmits data in 4K strings; the drive stores them as such.

When a 5k emulation disc drive combines eight chunks of 5k data from the host for writing to a 4K sector, it eliminates the eight separate sections of sync bytes, ECC fields and timing gaps that are at the end of each 5k data block. Instead, the 4K sectors on the drive feature one single section of ECC data at the end, with all other supporting information in one section at the beginning.

Flexibility in Using 4K and 5k Block and Sector Sizes

For easier reference in this paper, 4K and 4xxx refer to the new sector formats on Seagate SAS drives, which can actually comprise 4096 bytes, 4160 bytes, 4192 bytes, and 4224 bytes. ➡ Customers will have the flexibility to use these drives in a variety of RAID configurations or use the additional sectors to add their own ECC to Seagate's. Given limitations due to their interface protocol and command structure, 4K SATA disc drives will only be available with 4096-byte sectors. Similarly, 5k and 5xx refers to SAS drives with host block sizes of 512 bytes, 520 bytes, 524 bytes, and 528 bytes, whereas 5k SATA drives will only come with host sectors of 512 bytes. ➡ System builders should keep in mind that the operating system needs to be aware of the drives' exact sector size, or data and sector alignment issues may result.

All sector sizes in Seagate drives will support the protection information (PI) feature with a common LBA count. PI allows the host to attach additional 8 bytes of error detection capability to a data string. Seagate supports this feature without reducing the capacity for user data.

Standard Configurations for 4K Disc Drives

Today already, Seagate offers a 4K native, high-capacity, 3.5" disc drive called 'Megalodon' to give customers an opportunity to try out the new sector format and establish their own best practices in integrating 4K native drives successfully into their systems. Between now and the second half of 2015, the complete product lines of Seagate's mission-critical, high-capacity, high-performance and near-line disc drives will shift to 5k emulation and 4K native designs. In the same time frame, Seagate will phase out the 512k-native drives and eventually stop supporting them. Once the transition is complete, just four main drive designs will remain for all drive capacities:

- **Configuration 1** – SAS. 5k emulation; supports all write performance and power-loss features.
- **Configuration 2** – SAS. 4K native; performance is comparable to configuration 1, but features additional power-loss features.
- **Configuration 3** – SATA. 5k emulation; supports all write performance and power-loss features.
- **Configuration 4** – SATA. 4K native; performance is comparable to configuration 3, but features additional power-loss features.

➡ System builders can change drives in configurations 1 and 2, reformatting from 5k emulation to 4K native or the reverse. To accomplish this, it will be necessary to download from Seagate a new bridge code as well as updated firmware before issuing the proper SCSI format command. The format command will need to be in the long form, because the entire drive from the first to the last logical block address (LBA) will have to be reformatted and verified. Such reformatting actions may take significant time—up to 8 hours. It is not possible to reformat SATA drives in the same way. Also, Seagate will not support reformatting drives from 5k native to 5k emulation or 4K native sector layout.

Purpose of the Emulation Design

The Seagate 5k emulation design eases the transition to the new format and minimizes the risk, cost and effort involved. 5k emulation provides backward compatibility with 5k native drives and the processes and applications they support. ➡ Seagate customers can, for example, use 4K drives to replace failed 5k native drives. Computer-manufacturing processes based on 5k formats can continue without time-consuming, disruptive adjustments. ➡ Storage solution providers can continue to deliver servers, storage arrays or other systems that are based on 5k formats, and do not need to convert all of their solutions immediately to 4K, which might mean shipping delays. Systems can easily adjust to the emulation format of eight concatenated 5k sectors, without making it necessary to redesign them for the 4K sector format. ➡In addition, without the emulation, storage technologists would have to build their own volume manager or partition manager tools to compensate for the misalignment of data and sectors in their applications. Seagate eliminates these concerns.

Understanding Alignment

In discussing native and emulation sector formats of hard disc drives, the notion of alignment is crucial. These are the basic alignment scenarios:

- **Full alignment:** The first logical block address (LBA) of data the host sends to the drive aligns to the beginning of a physical sector. The last LBA block sent by the host aligns to the end of the physical sector. There are no concerns related to performance or torn writes.
- **Front-end alignment:** The first LBA of data sent by the host aligns to the beginning of a physical sector, but the last LBA block does not align to the end of the sector. This requires a read modify write (RMW) operation.
- **Back-end alignment:** The first LBA the host sends to the drive does not align with the beginning of a physical sector, but the last LBA block aligns with the end of the sector. This, too, requires an RMW operation.
- **Front-end and back-end misalignment:** Both the first and the last LBAs of the data are misaligned with the beginning and end of the sector, requiring an RMW step to take place.

➡**Alignments and misalignments may require different intervention from systems administrators or storage managers.** To achieve best performance of the drives, storage managers should consider planning storage assignments within 4K boundaries or increments of 4K, a change in awareness and administrative practice from simply assigning space and relying on its proper allocation.

Data Flow and Data Protection in 5k Emulation Drive Configurations

In devices that make use of the 5k emulation, the host sends 5k data strings with an optional eight bytes of PI. In the Dynamic Random Access Memory (DRAM) flash memory of the drive, the firmware reviews the host command, analyzes the alignment of the transfer, determines whether an RMW step is needed, and verifies whether a cache or media cache hit is associated with the transfer. **The drive aggregates commands to reduce read modify write (RMW) operations.** During an RMW event, 5xx emulation drives first need to *read* the unaligned section of data from the media, combine (*modify*) it with the host data in the buffer, and then *write* the 4K-aligned sector to the disc.

In the drive's command queue management, the drive **retains the PI data on 5k-sector boundaries**. In addition, the 5k emulation drives also maintain an important data integrity feature on the 5k boundary, common to Seagate drives today. The emulation drives **generate the appropriate input/output error detection code (IOEDC) for the LBA**, an additional 2 bytes of data integrity information to ensure that data written to DRAM goes to the right logical block of the storage media and that data read back to the host comes from the proper logical block. When the drive reads the 5k data out of DRAM, it combines 8 formerly separate data transfers, packs them into a formatter for the write channel, generates the ECC, and writes the data correctly as a 4K physical sector.

In a random read request, when the host request extends to, for example, 10 LBAs instead of 8, these LBAs span two physical sectors. A read command to the read/write channel prompts the drive to read both physical sectors, move the data into DRAM, and check the ECC and LBA-seeded IOEDC. The six logical blocks not used remain in DRAM as read look ahead (RLA) data for possible cache requests from the host. The performance impact of reading two disc sectors is negligible.

When the host sends 10 LBAs to be written to media, the data likewise spans two physical sectors. The drive has to perform an RMW step to manage such writes correctly: It first reads all 10 host LBAs into DRAM, modifies them against existing data, and then writes the combined data to media into a 4K sector. In doing so, it also generates new ECC for the two additional LBAs.

The RMW operation requires an extra revolution of the disc for completion, resulting in a potentially significant performance impact. However, Seagate has developed a number of innovations to address performance concerns and avoid RMWs—discussed in the next section.

Systems Need to Become 4K-Aware

➨ OEMs, systems builders, and other storage technology providers making use of Seagate enterprise-level hard disc drives need to **understand the implications of changing from the traditional 512-byte sectors to 4K sectors**. As by and by the 4K drives replace their predecessors and Seagate customers acquire the new drives, they can take advantage of the 4K drives' new efficiencies and advanced performance to deliver more value and a better user experience to their own customers.

➨ To integrate 4K hard disc drives—either native or as emulation—most effectively into their solutions, systems builders need to make their entire systems 4K-aware. Some critical considerations for optimizing applications in the presence of drives with 4K sectors include the following:

- **Operating systems** must support either 4K-native sector sizes or alignment to 4K byte boundaries to avoid Read Modify Write events.
- **Third-party applications** need to be 4K-aware or support 4K-native sector sizes, or they may experience performance impacts under the new format.
- **Database systems** that bypass the file systems of operating systems, accessing disc drives directly and creating metadata based on drive sector size, may see a difference in performance.

- **Partitioning tools and replication software** must be 4K-aware for proper alignment and best performance.
- The **boot BIOS** has to be 4K-aware in order to read up the boot sector.

New Disc Drive Design Elements to Optimize 4K Operations

The main design innovations in Seagate's 4K-sector disc drives are the **non-volatile cache** and the **media cache**. This section introduces these drive components and provides an overview of how they operate to ensure optimal drive performance and information integrity.

How the Non-Volatile Cache Works

The non-volatile cache (NVC) is a portion of the DRAM that is **designed to facilitate performance mitigation and torn write protection in 5xxx emulation and 4K native disc drives**. The NVC, which incorporates NOR flash technology, is a logical, not a fixed, designation within the DRAM. It is rated for 100,000 write cycles, twice the capability of Seagate disc drive motors, and includes ECC to protect against single-bit errors.

When the host presents commands and data, the data goes into the command queue and into the NVC. The drive reports a positive status to the host, which can continue transmitting. Processing for any IO in the NVC takes place *after* the drive has returned the status response to the host. With the drive's write cache enabled (WCE) mode bit set to 0, storing data and commands in the NVC ensures that after a power interruption all the information will be available when the power comes back on—the torn write scenario.

The drive moves commands and data in the NVC along to the storage media only when the NVC is full, which is far more efficient than processing them as they come in. To maintain data integrity when powering down, the drives flush any data in the NVC to the more robust media cache (see below) or the main store, the drive's rotating media. After a *controlled* powering down, the NVC erases the data already written to the media when the system powers up again. When power comes back on after an *uncontrolled* interruption, the disc drive flushes all data still in the NVC to the media cache or the main store before erasing it. If the NVC is full of commands and data, this could add approximately 1.5 seconds to the drive's time-to-ready. Any failure of erase activity from the NVC is reported to the host as a SMART alert, and the NVC will then be disabled. The host can still issue read and write commands to the drive, but without the enhanced performance the NVC makes possible and without an available location to flush data to in case of an uncontrolled data loss.

For best efficiency, the NVC handles all writes that are smaller than 64Kb. Larger writes only pass through the NVC if they are *not* aligned *and* can fit into the available NVC space. Because of the tables and metadata present in larger write requests, it is more efficient to run them through the media cache or the main store. If the NVC is out of capacity, the drive will leave data and commands in DRAM outside of the NVC and not report their status to the host until the data is written to the main store.

Boosting Performance with the NVC

The NVC allows a performance enhancement through the management of queue depths. Under most operational conditions, the host issues commands with a queue depth of 1 to 4, sometimes reaching up to 8 or higher. As discussed, the drive stores commands and data in DRAM before processing commands in the disk side queue and writing the data to rotating media. *Without* the NVC, a low command queue depth would affect the efficiency of the firmware algorithm in determining the best location to store data and the most efficient seek operation. However, with NVC, the drive captures commands and data there, reports status to the host, and builds up a larger queue on the disc drive side.

The host, in the meantime, continues to send commands and data at the same low-intensity queue level, and the drive continues to gather them in the NVC. This intermediate step makes the algorithms for determining the next best seek and data location significantly more efficient. The drives can process commands based on locale, transfer length, and the sequential positioning of LBAs.

The NVC allows the drive to deliver a queue depth performance of up to 64, even if the host only operates at a queue depth of 1 to 8. This performance enhancement through queue depth management by the NVC has a strong impact on unaligned 5k writes, which require RMW operations and additional disc turns. Under most workloads, it eliminates the slowing of drive performance because of RMW events.

➡ Because the NVC builds up the queue depth for better efficiency while storing data and commands in a secure location, **the host can run in write cache disabled (WCD) mode and the drive still delivers the same performance as it would in WCE mode**. System builders can rely on the disc drive to play a more proactive role in maintaining data integrity, which might help to address concerns related to backup power, RAID controllers and costs on the host side.

Basic Media Cache Operations

The media cache (often abbreviated as MC) is a reserved area of space on the drive's rotating media that provides performance mitigation for 5k emulation drives to manage non-aligned requests from the host. 4K native drives don't need a media cache. ➡ Thus, when systems builders make their entire systems 4K-aware, data and commands will never move into the media cache. Instead, they will go directly from the NVC to the main-store media. That means the drive performs significantly fewer data-management tasks.

Typically, for efficiency's sake, the media cache is in the outer third of the rotating media, given that most users use approximately two thirds of their drives with data before they add more storage capacity. Media cache content is persistent across power cycles and resets. The writing of data to the media cache is sequential and very fast.

When the drive receives write commands from the host and the NVC becomes filled with commands and data, the disc streams additional commands and data to the media cache. When there is idle time, the drive searches through the metadata tables in the NVC and media cache, processes the commands present in the media cache, and writes data in the media cache to the main store.

The media cache does not minimize the user's LBA space—the available capacity remains the same. The selection of writes going to the media cache depends on alignment and transfer length. If an IO the host presents to the drive is at 64Kb or larger, *and* it is also 4K aligned, the drive will stream it directly to the main store. However, in case of a misalignment with a physical 4K sector, the drive routes those blocks through the media cache for later processing. In the case of such unaligned writes, if both the NVC and the media cache are full, the drive may have to process data and commands directly to the rotating media.

A component of the media cache, the **media cache table**, supports lookups to determine whether new writes and reads overlap with data already in the media cache. A working copy of the media cache table resides in the DRAM; the drive periodically saves a copy to its system area. The media cache table includes complete information on the data in the media cache, including time stamps, which host sent, host commands, and so forth. Through updates, the media cache table also tracks overwrites of new to already present data, writing only the latest copy to the main store. Metadata describing the write ranges in a given segment is written at the head and the foot of the segment, providing the essential journaling information that helps to reconstruct the media cache table after an unexpected power loss. If the media cache table in the DRAM becomes corrupted, the disc drive reconstructs the table and carries on with writing and reading data.

Maintaining Best Performance and Data Integrity with the Media Cache

Capabilities of the media cache include **power safe operation**, which involves recording and tracking the physical location of data stored in the media cache by means of metadata. This also provides a means to search the media cache for read hits or write overlaps when the host makes a request. Operating systems have handled these tasks for many years, but Seagate design allows for enhanced system performance by moving them to the level of the disc drive. In addition, the media cache invalidates obsolete data and refreshes them with new information, for example, in the case of multiple write requests to a certain host block.

In the media cache, user data and metadata sit next to each other. A failure in reading the metadata could result in the data of multiple sectors to become invalid. For that reason, Seagate designed the media cache with a number of features to provide uncompromised reliability. Special protection for the media cache includes:

- The media cache is a **separate zone with a relaxed areal density or BPI**. The probability of a media error in the media cache is even lower than in the main store. At the same time, the media cache uses the full effectiveness of advanced ECC, making it stronger because of lower areal density.
- The media cache access pattern results in a **single-sided squeeze** only, and there is no concern about adjacent tracks. Data is written to the media cache in sequential order on a first in, first out basis.
- In the media cache area**, super parity is implemented and always valid**. The disc drive only performs writes in blocks that are equivalent to the super parity block size.
- The **read after write (RAW) criteria for the media cache are more stringent than for other disc drive areas**. The disc drive scrubs the media cache sooner than main store locations to verify that data is still valid and does not present issues.
- The drive writes **two redundant copies of journaling metadata along with media cache user sectors**. This metadata only comes into play when the drive needs to recover from a rare, uncontrolled power loss. The regular media cache table exists in three copies on separate heads, just like all other system files and defect tables.

- As mentioned, **writes to the media cache happen only sequentially**, not at random, to minimize the impact of adjacent track interference (ATI) and side track erasure (STE). A guard band around the media cache protects adjacent user tracks from side track erasure.

The design of the IOEDC in Seagate's 5xx emulation and 4K native drives maintains this important data integrity feature in all enterprise drives. Today, in standard hard disc drives, the host sends a command, and the drive generates an LBA seed that it attaches to data as it goes into DRAM. When it reads the data out of the DRAM and into the main store, it checks the LBA seed to verify that it is pulling the right block of information from the DRAM before it writes it to media. In a read operation, the drives also checks the LBA seed to make sure it retrieves data from the proper disc location, then checks it again as it moves the data from the DRAM to the host.

With a media cache present in the drive, Seagate also introduces an **additional layer of data integrity validation**. As before, host commands result in an LBA seed, which moves into the media cache. When the drive reads data out of the DRAM, it checks the LBA seed in the media cache, checks it back into the DRAM, and verifies it one more time before committing it to the main store. When a read command from the host involves a hit of data in the media cache, the drive will read the data out of the media cache, checking the LBA seed before returning the data to the host.

Performance Assessments of Seagate 4K-Sector Drives Validate the Design Choices

In a series of challenging IOMeter and SPC-1C tests, Seagate verified that 5xx emulation and 4K native disc drives with an NVC can provide read and write performance beyond what today's drives can deliver. ➡ As storage managers, OEMs, and systems builders prepare the transition to the new drive format, they should keep in mind that drive performance depends on several changeable factors, including:

- **Percentage of alignment** of the host requests to the physical format of the drive
- **Queue depth** of host requests
- **Transfer size** of host requests
- Level of **workload** presented to the drive

In the IOMeter tests, Seagate engineers used the following alignments of host requests and physical sectors on the disc drives:

- Unaligned: 12 percent alignment, running IOMeter with no change to parameters
- 4K-aligned: 100 percent alignment to the physical 4K sector layout

➡ As the Seagate findings show, 5xx emulation and 4K native disc drives perform at their best when all the host commands and data strings align to 4K boundaries. As systems builders create solutions, in addition to drive-level alignments, they might also want to review the management resources they use to define partitions and volumes, and make sure that these tools can function in complete alignment with the beginnings and endings of 4K physical sectors. For example, even when IOs fully align to 4K sectors, but a storage technologist creates a partition that is offset from the physical sectors, every subsequent command could be offset by a certain amount.

Performance of Single Discs and Raid Configurations

Figure 1 illustrates a test of a 5k native drive (light blue), two different 5xx emulation configurations (red and dark blue), and a 4K native drive (brown). The Storage Performance Council (SPC) test[1] emulates a typical working environment with a combination of random and sequential read and write requests. Multiple streams of activity take place at all times during the test. The IOs concentrate on three disc drive locations. One of them receives mostly sequential commands, another one random commands, and the third one a mixture of both.

The diagram shows the average response times in milliseconds the drives achieved under increasing loads of IOs per second. **The emulation and 4K native drives accommodate significantly more IOs per second while still keeping average response times below 30 milliseconds.** For the three 5xx emulation and 4K native disc drives, **performance improvements amount to roughly 50 percent** over the 5k native drive.

[1] **Note**: SPC-1C tests performed by Seagate were not audited by the SPC.

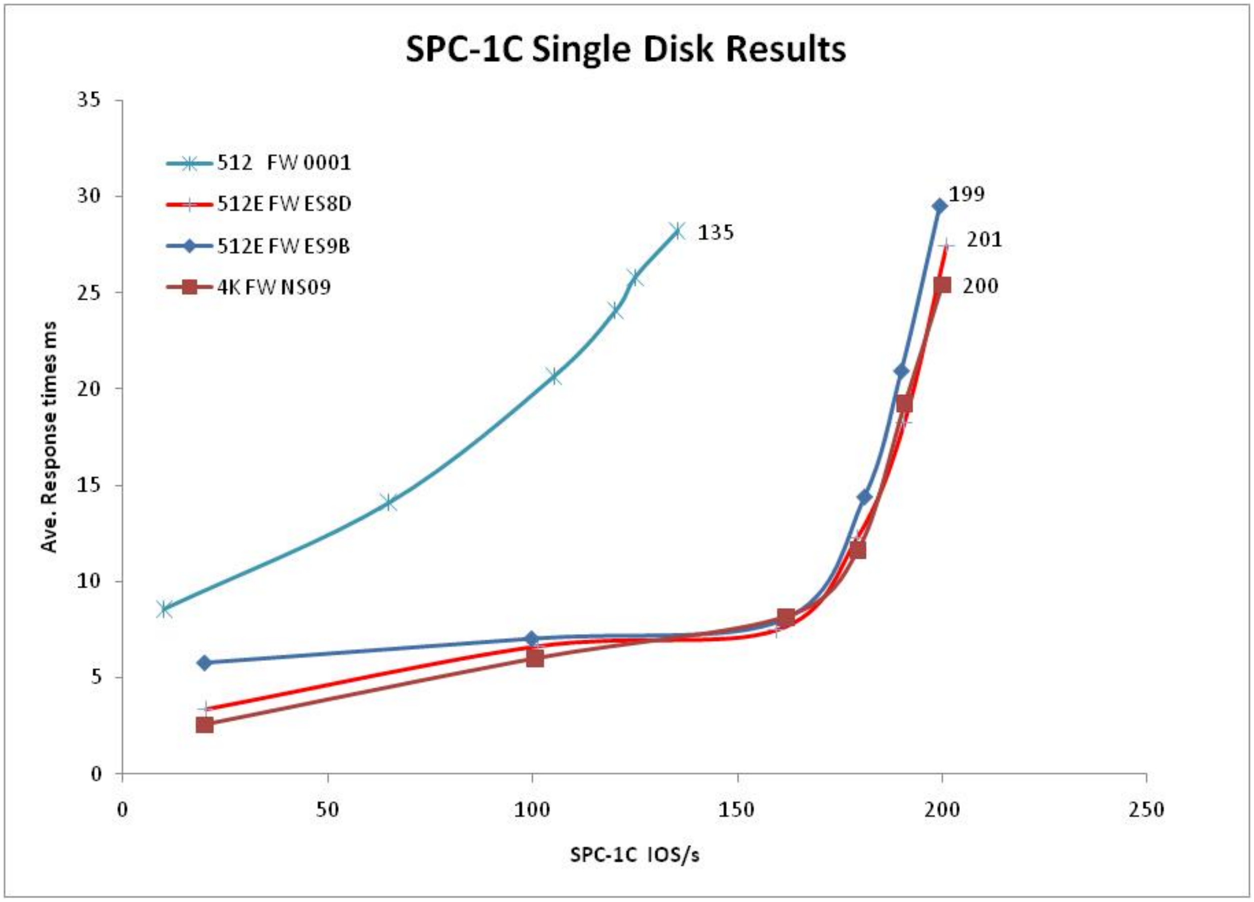


Figure 1. *Performance comparison of 5xx native, 5xx emulation, and 4K native disc drives.* ***Note:*** *Results were not audited by the SPC.*

In a similar SPC-1C test of RAID 0 systems with four-drive configurations of 5xx emulation and 5k native drives (see figure 2), the performance levels are very close, with minor improvement in the emulation drives' performance.

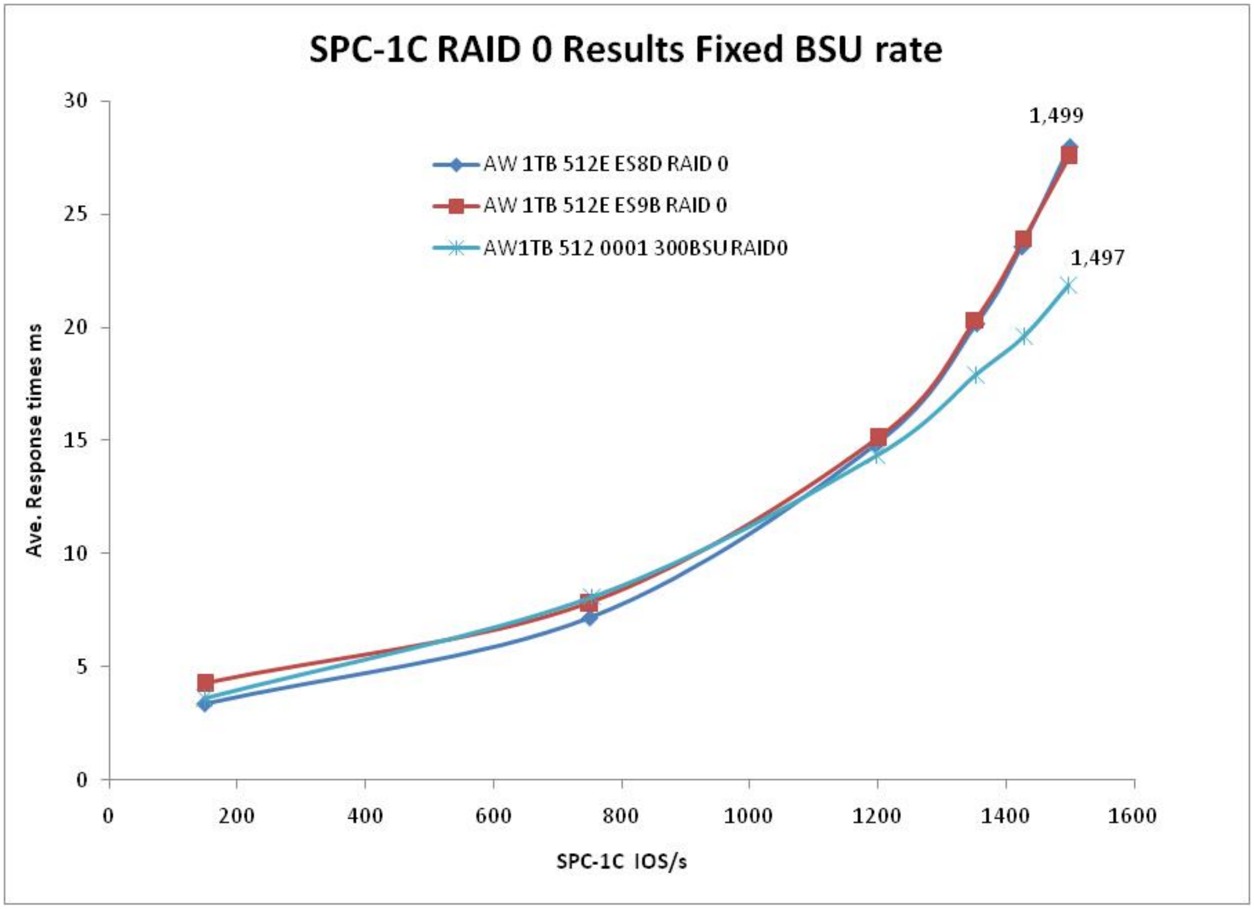
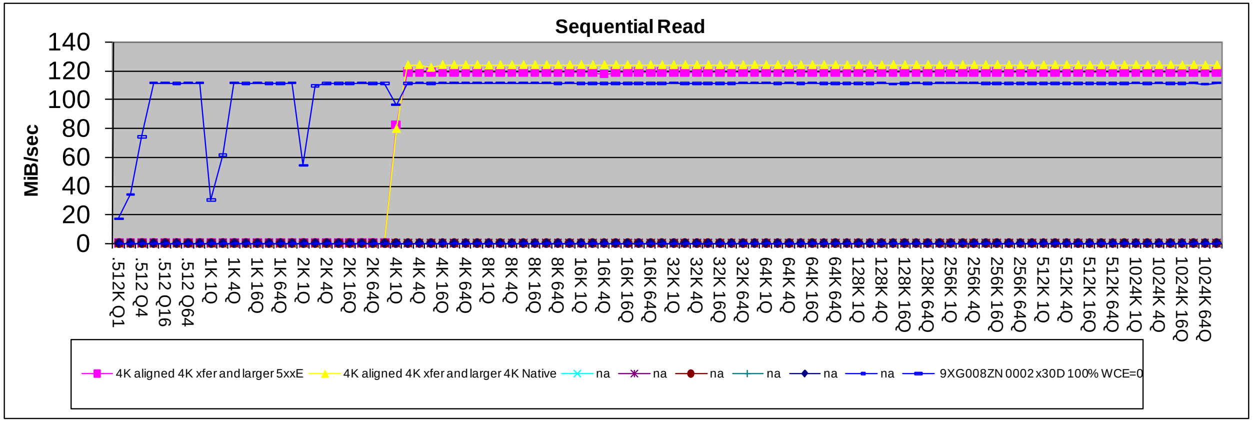


Figure 2. *Performance comparison of 5k native and 5xx emulation disc drives in a RAID 0 configuration.* ***Note:*** *Results were not audited by the SPC.*

4K Native Drives in Read Operations

Seagate also assessed the performance of 4K native disc drives in sequential and random read operations (see figure 3). In the test, both the 4K native configuration and the emulation configuration processed 4K aligned transfers and transfers larger than 4K. In the vertical dimension, the diagrams display MBs per second and IO operations per second, with transfer sizes increasing from left to right on the horizontal. The test demonstrates the **advantages of the enhanced format efficiency of the 4K-on-media drives, which translates directly into performance. In addition, the 4K native drives** (yellow line) **perform even better than the emulation configurations.**



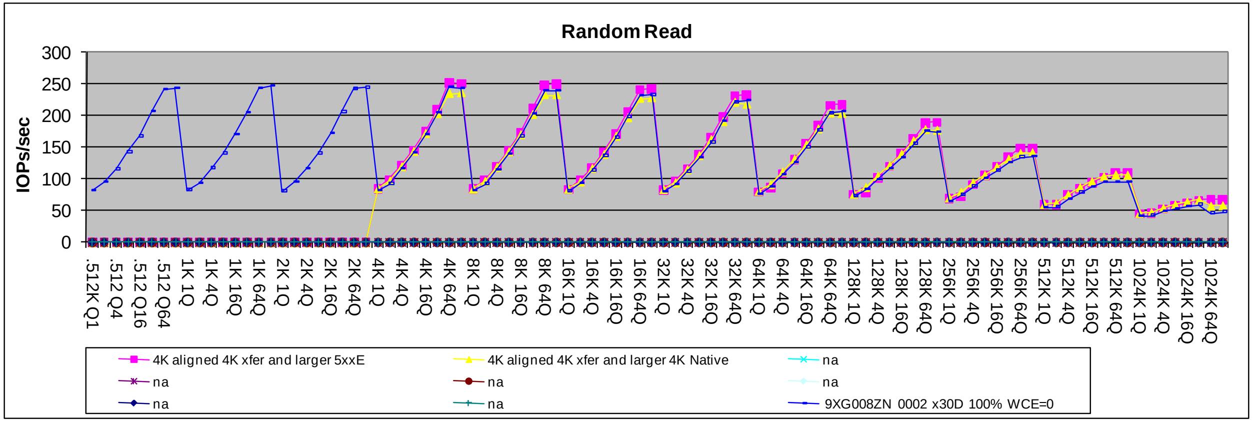
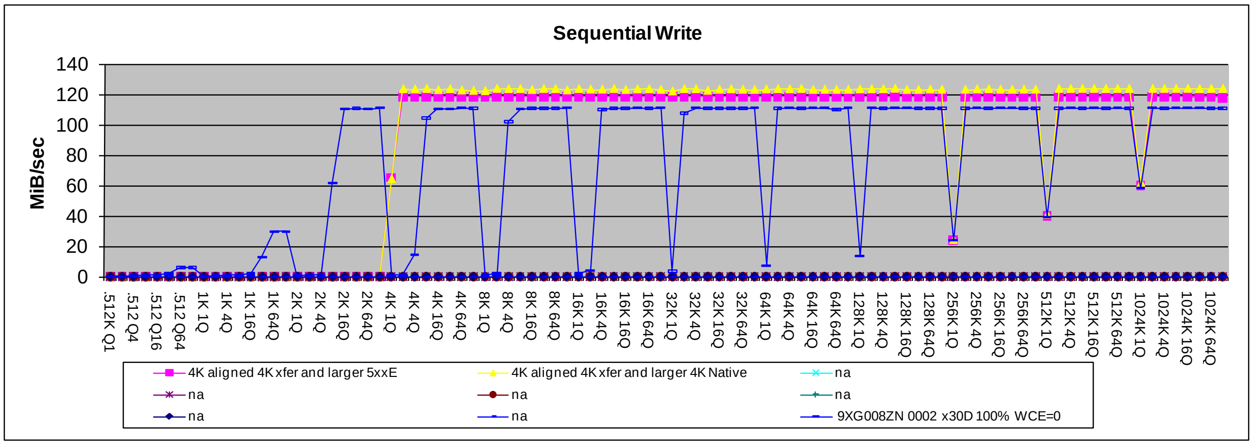


Figure 3. *Performance of 4K-on-media SAS drives in sequential and random read operations.*

4K-Sector Drive Performance in Write Operations

A similar test of sequential and random write operations of 4K-on-media SAS drives maintains comparable results (see figure 4). **The NVC, as discussed above, removes the inefficiency of low command queue depths from the host by aggregating commands and data before passing them to the drive media.**



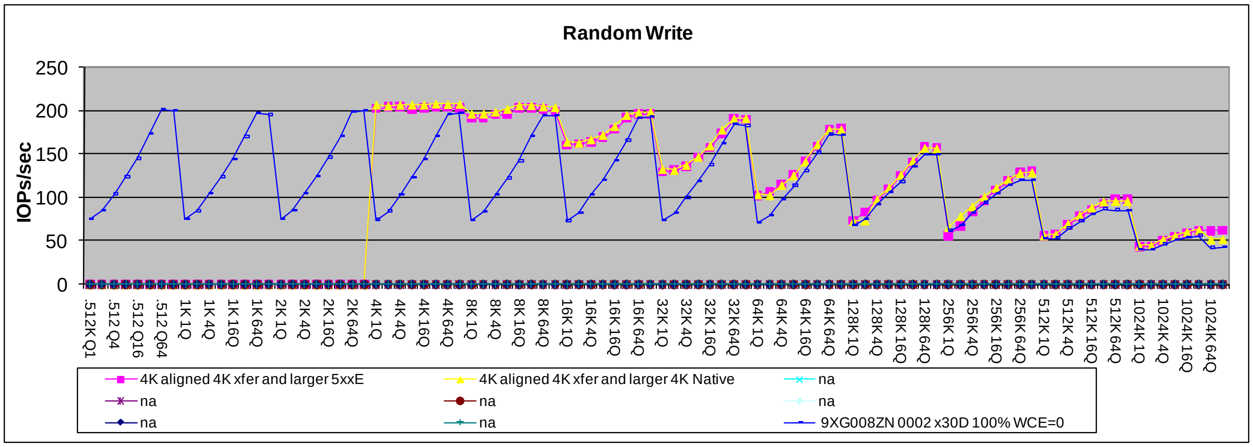
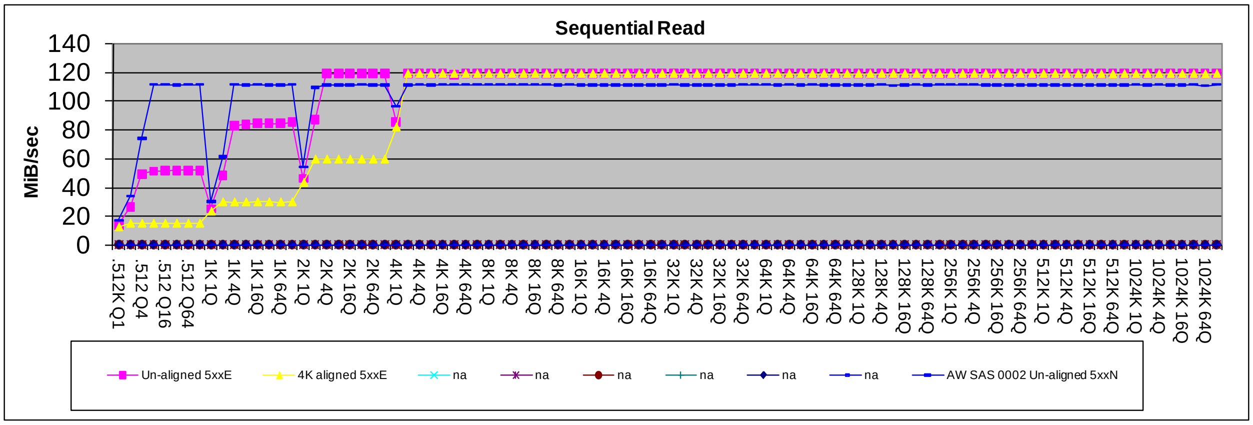


Figure 4. *Performance of 4K-on-media SAS drives in sequential and random write operations.*

5xx Emulation Drives in Read Operations

Figure 5 shows the performance record of 5xx emulation SAS disc drives in sequential and random read actions. **Performance is consistent for all alignment scenarios.** An increase in the sustained data rate is caused by the format efficiency of the 4K sectors. In this test, 4K-aligned transfers that are smaller than 4K result from hardware streaming. **4K-aligned IOs take best advantage of the drives' capabilities, delivering close to triple the conventional performance even at relatively low queue depths.** When transfers become large enough to make the rate of the data streaming a key factor in drive performance, the performance drops off to approximate what a 5k native drive would present.



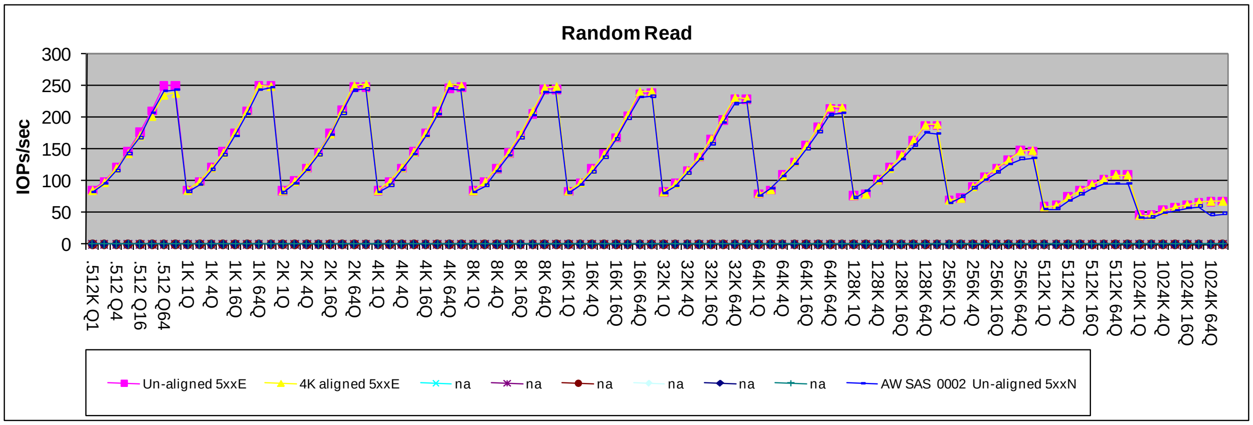
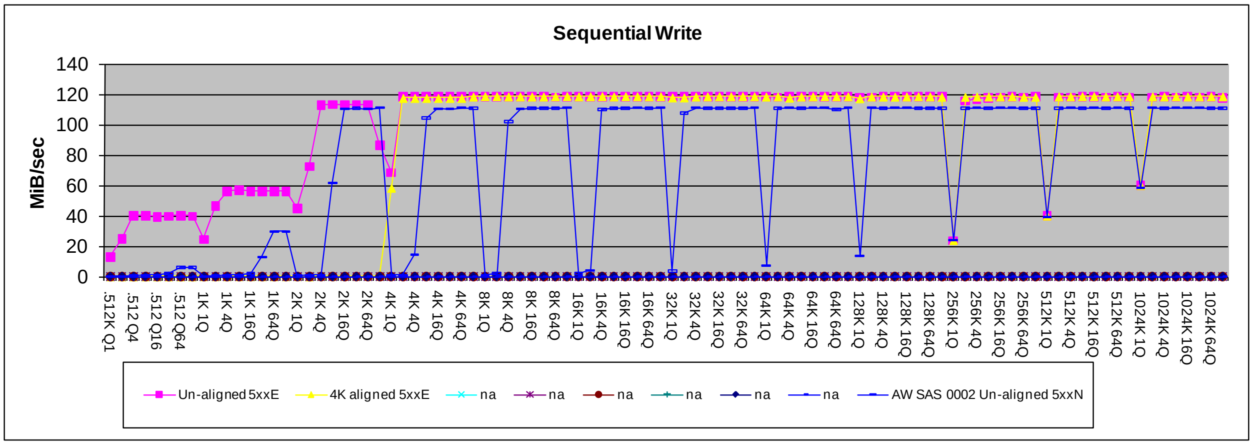


Figure 5. *Performance of 5xx emulation drives in sequential and random read operations.*

5xx Emulation Drive Performance in Writes

Sequential and random write operations of **5xx emulation SAS disc drives manifest similar performance benefits at low queue depths and deliver enhanced performance even with transfers that are smaller than 4K** (see figure 6). Even if the drive is presented with a large number of misalignments, it performs well above the 1 through 8 queue depth levels. ➡ In addition to 4K-alignment, understanding and managing queue depth is a key aspect in achieving best levels of system performance. Ideally, queue depths should not go above 16. With higher queue depths and misaligned data, drives have to perform many RMW cycles. As a consequence, first the NVC, then the media cache, may fill with data and commands.



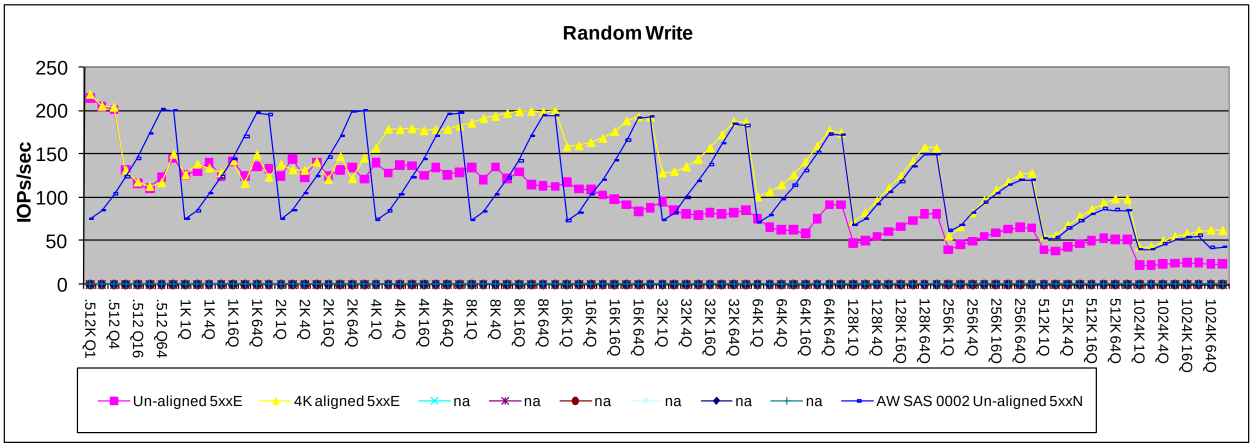


Figure 6. *Performance of 5xx emulation SAS disc drives in sequential and random write actions.*

4K-Sector Drive Performance and Duty Cycles

Seagate also tested the performance of the 5xx emulation drives at different duty cycles. With idle time after each data collection run, the drives were able to manage the media cache. The most likely real-world usage scenarios involve a 33 percent duty cycle. The tests used an unaligned IOMeter configuration with only 12 percent of alignment of host data to the disc drives' 4K format. Table 1 lists time delays and duty cycles used in this test series.

Time Delay (seconds)	Duty Cycle (percent)
0	100
7	74
10	66
15	57
20	50
33	37

Table 1. *Time delays and duty cycles Seagate used in duty-cycle performance testing of 5xx emulation disc drives.*

Figure 7 shows the performance of 5xx emulation drives in sequential write actions under various duty cycles. **The drives still deliver enhanced performance, and the duty-cycle workloads do not impact the sequential write performance.**

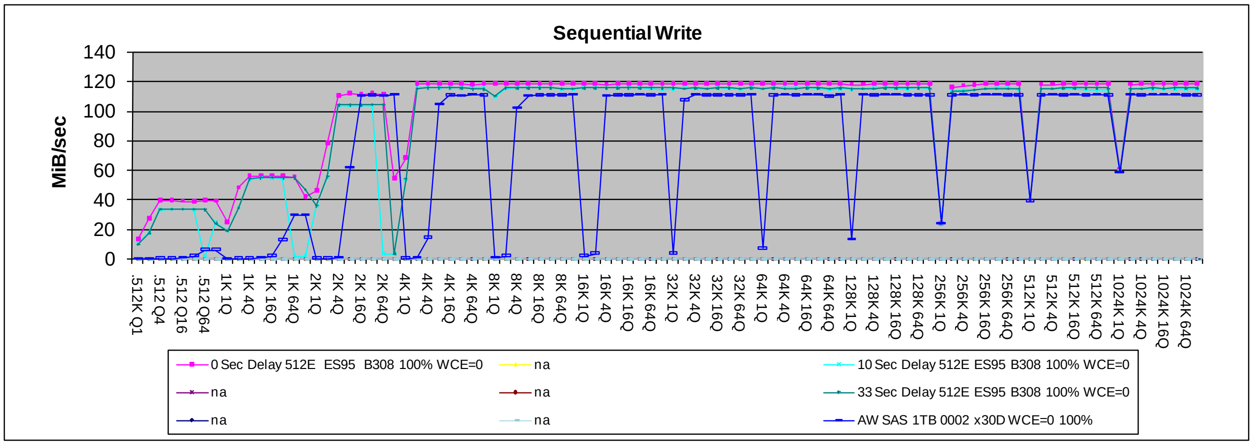


Figure 7. *Sequential write performance of 5xx emulation drives under varying duty cycles.*

As figure 8 illustrates, **the 5xx emulation drives also maintain enhanced performance during random write actions. Under a 37 percent duty cycle, the closest to the most common usage scenarios, the drives show significantly enhanced performance, with a consistent queue depth of 64.** At 66 percent, performance is still strong, and even at a 100 percent duty cycle, the drives present an advantage for many workloads, with host commands in the queue at all times.

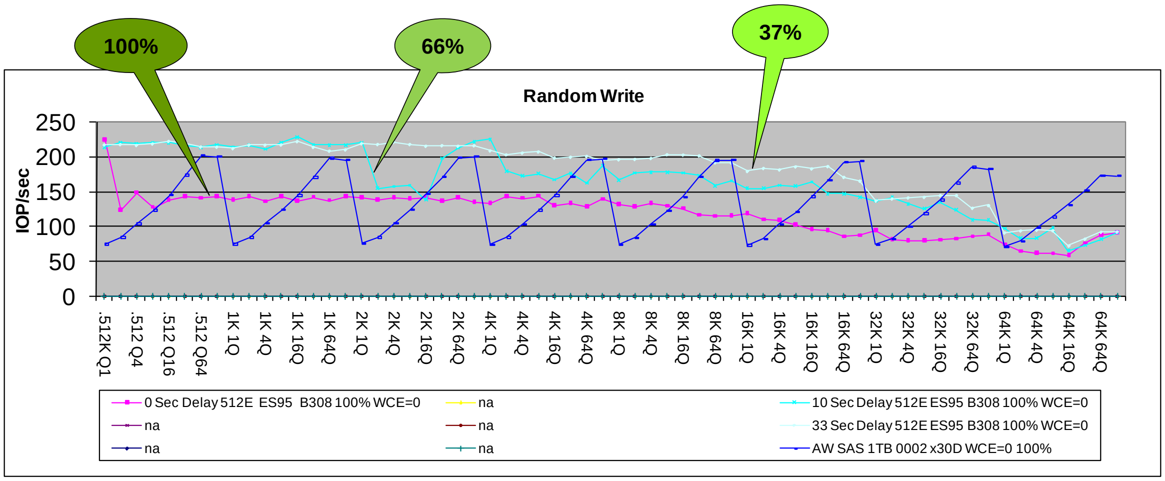


Figure 8. *Emulation drive performance under different duty cycles for random writes.*

➡ To make the most productive use of the 5xxemulation disc drives, OEMs and other solution designers need to **consider the actual duty cycles the drives are likely to experience under the most prevalent real-world usage conditions**. They should also keep in mind that the alignment or misalignment of the IOs presented by the host to a drive has a large impact in how different drives can perform their tasks.

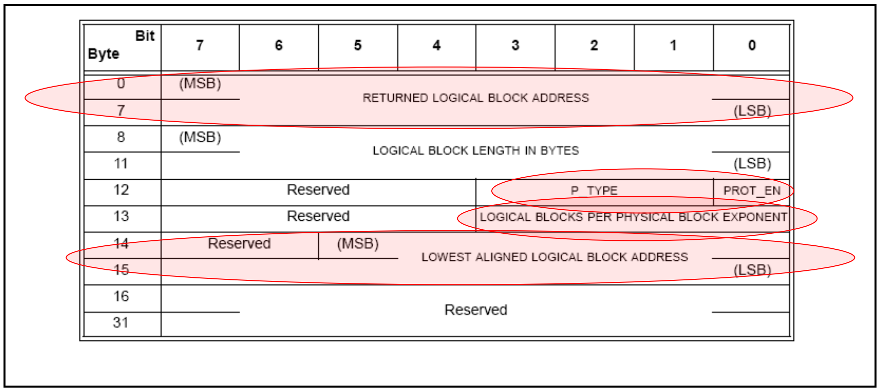
Alignment Tracking and Command Operations in the New Drive Designs

Design of the Seagate 4K-sector drives involves adding to and extending the capabilities of the firmware and hardware in many areas. This section presents more detail to highlight how 4K-sector disc drives allow the host to determine drive configurations, how SCSI commands are impacted by the new design, how log page 02h facilitates alignment tracking, and how the drives operate with low-priority commands.

READ CAPACITY Command – How the Host Determines the Drive Configuration

In SAS drives, **the READ CAPACITY command is essential when the host determines whether the drive is a 5xx emulation or a 4K native drive**. Seagate has consistently supported this same capability in its drives since 2004, and now takes it to the 5xx emulation and 4K native drives. In this process, the host initially receives an LBA of 8 bytes. It checks whether PI is enabled or not, and of what type it is if enabled. It reports multiple logical blocks per physical block in an exponential manner, so that a value of 3 means 2^3 = 8. It identifies the first LBA through alignment reporting, and returns data as shown in figure 9.

Mode pages 03h and 04h were traditionally used to report on the physical conditions of the drive to the host, but were made obsolete by the SCSI Standards Committee. Seagate still supports both mode pages, but does not guarantee the validity of the emulation in them.



Byte \ Bit	7	6	5	4	3	2	1	0
0	(MSB)	RETURNED LOGICAL BLOCK ADDRESS						
7								(LSB)
8	(MSB)	LOGICAL BLOCK LENGTH IN BYTES						
11								(LSB)
12	Reserved				P_TYPE			PROT_EN
13	Reserved				LOGICAL BLOCKS PER PHYSICAL BLOCK EXPONENT			
14	Reserved		(MSB)	LOWEST ALIGNED LOGICAL BLOCK ADDRESS				
15								(LSB)
16	Reserved							
31								

Figure 9. *How the host determines drive configuration.*

In ATA and SATA drives, the host uses the IDENTIFY DEVICE command, a 28-bit command, to determine the drive configuration. ATA word 106 is used to identify the format of the media and physical sectors. ATA word 209 is used to identify the offset value the drive is formatted with.

SCSI Commands Impacted by the New Disc Drive Designs

A number of SCSI commands are impacted by the new 5k emulation and 4K native drive formats, which deliver greater efficiency of drive operations. These commands include:

- **Format** – Using this command, it is possible to change the sector size in a full format action. The drive design supports PI with a common LBA count.
- **Block Translation** – In logical to physical translation, 8 host blocks can be mapped to a single physical block, and in physical to logical translation, the command will return 8 host blocks for a

single physical block. This departure from the previous one-to-one alignment is critical in error recovery.

- **Inquiry** – This command requires support of the NV_SUP bit. If there is a problem with the NVC, the drive’s firmware will cause this bit to be cleared out to tell the host that NVC is no longer enabled.
- **Read/Write** – It is now possible to use the FUA_NV and FUA bits in read and write commands. If the drive receives a write command with the FUA bit set, it writes the data to the main store before reporting completion of the write command. If the drive receives a read command with the FUA bit set, it will write the data to the main store if it was in the NVC or the media cache before delivering it to the host.
- **Synchronize Cache** – This command requires support of the SYNC_NV bit.
- **Verify** – The drives write all referenced LBAs to the main store before it can perform a verify operation. If the media cache houses data for any host block, the drives perform an RMW operation to read it, write it to the main store, and then perform the verify operation to make sure the data is valid.
- **Start Stop Unit** – This command supports the NO_FLUSH bit. With this bit set, if the host issues a start command, the drives power down without flushing data out of the DRAM. The drives always synchronize the NVC before executing on a stop unit command, because, when the drive goes into a power-save state or powers down, the drive heads may no longer be accessible to the media or there might not be enough EMF data available to write data to the flash memory.
- **Rezero** – This command forces all media cache to be flushed to the main store by means of the MCF bit. Seagate has added a bit to the zero command to enable the host to prompt the drive to complete all media cache cleaning and move all data from the media cache to the main store.

The **Skip Mask** and **XOR** commands are not supported by the 5k emulation and 4K native drives. Idle power and standby power events may now evidence small differences in behavior compared to the traditional drives.

Alignment Tracking and Checking with Log Page 02h

Seagate has **added fields to log page 02h**, the alignment counter log page, to track the alignment of host blocks to physical sectors and enable alignment checking for the first and last host blocks in each read or write command.

Whenever the host sends a write command to the disc drive, the following happens:

- One of the start LBA alignment counters (F800h through F807h of the physical sector) is incremented according to the alignment of the starting LBA. The drive checks whether this alignment is the case or not. The counters indicate the alignment of the host block to the first physical sector.
- One of the end LBA alignment counters (F810h through F817h) is incremented according to the alignment of the last LBA for that command. The drive increments these counters based on the alignment of the last logical and physical sectors.

- If *both* the start and end LBAs are aligned, then counter F820h is incremented as well. Thus, in case of a full alignment, three counters are incremented: F800h, F817h, and F820h.
- If a series of write commands is streaming (i.e., the *first* LBA of one command is one greater than the *last* LBA of the previous command), only the first LBA of the first command and the last LBA of the last command are counted in the counters. Sequential commands like this are merged by the drive and treated (internally) as if they were one command. Interrupting this process would reduce the write performance of the drive. Hardware streaming in Seagate drives lets the drives take over in detecting commands for sequential writes and LBAs on the drives. The drives report on the alignment of only the first and last host blocks. In sequential streaming, blocks between the first and last always fall on a physical sector, so only the first and last commands are important.

The LBA alignment counters in log page 02h *only* are reset by sending a LOG SELECT command to the drive with the:

- PCR bit set to one
- PAGE CODE field set to 02h
- SUBPAGE CODE field set to 00h.

➡This is a way to clear the counters and set them all to 0, a capability that can be useful when technicians make modifications or change alignments and need to test repeatedly to ensure that all of their IO requests align with the physical format of the drive. If the page code field is set to 00h instead, then *all resettable log counters in all log pages* will be reset.

Low-Priority Commands and 4K-Sector Disc Drives

Low-priority commands (LPCs) typically are diagnostic commands that can often be issued when the drive is in error recovery mode and performance is not the main concern. LPCs can claim large portions of the DRAM and thereby impact how quickly the drive can perform media cache cleaning. LPCs that need the entire DRAM must cause media cache cleaning to abort. ➡One can improve the cleaning performance by limiting such cleaning aborts to only the LPCs that truly require it. Here are some additional considerations regarding the interaction of the 5k emulation and 4K native disc drives with LPCs and the media cache:

- For **WRITE SAME**, the drive invalidates the WRITE SAME range (if present in the media cache) after the write completes successfully. If any blocks the drive is trying to access were in the media cache, it would invalidate all those blocks in the media cache because they would be in a main store location.
- For **REASSIGN BLOCKS** (SCSI), the drive reassigns the main store LBA no matter whether the valid copy is currently in the media cache. LBAs are moved between the media cache and the main store, and there is no assurance regarding which physical location is used for that LBA.
- For **SEND DIAGNOSTICS** for the Translate Address Page (SCSI), the drive returns the main store location. For any particular LBA where an address has to be translated, the drive translates to a main store, not a media cache, location.
- **WRITE VERIFY** (SCSI) always writes to the main store, invalidating data in the media cache.

- **VERIFY** pre-cleans and then verifies the main store. This command indicates that the drive is to write media cache data to the main store before completing the command.

Read Long and Write Long Behavior of 5xx Emulation Drives

Table 2 **illustrates how the 5xx emulation drives contrast with conventional drives when it comes to read long and write long behavior**. The first column lists the bits in mode pages that the host can control through the mode pages or a command field used by the drive. This list applies to SAS and SCSI drives only, not to SATA drives. Note that the drives can mark no more than 128 physical locations as bad (next-to-last row).

	Legacy Drive Behavior	Emulation Drive Behavior
COR_DIS=0, WR_UNCOR=0, Pblock=0	Host sends read long command	Host sends a read long command.
	Host sends write long command with data	Host sends a write long command with data.
	Drive writes the data to media	The drive ignores the data sent by the host. The drive reads the 4K sector from the media with a metadata value 'corrupt' and writes back to media.
	Drive could only get write errors	The drive could report read and/or write errors.
COR_DIS=0, WR_UNCOR=0, Pblock=1	Host sends read long command	Host sends a read long command.
	Host sends 5xx+ bytes of write long data for a 5xx legacy drive or 4K+ bytes of write long data for 5xxE and 4K native drives	Host sends 5xx+ bytes of write long data for 5xx legacy drives or 4K+ bytes of write long data for 5xxE and 4K native drives.
	Drive rejects command with check condition 05/2400.	The drive writes the data sent by the host to the media and creates a "correction mask" for the first 128 write long commands. Any request to create more than 128 recovered ECC errors results in the drive converting the request to an unrecovered error.
	Drive could only get write errors	Drive could report read and/or write errors.

- ***Table 2.*** *Response of 5xx emulation drives to write long and read long commands.*

Error Handling and Correction in 4K-Sector Disc Drives

With the NVC and the media cache, the new architectural elements on the emulation and 4K native drives, Seagate 4K-sector drives have new opportunities to mitigate errors and failures extremely efficiently. However, outside of these efficiencies, the error-handling sequences are very similar to the operation of today's 5k native disc drives.

When a 5xx emulation drive detects **defects on a physical sector**, it reports the defect to the host, which then will reassign that 4K sector, requiring reallocation of 8 host blocks. Several circumstances can prompt such reassignment:

- A read or write command from the host when the auto write reallocation (AWRE) bit or ARRE bit are enabled
- A reassign command from the host
- Background tasks performed by the drive (BMS, DOS)

When a reassignment happens, the reassignment command will always move all 8 host blocks to an alternate, spare physical sector, and copies the data to that sector if it is readable. Also, the BMS log page 0x15 reports only the first host block of the physical sector, and reassignment of a physical disc sector results in just 1 G-list entry.

The following tables, read from left to right in the sequence of events, note some more detail on how 5xx emulation drives handle errors.

Table 3 shows **how the drives respond to errors when the host issues write commands.** The status column indicates what the drive reports to the host. "Sent" means that the drive has reported to the host that it has received the command and data and is processing them through the NVC. "Not sent" means the drive has not responded to the host and still has the option to report an issue. This is a contrast to conventional, 5k native drives, where the host would receive updates through its visibility of the hard disc drive's internal status.

Host Action	HDD Internal Event	HDD Internal Status	AWRE	PER	Status	Error Code	Notes
Write command	Error on the read portion of RMW	Unrecovered read	0 or 1	0 or 1	Sent	No error reported	Reallocate the physical sector and "harden" (i.e., write filler data and keep a record for the failure location in the defect table) the host blocks for which data is not available. If host data is available for any of the host blocks for the disc sector in error, the data is written to the new (reallocation) location.
				1	Not sent	01/0C/01/01	
				0	Not sent	No error reported	
		Recovered read	0 or 1	0 or 1	Sent	No error reported	If the data for the full disc sector is available using error recovery on the disc sector, this would be a
				1	Not sent	01/0C/01/00	

				0	Not sent	No error reported	recoverable error. The sector is reallocated per the normal reallocation algorithm. No additional status is reported.
	Failure to write NVC-cached data to media	Unrecovered write	0 or 1	0 or 1	Sent	No error reported	Consider AWRE and DAR always ON, irrespective of mode-page values. In other words, while committing the NVC data to media, any write errors (or read errors on RMW) will result in reallocation, even if AWRE and DAR are off.
		Unrecovered Write - Fail to Reallocate	0 or 1	0 or 1	Sent	[SAS] 02/0400/00 [SATA] Device not ready - 0x10 instead of 0x50 (in FIS34).	If the reallocation fails (fail to write re-map table), the drive performs an internal reset, which causes the drive to retract the heads and write the NVC data to flash. After the power-on-reset, the drive performs a head (write) test. If the test passes, then the drive can recover the data from NVC normally and proceed. If the test fails, the drive comes up as not ready (02/0400/00) After three attempts the drive will report 03/3100/0A.

Table 3. *How 5k emulation drives handle errors subsequent to write commands.*

Table 4 shows **how the 5xx emulation drives make use of their NVC and media cache to manage errors that can take place during power-on and power-down sequences**. Error codes here are different for the SAS and SATA drives. Note how the drive addresses the error in the last row, when it experiences a power loss during the write phase of a RMW operation. It uses back-EMF from the drive motor to write the data to the NVC, providing effective torn write protection. Conventional drives would not be able to maintain data integrity in the same way, and, because of the drives' inability to read all the data, would report a new ECC error each time they attempted to read it.

Host	HDD Internal Event	HDD Internal Status	Error Code	Notes
Power-On	Error in NVC committed data	Checksum error on the user data	[SAS] 06/2901 and then 06/2904/03 [SATA] No error reported - log this in CE Log	Punch holes (mark these as uncorrectable in the defect table) for the sectors in error. Also, the sectors are marked for deferred auto-reallocation. Reads on these sectors, until they are overwritten by new host writes, will return unrecovered read error.

		Metadata is corrupt	[SAS] 03/3100/0x [SATA] Device fault condition (or see what we do with current format corrupt on SATA)	Drive will show a volatile format corrupt. The data is retained in NVC and the drive will try to recover on subsequent power cycles. If any of the subsequent attempts is successful, the format corrupt condition goes away and drive is usable. (See the FRU code document for the complete list.)
	NVC disabled	Erase-in-progress is present.	07/2700/00	If the erase on NVC on power-up fails, the drive writes 'erase-in-progress' status to disc. If an erase for the media cache or WCD data area fails, the drive cleans the media cache before coming ready.
	Fail to read media-cache data	Unrecoverable error on user data	None	Errors on the user sectors are ignored during power-up (will be handled later when cleaning the media cache). The sectors are marked for deferred reallocation.
		Unrecoverable error on metadata (both copies)	[SAS] 03/3100/0x [SATA] Device fault condition (or see what we do with current format corrupt on SATA)	If an unrecoverable error is encountered on the first copy of the metadata, the drive attempts to read from the copy. If the read from both copies fails, the drive will show a format corrupt.
Power-Down	Flash write failure	Not enough power to complete burn of flash	N/A	The firmware maintains a 'burn-complete' indicator that is written to the NVC. This will be detected at next power-on and handled per the 'power-on' failure description.
	Power loss during the write of RWM	Incomplete write	None	The write sectors that are in the pipeline (not validated by servo system as being written) are saved to the flash using back-EMF and are restored to the media at the next power-on (torn write protection).

Table 4. *How emulation drives use their NVC and MC to manage errors during power-on and power-down.*

Table 5 illustrates **what happens if the NVC were to fail**. There is no host action associated with NVC errors.

HDD Internal Status	Error Code	Notes
Temperature is below 5°C or above 65°C	01/0B06/00	There will be a significant performance degradation when the NVC is disabled. A warning is reported (similar to temperature warning) when the NVC is disabled due to temperature.
Erase-In-Progress is present.	07/2700/00	If the erase on the NVC at power-up fails, the drive writes 'erase-in-progress' status to disc. The drive posts a 'write protect' error code.
Running out of spare sectors (RST) for NVC	01/5D00/09	If the number of spare sectors falls below the required threshold, the NVC is disabled. Data in the NVC is flushed to the media cache. The media cache is effectively disabled, because no new data is moved into it.

Running out of margin to write full NVC	01/5D00/0A	The drive will determine if it is running out of margin to complete the NVC burn in a power loss. It will disable the NVC and generate a SMART trip if the critical threshold is reached.

Table 5. *How 5k emulation drives manage NVC errors.*

Table 6 summarizes the **actions taken by the 5xx emulation drives if the media cache fails**.

Host	HDD Internal Event	HDD Internal Status	Error Code	Notes
Write	Media cache write failure (from NVC)	Media cache reallocation - Success	None	The drive attempts reallocation within the media cache, regardless of AWRE or WCE settings.
		Media cache reallocation - Failure	[SAS] 02/0400/00 [SATA]Status field - device ready bit is cleared. Status will be 0x10 instead of 0x50 (in FIS34).	If the reallocation fails, the drive creates an internal reset, which causes it to retract the heads and write the NVC data to flash. After the power-on-reset, the drive performs a head (write) test. If the test passes, the drive can recover the data from the NVC normally and proceed. If the test fails, the drive comes up as not ready.
	Main store write failure (during cleaning)	Main store reallocation - Success	None	The drive attempts reallocation within the media cache, regardless of AWRE or WCE settings.
		Main store reallocation - Failure		If the reallocation fails, the drive creates an internal reset, which causes it to retract the heads. After power-on-reset, the drive tries to clean again.
Read, media cache cleaning	Media cache read failure	Recoverable error on user data	None	The drive recovers errors in the media cache with valid super-parity for all MC segment writes.
		Unrecoverable error on user data	[SAS] 03/1100 [SATA] status field=51 (error bit is set) ; error field = 0x40 (uncorrectable bit is set)	The drive hardens the failure, meaning it writes a fill pattern indicating an unrecoverable error on the main-store location and updates the MC table to indicate that the data is invalid in the media cache. The media cache range with the unrecoverable error is marked for reallocation so that on subsequent MC writes the location is skipped. A media defect most likely causes this failure; the location should not be used again).
		Unrecoverable error on metadata	None	The drive reads the metadata so that can DAR or reallocate the location – it is not used for media cache cleaning during normal operation. The meta-data is used only on power-up (see the power-on failure handling in table 2).
	Failure to write MCT	Unrecoverable error on write of MCT	01/5D00/64	The drive only attempts to write the first copy. It disables the media cache and forces a cleaning of the media cache. It uses the second copy of the data (if needed) during the next power cycle.

Table 6. *How 5xx emulation drives handle media cache failures.*