TOSHIBA
Leading Innovation

4K Sector Disk Drives: Transitioning to the Future with Advanced Format Technologies

By Michael E. Fitzpatrick,
Engineering Fellow, Storage Products Marketing
Toshiba America Information Systems, Inc.

Overview:

Trend: The computer industry is moving to a 4,096 (4K) byte sector size for hard disk drives—called Advanced Format. Some older operating systems that are still in use today expect a legacy 512- byte sector size, and thus need a bridge to the new sector size.

Today's Solution: Support the emulation of 512-byte sectors within the new larger 4K sector by allowing the 4,096 byte sector to be transferred as eight 512-byte blocks of data.

Future Solution: Operating systems will transfer 4K data blocks that match the 4K native sector size on the disk.

Some Historical Information:

In 2006, the storage industry celebrated the 50th anniversary of the first hard disk drive (HDD), the IBM RAMAC. The RAMAC HDD had a capacity of 5 megabytes (approximately 5,000,000 bytes); it was the size of a side-by-side refrigerator and weighed hundreds of pounds. The first RAMAC HDD used 50, 24-inch disks to achieve 5 megabytes of storage capacity. All for a price that only the largest companies could afford.

Compare the RAMAC to today's disk drives – consumers routinely require 500 gigabytes (100,000 times the RAMAC capacity) in a 2.5-inch form factor that fits in the palm of your hand, weighs about 4 ounces and is readily available in electronic stores for less than $100.

Over the past 20 years, we have seen the physical size of drives shrink from a 5.25-inch form factor to a 3.5-inch form factor to today's 2.5-inch, and even to the tiny 1.8-inch form factor. During this same time frame, we have also observed the capacity of HDDs doubling every 12 to 18 months.

With each new generation of HDDs, power requirements and weight have decreased, while overall performance and capacity have increased. One of the biggest changes has been in the price of today's HDD. Over this same period, the price has gone from dollars-per-gigabyte to today's pennies-per-gigabyte.

The previously mentioned changes have happened through technological innovations, some evolutionary, some revolutionary. Each new generation of HDDs has seen some change. This change could be a new technological implementation or an enhancement to an existing design in order to increase capacity, reliability or operational efficiency. What is the bottom line? The HDD industry is constantly looking for ways to improve how things are done.

How Did We Get Here?

In 1981, with the introduction of the first IBM personal computer (PC), the PC-DOS operating system with 512-byte sectors was introduced. Each HDD has a number of sectors based upon the formatted capacity of the HDD. Each one of these sectors has a unique address so that the operating system (OS) can write to or read from a known location to keep track of all the data files stored on the HDD. The file system part of an OS, such as Microsoft's Windows®[1], will keep track of all the file names and where each file is located. It will also keep track of the length of the file.

An HDD advertised as having a 500 gigabyte (GB[2]) capacity, will have 976,773,168 512-byte sectors, known as "Logical Blocks." Given how much HDDs have changed over the past 20 years, it is surprising that the size of the sector has remained the same. Since the introduction of PC-DOS and its 512-byte sector in the early 1980's, we are still using a 512-byte sector on the disk to store our data.

The Need for Change:

When writing data into a sector, HDD manufacturers are concerned with protecting data against corruption. Each physical sector, independent of data size, has some overhead associated with it. This overhead is used to: 1) provide unique identification for each sector; 2) provide information to identify the start of the sector; and 3) provide data protection information in the form of error checking and correction (ECC) code.

The ECC information for each block of data is derived from the data field in the sector and is therefore unique and automatically created when data is written to the disk. When the computer reads a stored file, the ECC of each block of data is checked to determine whether any errors are present. If an error is encountered, the ECC information can be used to correct the error before the data is passed to the user.

When necessary, data correction is performed by the processor on board the HDD. If an error is too large for the HDD to correct using the ECC information, then the HDD flags the data so the computer knows the data is possibly in error. See Figure 1 for a description of a sector.

Typical Sector Format

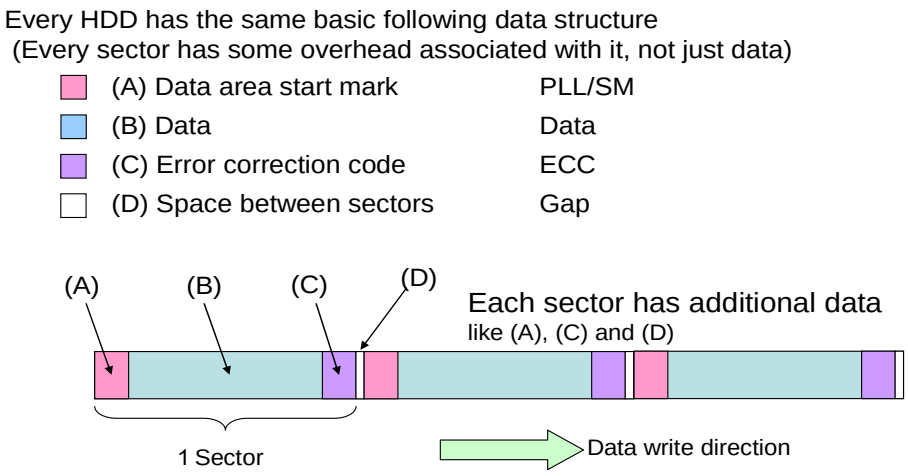

Figure 1

The size of the ECC data is based upon two factors -- the size of the data to be protected and the basic ECC algorithm being used. Over the years, ECC algorithms have improved. In today's ECC technology, an HDD is typically able to correct errors of up to 50 bits out of 4,096 bits per sector (512 bytes times 8 bits-per-byte). As the ECC algorithms have improved, their size has also increased.

Ideally the size of the ECC field should be less than 10 percent of the size of the data field. If the percentage gets too large, then the potential area available for data storage capacity will be traded for ECC protection data. If the ECC field exceeds 10 percent, the cost per gigabyte could increase. Why? Because this cost is calculated on the useable data storage capacity and ECC data is considered part of drive overhead; any increase in overhead has a direct impact to the useable storage area.

In order to maintain a very high level of data integrity over the years, while keeping the cost within reason, the HDD industry has continually improved the ECC algorithms used in hard drives. With a 512-byte data field in each sector, the ability to make additional improvements to the ECC algorithm has about run its course. However, if the data field to be protected in each sector is larger than 512 bytes, the ECC algorithm could be improved to correct for a higher number of bits in error, while at the same time reducing the percentage of overhead space required with respect to the size of the data field it is protecting. Starting as early as 1998, it was recognized that trying to maintain a reasonable ECC-

to-data field ratio and staying with the 512-byte sector size, would result in the stagnation of storage capacity on hard drives in the form factors that are used today.

Throughout its history, the HDD industry has pursued a growth in disk capacity. There are generally two ways by which the capacity of an HDD is increased. The first way is to add more storage disks. Sometimes, this is not the most efficient way to increase capacity, because it adds to the cost and physical size of the storage device, (more read/write heads and media disk). The second way to grow disk capacity is to increase the amount of data stored on each individual platter.

In the HDD industry, the term "areal density" refers to the number of bits contained in a square inch of media surface. In order to double the areal density, twice as many bits must be squeezed into the same space. So the area on the disk media used to store each bit is at least half the size it was before the density was doubled. As the areal density increases and the individual bits get smaller, smaller magnetic defects in the media impact a larger number of bits and the possibility to create data errors. This result is clearly unacceptable. Disk drive manufacturers want improvements in ECC capability to maintain data reliability and continue the growth of areal density.

Advanced Format (AF):

In a presentation by IBM at a 1998 Magnetic Materials Conference, it was noted that changes were needed in order to attain better ECC protection. It also was recognized that at some point in the not too distant future, the HDD industry no longer would be able to economically protect data in the traditional 512-byte sector format with the ECC algorithms that were projected to be available. Discussions began within the HDD industry to address this issue. Around the year 2000, the industry-wide "Long Data Sector Format" Committee was formed within the International Disk Drive Equipment and Materials Association (IDEMA®[3]).

With the proliferation of computers and a variety of hand held devices, the importance of maintaining consistency, compatibility and interchangeability of devices over time has become to be expected. The ability of the file saved yesterday to be retrieved today and any time in the future, or to be seamlessly shared with a co-worker or family member either by email or other methods, is not questioned. Likewise, new applications and OSs are expected to be compatible with existing OSs and possibly with all other applications installed on a system. For this reason, the data storage industry develops and maintains industry standards.

While the 512-byte sector size was a de facto standard for computer applications, other data storage applications may benefit from doing things differently. For example, when viewing a digital video, the video starts at the beginning of the video, or data file, and is watched while the host viewing system reads the data from the hard disk or digital video disc (DVD). This is called "streaming data" and lends itself to applications that could take advantage of larger (or longer) sector sizes. In other words, the computer or DVD player starts at the beginning of the file and reads the data in a serial mode until the end of the file. In a dedicated application, such as viewing a video or listening to music, it could be advantageous to store the data differently.

Toshiba was one of the first HDD manufacturers to design and produce HDDs configured with longer data sectors for usage in applications such as media and music players. Sector sizes of 1K and 2K were initially used. These larger sector sized drives were primarily targeted for usage in embedded systems instead of typical computer systems. The embedded systems could stream larger increments of data and therefore take advantage of the longer sector length. This also helped lay the foundation for the new generation of Advanced Format (AF) disk drives.

The fundamental question facing the HDD industry was simply, "How big do we make the ECC field with respect to the data field?" Between the initial formation of the IDEMA Long Data Sector (LDS) Committee in 2000 and 2007, industry meetings were held to present and discuss various concepts, approaches and ideas. Toshiba was and still is an active member of the LDS Committee, which undertook the endeavor of determining the new sector size. This consideration was important, since the change was recognized as difficult to implement across the entire computer infrastructure. As such, the HDD industry only wanted to make the change once. The industry focal point quickly centered on sector sizes options of 1K, 2K or 4K bytes in length.

Following a one day conference in 2007, attended by representatives from the HDD industry, system manufactures, software vendors, end users and other participants, the industry zeroed in on the data sector size of 4,096 (4K) bytes. As part of the IDEMA LDS Committee standards activities, the HDD manufacturers agreed in 2008 that the majority of new HDDs introduced into the "client" marketplace (consisting primarily of notebook and desktop computers) would use 4K data sectors starting in 2011. For those HDDs intended for the "enterprise" market" (comprised of servers and mainframe computers), this transition to 4K sectors would occur at a later date.

512 Emulation (512e) and 4K Native (4Kn):

In 2009, a proposal was introduced to the IDEMA Committee to clarify two things. First, it defined an HDD with a sector size of 1K bytes or larger to be classified as an AF HDD. Second, it defined how the OS could interrogate the HDD to determine the sector formatting scheme used on the drive. This enabled the host computer to identify whether the hard drive was formatted to the legacy 512 bytes-per-sector or the new AF 4K bytes-per-sector format.

While the 2009 proposal was accepted, a key challenge remained. Although some more recent OS versions anticipated the introduction of longer sectors, millions of PCs still used older Microsoft Windows OSs that were set up to read and write 512-byte blocks or multiple 512-byte blocks. A solution to make the OSs installed on computers in the field compatible with the 4K sector format was needed. Clearly updating all the OSs in use did not make sense. To solve this problem, the IDEMA committee devised the concept of 512 emulation (512e). This emulation allows the application software and the file system to continue to operate using 512-byte blocks and to communicate with HDDs formatted to the 4K-byte sector architecture. See Figure 2.

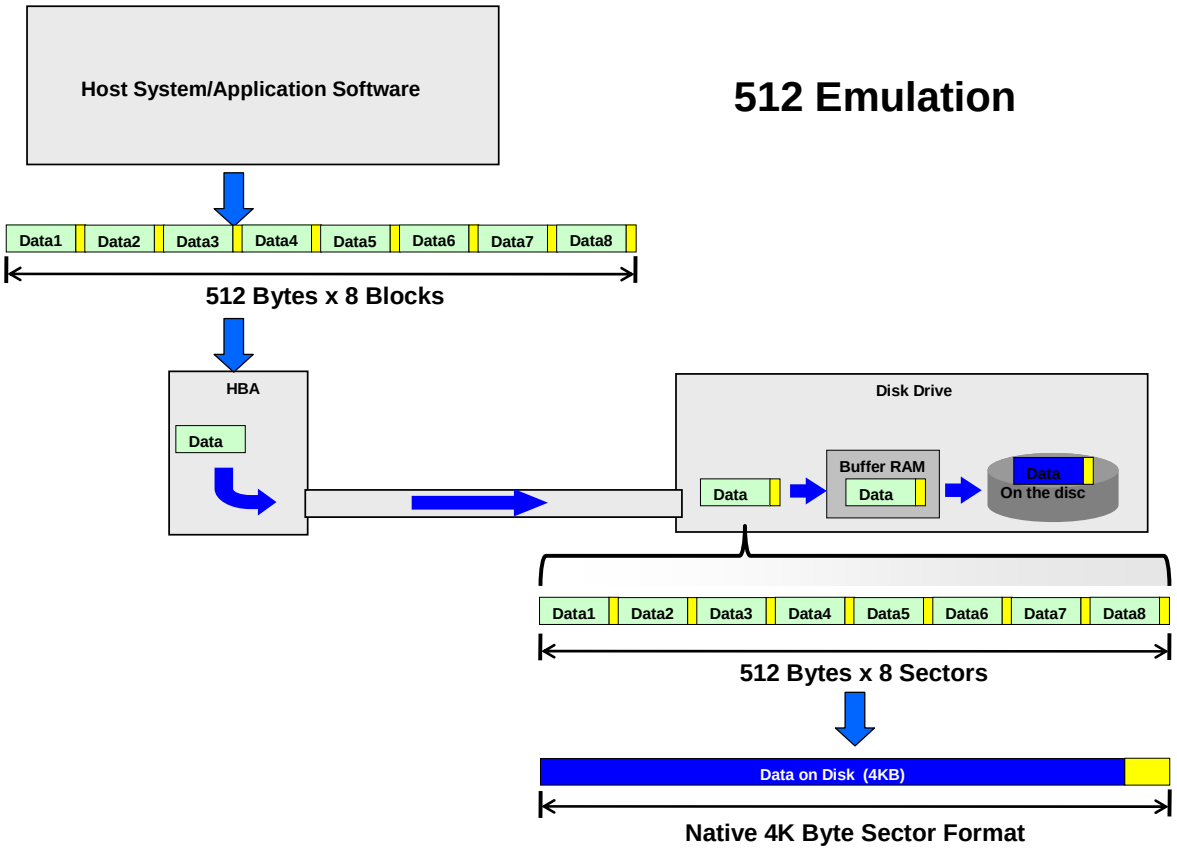

Figure 2

The new AF HDD storage devices read data from or write data to 4K-byte (4,096 bytes) physical sectors on the HDD media. Since the application and file system still want to transfer in 512-byte blocks, the AF 512 emulation allows the transfer between the host computer and the AF HDD to continue to use 512-byte blocks by transferring 8, 512-byte blocks as a single read/write request. For a write command, the AF HDD will take 8, 512-byte blocks from the host and treat it as a single block of 4K-byte data, where it will then be written to the media in a 4K-byte sector. For read commands, the AF HDD will read a 4K-byte sector from the media and transfer it as 8, 512-byte blocks back to the host system.

Independent of data size, every physical sector has some architectural overhead associated with it. This overhead provides: 1) a unique header to identify each sector; 2) information to identify the start of the

sector; and 3) data protection information in the form of the ECC code. In the 512 emulation case, the equivalent of 8, 512-byte sectors of data are grouped into a single, 4K physical sector on the media.

Because the overhead associated with 1, 4K-byte sector is less than the total overall overhead associated with 8, individual 512-byte sectors, there is a recording efficiency improvement of about 9 percent on an HDD using 4K sectors versus an HDD with 512-byte sectors. See Figure 3 below.

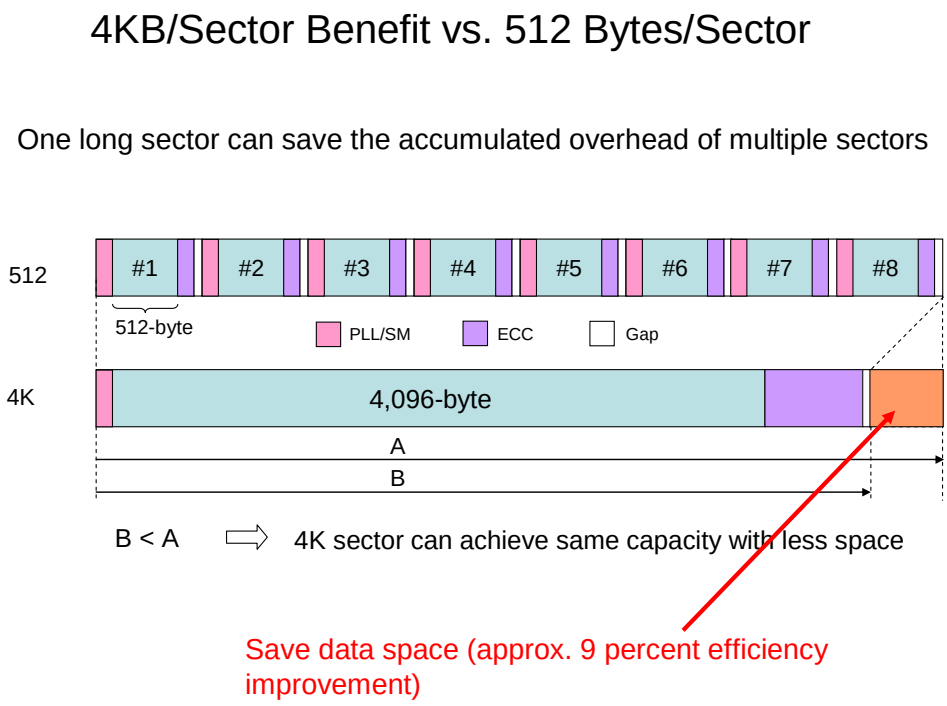

Figure 3

In single HDD systems, such as notebook computers, the OS plus the user data and various applications will be stored on the HDD. The OS will structure the HDD into various areas that serve different purposes, e.g. the OS area, protected areas, user data area, etc. The OS generates at least one partition, typically called the "primary partition," which is where most user files reside. In most cases, the primary partition represents the majority of the formatted capacity of the HDD.

The OS typically controls the placement of the beginning of this primary partition. For example, the Microsoft Windows XP OS will begin the primary partition at physical sector 63. However, the start of the primary partition under Microsoft Vista®[5] OS with Service Pack 1 and Microsoft Windows 7

begins at sector 1M. The IDEMA Committee's activities with the major OS suppliers resulted in Microsoft Vista® and all ensuing Microsoft OSs to be equipped with the capability to interrogate the HDD to determine if an AF HDD or a legacy (512-byte sector formatted) HDD is present in the system.

If Windows Vista or a later OS detects a legacy HDD, the OS will set itself to read and write 512-byte blocks to support the 512 byte-per-sector drive. If Windows Vista or a later OS detects an AF HDD with 4K sectors, the OS will set itself to read and write 4K-bytes worth of data in the form of 8, 512-byte blocks using the "512 emulation" standard, as shown previously in Figure 2.

An AF HDD supporting 512 emulation is denoted as "512e," referring to 512 emulation. The older Microsoft OSs, including Windows XP, are not designed to interrogate the HDD to determine the sector configuration of the HDD; as a result all HDDs are treated as legacy HDDs using the traditional 512 byte-per-sector format.

Now the fun begins. Until 2011, the majority of HDDs were formatted to 512-byte sectors. At some time in the future, OSs and applications will transfer 4K blocks of data directly to and from the HDD, which are referred to as "4K native" (4Kn) HDDs. However, until 4Kn HDDs and operating systems are widely available, emulation will serve as the bridge between older Windows OSs and the newer AF HDDs. As of 2011, HDD manufacturers have provided samples of 512e HDDs to their large account customers. The HDD industry will start the transition from 512-byte sector to the new 4K sector formats on most new generation 512e HDDs shipping into the client marketplace.

Misalignments and Proper Alignments:

Ideally the 4K HDD can be thought of in a similar way as a 512-byte sector HDD. A 4K sector on the media represents 8, 512-byte data segments. In Figure 4 below, the 4K sectors (X-1, X, X+1 and X+2) each represent 8, 512-byte segments of data (labeled 00-07). On the top line of Figure 4 (Figure 4A - Alignment 0), the 8, 512-byte segments of data perfectly align with the boundaries of the 4K sectors, e.g., sector X-1, X, X+1 and X+2. This is typically referred to as an "Alignment 0" configuration, or no offset in one direction or the other. Generally Alignment 0 is the ideal case and typically yields the best performance, but there are some applications that prefer other arrangements.

Alignment Scenarios

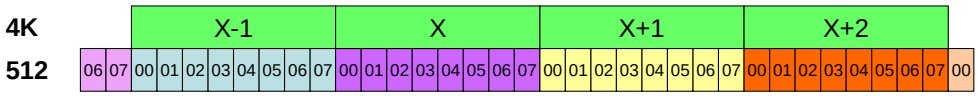

Figure 4A - Alignment 0

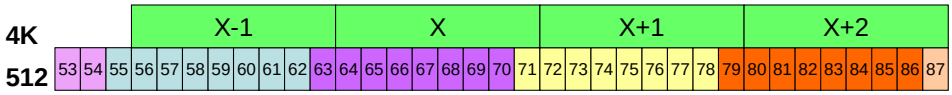

Figure 4B - Alignment -1

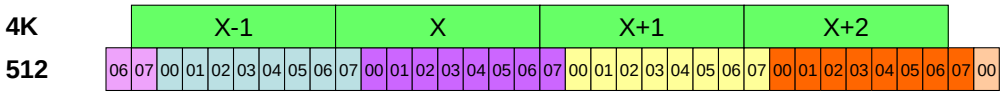

Figure 4C - Alignment +1

Figure 4

In the event of a misalignment, the first 512 byte data segment of an 8, 512-byte transfer is not directly written into the first 512 bytes of the 4K sector. Misalignment basically means that the 8, 512-byte data segments are spanning parts of two separate 4K-byte physical sectors and not just one single 4K-byte sector. In Figure 4B, Alignment - 1 (above), the purple 512-byte data segment 00 is located in 4K physical sector X-1, while the remaining purple data segments (purple 01-07) are contained in 4K sector X.

Let us consider rewriting all 8, 512-byte data segments in the purple sector, or a total of 4,096 bytes. In the case of Alignment 0 (Figure 4A), the system would simply issue a “write sector X” command and the data in sector X would be rewritten.

If the same purple sector on a misaligned HDD is rewritten, for instance the “Misaligned – 1” configuration shown in Figure 4B, the HDD would have to first read sectors X-1 and X into cache memory. In this misalignment scheme, sector X-1 contains data from both the blue data group segments 01-07 as well as segment 00 of the purple data group. Likewise, sector X contains the purple data group segments 01-07 and segment 00 of the yellow data group.

Once the data is in cache memory, the purple data group can be overwritten. The remaining portions of data in cache, specifically the blue data segments in sector X-1 and the yellow data segment in sector X must remain unaltered. Once all the purple data segments have been updated in the cache, sectors X-1 and X will be rewritten to the media with the ECC recalculated to account for the new data contained within these sectors.

This three-step process of reading data from the media into HDD cache memory, altering the data in HDD cache memory, and then writing the data back to the media is typically referred to as a "read-modify-write" (RMW) operation. This process affects performance. Once the data is read from the media into cache memory, the rotating HDD media will make at least one revolution before the data can be written back to the media. For a 7,200 RPM HDD, there is a minimum delay of 8 milliseconds before the data in the cache memory can be written to the media.

It does not matter by how much or in which direction the data fields are misaligned—any misalignment between the data fields and the physical sectors will trigger a RMW operation and a delay will occur. This delay will most likely occur on every write command. On a typical system, because of the difference between file systems and the ratio between reads and writes, the overall system performance could decrease by as much as 20 percent due to misalignment.

As stated previously, for both Windows Vista with Service Pak 1 and Windows 7 the primary partition starts at the beginning of sector 1M (sector 1,048,576). By starting at this predetermined sector, this 4K sector number is evenly divisible by 8 with a whole number resulting, e.g., 1,048,576 divided by 8 equals 131,072.0. Why is this important? If the OS data partition is properly aligned with the layout of the physical sectors, you will have "Alignment 0," or no offset with respect to the logical sectors and the physical sectors. Thus with Alignment 0, RMW operations will not be required and misalignment will not negatively impact overall system performance.

With the Windows XP OS, the OS supports only 512 byte sectors. Further, Windows XP uses the DOS format, which means the "primary data partition" starts at sector 63. See Figure 5A.

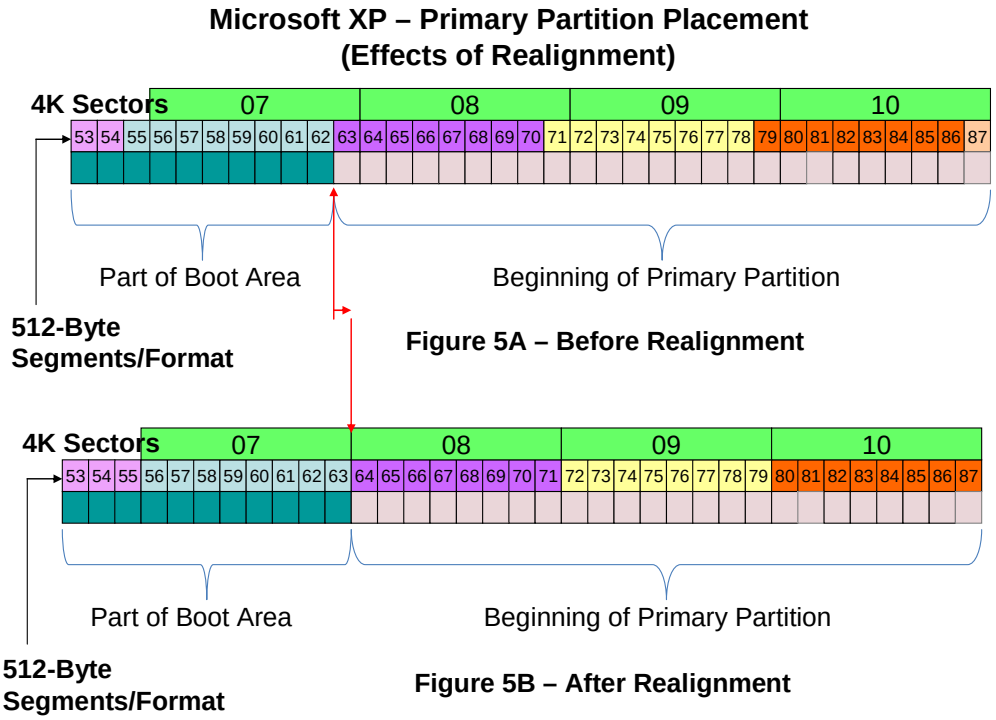

Figure 5

Because the Windows XP OS starts the primary partition at sector 63, there is an automatic misalignment between the file system and the 4K physical sectors. As discussed previously, this misalignment can have a significant effect on performance.

To address the misalignment performance penalty associated with the Windows XP OS, HDD manufacturers, like Toshiba, worked with disk utility software vendors to provide a solution to the misalignment and performance penalty problem. This allows Toshiba and other HDD manufacturers to provide a means to realign the data fields to the physical sectors and mitigate exposure to misalignment.

In Conclusion - The Toshiba Solution:

Toshiba has collaborated with Paragon Software Group™[6] to provide Toshiba AF 512e HDD users a utility program optimized for Toshiba drives. The Paragon Alignment Tool (PAT[7]) for Toshiba AF 512e Hard Drives will take the primary partition and move it back one 512 byte sector (to sector 64) as shown in Figure 5B above.

PAT for Toshiba AF 512e Hard Drives first will poll the HDD to determine the starting location of the primary partition. If the primary partition is properly aligned, the utility will not do anything further. If there is misalignment, PAT will move the partition so it is properly aligned for the Windows OS installed in the computer system.

The PAT utility offered by Toshiba is optimized to work exclusively with Toshiba HDDs. The PAT utility is offered as a free download and is specifically geared for single-licensed system users. For access to the PAT utility for Toshiba AF 512e Hard Drives, visit the Technical Support section of the Toshiba Storage Device Division web site at www.toshibastorage.com.

[1] Windows is a registered trademark of Microsoft Corporation in the United States and other countries. All rights reserved.

[2] One Gigabyte (1GB) means 10^9 = 1,000,000,000 bytes using powers of 10. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 2^{30} = 1,073,741,824 bytes, and therefore shows less storage capacity. Available storage capacity will also be less if the computer includes one or more pre-installed operating systems, pre-installed software applications, or media content. Actual formatted capacity may vary.

[3] IDEMA is a registered trademark of The International Disk Drive Equipment and Materials Association. All rights reserved.

[5] Microsoft, Windows and Windows Vista are trademarks of Microsoft Corporation in the United States and/or other countries. All rights reserved.

[6] Paragon Software is a trademark of Paragon Software Group. All rights reserved.

[7] The Paragon Alignment Tool was developed and is licensed through Paragon Software Group.

While Toshiba has made every effort at the time of publication to ensure the accuracy of the information provided herein, information, including specifications, configurations, system/component/options availability, is subject to change without notice.