IDEMA®

# Hard Disk Drive Long Data Sector White Paper

**Authors: P. Chicoine, Phoenix; M. Hassner, Hitachi GST; M. Noblitt, Seagate; G. Silvus, Broadcom; B. Weber, LSI Logic; E. Grochowski, Consultant**

**April 20, 2007**

## Abstract

This paper summarizes the results of the Long Data Sector IDEMA Committee that was started in 2000. The activities of this committee were motivated by a 1998 NSIC Paper, in which the incompatibility of HDD Industry areal density growth at the existing 512-Byte Sector Format and maintaining data integrity was recognized.

Editor's Note: In this white paper, the term sector usually refers to the space on a disk containing user data and overhead; block refers to a group of bytes, so that data block is the space for data bytes, ECC block contains error correction code bytes. While the term 512-byte or 4096-byte sector is technically incorrect since a sector contains more than data and is therefore longer, the term data sector and data block will be used interchangeably. The text should provide the exact definition.

## 1. Motivation, Scope and History

The motivation behind the IDEMA LDS-Committee activity lies in the realization that continued HDD areal density growth, while maintaining the data integrity required by HDD storage customers, at the current 512-byte sector format, are incompatible in the long run. As a result of this realization, Seagate, Maxtor, Hitachi GST, and Fujitsu agreed to form a committee, under IDEMA's sponsorship, to address long data sectors. In 2003 this Committee jointly requested the Microsoft Corporation to support up to 4096-byte sector format in their next generation OS. In 2004, Microsoft agreed to support this request by the HDD developers. Following this milestone, the IDEMA Committee activities were focused on implementation of long data sectors.

This document contains sections that represent different aspects of this implementation. Section 2.1, written by S. Jenness, Microsoft, describes the WINDOWS OS Support. Section 2.2, written by P.Chicoine, Phoenix, describes the BIOS implementation. Section 3, written by B. Weber, LSI Logic, provides the viewpoint of the Storage System User. Section 4, written by M. Hassner, Hitachi GST in collaboration with G. Silvus, Broadcom, describes the Soft Error, Format Efficiency and Hard Error Gains due to a Long, 4K-Block, Data Sector Format.

The transition to a Long Data Sector requires backward compatibility with the 512-byte existing sector format. While the HDD developers are in agreement as to the need for transitioning to a Physical Long Data Sector, currently, there is no

unanimous agreement on how to manage the transition period. To achieve the ultimate goal of Physical Long Data Sector, on which there is agreement, it may be necessary to formulate a Long Data Sector HDD Industry Implementation Standard; this is left for future work. This document should be viewed as a stepping stone toward this goal.

# 2. Impact on OS/Software Applications

With the introduction of drives that support larger than 512-byte logical sectors the impact falls on the BIOS firmware, the OS Installation Software, the OS Bootloader and the OS kernel IDE driver. Since systems can be populated with multiple drives, some of which may support 512-byte logical sectors while others may support greater than 512-byte logical sectors, it becomes the responsibility of the software mentioned above to properly detect, identify, configure and access the each installed drive, using its native logical sector size.

## 2.1 Windows Vista and Longhorn Server Support

Windows Vista and Longhorn Server support both emulated and non-emulated long sector hard disk drives.

- **Startup support**

Windows Vista and Longhorn Server will allow a start from emulation drives that have a logical sector size of 512 bytes. BIOS vendors and other hardware vendors may have to update their firmware to correctly start together with drives that expose a long logical sector in addition to a long physical sector. Microsoft will test Windows Vista and Longhorn Server for startup support for both emulated and non-emulated drives when the necessary hardware support is available.

- **API support for querying hard disk drive properties**

To act appropriately with long sector drives, some applications and other components may query for physical sector size, for logical sector size, and for sector alignment. Applications may have strict requirements for the sector size of a hard disk drive. The IOCTL_STORAGE_QUERY_PROPERTY request has been updated to include sector information in the STORAGE_ACCESS_ALIGNMENT_DESCRIPTOR request. This document is included in the Reference section of this paper.

### 2.1.1 Storage Stack, Partition, and File System Alignment

The file system, the volume manager, and other parts of the storage stack in Windows Vista have been updated to accommodate hard disk drives that have a long sector size. In earlier versions of Windows, the default starting offset for the first partition on a hard disk drive was sector 0x3F. Because this starting offset was an odd number, it could cause performance issues on long sector drives because of misalignment between the partition and the physical sectors. In Windows Vista, the default starting offset will generally be sector 0x800. However, the starting offset might be different for drives that

have special alignments. For emulation drives that have special alignments, dynamic disk operations were not supported in the initial release version of Windows Vista. This support has been added for Windows Vista SP1 and Longhorn Server. When the file system is formatted, only cluster sizes that are larger than or equal to the underlying physical sector size will be supported.

### 2.1.2 Windows Applications

Support for these new hard disk drives in Windows Vista does not mean that the drives are automatically supported by applications. Applications that care about the underlying structure of their data storage and the atomic write size to stable storage must be updated to reflect the new hard disk drives. Independent software vendors (ISVs) must update their applications.

## 2.2 BIOS Support for Long Sector Sizes

When power is applied to a system, the first software that executes is the BIOS firmware on the motherboard. When BIOS executes, POST (Power On Self Test) routines are responsible for detecting, identifying, and configuring the hardware devices and peripherals in the system. In addition, BIOS provides public services for other run-time software to access these devices, by providing ISRs (Interrupt Service Routines) to handle IRQs (Interrupt Requests) and DSRs (Device Service Routines) in the form of INT calls. For hard drives the DSR provided are the INT13 functions. When BIOS has completed, it then loads and invokes the OS bootloader from the MBR (Master Boot Record). The BIOS ISRs and DSRs remain available during run-time for the OS to utilize. After execution has been passed to the bootloader, BIOS is no longer responsible for booting the system.

With the introduction of ATA hard drives that support larger than 512-byte physical and logical sector sizes now requires some adaptation to the traditional PC-compatible BIOS. Currently there are two proposed methods of increasing the physical sector size with relation to the logical sector size. One method maintains 512-byte logical sector size backward compatibility by emulating the logical-to-physical translation within the drive itself. Internally within the drive the physical sector size can be any size (512-bytes, 1K, 2K, 4K), because the BIOS sees it externally as 512-byte sectors regardless of the actual internal physical sector size. For this method there are no required BIOS changes. The other method utilizes a logical sector size that is greater than 512-bytes and can be the same size as the underlying internal physical sector size, such as 4K. This method will require software changes, including the BIOS, to accommodate the longer than 512-byte logical sector size.

When BIOS detects and identifies the hard drive, it now must keep track of the logical sector size. During configuration, BIOS must now account for the logical sector size: when reporting the drive size, it must account for the max LBA (Logical Block Address), the size reported in SETUP and by the DSR functions. When BIOS ISRs/DSRs

transfer data the BIOS must now ensure that the buffer that it is transferring to/from accounts for the longer than 512-byte logical sector size, in order to eliminate possible data overruns. For software calling the BIOS INT13 DSR, the data read or written per LBA may be different for each drive installed. It is the responsibility of the calling software to know the drives logical sector size before transferring data. The recommended method is to issue an IDENTIFY_DEVICE ATA command directly to the drive.

### 2.2.1 OS Installation on Long Sector Sizes

When the OS installs on a system, it must now determine whether the hard drive it is installing on supports longer than 512-byte logical sectors. If it does the installer must adjust any internal buffer sizes to accommodate the larger LBA transfers.

### 2.2.2 OS Bootloader Support for Long Sector Sizes

Once BIOS passes execution to the OS bootloader it is now responsible for loading the OS into memory and passing execution to the kernel. Since the bootloaders run in real mode, they will rely on the BIOS INT13 DSR services to load enough of the kernel so that it can execute. In some OSes this is both the real mode and protected mode portion of the kernel and perhaps a ramdisk. Once execution passes to the kernel the BIOS INT13 DSR services are typically not used (except OSes like DOS). The bootloader must now accommodate drives that support longer that 512-byte logical sectors.

### 2.2.3 OS Kernel and Driver Support for Long Sector Block Sizes

After the bootloader passes execution to the kernel, the next time the drive hardware is accessed is when the IDE driver loads. If the driver detects devices that it will claim then the drivers will install and hook the appropriate IRQ. It now owns the device. The OS IDE driver must now accommodate drives that support longer than 512-byte logical sectors.

## 3.0 Storage System Support Required for Long Sector HDD

An external Storage System Controller is essentially a virtualization device. The controller will typically have the capability of virtualizing anywhere from 1 to 400 SAS, SATA, or Fibre Channel disk drives. Using the storage controller utilities, the System Administrator will configure drive groups, volumes, and raid levels from the pool of available physical disk drives. These resulting volumes are then mapped to host machines as LUNs (logical units). Because of this virtualization capability, the controller essentially isolates the drives from the host. From the host standpoint, all that is seen is the storage controller device which will display the previously configured LUNs. Today's controllers will typically use a 512 byte logical block size as reported from a SCSI Read Capacity command.

As disk drives begin supporting longer physical sizes, it will now become the controller's responsibility to work with these drives of different logical block sizes, while still presenting a consistent logical block size to the host.

There are two obvious ways that an external storage subsystem can handle this.

1) Do emulation of 512 byte or 4K logical blocks to the host, while supporting disk drives of either native block size behind the storage controller
2) Require that a single Drive Group use disk drives that have a consistent logical block size and project that same block size to the host for that particular LUN

If the controller chooses to implement the first option, from a host server standpoint, the interface to an external storage subsystem will look essentially unchanged. If the host operating system only works with 512 byte logical blocks, then the controller would mask the fact that 4K sector disk drives were being used, and the operating system would continue to work normally. In this case, the controller cache algorithms would help to mask any performance impact due to unaligned accesses to the controller that did not match the drive 4K sectors. This method would also ease the case of multi-initiator access where some servers have not yet been modified to work with the larger block sizes. In this case you could expose a 512 byte logical block that all servers could work with.

For the second option, restrictions are imposed on the RAID configurations. In this case, all drive groups will have the restriction that they have to be made up of drives having the same logical block size. This would prevent users from mixing drive types in a particular drive group. It would also put restrictions on which drives could be used as hot spares on which drive groups. Any LUNs that are mapped to these drive groups would be locked to the logical block size of the drives that make up the drive group. This will open up the possibility that a single controller device could have LUNs of different base logical block sizes as reported in the Read Capacity command. In this case, it is also important to make sure that the HBAs and Drivers that are used will support LUNs of different logical block sizes behind the same physical device.

In conclusion, external storage controllers will also have to do work to adapt to disk drives of different logical block sizes. In some cases, these controllers have the capability to mask some of the issues to the host, so that the host server can continue to use 512 byte logical blocks and will not know the difference. In the case that the storage controller chooses to limit drive groups and report the native sector size, then the operating systems will have deal with same issues as Direct Attached Storage, as well as account for the cases of multi-initiator.

## 4. Large Block Error Correction

In this section, we will distinguish between the following types of blocks:

- Physical Block: These are the physical units of storage on the surface of the disk, and the smallest unit of data which can be physically written to or read from the disk.

- Logical Block: The disk drive presents itself to the outside world as a linear address space of logical blocks, whose size may be in principle be different from that of the physical blocks.

- ECC Blocks: The Error Correction Coded (ECC) block size is determined by the extent of the ECC code, which could in principle span more than one physical block or more than one logical block.

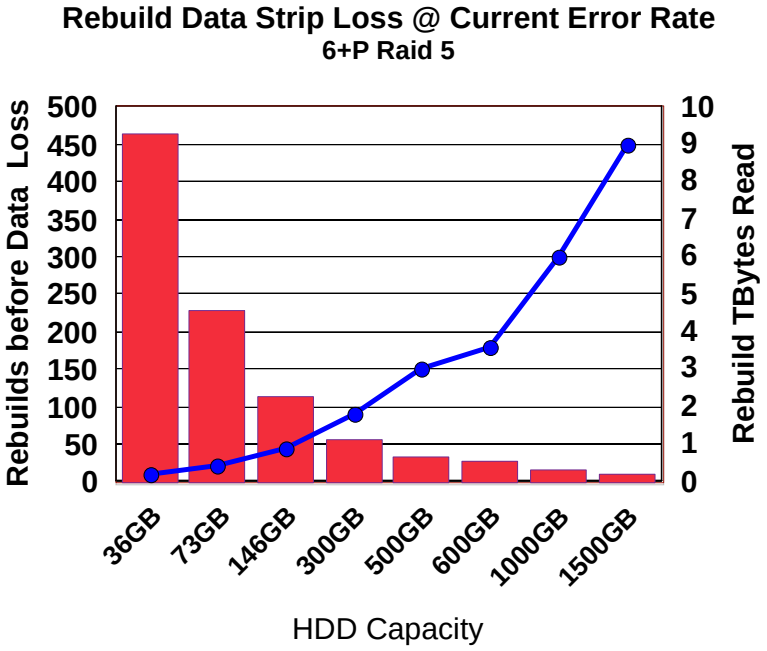
Of course in current disk drives these three definitions coincide: the physical, logical and ECC blocks each consist of 512B of user data followed by ECC.

## 4.1 The Need for Large Block Sizes

There are a number of strong arguments in favor of larger ECC block sizes. First, coding efficiency increases with ECC block size. For very large ECC block sizes, there are on average many errors per block, and the variance in the number of errors per block is small compared to the mean. With such a precise estimate of the number of errors per block, one need not over design the code--the error correction power required to correct almost all code words will be only slightly greater than the overhead required to correct the average number of errors per block. On the other hand, with small ECC blocks, one must over design the code to account for the relatively larger uncertainty in the number of errors per block.

Second, there are simple scaling rules which suggest that hard errors and defects in disk drives are likely to get worse in future rather than better. As areal densities in disk drives continue to increase, the physical size of each sector on the surface of the disk becomes smaller. If the mean size and number of disk defects and scratches does not scale at the same rate, then we expect more sectors to be corrupted, and we expect the resulting burst errors to more easily exceed the error correction capability of each sector.

Third, as areal densities are expanded, the signal processing and detection problems become more difficult, due to either decreasing signal to noise ratio (SNR) or increased inter-symbol interference (ISI). So far, this trend has been accommodated through the use of improved signal processing, without increasing the ECC block size. Recently, progress in these areas has slowed, and there is a general consensus that use of large ECC block sizes will open up new opportunities in this field.




A dramatic illustration of the urgent need for Large Sector ECC in HDD is provided by the impact of Hard Error Rates of HDD, due to head-disk interface defects and scratches which are uncorrectable at the current sector format, on RAID Storage System Reliability. In the attached graph, from a 12/07/06 IDEMA PMR-Symposium presentation by IBM-Storage Systems, describing the Number of RAID-5 Rebuilds before Data Loss as a function of HDD-Capacity, it is projected that for 1.5 Terabyte HDD this number will increase to 1-Data Loss/10-RAID-5 Rebuilds, unless the capacity increase is matched by HDD Hard Error Rate improvement. At the current 512-Byte Format it is impossible to correct larger defects, as required. This is possible only at a Large Sector Format! The Hard Error Rate Gain from 4K-Block ECC has been estimated through use of histograms of defect lists logged during Surface Analysis Tests. Current Hard Error Rate HDD-Specs claim 1 Hard Error/10^15 bits read. The logged defect size data provides an estimate of at least 1 order of magnitude gain in Hard Error Rate from 4K-Block ECC, at current linear densities. This gain is expected to grow proportionally to the projected HDD-linear density growth rate.

In addition to this quantitative argument in support of larger ECC block size, there have also been other quantitative arguments made in favor of larger physical sector sizes. Use of larger physical sector sizes would reduce the amount of overhead required for synchronization fields, inter-sector gaps, etc., and so lead to more efficient use of disk real estate. Furthermore, at a Large Sector Format there is an SNR Margin Gain, i.e. the linear density can be increased, as a lower SNR becomes correctable by the Large Block ECC. Both arguments result in Capacity Gains that have been quantified by actual measurement, as described in the sequel.

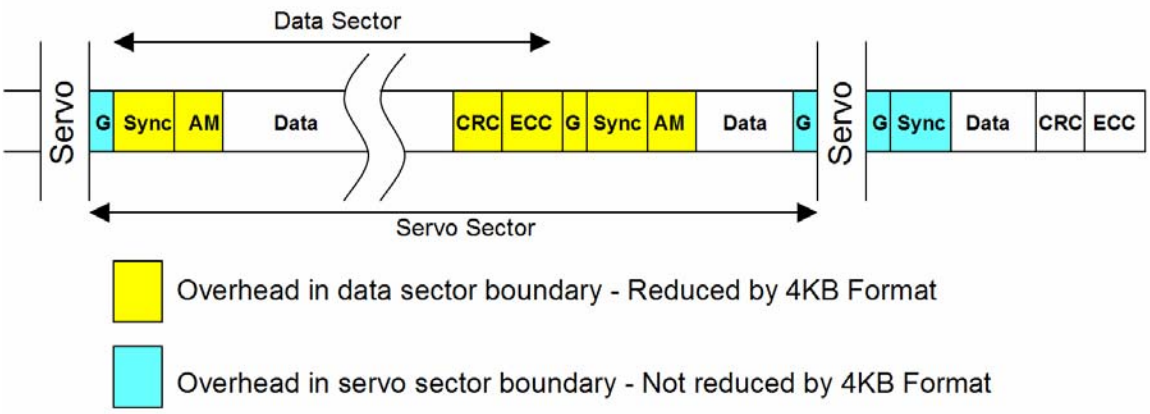
## 4.2 What Determines the Optimum Block Size?

From the perspective of an ECC designer, larger ECC block sizes are in general better, for reasons of code efficiency as described above. As the ECC of choice, Reed-Solomon Codes are effective as long as their Symbol Error Rate is smaller than 10^-2. This is effectively achieved by 12-bit Symbols and 4K-Block Codeword Size.

From a throughput perspective, the ideal logical block size should be roughly equal to the characteristic size of a typical data transaction. Excessively small logical block sizes will impose too much transaction overhead. For excessively large logical block sizes each transaction will involve the transfer of a large amount of unnecessary data.

The consensus of opinion in the hard disk drive industry has been that physical block sizes of 4K-Block would provide a good compromise. This block size corresponds to the paging size used by operating systems, such as WINDOWS, and processors. It is also the size of a standard transaction in Relational Data Base Systems. Regarding the logical sector size, ideally it should match the physical sector size; however, the need for backward compatibility poses a severe constraint in this regard. On the other hand having different logical and physical sector sizes requires Read-Modify-Write. It has been agreed among the HDD Manufacturers to leave the logical sector size up to the decision of each individual HDD Company.

## 4.3 4K-Sector Format Efficiency/SNR Gains




Based on actual Product Format Calculation, the estimated Format Efficiency Gain from 4K-Block Sector is 12-14%, due to the overhead reduction in the data sector boundary, as described in the diagram above.

Using a Drive-Tap Experimental Tool that permits to vary the Linear Density as well as Offtrack and measure actual 2.5" HDD Error Events,

the 4K-Block ECC SNR Gains are summarized below. The linear density gains for 12-bit Symbol 4K-ECC are calculated for different Sector Failure Rates and varying amounts of Check Overhead/4K-Sector. The results below show these gains for 4x512B,6x512B and 8x512B-ECC, where in this notation the amount of check overhead/4K-Block is that of 4-512B, 6-512-B and 8-512-Byte Sectors respectively.

| Operation log10(SFR) | Measured linear density gain (over 512B ECC) |
|---|---|
| | 12-bit symbol (6x 512B Corrections) |
| -6 | 8.4% |
| -12 | 12.8% |
| -15 | 14.9% |

| Operation log10(SFR) | Measured linear density gain (over 512B ECC) |
|---|---|
| | 12-bit symbol (4x 512B Corrections) |
| -6 | 9.1% |
| -12 | 11.4% |
| -15 | 13.4% |

| Operation log10(SFR) | Measured linear density gain (over 512B ECC) |
|---|---|
| | 12-bit symbol (8x 512B Corrections) |
| -6 | 8.2% |
| -12 | 12.8% |
| -15 | 14.9% |

## 5.0 Summary

A new sector standard for magnetic hard disk drives, 4096 bytes, has been proposed replacing the previous 512 byte sector length. The disk drive industry recognizes the need for a longer sector to maintain data integrity as areal density increases in future drives. Progressively more error correction code bytes are required at these higher areal densities which would increase overhead and reduce efficiency if the industry remained with 512 byte sectors. A longer sector results in distributing these ECC bytes over more data contained in a longer sector on the disk. IDEMA's mission of standard development for the storage industry prompted the formation of committee comprised of HDD developers, component and software suppliers from which this new sector length of 4096 bytes resulted. A longer than 4K-byte sector length could entail transfer of large and unnecessary data.

Drive interfaces and operating systems have different requirements for recognition of sector lengths. The Microsoft Corporation participated in the Committee and developed the next generation OS, Windows Vista and Longhorn server to be compatible with 512 byte to 4096 byte sectors. In Windows Vista the file system, volume manager and other parts of the storage stack have been updated to accommodate longer sector lengths. In the ATA interface, the OS specifies sector length, SCSI interfaces will operate with varying sector lengths.

There is agreement within the industry on increasing the physical sector length. However, there are different aspects for the optimum implementation of this standard. Backward compatibility with existing 512 byte disk drives and systems requires a 512 byte emulation in which legacy operating systems could be used. Storage systems comprised of 512 byte and 4096 byte sectored drives will function as 512-byte drives. In this case, existing BIOS and hard disk controller designs can be used. Emulation requires alignment of 512 byte sectors within a 4096 sector drive and the possibility of a read-modify-write occurrence exists. A second implementation concept uses a 4096 logical OS with a 4096 byte physical sector length, requiring new BIOS designs. In this case, storage systems may specifically use drives with the longer sector length. Windows Vista and Longhorn Server support both emulated and logical/physical drives. There is a need for further consideration within the industry of implementation strategies, and IDEMA may participate in this activity.

The use of 4K-byte sectors allows the use of a more robust 12 bit ECC which will give drive designers more capability to maintain reliability at very high, future areal densities. In addition, it has been postulated that longer sectors will also reduce the overhead requirements for synchronization fields, inter sector gaps and allow a SNR gain. Today, drive and component developers are building new hardware to begin testing of longer sectors. Testing of Windows Vista and 4K-byte BIOS designs will be a principal activity in the near future.

Finally, the disk drive industry fully realizes that increasing sector length is a significant change for storage users, and compatibility with existing 512-byte drives, OS and storage systems is a considerable concern. The IDEMA Committee believes this change is warranted by the valuable capacity, performance and reliability gains. Longer sectors allow the continuance of magnetic disk drive progress.

# 6.0 References

## 6.1 Windows Vista Support for Large-sector Hard Disk Drives, Microsoft article 923332, March 5, 2007

### Introduction

Hard disk drive manufacturers will soon start producing hard disk drives that contain physical sector sizes that are larger than the traditional 512 bytes per sector. For example, sectors may be 1 kilobyte (KB), 2 KB, or 4 KB. This change will enable manufacturers to improve the capacity, the performance, and the reliability of their hard disk drives. This article discusses Windows Vista support for large-sector hard disk drives.

### Background

When the hard disk drive industry makes this shift in the underlying structure of their drives, two kinds of hard disk drives will appear on the market. The simpler kind of hard disk drive will use a large physical sector size internally and expose that same size to the system. This kind of hard disk drive will first appear in enterprise-class drives, such as SCSI drives or Fibre Channel drives. The main drawback of this kind of hard disk drive is backward compatibility. BIOSes, host adapters, operating systems, and business applications may have to be updated to correctly operate together with this kind of hard disk drive.

To reduce backward compatibility issues, some manufacturers will produce hard disk drives that use a large physical sector size internally, but expose only a logical sector size of 512 bytes to the system. These hard disk drives are referred to as emulation devices because of the method that the drives use to write data. This method is frequently called "read-modify-write." For writes that are smaller than a physical sector, the drive must read the physical sector, modify the small, changed part of the sector, and then write the whole physical sector. The main drawback of this kind of hard disk drive is decreased performance. The extra read operation that must occur for writes that are smaller than the physical sector may decrease performance.

### Windows Vista support

Windows Vista supports both kinds of hard disk drives if the underlying hardware in the system also supports the drives.

- **Startup support**
  Windows Vista will let you start from emulation drives that have a logical sector size of 512 bytes. BIOS vendors and other hardware vendors may have to update their firmware to correctly start together with drives that expose a large logical sector in addition to a large physical sector. Microsoft will test Windows Vista for startup support for these kinds of drives when the necessary hardware support is available.
- **API support for querying hard disk drive properties**
  To act appropriately with large-sector drives, some applications and other components may query for physical sector size, for logical sector size, and for sector alignment. Applications may have strict requirements for the sector size of a hard disk drive. The IOCTL_STORAGE_QUERY_PROPERTY request has been updated to include sector information in the STORAGE_ACCESS_ALIGNMENT_DESCRIPTOR request. For more information, visit the following Microsoft Web site:
  http://msdn2.microsoft.com/en-us/library/ms803642.aspx
- **Storage stack, partition, and file system alignment**
  The file system, the volume manager, and other parts of the storage stack in Windows Vista have been updated to accommodate hard disk drives that have a large sector size. In earlier versions of Windows, the default starting offset for the first partition on a hard disk drive was sector 0x3F. Because this starting offset was an odd number, it could cause performance issues on large-sector drives because of misalignment between the partition and the physical sectors. In Windows Vista, the default starting offset will generally be sector 0x800. However, the starting offset might be different for drives that have special alignments. For emulation drives that have special alignments, dynamic disk operations will not be supported in the initial release version of Windows Vista. This

support will be added in a future service pack.

When the file system is formatted, only cluster sizes that are larger than or equal to the underlying physical sector size will be supported.

Support for these new hard disk drives in Windows Vista does not mean that the drives are automatically supported by applications. Applications that care about the underlying structure of their data storage must be updated to reflect the new hard disk drives. Independent software vendors (ISVs) must update their applications.

## 6.2 Position Statement -4K Data Blocks Hitachi Global Storage Technologies

**Purpose:**

The purpose of this letter is to present the position of Hitachi Global Storage Technologies regarding a new data block standard, longer than the present 512 byte block standard. This new long data block standard is considered necessary to store data on future hard disk drives with higher areal densities at acceptable error rates. This letter also requests that all operating system developers, storage systems developers, electronic circuit designers as well as all users of hard disk drives work with Hitachi GST, and all disk drive producers, study disk drives formatted for a longer data block. This work would precede the adoption of this new standard, and be completed before year end 2004.

**Sponsorship:**

Under the sponsorship of the International Disk Drive and Equipment Association (IDEMA) a committee was formed to study long data blocks, recommend a new standard data block length if required, formulate an implementation strategy for the industry and seek support from component producers, operating system developers and hard disk drive users. This committee had representation from a majority of disk drive producers, principal component suppliers and worked in contact with the T13 ATA interface standard committee and the T10 SCSI interface committee. In addition, the committee has discussed progress in adopting a longer data block standard with selected disk drive users as well as operating system developers. Hitachi GST participates in the IDEMA committee and supports its results.

**Justification:**

The committee has made a recommendation that the new data block standard be 4096 bytes in length. It is understood by the committee that migration to a longer data block length would necessitate modifications in operating systems for some SCSI (the current SCSI interface/protocol does support 4K data blocks) and most ATA interface drives, and would require system/user changes. While this recommendation is not made without adequate justification, Hitachi GST proposes the adoption of a 4K standard based on

areal density studies which relate bit lengths to defect lengths. (A detailed description of the rationale for adopting 4K byte data blocks is found in the appendix to this letter, and a brief summary is included here.)

The principal need for a longer data block originates from error correcting code (ECC) efficiency. Increasing areal density is a trend in the disk drive industry, the rate of this increase may slow in the future, but it will continue to increase. As this areal density increases, the signal-to-noise ratio will decrease and serial error rate will increase, that is, defect sensitivity increases. The influence of long burst errors originating from the media will also increase and the present 512-byte data block does not provide adequate ECC to assure a reliable disk drive operation.

There are two fundamental solutions to this trend. The first is a trivial one, that is, to invest more disk space to ECC bytes that would, hopefully, assure continued data reliability. A result of this course of action is less disk areal efficiency. As areal densities grow, it would seem that disk space is cheap enough to allocate larger proportions of available space to ECC without causing serious impacts on disk drive capacity. Unfortunately, this solution suffers from the realization of diminishing returns. A high percentage of ECC bytes, added to assure reliable operation of 512 byte blocks at increasing linear density, would themselves increase the soft error rate due to their length, defeating their very purpose.

The second solution is to increase the data block size, while slightly increasing the percentage of ECC bits. The committee chose the new data block size to be 4096 bytes, a multiple of the existing 512-byte block standard. An optimum selection of ECC bytes would assure data integrity of an entire 8 sector group of 512-byte sub sectors, or a 4096-byte block. This is the committee's recommendation. The general opinion of the committee is that 4096 bytes would be the industry standard for some time but that the future could bring another increase beyond 4096 bytes to 8192 or 16384 bytes based on ECC requirements. This further migration was regarded as too distant to be realistically considered by the committee in detail.

Adopting the 4K standard is straightforward for SCSI interface drives based on the variable block size of this interface. For ATA interface drives, more involved modifications in either the drive or operating system, or both, could be necessary. This would be dependent on the type of operating system used.

**Timeframe:**

It is likely that after the year 2004 the migration of disk drive products from 512-byte to 4096-byte data blocks will become increasingly more necessary. Some server drives have already been made available with a 4K-byte capability using the SCSI interface. It is expected that desktop and mobile drives with a 4K-byte block/ATA interface could be available after 2004. During this time period legacy 512-byte drives using existing operating systems would continue to be in production while the newer 4K-block drives, probably at a higher areal density beyond 100 Gigabits/in$^2$, would begin to become available. It would be the implementer's option to use drives with either data block length during this period. Eventually SCSI and ATA interface

drives operating with 4K-block sectors would be available for the entire industry and a majority of disk drive users would implement these 4K drives.

Disk drives could be identified as 4K-byte by contact with a specified pin in the connector that is unused today.

**Implementation:**

The availability of a modified operating system that includes write commands which permit aligned 4K block writes would be the optimum implementation approach. This operating system should be available at the time of production for these new 4K drives. Users who elect to continue with legacy 512-byte operating systems at his time would have these choices: (1) sustain a performance penalty based on read-modify-write occurrences, which is in all likelihood an unacceptable option (2) adopt an approach as Integrated Sector Format using a multilevel ECC code. This approach is based on the need for an ECC that can accommodate both 512-byte and 4096-byte data blocks, i.e. to function with existing legacy drives and take advantage of the enhanced data integrity though the use of 4K-byte blocks. Other options may also exist, but these have not been discussed in the industry to date.

Hitachi Global Storage Technologies supports the recommendation of IDEMA Committee that disk drive users work with disk drive producers to study the results of adopting a 4K-byte operating system, anticipating that a majority of drives in production would operate with 4K-block sectors after a selected date. For ATA drives, the recommendation is that a suitable operating system be developed which will allow direct mapping of physical and logical 4K byte sectors.

**Conclusion:**

The intent of this document is to communicate Hitachi GST's position as a disk drive producer to migrate to 4K byte data blocks for products available in the post 2004 timeframe. Also, it is the intent of Hitachi GST to work with operating system developers as well as all disk drive users to investigate the migration to a longer data block standard. It is recommended that the necessary OS software changes, as well as other software, be implemented to allow hard disk drives with long data blocks, as 4K bytes, to be applied in an optimum way.

## 6.3 Fujitsu Open Letter to the Storage Industry, Oct 27th, 2003

This letter describes Fujitsu's position on extending the length of data blocks on Hard Disk Drives (HDD). The HDD vendors are converging on a plan to increase the physical data sector length from 512 Bytes to 4,096 Bytes. We are driven to this decision by the need to improve areal density without increasing the data error rates.

Over the past 50 years the HDD industry has succeeded in overcoming many hurdles in order to deliver cutting edge technology. We have focused on

improving capacity, with greater reliability and lower costs. Through this period, we have managed to keep the physical sector size stable.

Increased capacity is not simply about storing more movies on one hard disk drive. The HDD industry has taken advantage of capacity improvements to reduce form factors, reduce acoustics and reduce power; we want to continue to do this. In addition, error rates have remained constant over the years, even as the actual bit density has multiplied.

We have managed to maintain the error rates by improving all of the fundamental hard disk technologies and delivering a continuously improving signal-to-noise ratio(SNR). However, the error correction code technology (ECC), applied to each physical sector to enable multiple bits in error to be corrected on the fly, is approaching the point where the correction bytes needed are causing a significant loss of track utilization efficiency.

We are currently shipping hard disk drives where the ECC and other disk recovery data is approximately 15% of the sector size. Over the next few years, we expect to see the greatest opportunity in SNR improvement coming from adding more ECC bytes. This will push ECC to >30% of the physical sector. We expect to reach this point around 2006 if the industry follows its current development schedule.

We will, of course, continue to improve SNR using all available technologies such as RLL coding, Recovery Algorithms, etc. and it is possible that one or all of these techniques may delay the increased ECC requirement by another generation. Regardless of the exact timeframe, it is Fujitsu's position that expanding the physical sector size is inevitable in the next 3-5 years.

The most logical step appears to be to a 4K or 8K block size. 4K bytes has been accepted by the hard disk drive vendors as the optimum size. Fujitsu is requesting that all software suppliers evaluate this change and investigate the impact on their systems. The hard disk drive vendors would like to enter into negotiations to resolve the issues and agree on a solution that does not cause any slowdown in the development, and improvement of hard disk drives.

## 6.4 Seagate Position on Large Blocks, October 06, 2003

### Statement

Seagate endorses the need for long blocks to continue advancement in magnetic disc drives.

### Motivation: The Need and the Consequences

As areal densities have climbed to reach the almost inconceivable level of recording density we are at today, improvements have come from numerous sources to ensure that each generation of hard drives stores data with reliability and speed. These sources have included improved media, heads, data detection, and error control coding (ECC). The changes have sometimes been gradual and sometimes breakthrough, but one constant has been that next-generation components must recover smaller bits with lower signal-to-noise ratio (SNR). Lower SNR results in worse error rate, and worse error rates make reliability a more difficult goal.

Customers tell us that reliability is paramount. One near-term solution to improve reliability is advanced channels, but these need to be implemented in a way that does not excessively impact format efficiency. This progress will require larger physical block sizes. Furthermore, large blocks afford greater efficiency for data transfer (less overhead per block written and read); reduced format time; and faster drive maintenance (scan disc and defrag run times). The hard drive magnetic recording industry under the lead of the International Disk Drive Equipment and Materials Association (IDEMA) needs to adopt a unified position to drive toward a standard large-block format. The primary roadblock is legacy software that requires a 512-byte block size.

**The Large Block Solution**

We believe the best solution is an industry standard for large-block sectors. A large block standard would provide needed freedom of action for the hard drive manufacturers, permitting them to continue to develop higher capacities and differentiate their products on a reliability basis. Large blocks also provide a clear path for future gains through further increases in block size. Most importantly, large blocks meet our customer's requirements, simultaneously providing greater reliability and greater data transfer performance.

This standard must be accepted by the software industry and implemented in such a way that it will be easily updateable and backward compatible. Certainly, legacy software is an issue in this solution. Unfortunately, we do not have the advantage of an industry wide driver that the Y2K potential threat presented to get the industry motivated to adopt this admittedly difficult change. To minimize the impact to all parties involved, this standard needs to be put into place TODAY, so that when it is rolled out in products, software manufacturers will be prepared with new code; hard drive manufacturers will be able to provide enhanced reliability and customers will notice no change.

Seagate supports an industry-wide transition to a large, 4 Kbyte, sector-size standard with product introduction times in 2006. No other solution will meet the requirements of our customers while allowing us to improve reliability.

## 6.5 IDEMA Long Data Block Committee Letter to Nathan Obr, Microsoft, Corp., November 19, 2003

Mr. Obr:

In December 2002 you attended a teleconference of the IDEMA Long Data Block Committee and listened as several Hard Disk Drive vendors discussed our desire to investigate increasing the size of the data block in the various Disk Drive interfaces (IDE/ATA and SCSI). At the conclusion of that call you expressed a concern that it did not appear to you that all of the HDD vendors were supporting this investigation and, speaking for Microsoft, you were reluctant to engage in this effort until the various HDD vendors communicated their commitment to this project. We are writing to you today to confirm the commitment of four HDD vendors to this investigation. Attached to this letter are 'Position Statements' from Hitachi, Fujitsu, Maxtor and Seagate stating their corporate positions regarding the investigation of longer data blocks. We forward these to you in the hope that they will effectively communicate the

various corporate positions and address your earlier concerns regarding commitment.

In addition to these position statements we are writing to request your participation in a meeting with representatives from the HDD vendors to discuss the many issues related to increasing the data block size. We would like to hold this meeting at your location in order to facilitate as much Microsoft participation as possible without causing you travel overhead. We do not propose the migration to longer data blocks without recognition of the serious issues it may cause to Operating System and Application development, Operating System and Application support cost, product compatibility, and system integration. The IDEMA committee feels the time is right to begin exploration of these issues and we request Microsoft's help and input.

We ask that you review the corporate position statements and respond to Ed Grochowski, chairman of the IDEMA Long Data Block Committee, with any questions or concerns. Likewise we ask that you coordinate with Ed regarding the timing and location of a meeting where we can discuss the many aspects of longer data blocks with Microsoft.

We appreciate your participation in this effort and look forward to working with Microsoft on this project.

Regards,

The IDEMA Long Data Block Committee representing Hitachi GST, Fujitsu, Maxtor, Seagate

## 6.6 Maxtor Position Statement for the OS, Application & System Industry, Advanced Technology Memorandum

**Purpose:**
The purpose of this letter is to communicate Maxtor's position regarding the migration to longer data blocks at the host and physical layers in Hard Disk Drives. Maxtor supports continuing work on longer data blocks, and we believe that the next crucial step is to begin a dialogue with the Operating System, Application, and System Integration providers to address the problems brought about by increasing the block size and to point out the potential benefits of longer blocks.. We believe that, to date, the system-level obstacles caused by longer data blocks have not been fully enumerated, keeping us from understanding the full effect of any increase to today's block sizes. It is through discussions with OS, Application, and Systems Integration providers that these issues can be brought to light, discussed, prioritized, and ultimately solved allowing for the migration to longer data blocks.

**Sponsorship:**
Under the sponsorship of the International Disk Drive and Equipment Association (IDEMA), a committee was formed to study long data blocks, recommend a new standard data block length if required, and seek input from component producers, operating system developers and hard disk drive users. This committee had representation from a majority of disk drive producers and principal component suppliers and worked in contact with the T13 ATA interface standard committee and the T12 SCSI interface committee.

**Justification of Position:**

The Hard Disk Drive industry has been using the same 512-byte physical data block since nearly the very beginning of the industry. In the past there have been several attempts to lengthen the data block size. These attempts were unsuccessful because the drive- level advantages available with longer blocks were not significant enough in the face of the system- level difficulties to warrant the migration. While the system- level issues have not changed, we believe that drive-level circumstances have changed enough to make the advantages of longer blocks desirable. We have now reached a point where a better understanding of the scope and severity of the system- level issues is needed before we proceed with longer block development. Direct technical discussions need to be held with the OS and application developers, the system integration specialists, and the system providers to list and address these issues and come to a consensus regarding the migration to longer data blocks. Since this appears to be yet another in a long line of block-lengthening efforts, it is important to state what changes have occurred that now make longer block advantageous. We begin with the industry-demanded characteristics of a disk drive: data integrity and data transfer rate. Achieving the industry demands in these areas affects the overall cost of the drive. In the past the Hard Disk Drive industry has been successful balancing these design factors with product cost, but this is getting increasingly difficult as drives evolve. Fundamentally this is an issue of scaling. We have reached a point in areal density where the standard block size (512 bytes) represents an infinitesimal fraction of the available storage capacity of a disk surface – nearly 1 part in a Billion (at the current 40 GB/surface capacity point). To meet the demands for data integrity and transfer rate each of these physical blocks requires its own overhead – extra data written on the disk in order to correctly read and recover the block. Greater industry demand for improvements in capacity, integrity, and transfer rate cause the required overhead to increase, but now, due to the small block size, the overhead is increasing at a more rapid rate. This rapid increase in overhead affects the cost of the drive in two ways:

1) Data Integrity gets more costly: On a 512 byte block the industry has reached the ‘peak’ of error correction capability. This means that increases in ECC capability come at a net *loss* in areal density. This is a new phenomenon. In the past we were able to increase ECC and achieve a subsequent net *gain* in capacity. Maintaining industry-demanded data integrity will require higher actual densities (and thus higher costs) for a given capacity point. By spreading out the ECC over fewer physical blocks (by making the blocks longer) we can return to the era of net *gains* in density with increased ECC capability. Moving to longer blocks allows us to maintain, or improve, data integrity in a cost effective manner.

3) Data transfer rate gets more costly: As capacity and spindle speed increase, the amount of overhead required to successfully write and read the data increases in order to maintain industry-demanded transfer rates, thus increasing overall cost. While not new, this phenomenon is accelerating as capacities, spin speeds, and transfer rate criteria reach new heights. As with ECC, we can mitigate this increase by moving to longer blocks and spreading out the overhead over more of the data. Longer blocks allow us to maintain, or even improve, data transfer rates in a more cost-effective manner. Since data integrity and data transfer rate requirements show no signs of receding, the effect of maintaining today’s block size will increase the cost of the drive.

Moving to longer physical blocks will allow the drive manufactures to maintain the balance between cost, transfer rate, and data integrity. It is estimated longer blocks will enable >10% gain in capacity and transfer rate (with constant areal density)

**Problems:**

As described above, the landscape at the drive level has changed making longer blocks extremely advantageous. What have not changed are the system- level obstacles that make moving to longer blocks extremely difficult. It is these issues that we wish to discuss with the OS, Application, and system providers in order to properly weigh our desire to move to longer blocks against the system-level difficulties associated with it. Among the issues we know exist, but do not have enough experience to appropriately measure, are the following: We are aware that longer blocks will cause difficulties at the Operating system (especially "legacy" OS's), file system, and device driver level. Thus we wish to talk with Operating System manufacturers to discuss the nature and severity of these issues. Likewise, we are aware that longer blocks pose problems for certain applications that maintain direct access to the disk drive, and may even rely on physical placement of data in 512 byte blocks. We wish to talk with application providers that have such constraints to understand the nature and severity of these issues as well.

Finally we are aware that the system- level integrators will have manufacturing issues associated with a 'new' drive with longer blocks. We, as Hard Disk Drive manufacturers, will also share similar manufacturing issues, including backward compatibility, management of a transition period between 512 byte and longer block drives, and post-sale support. We wish to discuss these manufacturing issues with the system providers to understand the scope of these difficulties.

We are confident that in addition to these issues there are others that have not yet been raised, but certainly need to be addressed before longer blocks are commercially viable. These issues can only be effectively addressed, and potentially solved, through open technical dialogue between the disk industry and the various manufactures and providers mentioned. Maxtor would like to see those discussions take place and is eager to be an active participant in any such discussions. Only by having a thorough understanding of the issues at hand – at both the drive level *and* the system level – can the industry move to more advantageous block sizes.