Large Block Size -- Draft of NSIC White Paper

Paul Hodges and David Cheng, IBM Corp.

The National Storage Industry Consortium is pursuing a study of Extremely High Density Recording for disk drives, aimed at developing the technology needed for recording data at 40 - 100 GBits/sq-in. One part of that study proposes the use of native block sizes larger than the current de facto standard of 512 bytes. This paper describes the proposal, sets forth the rationale, and provides a road map for getting to the larger block size.

1. Proposal

It is proposed that the native block size recorded on disks be increased from the currently pervasive 512 bytes to a larger size, initially 4096 bytes. The change is motivated by the expectation that, as we push the technology, an increase in recording density will not result in an equivalent increase in user capacity on the disk.

The rationale for needing larger block sizes to mitigate this problem is set forth in section 2, with further technical details in Appendix A.

While there is no immediate problem that would require larger block sizes, it makes sense to adopt a large-block-size strategy now, in order to provide sufficient lead time for both hardware and software companies to make the necessary changes.

A road map for moving to larger block sizes is described in section 3. A key part of that road map is the decoupling of hardware and software changes by including a lengthy period during which either block size can be used.

The International Disk Drive Equipment and Materials Association (IDEMA) proposes to write a standard for larger block sizes (see Appendix B). A draft will be included in this paper, when it is available.

2. Rationale

As recording density increases, it becomes more and more difficult to reliably retrieve data from the surface of the disk.

For many years, disk technology has advanced at a rate that allowed increased density with little degradation of the read process, so that the reliability of retrieving data has been relatively constant. Where there was degradation, the reliability of data was restored by means of more powerful error-correcting codes (ECC). Since the redundancy associated with these codes was modest, the overall effect has been that the capacity of the disk as seen by the user has increased in proportion to the recording density, at a compound growth rate of about 60% per year, with commensurate improvement in cost/megabyte.

Large Block Size -- Draft of NSIC White Paper

Paul Hodges and David Cheng, IBM Corp.

In not too distant future, density increases will stretch the limits of the technology, so that reliable reading must depend strongly on enhanced ECC and there must be significantly more information in a block for proper initiation of the read process. The overhead needed for clocking data, decoding data, and correcting errors will be substantial. Since much of this overhead is "per block", rather than "per byte", short blocks will be much less efficient than longer blocks. The result is that users could see the rate of increase in capacity slowing, even if the density continued to increase on the 60% per year basis. Eventually the rate of technological improvements may be less than 60% per year, which will further slow the rate of improvements in capacity and cost per megabyte.

At today's densities, changing the native block size on a drive from 512-bytes to 4096-bytes would increase the capacity seen by a user by 7 to 10%. Since that represents about one quarter's advance in technology, a change in the common block size has relatively little benefit today. However, in the middle of the next decade, when recording density is expected to reach 100 Gbits/sq-in. with significantly degraded raw error rate, changing the native block size on a drive from 512 bytes to 4096 bytes is estimated to increase the capacity seen by a user by 25 - 50% (depending on the assumptions made about future disks). At the current rate of advance, that leads to two to four quarters delay in achieving a particular level of cost/capacity. If, as must eventually happen, the rate of density increase slows, the delay is even greater. At that point, a change in the native block size is justifiable.

3. Road map

A transition to larger block size at some time in the future is seen as inevitable. In order for the industry to be able to prepare for that change, it makes sense to provide enough lead time for both hardware and software providers. It is intended that hardware and software evolution be decoupled, so that the change can be done in a orderly manner. The key points are

- a lengthy period in which data can be addressed as either 512-byte blocks or 4096-byte blocks,

- an architecture such that legacy software based on 512-byte blocks will continue to work with no change, and

- a transition time after which performance may suffer if data is addressed on a 512-byte basis.

Stage 1:

As a near-term strategy it is proposed that 512-bytes be retained as the native block size, and that new code based on 4096-byte blocks be accommodated. Thus designers of new systems and software would immediately be able to design with the 4096-byte blocks in mind.

For SCSI, disk drives would allow LBA addressing based on either 512-byte blocks and 4096-byte blocks with equivalent performance.

For IDE, systems and software designers would be encouraged to access all data on 4096-byte boundaries.

Stage 2:

In the mid-term, it is proposed that disks begin to use 4096 bytes as the native block size. Access on a 512-byte basis would continue to be supported, but performance would be inferior to that in which access is done on a 4096-byte basis, and might well be inferior to that of previous drives with 512-byte native block size.

Stage 3:

In the long term, it will be appropriate to consider even larger block sizes (in powers of 2). For example, DVD currently accrues significant advantage form using 32K-byte blocks.

4. Some system and software considerations

It is assumed that there will be little difficulty in using longer blocks for newly written programs, but it is recognized that much of the current software depends on 512-byte blocks. It is for that reason that a lengthy transition period is proposed.

In existing software, it may be difficult to find all instances in which a program depends on 512-byte blocks. Therefore, it can be expected that new versions of existing programs will have most, but not all, accesses converted to 4096-byte blocks. In Stage 1, the residual 512-byte accesses are expected to work as they do now. In stage 2, the residual 512-byte accesses may have relatively poor performance. Obviously, it will be advantageous to convert all accesses that are critical for performance.

Informal contacts in the industry have indicated that 4096-byte blocks could be accommodated with sufficient lead time, but that very large blocks (e.g., 32KB) would require significantly more effort.

Some of today's systems use blocks that are not precisely 512 bytes. At their request, disk manufacturers build variations ranging from 504 bytes to 524 bytes. There are two potential approaches for those systems that cannot convert to 4096-byte blocks:

- retain the currently used block sizes, taking a penalty in cost/megabyte of up to 50%, or

- convert to longer blocks that are a multiple of the current block size, i.e., 4032 to 4192 bytes.

Appendix A

Today's disk drives have raw error rates that permit the use of 512-byte sectors with acceptable overheads in data sector, track servo and error correction code (ECC). Over the next several years, it is expected that fundamental physical phenomena such as superparamagnetism will result in a decrease in signal-to-noise ratio (SNR) and a corresponding increase in raw error rate. The overheads must be increased in future products in order to maintain the current level of reliability and user error rate (rate after error correction).

The amount of overhead required depends heavily on the statistical nature of the errors. Media with bursty errors (error statistics with strong correlation) require a large amount of overheads at high raw error rate. Overheads for this type of error statistics can be significantly reduced by using large sectors. On the other extreme, media with random errors (error statistics with no correlation) require the least amount of overheads. The overhead gain from larger sector size is also less significant for random errors. Figure 1 [chart of format efficiency, to be distributed by Gene Milligan] shows the trend of format efficiency at high density based on a particular set of assumptions. The efficiency for 512-byte sector format drops rapidly at very high density, and this efficiency loss can be recovered with larger sectors. The exact density at which the format efficiency drops rapidly depends on the assumptions of the model. But the general trend shown in figure 1 is relatively insensitive to assumptions of the model. The increase in overhead and the corresponding reduction in user capacity will become excessive at some point in the future. The disk drives must then be designed with larger sector size in order to maintain an acceptable format efficiency.

Optimistic Case:

In order to estimate the lower bound of the benefits expected from larger sector size, we use a model of random error statistics with the following conditions:

③Signal-to-noise ratio 3 dB lower than current products
③Worst case raw byte error rate of 10^{-2}
③ECC corrected byte error rate of 10^{-15} or lower
③Reed-Soloman error correction encoder-decoder
③100 Gb/in^2 areal density

The major contributors to format efficiency are: ECC, data sector (synchronization field and preamble), zoning, and track servo. The error-correction overhead is estimated by using the Poisson approximation to the binomial probability of decoded symbol error for the Reed-Solomon decoder, which is given by:

[Formula: binomial probability of uncorrected error; could not export to MS Word]

where **p** is the probability of raw errors at the input to the decoder,
N is the number of bytes in a sector,
t is the maximum number of errors the decoder is allowed to correct, and
the equation sums over all values of **j > t**.

Based on considerations from signal processing theory, the length of the synchronization field (in bits) is scaled inversely proportional to the power signal-to-noise ratio, and the length of the preamble is also scaled inversely proportional to the signal-to-noise ratio. The length of the servo burst (in bits) also varies inversely with signal-to-noise ratio, but does not change with sector size except for the case in which some sectors must be split to accommodate the servo bursts. Using these scaling rules, we derived the format efficiency for different sector sizes shown in the following table:

Block Size	**512 bytes**	**1 KB**	**2 KB**	**4 KB**
ECC	0.889	0.916	0.936	0.960
Data (Sync)	0.855	0.919	0.957	0.978
Zone	0.960	0.966	0.978	0.978
Servo	0.843	0.843	0.843	0.843
Sector split	1.000	1.000	1.000	0.996
Rel. capacity	**100%**	**111%**	**120%**	**125%**

More realistic case:

The results of the above computation show that a disk drive with random byte errors would have low format efficiency for 512 byte sectors. Part of the efficiency loss can be reclaimed by using larger sectors: Disk drives with 4 kilobyte sectors have 125% capacity compared to drives with 512 byte sectors.

Since error distributions in disk drives are generally bursty, rather than random as assumed in the model, one can expect an even higher level of recovery in capacity for real products. In a series of experiments in which a high level of raw error rate was simulated through stress testing in off-track reading and BPI (high linear density), we have observed a range of capacity improvements. The minimum improvement observed is consistent with the random error model above, and the maximum improvement of 151% observed is consistent with a bursty error model.

Appendix B -- Proposal from IDEMA for draft standard

Proposed IDEMA Standards Committee Topic -- Increase in Disk Drive Data Block Size

Today's disk drives use an error correcting scheme which allows for acceptable operation at a worst case raw error rate as low as 10^{-6}, and with an acceptable overhead of disk space consumed by code redundancy. Over the next five years, it is expected that fundamental physical phenomena will result in a lowering of the disk signal-to-noise ratio, and a required increase in the size of the error correction code overhead. As a percentage of raw capacity, this overhead would be excessive for small data blocks. Today's drives use a 512-byte minimum block size for reasons of compatibility -- some operating systems still call for records in multiples of this size. Although some modern operating systems (such as Windows NT) use a 4K block size, it is still impossible to market a drive today that doesn't efficiently service 512 byte records.

We would like to promulgate a named interface standard that explicitly uses a 4K (or larger) block size, and which works in older systems only by an emulation technique (read before write) that involves a substantial performance penalty for the use of blocks smaller than 4K. We would then like IDEMA to work with operating system companies to make sure that future operating system releases will have a default block size of at least 4K bytes. It is expected that this requirement will be important within the next five years.

The benefit to the disk drive industry of this proposal will be the ability to develop and market higher capacity drives sooner than would be the case under the present AT and SCSI formats.

. . . The committee's objectives would be to reach agreement on the preferred block size for these future disk drives, and to approach the software houses with the message that future systems will be more cost effective if they plan for this change in format.