

Wise Drives

Manufacturers may soon be unlocking the computing power hidden in your hard-disk drive

If disk drives were enabled to perform a host of tasks that have traditionally been the work of CPUs, dramatic low-cost improvements could be made in computing environments ranging from laptops to storage networks and drive makers would gain new profit centers. But this will only happen if computer and drive manufacturers seize the opportunity and create new interfaces for these so-called intelligent disk drives.

Disk drives currently come equipped with 32-bit internal microprocessors, several megabytes of internal RAM, and tens of megabytes of disk capacity reserved for internal drive purposes. This onboard computing power has increased alongside improvements in the speed and capacity of disk drives, and has already allowed them to take on several intelligent tasks beyond basic data storage and

retrieval, such as detecting imminent drive failures. This power could be exploited for even more tasks, and the addition of some relatively inexpensive extra processing power would open the door to a huge array of functions, such as searching and encryption, and so shift some of the processing burden from CPUs and networks to disk drives.

Modern drives are nearly all standardized: they have common physical dimensions, electrical and data connections, and operational commands—all defined by ANSI standards. Interfaces between computers on one side and

drives on the other constitute an agreed contract on the rules for storing and retrieving data. Adding intelligent drive features may require changing these entrenched standards, and may also require changes in operating system and application software. High-level control of intelligent features may best reside in user applications, allowing them to tune a drive's performance to their specific needs.

It is my hope that after reading this article, developers of such software would become involved in the interface committees of organisations such as the

BY GORDON F. HUGHES
University of California
(San Diego)

Storage Network Industry Association (SNIA), which currently consist primarily of drive makers and drive buyers. But why should intelligent features appear now, when for decades disk drive development has been fixated on faster access to more data?

For 50 years hard disks have acted as relatively dumb devices, simply serving and storing blocks of data. The first generation of hard-disk drives were the size of refrigerators and stored a few megabytes for mainframe computers. Since 1992, drive storage capacity has been increasing at an annual rate of 60 percent, and over 100 percent in recent years—a rate exceeding the 18-month doubling of Moore’s law for IC complexity [see graph below]. Single drives today can store up to 200 GB and standard 3.5-inch disk drives are in everything from PCs to supercomputers (laptop computers use smaller 2.5-inch drives).

But the drive manufacturers have to some extent become victims of their own success. The industry has matured and is delivering highly reliable low-cost storage, at capacities that exceed the real needs of the majority of today’s end

users, typically desktop PC users. All drive manufacturers offer similar-performance mass-market products, which are seen as commodity items.

In my position as associate director of the Center for Magnetic Recording Research (CMRR) at the University of California, San Diego, I see how this commoditization shrinks the number of drive manufacturers and drive component makers, and reduces the number of research jobs for our students—hampering the very development that produced the remarkable drive technology advances in the first place. Customizable intelligent features would allow drive makers to distinguish their products and provide a competitive edge and an impetus for innovation.

The most common drive interface in consumer desktop computers is the advanced technology attachment (ATA) interface, also known as integrated drive electronics (IDE). The small computer system interface (SCSI) is prevalent in high-performance systems. Each interface consists of an adapter integrated into the host at one end connected via a cable to the disk drive(s) at the other end.

In both ATA- and SCSI-based systems, computers commonly divide a data file into a set of so-called logical blocks whose size ranges (depending on the platform and operating system) from 512 bytes to 4096 bytes and up. The logical blocks are passed over the interface to the drive, which stores them as physical blocks, usually of the same byte size, by magnetically recording bits on a cobalt alloy magnetic film coated on the disk surfaces. The physical blocks are stored in concentric tracks subdivided into angular sectors on the disks, which are constantly rotating during operation.

Conversely, in accessing stored data, the drive reads back each physical block magnetically as a noisy analog electronic signal and translates it into digital bits to reconstitute the physical block.

Already pretty smart

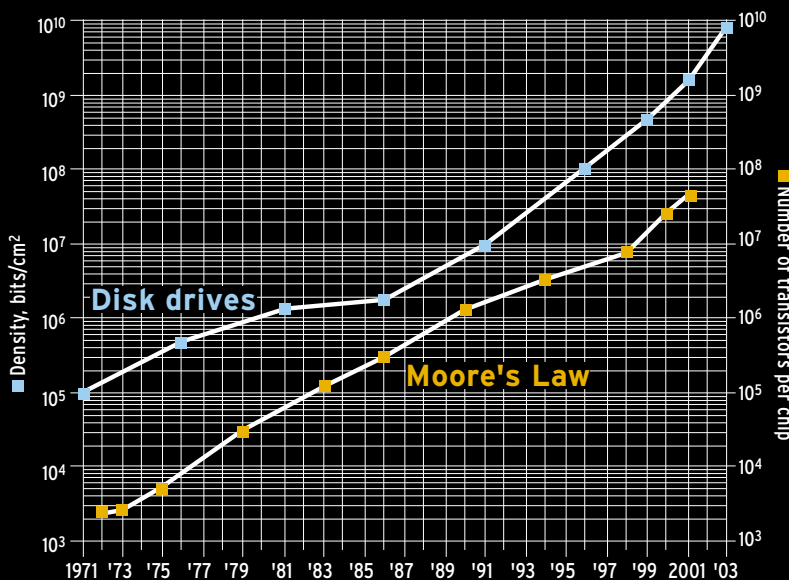
To do all this, disk drives already perform complex tasks hidden from the host computer. Errors can be detected and corrected. An impending crash can be anticipated. An array of disk drives can even recover data on their own from a disk that has crashed, without involving the system at large.

Data integrity is ensured by sophisticated algorithms that clean up the raw data from the disk using digital error checking and correction codes (ECC). ECC code bits are recorded along with user data in the physical blocks. A first level of error correction is built into the drive control electronics, to deliver corrected user data without delays while decreasing the probability of an incorrect bit from about one in a million at the playback magnetic bit detector to less than one in 10 trillion for the delivered logical block.

If this first line of error correction fails, additional levels of off-line error recovery are invoked. In disk drives, off-line means that the drive stops delivering data at full speed as time is needed to correct a data error. Typical off-line error-recovery techniques include: attempting to reread the data on a second disk revolution; rereading the data by positioning the read/write head slightly off track; and using a higher-level ECC algorithm in the drive microprocessor firmware. Hard-to-read phys-

An Embarrassment of Riches

The bit density of disk drives is growing even faster than the transistor population of microprocessor chips (which still obeys Moore’s law). With disk capacity no longer an issue for most users, drive manufacturers would do well to add features to stay ahead of the competition.



Source: IBM, Intel

ical blocks are reassigned (re-recorded) to alternative disk locations and the original locations are not reused.

The philosophy here is that returning error-free user data is so important that delays are acceptable. For some applications, though, this assumption may not be correct—users of a multimedia player showing a digital video clip may find a pause more distracting than the occasional dropped frame. Currently there is no good way to allow a drive to distinguish between different types of user data.

Another intelligent feature of today's disk drives reduces the chances that a hard-disk crash will leave a user without data. Disk drives have failure rates of

higher Smart warning accuracy is possible with improved drive firmware.

A third intelligent drive task today helps high-end storage systems, those using a RAID-5 disk organization (redundant array of independent disks). These systems store data blocks over multiple drives to gain the speed of parallel access in such a way that drives can fail and be replaced without loss of user data. Such computer systems mostly use the SCSI interface, which allows drives to recover the data autonomously from the failed drive, without burdening the host computer system with the work. So drives are already networking without computer help.

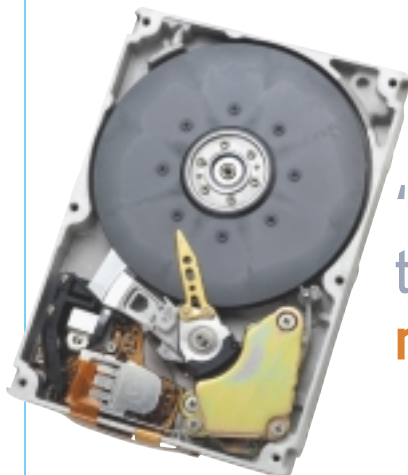
Better, faster than before

But if more intelligence is to be added to disk drives in the form of new firmware

deemed excessive, or to optimize a group of data accesses. Today, drive QoS refers to nothing more precise than the average data-access speed of a group of drives.

Off-line error correction also introduces unexpected data delays. Currently, drives sold for video streaming applications may simply have off-line error correction disabled to guarantee throughput, but QoS commands could allow computers to turn the off-line correction on and off depending on the file type—off for MP3 music data files and on for application software files, for example.

Sending data over a network to a drive or array of drives can also cause data errors, which must be detected. Currently, drives calculate a data-error detection code (called a CRC polynomial code) before sending blocks across the interface to the computer, intended to allow



“Commodization [of disk drives] shrinks the number of manufacturers and reduces the number of research jobs”

only about 1 percent per year, but the loss of a user's data can easily be more costly than the drive itself. If they were forewarned of an impending drive failure, users could take steps to back up their data onto another storage device.

To this end, in 1994 the drive industry adopted a standardized specification for such failure warnings called Smart (self-monitoring and reporting technology) at the request of Compaq Computer Corp. Smart is based on monitoring for excessive internal drive errors, such as bit-read errors and track-seek errors. A failure-warning algorithm running in the drive's microprocessor checks whether error rates exceed a threshold value and if they do, produces a warning that is sent over the drive interface to the host CPU.

Virtually all disk drives today have this failure-warning system built in. All the same, some computers simply ignore the warning. A research program on Smart here at the Center indicates considerably

and/or extra onboard processing power, computer manufacturers must be convinced of the benefits of the new features to them and to application developers and end users. An example would be in-drive tools for handling data delays, integrity, and corruption control. These would allow so-called disk-drive quality-of-service (QoS) features.

In retrieving user data, drives have variable delays due to track-seek and disk-rotation times. A drive may have to wait several milliseconds for its read/write heads to travel back and forth over a span of tracks, and then wait a few more milliseconds for disk rotation to bring the desired data block under a head. When beginning an access, drives can internally determine these motion delays, but there is currently no way of informing the application using the disk drive—even though applications like streaming video cannot use data that arrives too late.

Interface QoS commands could be added to provide precise information about delays and allow applications to abandon a data access if the delays were

checking of storage device connectors and cables. The CRC system could be extended to assist storage network administrators, allowing fault tracing down to individual hardware such as routers, disk and tape interfaces, and drives, and to environmental factors such as the power supply and ambient temperature.

Orderly storage

Storage systems such as file or database servers offer a fertile ground for other improvements. Many of them cache logical blocks in their RAM during drive accesses. This helps because consecutive logical block operations are often made on the same or the next logical block in the sequence, forming a complete file or database record. Simultaneously, drives perform physical block caching (such as reading all the sectors on a track into the drive's on-board RAM when a single sector is read) to speed access because the next logical block needed is often also the next physical block.

These two caching operations are unaware of each other and may be redun-

dant. Worse, they may even mutually interfere, as when the storage system demands a chain of logical blocks to cache, which may be scattered over several disk tracks, while the disk drive is trying to cache all the physical blocks in each track. Intelligent drives could allow servers to control drive caching on the level of entire files. To this end, an Object-Based Storage Devices (OSD) standard has been proposed for SCSI systems by the SNIA.

A storage object may be any number of things. It can be a file consisting of an ordered set of logical data blocks, a database containing many such files, or just a single application record such as a database record of one transaction. Information about the data is also stored in an object, which can include QoS, security, caching, and backup requirements. OSD disk drives could perform data pre-access, which could put sequential logical blocks in the drive's cache memory, instead of sequential physical blocks. Drives could transmit them at the full interface transfer rate without waiting for disk rotation or track seeks. And once drives start dealing with data objects instead of their individual logical blocks, even more possibilities open up. Storage systems, which must handle such tasks as data mining, backup, and com-

pression, could be implemented and scaled up much more efficiently.

Object-aware drives could speed up the process of making backups by maintaining internal lists of which object blocks have been rewritten since the last backup (currently a housekeeping task managed by host computer backup software). Only altered blocks would need to be transferred to an incremental backup system, instead of entire files.

Data compression on individual files could also be performed by OSD-based systems. Without OSD, compressing data at the drive level can generate a block consisting of data from multiple files. If any errors occur during compression, all those files are at risk. Another risk is storage capacity overflow, because the size of a compressed file isn't known until its compression is finished.

Spreading the load

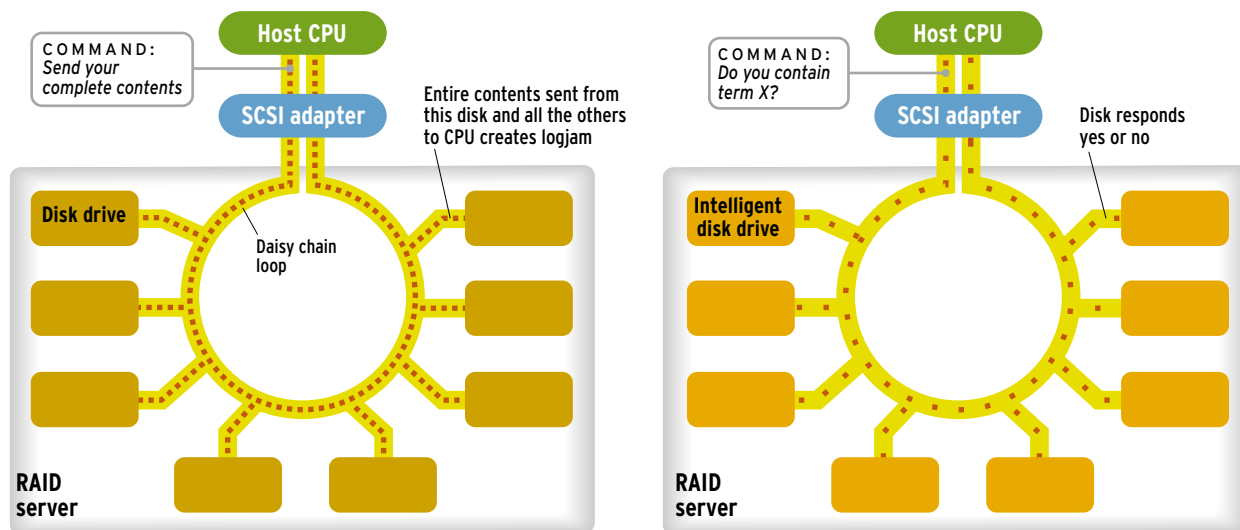
But data mining has perhaps the most to gain from object-aware drives. Data mining searches are a major bottleneck for the data farms that provide storage for such applications as e-commerce and scientific analysis of very large data sets such as the human genome.

Designers have long known that these intensive data mining searches

benefit from parallel or distributed architectures. Dealing with such applications has given rise to many solutions in which storage intelligence has been moved as near to the drive as possible, but ultimately still resides on the computer side of the drive interface.

Since the 1980s, specialized database systems have been built that use a main computer to divide search jobs into independent tasks that are sent to sub-processors for execution, each having its own memory and own local disk drives. The parallelism allowed searches of terabyte-sized databases with an order of magnitude performance increase over normal serial computer systems. But these specialized systems were expensive—and they still just used standard disk drives.

Today database computers more commonly use storage systems based on RAID systems using fiber-channel SCSI drives. Damaged disk drives in these redundant arrays can be replaced without losing data. Up to 127 drives can be daisy-chained in a loop that can accommodate two simultaneous drive commands. Because many drives are running in parallel, an array is also fast. Yet the maximum data transfer rate of the array can be far below the combined



● Easing Traffic on the High-Performance Data Highway

Currently, data searching on a RAID server is inefficient. To find records containing a given term all the disk drives must send their entire contents back to the host CPU for examination, causing data congestion in the daisy chain loop [left]. Intelligent drives could be sent the search term and autonomously return just those records that contain it, allowing faster searches with less congestion [right].

transfer rate of all the drives in a loop—because all the drives share the bandwidth of a single interface connection. For example, a database query requiring searching through all the drives could be as little as 2/127, or 1.6 percent, as fast as the maximum data transfer rate of all the drives operating independently.

As a consequence, in storage systems where speed is a concern, loops are limited to five or six drives each, trading off cost for performance. A technique known as fiber-channel switched fabric can also address the bottleneck by establishing direct independent connections between the storage computer and the drives, but again at increased cost.

Putting low-level search intelligence inside drives could be faster and far less expensive. The database application would break searches into individual commands, which would be sent out simultaneously

on a drive, including any disk flaw reassignments. Still, it can take over an hour to repeatedly overwrite all the blocks on a modern 180-GB drive—if data is only overwritten once, it may be recovered by scanning the disk platter with a magnetic force microscope. Also, secure erase is still an optional feature in the specs and not always implemented. Nor does secure erase eliminate the problem of securing data on disks in active use.

To protect data in active use, operating systems such as Microsoft Windows XP and Linux provide file encryption. However, encryption at the software level is complex to manage. Implementing file encryption keys and algorithms in disk drives would be easier, with the drives transparently encrypting the data. (Decrypting data from drives would remain an external software task.)

In short, there must be an industry consensus that the task is of general appeal and offers market opportunities for multiple computer and drive companies. A recent example is the addition of the specialized requirements for flash memory PC data-storage cards to the ATA interface specs.

Computer application developers, such as large database search engine designers, may need to first see potential benefits from QoS features demonstrated. University research projects may be the initial step (RAID storage systems were first studied at the University of California, Berkeley). Computer simulation and monitoring of storage networks can also allow assessing performance improvements offered by intelligent features in advance of actual deployment.



TO PROBE
FURTHER,
SEE PAGE 63

“The encrypted contents of a disk could be **destroyed in an instant** ... imagine such a feature onboard the EP-3 spyplane forced to make an emergency **landing in China**”

to all drives. Only successful search results from drives would be returned. [See illustration, opposite page.]

Inner security

Intelligent disk drives could also make computers much more secure and alleviate some of the concerns that have arisen in recent years in military, governmental, and business circles when laptops or hard disks themselves have gone missing.

With current systems, normal erasure of disk data does not necessarily prevent later retrieval of that data. Indeed, drives are designed to resist accidental erasure by a user. Special programs are widely available for retrieving apparently erased files, forcing security-conscious organizations to take precautions to ensure that disk drives have been completely wiped before being disposed of.

Commands that aid in this disk sanitizing were added to both the ATA and SCSI drive specifications at CMRR's request. The secure-erase process is similar to whole-drive formatting because it overwrites all user-accessible data areas

The long execution time penalty of secure erase would vanish, because the encrypted contents of an entire disk could be effectively destroyed in an instant by simply deleting the external software decryption keys (securely, of course). One can imagine the utility of such a feature onboard, for example, the EP-3 spy plane that was forced to make an emergency landing in Hainan last year as the crew hastily destroyed disk drives and other sensitive equipment in a bid to prevent information falling into Chinese hands.

Contingent arrival

While intelligent drive features can be added without interface specification changes by using the “vendor reserved commands” already defined in ATA and SCSI, this limits the features to just those agreed on between a single pair of drive and computer manufacturers. Such an arrangement may be a useful proving ground for a new feature, but for widespread use its input/output and command requirements need to appear in the interface specification.

Object storage devices have been studied at Carnegie Mellon University, and are under study by the SCSI spec committee. Interested members of the SNIA trade group are moving this proposal forward by working on the definition of an OSD standard.

Customers always determine the acceptance of new technologies, and many of these intelligent drive tasks are of primary benefit to final computer users, who are not the direct buyers of disk drives. For example, although Smart is advertised by PC makers Dell and Compaq as a principal hard-disk feature, PC makers are not currently asking for the higher-accuracy Smart warning discussed above, although it adds no drive cost. This may offer a market opportunity window for the first drive and computer makers to offer it. When developers and users become aware of potential new intelligent features, and ask computer and storage system makers for them, then drive makers will design them in. ●

Stephen Cass, *Editor*