PART

2

STORAGE TECHNOLOGIES

CHAPTER

4

THE PERPETUATION OF A DISK-BASED STORAGE STRATEGY

Data storage, of course, is a generic term, encompassing a broad range of devices and media. Current popular storage media include random access memory (RAM), solid state disk, hard disk drives, magneto-optical (MO) disc, Near Field Recording (NFR) drives (a hybrid magnetic optical disk technology), CD-ROM and DVD, and magnetic tape. Each technology has capabilities and limitations that need to be understood when designing a storage architecture or managing storage deployments that are already in place.

Given the preference in modern business organizations to keep all data highly accessible to end users, customers, and decision makers, the hard disk drive has become the preferred data storage device in production systems. To facilitate an understanding of current hard disk technology, a brief overview of its evolution may be useful.

EVOLUTION OF HARD DISK TECHNOLOGY

The earliest mode of data storage, dating back to the 1800s, was the punch card and paper tape.[1] By the early 1950s, card decks or punched tapes, as

well as magnetic tape, were the predominant storage technology for mainframe and minicomputers. Cards or paper tapes were loaded into reader peripherals so that their sequentially stored contents could be transferred into the random access memory of the computer.

This cumbersome storage technology underwent a significant change in May 1955, with the announcement of a new product by IBM: the RAMAC disk drive. RAMAC offered random-access storage of up to 5 million characters, weighed nearly 2000 pounds, and occupied about the same floor space as two modern refrigerators. Data was stored on fifty, 24-inch diameter, aluminum disk platters that were coated on both sides with magnetic iron oxide. The coating itself was derived from the primer used to paint the Golden Gate Bridge in San Francisco, CA.

Before the release of RAMAC, hard disk drives were considered an impossible dream. One engineering problem was a head fly height barrier. For a hard drive to write information to its media (typically a metallic disk coated with iron oxide), the media needs to rotate under the read–write head, which is positioned at the end of an actuator arm. Data is written to the media using an electrical field generated in the head, which, in turn, produces a magnetic pattern on media surface. Before RAMAC, engineers believed that to accomplish this function, heads needed to be spaced (to "fly") within 800 microinches of the media surface. This seemed an impossibility because irregularities in the media itself, and "runout" or "wobble" of disk media when rotating at 1200 rpm, necessitated that heads fly at least 20,000 microinches above the platter. Clearly, the head fly height required to make a drive work would also cause the heads to crash into the disk media during operation.

A group of IBM engineers, known as the "Dirty Dozen," succeeded in surmounting this barrier and the RAMAC drive was the result. IBM became the sole supplier of disk drives for several more years.

The cost for the new storage medium was steep. At $35,000 for a year's lease of a RAMAC 305, companies embracing the new technology had to spend about $7,000 per megabyte of storage for the privilege.[2] Despite the costs, RAMAC and its successors at IBM defined a new market. In a fascinating chain of events, the "Dirty Dozen" left IBM in 1967 and created a new company, Information Storage Systems, with the objective of creating competitive, plug-compatible, disk-drive products for IBM mainframes. Two years after their exodus, another IBM manager, Alan F. Shugart, left IBM, first to work with Memorex, then to form his own company dedicated to the perfection of the hard disk drive. Ultimately, his company, Shugart Technology, was renamed Seagate Technology.

Over the years, the hard disk drive went from being an expensive peripheral found only in mainframe data centers to becoming an integral, inexpensive, and ubiquitous component of server and desktop systems. Table 4–1 provides a summary of some of the firsts in hard disk drive technology.

The modern hard disk drive enjoys the distinction of being one of the most quickly advancing technologies in computing. Generational improvements have been made in hard disk technology on an almost annual basis since the early 1980s.

Introduced for the PC market by Seagate Technology in 1980, the first 5.25-inch hard disk drives offered 5 to 10 MB of storage—equalling, then doubling, the capacity of the RAMAC—in a form factor that was a small fraction of RAMAC's massive dimensions. The adjective "hard" became a part of the device description "disk drive" because the product needed to be discriminated from "floppy" disk drives already used in personal computers. Floppy drives derived their name from the metal oxide-coated Mylar used as a storage medium. The platter inside a Seagate hard disk drive (and most drives since that time) was a coated, rigid, aluminum alloy.

Another point of departure of the hard disk drive from earlier PC floppy drives was head design. In floppy disk drives, read–write heads made direct contact with media. The Seagate hard disk represented a return to the RAMAC design, since heads did not contact the media, but flew above it. This feature, combined with a durable disk media, gave hard disks a longer life expectancy than floppy media and soon made them a trusted medium for long-term data storage.

After the Seagate innovation, the market for disk drive technology exploded and product prices declined dramatically. Competition drove innovation—in head technology, platter coating, actuator arm control, and form factor. Within five years, 5.25-inch form factor drives were only three inches high and weighed a few pounds, while lower capacity "half-height" drives measured only 1.6 inches in height.

The 3.5-inch form factor hard drives appeared in 1987. Smaller than a paperback book and weighing only a pound, they enabled the laptop computer and later became standard components in both desktop and portable PCs, offering up to 500 MB of storage. They were soon challenged by 2.5-inch form factor drives weighing only four ounces and offering the same capacity as their larger cousins. Eventually, this miniaturization spiral resulted in IBM's 1998 announcement of a microdrive, which offered a capacity of 340 MB on a single disk platter the size of a small coin, weighing only 20 grams.[3]

Table 4–1 Firsts in Disk Technology

Vendor	Model	Name	Year	MB	Mb/in2	MB/s	Fly Ht	Comments
IBM	RAMAC	RAMAC	1956	5	0.002	0.0088	800	First disk drive
IBM	1301	Advanced Disk File	1962	28	0.026	0.068	250	First disk drive with air bearing heads
IBM	1311	Low Cost File	1963	2.68	0.05	0.068	125	First disk drive with removable disk pack
IBM	2310	Ramkit	1965	1.024	0.111	0.155	125	First disk cartridge drive
IBM	2311		1965	7.25	0.111	0.156	125	First disk pack drive
IBM	2314		1966	29.2	0.22	0.3125	85	First disk drive with ferrite core heads
IBM	3330-1	Merlin	1971	100	0.776	0.806	50	First track-following servo system
IBM	3340	Winchester	1973	70	1.69	0.886	17	First disk drive with low mass heads, lubricated disks, sealed
IBM	43FD	Crystal	1976	0.568	0.163	0.031	0	First flexible disk drive with two-sided recording
Shugart Associates	SA400		1976	0.2188	0.248	0.031	0	First 5.25″ flexible disk drive
IBM	3370	New File Project	1979	571.4	7.7	1.859	13	First disk drive with thin film heads
IBM	3310	Piccolo	1979	64.5	3.838	1.031	13	First 8″ rigid disk drive
Seagate Technology	ST506		1980	5	1.96	0.625		First 5.25″ rigid disk drive
Sony	OA-D30V		1981	0.4375	1.027	0.062	0	First 3.5″ flexible disk drive
Fujitsu	F6421	Eagle	1981	446	11.3	1.859		First 10.5″ rigid disk drive

SyQuest Technology	SQ306F		1982	5	5.2	0.625		First 3.9″ disk cartridge drive
Control Data	9715-160	FSD	1982	150	5.5	1.2		First 9″ disk drive
Rodime	RO352		1983	10	6.6	0.625		First 3.5″ rigid disk drive
Maxtor	XT-1140		1983	126	9.678	0.625		First 8 disk 5.25″ disk drive with in-hub motor
DMA Systems	360		1984	10	6.7	0.625		First 5.25″ disk cartridge drive
Hitachi	DK815-5		1984	460	12.9	1.8		First 8.8″ disk drive
Quantum	Hard-card		1985	10.5	11.3	0.625		First disk drive mounted on card
Conner Peripherals	CP340		1986	40	21.4	1		First voice coil actuator 3.5″ disk drive
Conner Peripherals	CP3022		1988	21	24.8	1.25		First one inch high 3.5″ disk drive
PrairieTek	220		1988	20	25.9	0.625		First 2.5″ disk drive
Hitachi	DKU-86I		1988	1890		3		First 9.5″ disk drive
IBM	681	Redwing	1990	857	50.8	3	7.6	First disk drive to use MR heads and PRML
IBM	3390-3	Pacifica	1991	5676	89.5	4.2		First IBM mainframe drive with thin film disks
Integral Peripherals	1820	Mustang	1991	21.4	89.5	1.9		First 1.8″ disk drive
Integral Peripherals	1841PA	Ranger	1992	42.5	140.9	2.2		First 1.8″ PCMCIA card disk drive
Hewlett Packard	C3013A	Kittyhawk	1992	21.4	134.4	1		First 1.3″ disk drive

(*cont.*)

Table 4–1 *Continued*

Vendor	Model	Name	Year	MB	Mb/in2	MB/s	Fly Ht	Comments
SyQuest Technology	SQ3105		1992	110	84	1.95		First 3.5″ disk cartridge drive
IBM	3390-9	Eureka	1993	17028	268.6	3.9		First large diameter drive with MR heads
Seagate Technology	ST11950	Barracuda	1993	1689	159	7.1		First 7200 RPM disk drive
Hitachi	H-6588-314		1993	2920	119.2	4.2		First 6.5″ disk drive
IBM	DPRA-21215	Presto	1995	1215	700.4	7.1		First 2.5″ drive over 1 gigabyte
IBM	DSOA-21-9	Sonata	1995	1080	923.1	6.7		Highest areal density for any drive
SyQuest Technology	SQ1080		1995	80	230.4	1.3		First PCMCIA drive with removable disk
IBM	DLGA-23080	Legato	1996	3080	1358	10.0		Highest 1996 areal density for any drive
IBM	DGHS-318220	Marlin	1997	18,220	1253	22.4		First 18 gigabyte 3.5″ server drive
IBM	DYKA-23240	Yukon	1997	3240	3123	11.7		Highest 1997 areal density for any drive
IBM	DTTA-351680	Titan	1997	16,800	2687	20.5		Highest 1997 areal density 3.5″ disk drive First drive with GMR heads
Calluna Technology	CT-520RM		1997	520	709.7	6.4		First 520 MB 1.8″ disk drive

Seagate Technology	ST19101	Cheetah 9	1997	9100	935.8	22.1	First 10,000 RPM disk drive
Seagate Technology	ST446452	Elite 47	1997	47,063	1491	23.0	Highest capacity disk drive in 1998
IBM	DBCA-206480	Biscayne	1998	6400	5693	14.8	Highest 1998 areal density for any drive
IBM	DCYA-214000	Cypress	1998	14,100	4976	15.7	Highest current capacity for 2.5″ disk drive
Calluna Technology	CT-1040RM		1998	1040	1403	6.6	First 1 gigabyte 1.8″ disk drive
Seagate Technology	ST118202	Cheetah 18	1998	18,200	1518	28.4	First 10,000 RPM drive with 3″ disks
Hitachi	DK3E1T-91		1998	9200	2490	27.3	First 12,000 RPM disk drive
IBM		Micro-drive	1999	340		6.1	First 1″ disk drive
Seagate Technology	ST150176	Barracuda 50	1999	50,000	3225	25.7	Highest capacity 3.5″ disk drive announced to date
Hitachi	DK229A-10		1999	10,000	6299	16.6	Highest current areal density for any drive

Source: DISK/TREND, 1999.

ENABLING TECHNOLOGY FOR DISK DRIVE ADVANCEMENT

Form factor miniaturization is not, in and of itself, a meaningful gauge of disk drive advancement. What is important are the underlying technologies that enable more data to be stored and accessed reliably in the same recording area of the magnetic media.

In 1956, the areal density (a measurement of the amount of data that can be stored in a media recording area) of the RAMAC disk was 0.002 million bits per square inch (.002 Mb/inch^2). On February 3, 1999, Seagate Technology announced that it had achieved a world record in disk drive technology by demonstrating areal density of greater than 16 billion bits per square inch (16 Gb/in^2). This areal density was achieved using merged read-write giant magneto-resistive (GMR) heads and an ultra-smooth alloy media designed and manufactured by Seagate. According to the vendor, "Recording information at an areal density of greater than 16 Gb/in^2 would allow a user to store more than 2500 copies of Homer's *Iliad* in the space of a thumbprint."[4]

Historically, the hard disk industry had been increasing the areal density storage capacity of hard drives at a rate of roughly 27 percent per year. In recent years, the growth rate has increased to as much as 60 percent per year, resulting in a current generation of drives that store information in the 600 to 700 Mb/in^2 range. By the year 2000, according to Seagate Technology, the areal density capabilities are expected to reach 10 Gb/in^2.[5]

READ–WRITE HEADS

A major contributor to meeting the areal density objectives of disk makers is read–write head technology. The current generation of disk drive products uses an inductive recording head design, with the reading and writing of data accomplished through the interpretation of the inductive voltage produced when a permanent magnet—the disk platter—moves past a wire-wrapped magnetic core—the head.

To write information to a hard drive, an electrical current flowing through an inductive structure in the read–write head produces a magnetic pattern in the coating on the media's surface corresponding to the data being written. To be more specific, microscopic areas, called domains, in the magnetic coating on the media platter are oriented to positive or negative states (corresponding to binary 1s and 0s) by the action of the read–write head when data is recorded to the disk. To read data back

from the disk, the read–write head converts these magnetic states into an electrical current, which is amplified and processed to reconstruct the stored data pattern.

Over a forty-year period, many variations and enhancements in inductive recording heads were introduced to improve areal density and other disk performance characteristics. Monolithic ferrite heads—composed a single block of ferrite, a magnetic ceramic material—were part of the earliest disk drives. Gradually, this design was improved through the use of "composite" heads, consisting primarily of nonmagnetic material with a small ferrite structure added. Building on composite designs were metal-in-gap, or MIG, heads, which featured very thin metal layers added inside a small gap in the head to improve magnetic performance. Finally, head technologies were enhanced through the application of thin-film technology.

With thin-film heads, the structures of an inductive head are deposited on a substrate material through a process closely resembling the manufacture of microchips. Thin-film technology allows head vendors to achieve much smaller physical dimensions and to control the fabrication process more exactly, both of which result in higher-performance head products.

More than 500 million thin-film heads were produced in 1997 to meet the enormous demands of the computer industry for data storage. However, industry sources claim that further improvements in processes by which thin-film inductive heads are manufactured, as well as refinements in the technology to support increasing areal density objectives, cannot be made cost-effectively. This prediction sent most major drive manufacturers back to their engineering departments in search of an alternative technology in the early 1990s. As the IBM microdrive and Seagate areal density records demonstrate, the new direction targeted by vendors is magnetoresistive (MR) head technology.

In a white paper on MR technology, disk maker Quantum Corporation summarized the situation succinctly. According to the author, the recognition of the need for a new head technology was

> spurred by a simple fact: As areal densities increase, the bit patterns recorded on the surface of the disk grow smaller, which weakens the signal generated by the head. That, in turn, makes it difficult for the disk drive's read channel electronics to identify the bit patterns that produced the signal.[6]

According to the white-paper authors, drive designers tried to circumvent the problem by harnessing several techniques to produce a

stronger read signal. Techniques included flying the head closer to the disk and adding "turns" (the number of wire wraps coiled around the magnetic core of the head).

Reducing head-fly heights to better detect and read bit patterns was a typical enhancement method for a time. Shrinking the current fly heights of two to three microinches, however, seemed a dangerous proposition. In the words of one observer, it is already like flying a 747 jet at a very high speed just inches above the ground. Attention turned to the practicality of adding turns to the inductive head.

Adding turns increased both the read signal—a good thing—and the inductance of the head itself—a not-so-good thing. Limits exist in the amount of inductance a head can tolerate and still perform read–write operations reliably. This was especially true for thin-film heads, which used the same head element for both reading and writing data.

Disk drive engineers also examined the feasibility of increasing signal strength by increasing the linear speed at which recorded data bits moved under the head. With thin-film inductive heads, the faster the speed, the stronger the signal. However, faster rotational speeds also increased the data frequencies—the rate at which the magnetic transitions, which encode the data, pass the head—beyond the capabilities of inductance heads and channel electronics to handle.

Like Quantum, many other vendors, including Seagate Technology and IBM, came to the conclusion that MR technology held an answer. At the heart of the technology is a magnetoresistive material (a ferromagnetic alloy) whose electrical resistance changes in the presence of a magnetic field.[7] By applying such a material to a disk drive read head, it is possible to detect and read weaker magnetic fields generated by bits stored at greater areal densities.

The concept took shape in vendor laboratories. Soon, most major vendors had designs for MR heads that supported Gb/in^2 areal storage capacities on hard disks.

A signature of MR head technology is a "two-heads-in-one" design, claimed by Seagate Technology. Prepared as a combined head component, the read head element contains a small stripe of MR material. As it passes over the magnetic patterns on the disk, it senses the strength of the magnetic field and creates electrical pulses corresponding to the flux reversals.

Since MR heads cannot be used to write data to hard disk media, heads employing this technology have a second thin-film inductive element, closely spaced but separate from the MR read element. This element is used to write data bits onto the disk surface. According to

Seagate, "This fundamental change in read–write technology will enable advances [in areal density] capable of carrying the disc drive industry well into the 21st century."[8]

OTHER DRIVE IMPROVEMENTS

From the previous discussion of MR heads, it should be clear that the overall capacity of a drive depends on how densely information (i.e., bits) can be recorded on the disk media. This is a function of many components of a disk drive operating in concert with each other. Figure 4–1 depicts typical disk components.

One can more clearly see the relationships between components by considering the many factors that contribute to the areal density of a disk drive. The areal density of a drive, its bits per square inch, is calculated by taking the number of bits per inch (BPI) that can be written to and read from each track and multiplying that number by the number of tracks per inch (TPI) that can be "etched" on the disk media.

The bits per inch (BPI) possible on a disk depends on the read–write head, recording media, disk RPM, and the speed at which the electronics

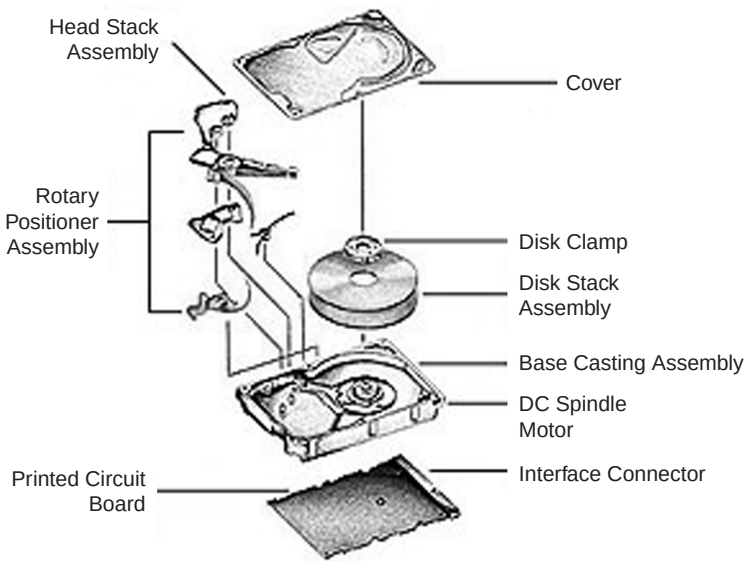

Figure 4–1 Components of a Hard Disk Drive. (*Source:* Quantum Corporation, Milpitas, CA.)

can accept bits. Similarly, tracks per inch is a function of the read–write head, recording media, the mechanical accuracy with which the head can be positioned on its actuator arm, and the ability of the disk to spin in a perfect circle. An increase in areal density is accomplished by increasing either or both BPI and TPI.

MR head technology enables the use of high areal density media by providing the capability to read information from more densely packed disks. To understand how increased storage densities are achieved requires more information about disk components.

Most of the current generation of hard disk drives feature two or more platters configured as a disk stack assembly with a common spindle. A spindle motor rotates the platters counter-clockwise at speeds of between 3600 to 10,000 revolutions per minute (RPM).

As previously mentioned, data stored on a disk is actually recorded as a magnetic pattern of bits in the magnetic coating on the platter. Read–write heads generate these patterns when writing data to the disk platters. When reading from the disk, the read–write head converts the stored magnetic patterns into electrical signals to represent stored data.

The writing of data to the disk platter occurs in accordance with the format geometry of the disk drive. Hard disk platters are divided into tracks, cylinders, and sectors to create a structure for data storage. A track is a concentric ring around the platter. In a disk stack assembly, tracks on each platter surface are identically positioned about 300 microinches apart. Identical tracks on multiple platters create cylinders.

Additionally, each track is subdivided into sectors that aid in expediting read–write operations. Sectors are given their identities during formatting when sector numbers are written to the beginning (prefix) and the end (suffix) of each sector. These identities consume hard disk space, accounting for the difference between a hard disk's formatted and unformatted capacity.

During write operations, data is recorded on the outermost track of all platters first. Once the outside track is filled with data, the heads move inward and begin writing on the next free track. This recording strategy greatly boosts performance since the read–write heads have considerable room to write data before they must be repositioned.

Since most hard disks enable the storage of data on both surfaces of a platter, drives are usually equipped with a read–write head for each platter face. Each head is held at an optimal head fly height by an actuator arm. Both the actuator arms and the read–write heads are moved over the platters by a positioning motor, which is, in turn, controlled by the disk controller.

The disk controller "knows" where to move heads to retrieve information by referring to the formatting information and sector addresses. Without formatting instructions, neither the controller nor the operating system would know where to store or retrieve data.

Early hard drive designs fixed a limit on the number of sectors per track. This, in turn, limited storage capacity since the number of sectors that would fit on the innermost track constrained the number of sectors that could be set for outer tracks that had a larger circumference. To address this issue, a number of vendors adopted a formatting technique called Multiple Zone Recording that allowed the number of sectors per track to be adjusted. By dividing the outer tracks into more sectors, data could be packed uniformly across the surface of a platter. Disk surfaces could be used more efficiently, and higher capacities could be achieved with fewer platters. With Multiple Zone Recording, effective storage capacity increased by as much as 25 percent. Additionally, disk-to-buffer transfer rates improved: With more bytes per track, data in the outer zones can be read at a faster rate.

The ability of read–write heads to fly directly to a new location once the CPU provides an address is central to its ability to randomly store and retrieve data. This capability, more than any other, explains the rapid displacement by hard disk of tape media as the primary computer storage technology after 1956.

Precise control of head positioning is a major factor in the areal density that can be achieved with disk drives. Drives are subject to many factors, including temperature-induced platter expansion and sudden movements that militate against precise head alignment. To counter these factors, most modern drives incorporate an electromechanical technique called embedded servo positioning.

Embedded servo positioning is a method that uses special data patterns, prerecorded on each track of a platter, to update drive electronics regarding head position. In fact, the strength of the signal is interpreted by the drive electronics to indicate how far the head has strayed from the center of the track. In response to the signals, the drive electronics adjust the position of the actuator motor, which repositions the heads until the maximum signal is from the bursts. This technology provides the most accurate, error-free, and cost-effective head positioning technique for small form factor drives.

Some difficulties have arisen when applying embedded servo technology in drives that use Multiple Zone Recording techniques. The varying number of sectors on tracks in different zones complicates the task of reading servo data. Quantum Corporation claims to have addressed this

problem effectively by developing specialized servo feedback and controller ASICs that efficiently handle the task of separating the servo information from the user data.

The signals that a read–write head picks up from a platter are very weak. The read preamplifier and write driver circuit, typically mounted on a flexible circuit inside the hard disk assembly, increases the strength of the signals so that drive electronics can convert electrical impulses into digital signals. Drive electronics themselves are typically contained on a printed circuit board attached to the drive itself and include

- A digital signal processor (DSP) to convert the incoming electrical signals to digital signals
- Motor controllers for the spindle and actuator motors that ensure that the platters spin at the correct speed and that the actuator arms precisely position the read–write heads over the platter
- Interface electronics to communicate with the central processing unit of the system where the drive is installed

Many drives also have a microprocessor, ASICs, and memory chips on the printed circuit board to support functionality such as drive caching, embedded servo head positioning, and multiple zone recording. The circuit board also provides the physical connector for the drive, enabling the connection of the drive to the I/O bus of the system (PC, server, NAS, intelligent array, etc.) in which it is installed.

NEW INITIATIVES IN DRIVE TECHNOLOGY

Before examining interfaces, it is important to reiterate that the components of the hard drive, operating in concert, are what account for the capacity and much of the performance delivered by the drive itself. The storage manager needs to keep up-to-date with the initiatives of vendors that are constantly seeking to enhance disk drive components and to enable greater product performance and capacity.

For example, a number of vendors are working on ways to increase areal density by further reducing head fly heights. One approach, suggested by Seagate Technology's areal density record, is to create a smoother platter with fewer flaws that could cause a disk crash. Seagate used a homegrown alloy in its demonstration drive. Others are looking at

alternative platter substrate materials, including glass, as a replacement for aluminum alloys in the disk substrate.

Another approach being pursued by Quantum and others is to eliminate flying heads altogether. Some researchers are looking at contact recording, a technology in which the head rides directly on the surface of the platter but does not generate friction that would quickly destroy both the head and platter surface at normal operating speeds. Liquid lubricants, "wet disk" technology, and low-friction, wear-resistant platter materials are all areas of research that may yield tomorrow's drive capacity breakthroughs.

PRML CHANNEL TECHNOLOGY AND ENHANCED MANAGEMENT FEATURES

Until the research initiatives above yield practical results, MR head technology will continue to provide a product development path for many vendors, supporting their efforts to deliver drives with increasing areal densities. Enhancing the capabilities of MR head technology is Partial Response Maximum Likelihood (PRML) technology. Briefly, PRML is a replacement for a read technique common in pre-MR head drives called peak detection. Peak detection, which served well for many years as a means for interpreting data from the read head, is less effective as bit density increases, read signal strength diminishes, and background noise begins to confound the drive electronics' efforts to distinguish individual bits by their voltage peaks.

PRML technology first converts the MR head's analog "read" signals into digital signals, then samples and filters the digital signals using sophisticated algorithms to eliminate noise and detect data bits. The result is that PRML can properly interpret more densely packed bits than can peak detection.

The efficiency of PRML, especially when used in drives with MR head technology, contributes directly to faster transfer of data as well as more accurate data. According to evaluations of PRML-enhanced MR head drives from Seagate, Quantum, and IBM, the technology is a must-have for anyone using high density drives.

Other enhancements being made to drives have little to do with areal density, but contribute a great deal to drive monitoring, management, and longevity. IBM and others have already begun to add temperature sensing capabilities to their drives that will allow storage managers to identify potentially damaging conditions before data is lost. IBM, Seagate, and

others have also added head parking capabilities to the drives, ensuring that heads will not crash platters if power is discontinued suddenly.

Disk drives have been termed commodity products by some analysts, but they are in fact among the most complex components of the computing infrastructure. The definition of a reliable and manageable mass storage architecture begins with an understanding of drive technology. The enhancements and safeguards built into some disk drives by their manufacturers make them excellent choices to store mission critical information assets. Conversely, deploying disk drives that lack the features that contribute to stable, long-term storage can be an Achilles heel for any storage-management strategy.

INTERFACES: CONNECTING STORAGE FOR EFFICIENT USE

For a disk drive to become a part of a storage strategy, it must be interfaced to the bus of an intelligent system. In a "captive storage" configuration, the disk device is cabled directly to the I/O bus of a server, PC, or workstation. This is typically accomplished by cabling the bus connector on the drive's electronics (usually a printed circuit board mounted to the hard disk assembly) to a I/O interface adapter installed on the bus of the "host" system.

In such configurations, the drive is accessed by the CPU of the host as part of a cyclical process of CPU interrogations. I/O requests transfer along the bus of the host system and are passed through the I/O interface adapter to the drive electronics. Responses from the drive take the same path back to the CPU.

The performance of the disk drive itself is only one factor in the performance of I/O processing in this configuration. The speed of data transfers along the host system bus, which are determined by the bus width and cycle time, have the greatest effect on overall I/O performance. The bus cycle time is proportional to the number of "words" of data that can be transferred per second. Bus width determines the width of the transfers and whether words of data are transferred in a single cycle or multiple cycles.

Older system bus architectures had bus widths of 4 or 8 bits and transferred data at rates of up to 1 MB/s. Today, system bus architectures typically feature 16- or 32-bit bus widths and data transfers speeds of 10, 20, and up to 132 MB/s in Peripheral Components Interface (PCI) bus architectures are very possible. The next logical development for the server system bus is a 64-bit wide interface, which will allow drives and other peripherals to reach even higher data transfer speeds.

The rate at which the disk can transfer data onto a system bus is a function of its interface. A number of interfaces have been offered for hard disk drives over the years, but since the early 1990s, only two have emerged as industry leaders: IDE and SCSI.

ADVANCED TECHNOLOGY ATTACHMENT/INTELLIGENT DISK ELECTRONICS (ATA/IDE) INTERFACE

When the Intelligent Disk Electronics (IDE) interface, which is now called the Advanced Technology Attachment (ATA) interface, was first released, it was welcomed by PC end users as a "high-speed" replacement for a "motley crew" of competing interfaces for earlier PC disk drives. The ATA interface was designed specifically for disk drives and for Intel AT/IDE bus PCs and delivered disk buffer-to-host data transfer rates of 4.1 MB/s.

Over time, the ATA industry standards committee extended the capabilities of the interface to keep pace with other computer platform advancements. Among those improvements was the introduction of Fast ATA in late 1993, which supported an accelerated data transfer rate to capitalize on the new, faster, local bus architecture in Intel PCs.

Fast ATA enabled a disk drive to be connected directly to the CPU bus in the new Intel PC bus architecture, completely bypassing the slower expansion bus, held over from the days of the PC/AT. End users and the industry applauded the change, which provided for data transfer speeds limited only by the speed of the local bus and the disk drive itself. Approximately 90 percent of desktop PCs used ATA or Fast ATA disk interface adapters in 1996.

The applause, however, soon quieted as desktop system application requirements exceeded the support provided by Fast ATA. According to Quantum, an early supporter of Ultra ATA, end users of Fast ATA drive interfaces were encountering bottlenecks during sequential transfers of large files such as system boot-up, the loading of increasingly large programs, and especially desktop video applications. Stated simply, the faster internal transfer rates in newer disk drives combined with the poor utilization of the ATA bus by PC CPUs were causing disk drives to fill their data transfer buffers much faster than system CPUs could unload them. The result was I/O bottlenecking and a need for a data transfer rate doubling Fast ATA's burst data rate of 16.7 MB/s.

Recognizing that part of the bottleneck problem was beyond the ability of disk makers to control, Quantum and other vendors assisted in the

refinement of the Fast ATA protocol within a set of known constraints. The Ultra ATA interface employed a new signaling and timing method that increased the speed of data buffer unloading and added a raft of additional features (plug-and-play support, CRC checking, etc.) Ultra ATA interface became an industry-recognized standard virtually overnight and drives tailored to the interface began shipping in late 1996 and early 1997.

While ATA continues to enjoy tremendous success in the PC market, the need for higher performance and multiple device support have driven many PC users to the Small Computer System Interface (SCSI). Today, SCSI is the second most common disk drive interface in Intel PCs, but it is the most widely used interface in both the workstation and server markets.

SMALL COMPUTER SYSTEM INTERFACE (SCSI)

Like ATA, SCSI is a family of protocols. The various implementations share in common a parallel interface definition used to connect host systems and peripheral devices, including disk drives. From a simple, twenty-page specification introduced to the American National Standards Institute (ANSI) in 1980, SCSI has grown into a 600-page specification for a veritable hydra of alternative implementations (see Table 4–2).

In 1985, the handwriting was already on the wall. Just as the first SCSI draft was being finalized as an ANSI standard, a group of manufacturers approached the X3T9.2 Task Group seeking to increase the mandatory requirements of SCSI and define further features for direct-access devices. Rather than delay the first iteration of the standard, the Task Group formed an ad hoc group to develop a working paper that was eventually called the Common Command Set (CCS).

The main problem with SCSI-1, according to some observers, was that the standard was too permissive, and allowed too many "vendor specific" options. It was feared that variations in implementations would result in serious compatibility problems between products from different vendors. The Common Command Set (CCS) was proposed in an effort to address compatibility problems before they created havoc, mainly for tape and disk drive products. It became a de facto component of the SCSI-1 standard for anyone serious about deploying the interface.

SCSI-1 and the CCS defined a number of basic command operations, an 8-bit wide bus with transfer rates of up to 5 MB/s, and a cable with several connector options. According to the initial "CCS-enhanced" SCSI-1 standard, up to seven devices could be connected to the bus, not

Table 4–2 SCSI Standards and Drafts and Key Features

SCSI Type	Bus Speed, Mb/s Max	Bus Width Bits	*Maximum Bus Length, Meters* [1]			Maximum Device Support
			Single-ended	Diff.	LVD	
SCSI-1[2]	5	8	6	25	[3]	8
Fast SCSI[2]	10	8	3	25	[3]	8
Fast Wide SCSI	20	16	3	25	[3]	16
Ultra SCSI[2]	20	8	1.5	25	[3]	8
Ultra SCSI[2]	20	8	3	25	[3]	4
Wide Ultra SCSI	40	16	—	25	[3]	16
Wide Ultra SCSI	40	16	1.5	—	—	8
Wide Ultra SCSI	40	16	3	—	—	4
Ultra2 SCSI[2,4]	40	8	[4]	25	12	8
Wide Ultra2 SCSI[4]	80	16	[4]	25	12	16

Notes:
[1] The listed maximum bus lengths may be exceeded in point-to-point and engineered applications.
[2] Use of the word "narrow", preceding SCSI, Ultra SCSI, or Ultra2 SCSI is optional.
[3] LVD was not defined in the original SCSI standards for this speed. If all devices on the bus support LVD, then 12-meter operation is possible at this speed. However, if any device on the bus is single-ended only, then the entire bus switches to single-ended mode and the distances in the single-ended column apply.
[4] Single-ended is not defined at Ultra2 speeds.
Source: SCSI Trade Association, San Francisco, CA.

including the host system. Asynchronous data transfers between the host computer and a given peripheral could occur in at speeds up to 2 MB/s, while synchronous transfers were supported at speeds of up to 5 MB/s.

Work on the new SCSI-2 standard began while ANSI was preparing to publish the standard it had ratified in 1986 (ANSI X3.131-1986, commonly referred to as SCSI-1). The original premise of SCSI-2 was to create a superset of SCSI-1 and the CCS. Later, the scope of the effort expanded and by the time that the draft standard for SCSI-2 was submitted for ANSI approval in 1990, the document had grown to more than double the size of SCSI-1. (The final draft was nearly 600 pages when issued in 1993.)

Nevertheless, SCSI-2 advanced several meaningful improvements to the SCSI-1 standard, including:

- Higher performance
- Increased data transfer rates
- Lower overhead
- New definitions for single-ended and differential interfaces
- New bus definitions
- Support for new peripheral types
- Support for new functions such as command queuing and disconnect
- Enhanced reliability through the implementation of functions such as parity and error checking and arbitration

Backward compatibility was a touchstone of SCSI-2 standards development, as with later iterations of the standard. "Single-ended" SCSI devices defined in SCSI-2 were backward compatible with single-ended devices conforming to the SCSI-1 standard in order to facilitate a smooth transition between the standards. Single-ended refers to an implementation of signal transmission wiring in which all data and handshaking signals to draw necessary current through common ground.

The SCSI-2 specification also defined a differential signaling implementation that was not backwards compatible with SCSI-1, but which did promise improvements such as noise reduction and longer bus cable lengths of up to 25 meters.

SCSI-2 also established two interface variations that have become synonymous with the standard.

- Fast SCSI-2 allows faster bus timing (10 MHz instead of 5 MHz in SCSI-1). The theoretical result on an 8-bit wide bus is a data transfer speed of up to 10 MB/s.
- Fast Wide SCSI-2, another variant, enables still faster data transfer rates through the use of 16-bit or 32-bit cables. Transfer speeds of up to 20 MB/s for a 16-bit bus, or 40 MB/s for a 32-bit bus are theoretically possible. Up to fifteen devices may be connected concurrently to the host under this configuration.

SCSI-2 became an official ANSI standard in 1994 (ANSI X3.131-1994)—almost a year after development had begun on a SCSI-3 standard.

Before one has a bout of déjà vu, it should be pointed out that SCSI-3 was actually intended to work on a number of separate enhancements to SCSI-2 rather than provide an entirely new standard. With SCSI-3, the draft standard was broken up from a single document into several smaller documents focused on different objectives. This was done, in part, to facilitate efforts to specify SCSI implementations over different physical transport layers, including Fibre Channel and IBM's Serial Storage Architecture. It was also believed that breaking the standards development effort into smaller projects would result in faster completion. Subprojects included

- *SCSI-3 Parallel Interface (SPI):* This project sought to further define the mechanical attributes, timing, phases, and electrical parameters of the parallel cable. Some of the electrical and cable parameters were tightened and improved from SCSI-2 specifications.
- *SCSI-3 Interlock Protocol (SIP):* New messages were added to the existing definition.
- *SCSI-3 Architectural Model (SAM):* This project endeavored to define a common set of functions, services, and definitions to explain a physical transport handles commands, data, and status exchanges between two devices. Error handling and queuing are also described.

The balance of the projects dealt with refining specific command sets associated with disk, tape, RAID, and CD-ROM devices, and with the use of SCSI over different physical transports including Fibre Channel, IEEE 1394 High Speed Serial Bus ("Firewire"), and the Serial Storage Architecture from IBM.

ULTRA SCSI

Significant outcomes of SCSI-3 development efforts included industry proposals for "narrow" and "wide" Ultra SCSI implementations. In 1996, Quantum Corporation termed Ultra SCSI "the next major performance advancement to the Small Computer System Interface and . . . a lower cost alternative to serial [interfaces]." According to the vendor, Ultra SCSI's proposed doubling of data transfer rates from the specified limits of 10 MB/s for Fast SCSI-2 and 20 MB/s for Fast Wide SCSI-2 promised to "create the bandwidth necessary to support the data intensive applications to be used in the coming generations of servers, workstations, and high-end personal computers."[9]

Even as Quantum's prognostications were hitting the trade press, Ultra2 SCSI proposals began to be heard. Ultra2 and Wide Ultra2 draft specifications were driven by growing interest in differential signaling technology and increased popularity of serial interfaces, such as Fibre Channel, as a potential replacement for SCSI.

The use of low voltage differentials (LVD) to optimize bus throughput had been addressed initially in SCSI-2. With Ultra2 SCSI, LVD transceivers were being applied to the task of providing better use of SCSI bus media. Testing suggested that LVD-enhanced interfaces could deliver data transfer speeds of 80 MB/s—doubling the fastest SCSI-2 transfer speed of 40 MB/s. Increased bandwidth translated to improved server performance as large files are moved between devices quickly and effortlessly.

As vendors readied Ultra2 devices for market, yet another proposal—Ultra160/m—was placed before the ANSI standards T10 Committee by seven vendors "representing a broad cross-section of the computer and storage industry." Like its predecessor, Ultra160m again doubled the data transfer speed possible with the SCSI interface. Ultra160/m raised the bar to 160 MB/s by furthering the use of LVD, combined with improved timing techniques, cyclical redundancy checking (CRC), and "domain validation" (i.e., verification of network integrity on the SCSI bus).

Taken together with the earlier SCSI interface alternatives, Ultra160/m SCSI helps to complete a picture of SCSI as a robust, evolutionary storage bus technology with adequate room for growth to meet the storage scalability requirements of many companies into the next millennium (see Figure 4–2). However, SCSI does have its limitations and it presently confronts challenges for hegemony in the multidrive interface market from two rival interface standards. One is IBM's Serial Storage Architecture (SSA). The other is an open standard gaining particular attention because of its use in storage area networks: Fibre Channel.

SERIAL STORAGE ARCHITECTURE (SSA)

According to IBM, the present growth of data bases and data-intensive applications within corporations signals a need for a storage technology that is robust, reliable, and scalable. Acting on this premise, the company introduced one of the first serial storage technologies in 1992, called Serial Storage Architecture (SSA). Products based on the technology began shipping in 1995, while IBM engaged in efforts to have the technology approved as an ANSI standard.

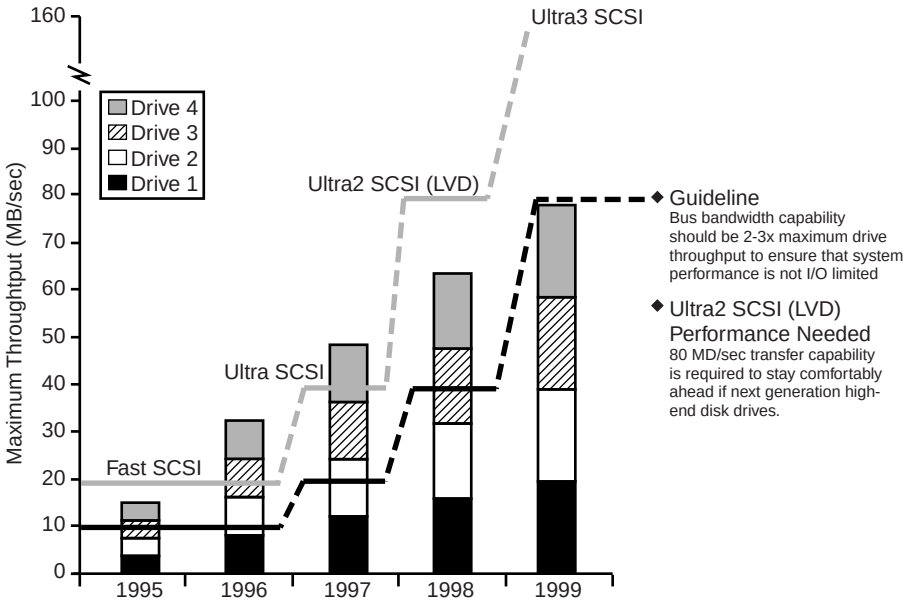

Figure 4–2 Evolutionary SCSI Capabilities Growth. (*Source:* Quantum Corporation, Milipitas, CA.)

By November 1996, ANSI ratified proposed standards covering the SSA Physical Layer (ANSI X3.293-1996), Upper Level Protocol—SCSI-2 Mapping (ANSI X3.294-1996), and Transport Layer (ANSI X3.295-1996). By February 1997, the success of SSA appeared to be confirmed with an announcement by IBM that more than one petabyte of Serial Storage Architecture (SSA) products had been shipped to customers in only 18 months.[10] Additional components of SSA were made ANSI standards in December of the same year (see Table 4–3).

SSA is the first serial storage technology to achieve such market acceptance, and IBM has leveraged its early success to argue the case for deploying SSA as a replacement for SCSI. The company also insists that its serial interface offers advantages that make it superior even to emerging Fibre Channel interfaces for connecting storage devices, storage subsystems, servers and workstations.

SSA provides serial disk interface using bidirectional cabling to establish a drive "loop." Data and commands sent from the SSA adapter can travel in either direction around the loop to their destination devices, or nodes. If interruptions are detected in the loop (e.g., a hard disk node

Table 4–3 ANSI Standards for SSA

STANDARD	TITLE	SUMMARY	APPROVAL DATE
ANSI X3.293-1996	Information Technology-Serial Storage Architecture-Physical Layer 1 (SSA-PH1)	Defines the physical layer of the Serial Storage Architecture (SSA). SSA defines a serial interface hierarchy to be used for purposes within its distance and performance characteristics, including, but not limited to, storage subsystems. This standard is intended to be used with an upper layer protocol (e.g., SCSI-2 Protocol (SSA-S2P)) and a transport layer (e.g., SSA Transport Layer 1 (SSA-TL1)). A major goal of the SSA-PH1 standard is to define a physical layer acceptable to device vendors, looking for an evolution from parallel SCSI, and systems designers looking for opportunities to more fully exploit the capabilities inherent to a serial bus.	11/29/96
ANSI X3.294-1996	Information Technology-Serial Storage Architecture–SCSI-2 Protocol (SSA-S2P)	Describes an upper-level protocol of Serial Storage Architecture. SSA-S2P is a mapping of the existing SCSI-2 protocol, described in American National Standard for Information Systems—Small Computer Systems Interface-2 (SCSI-2), ANSI X3.131-1994, with extensions to map SCSI-2 to the SSA serial link.	11/29/96
ANSI X3.295-1996	Information Technology-Serial Storage Architecture-Transport Layer 1 (SSA-TL1)	Defines the transport layer of the Serial Architecture (SSA). SSA defines a serial interface hierarchy to be used for purposes within its distance and performance characteristics, including, but not limited to, storage subsystems. This standard is intended to be used with an upper layer protocol (e.g., SCSI-2 Protocol (SSA-S2p)) and a physical layer (e.g., SSA Physical Layer 1 (SSA-PH1)). A major goal of the SSA-TL1 standard is to define a transport layer acceptable to vendors, looking for an evolution from parallel SCSI,	11/29/96

		and systems designers looking for opportunities to more fully exploit the capabilities inherent to a serial bus.	
ANSI NCITS 307-1997	Information Technology-Serial Storage Architecture-Physical Layer 2 (SSA-PH2)	The SSA-PH2 standard defines a physical layer that support the SSA transport layer 2 (see SSA-TL2) and any protocols supported by SSA-TL2. The goals of SSA-PH2 are: (a) extend the cable distance; (b) copper cable operation at 40 MB/s; (c) full duplex operation to achieve an aggregate 80 MB/s between two ports; (d) and other capabilities that fit within the scope of SSA-PH2 that may be proposed during the development phase by the participants in the project. This document defines the physical layer 2 (SSA-PH2) of the Serial Storage Architecture (SSA). SSA defines a serial interface hierarchy to be used for purposes within its distance and performance characteristics, including but not limited to storage subsystems. This standard is intended to be used with an upper layer protocol [e.g., SCSI-2 Protocol (SSA-S2P) or SCSI-3 Protocol (SSA-S3P)] and a transport layer [e.g., SSA Transport Layer 2 (SSA-TL2)]. A major goal of the SSA-PH2 standard is to define a physical layer acceptable to device vendors, looking for an evolution from parallel SCSI or SSA-PH1, and systems designers looking for opportunities to more fully exploit the capabilities inherent to a serial bus.	12/2/97
ANSI NCITS 308-1997	Information Technology-Serial Storage Architecture-Transport Layer 2 (SSA-TL2)	Defines a transport layer of the Serial Storage Architecture (SSA) that runs SSA-S2P and SSA-S3P (BSR NCITS 309) while running on SSA-PH2 (BSR NCITS 307). The goals of SSA-TL2 are to: (a) provide an Extended Distance Option; (b) provide support for higher data rates in the physical layer 2 (SSA-PH2); (c) enhance packet formats and addressing methods; (d) define a transport layer acceptable to vendors looking for an evolution from parallel SCSI and systems	12/12/97

(*cont.*)

Table 4–3 *Continued*

STANDARD	TITLE	SUMMARY	APPROVAL DATE
		designers looking for opportunities to more fully exploit the capabilities inherent to a serial bus; and (e) cover other capabilities that fit within the scope of SSA-TL2 that may be proposed during the development phase by the participants in the project.	
ANSI NCITS 309-1997	Information Technology-Serial Storage Architecture-SCSI-3 Protocol	Defines a protocol layer of the Serial Storage Architecture (SSA) that runs on SSA-TL2 (BSR NCITS 308) while running on SSA-PH2 (BSR NCITS 307). The goals of SSA-S3P are to: (a) map the SAM services and terminology to SSA; (b) define the data field format of the SSA-S3P SMSs; (c) support for dual port and alternate pathing; (d) provide support for auto-sense; (e) provide support for third-party operations; and (f) cover other capabilities that fit within the scope of SSA-S3P that may be proposed during the development phase by the participants in the project.	12/1/97

fails), SSA can automatically reconfigure the system to maintain connectivity with other nodes until the interruption is repaired.

SSA's loop configuration enables the addition and replacement of disk drives without interrupting normal operations. Users can configure the SSA array initially with only a few hard disk drives, then add more drives to the loop when needed. SSA hard disk drives are described as self-configuring. This capability obviates the addressing limitations and complexity of SCSI drive installation.

An SSA Adapter supports up to 192 hot-swappable hard disk drives per system. If desired, drives can be preconfigured for use as "installed spares" to be used by the host only if a failure occurs. Drives with capacities of 4.51 GB, 9.1 GB, and 18.2 GB capacities are available for use in SSA arrays. Storage capacities of up to 291 GB per tower or drawer or 1.752 TB per host adapter are possible (see Figure 4–3).

SSA provides for a maximum distance of 25 meters between hard disk drives and server host(s). Cabling consists of thin, low-cost copper wires. With a fiber optic extender, arrays can be positioned up to 2.4 kilometers distant from the server, if desired. SSA arrays can also be attached to up to four concurrently attached SCSI-based servers via an Ultra SCSI Host to SSA Loop Attachment. The host's standard SCSI driver and hard-

Figure 4.3 IBM 7133 Serial Disk Systems. (*Source:* IBM.)

ware are used to make the attachment, so no modification to server hardware or software is required.

The SSA subsystem supports both RAID and non-RAID configurations. Disks can be mirrored across servers to provide a hedge against unplanned downtime and data loss. According to the vendor, the subsystem provides up to 80 MB/s of maximum throughput, which is described as sustained data rates as high as 60 MB/s in non-RAID mode and 35 MB/s in RAID mode.

IBM positions SSA as an open, standards-based product and emphasizes the support that SSA subsystems offer for attachment to a broad range of system hosts. Critics point to the fact that both the array and disk drives bear the IBM moniker to suggest that this is IBM's proprietary architecture. While SSA has experienced substantial success, concerns about its proprietary nature do not appear to have been offset by ANSI standard approvals. SSA represents a very small segment of an interface technology market that is dominated by variants of SCSI.

As a serial storage architecture, SSA does offer many advantages over even Ultra SCSI. Tests of Ultra SCSI and SSA published on the IBM Storage Division website show that SSA clearly provides greater performance under identical loads. An SSA array also scales better than an Ultra SCSI, providing consistently high performance numbers while Ultra SCSI performance declines. IBM claims, based on the tests, that SSA's serial bus has between 18 and 33 percent less overhead than SCSI's parallel bus and that this is a major factor in the superior SSA performance numbers.

Head-to-head comparisons must top out, however, at sixteen drives, which is the maximum number of drives that can be connected to an Ultra SCSI bus. SSA's connectivity capabilities exceed this number by a factor of 10.

SSA, says IBM's marketing literature, "addresses SCSI's throughput bottleneck with its fundamental building block—the SSA connection, or node." An SSA node (an IBM SSA disk drive) has two ports that can each be used to carry on two 20 MB/s conversations at once (one inbound and one outbound), thereby enabling a total of 80 MB/s of throughput. By contrast, a single SCSI bus, "can easily be saturated by just one high-performance disk running at 12 MB/s."

Even though SSA is an entirely different architecture than SCSI, it still maps the SCSI command set, observes the vendor, so existing applications can migrate seamlessly to SSA-based subsystems. This saves the user the time and cost of rewriting applications while giving applications a performance boost.

Table 4–4 provides IBM's suggested review criteria for prospective customers who are evaluating SSA and other drive interface technologies.

Table 4–4 A Comparison of Interface Technologies

	EIDE/Ultra ATA	SCSI	SSA
Environment	Offers low cost and performance equal to SCSI in most desktop and mobile environments.	Delivers excellent performance for network computers with Intel-based processors. Data-intensive applications and large numbers of users in LAN environments.	An ideal solution for PC servers: combination of high storage capacity, data protection, extensibility, and affordability.
Performance	In a single-user environment, EIDE (Ultra ATA) and SCSI perform comparably.	A SCSI interface offers the performance edge, especially when coupled with Windows NT.	Greatly reduces the risk of downtime from communication failure. 80MB/s maximum throughput ensures data transfer rates will not be a problem.
Price	Generally least expensive. The controller is standard on the system board chipset.	Drives are a little more expensive for the same capacity and rotational speed as EIDE. May also require a SCSI adapter.	
Expandability	Can support high hard disk drive capacity, as well as CD-ROM and tape devices—up to four devices in all.	Holds more than 9 GB. Offers high capacity and performance for multiple hard disk drives, a wide variety of devices and long cable connectors for more convenient attachment of external devices. Backward-compatible.	Adapter supports allow for up to 192 hot-swappable hard disk drives per system. Hard disk drives are available in 4.51 and 9.1GB sizes.
Ease of Installation	Although most PCs ship with an EIDE (Ultra ATA) interface, ensure that your system's EIDE (Ultra ATA) interface and BIOS support all the functions of the new hard disk drive.	You may need to install a SCSI adapter.	SSA makes hard disk drives self-configuring, avoiding SCSI addressing limitations and complexity.

Source: IBM.

FIBRE CHANNEL

SSA, like SCSI, are more than drive interface technologies. They are also an interconnect specifications for mass storage disk arrays. Fibre Channel is another serial interface/interconnect technology.

Fibre Channel is a 1 GB/s data transfer interface that maps several common transport protocols including IP and SCSI, allowing it to merge high-speed I/O and networking functionality in a single connectivity technology. Like SSA, it is a standard ratified by ANSI (ANSI X.3230-1994 is the core standard) and operates over copper and fiber optic cabling at distances of up to 10 kilometers.

However, Fibre Channel is different from SSA in its support of multiple interoperable topologies, including point-to-point, arbitrated-loop, and switching. Additionally, Fibre Channel offers several qualities of service for network optimization. With its large packet sizes, Fibre Channel is ideal for storage, video, graphic, and mass data transfer applications.

Fibre Channel's developers have achieved the majority of their original goals in defining the technology, which included:

- Performance from 266 MB/s to over 4 GB/s
- Support for distances up to 10 kilometers
- Small connectors
- High-bandwidth utilization with distance insensitivity
- Greater connectivity than existing multidrop channels
- Broad availability (i.e., standard components)
- Support for multiple cost/performance levels, from small systems to supercomputers
- The ability to carry multiple existing interface command sets, including Internet Protocol (IP), SCSI, IPI, HIPPI-FP, and audio/video

Today, Fibre Channel stands out as a leading contender to become the dominant channel/network standard. As a drive interface technology, it combines high speeds with SCSI-3 command language support. As a network interface, it provides the required connectivity, distance, and protocol multiplexing. It also supports traditional channel features for simplicity, repeatable performance, and guaranteed delivery. And Fibre Channel also works as a generic transport mechanism.

As a true channel/network integration, Fibre Channel supports an active, intelligent interconnection among devices. All that a Fibre Channel

port must do is to manage a point-to-point connection. The transmission is isolated from the control protocol, so point-to-point links, arbitrated loops, and switched topologies are used to meet the specific needs of an application. The fabric is self-managing. Nodes do not need station management, which greatly simplifies implementation.

Most current thinking in storage area networking involves the deployment of Fibre Channel between end stations (hosts and devices) in an arbitrated loop configuration or a switched environment containing several loops. Fibre Channel-Arbitrated Loop (FC-AL) was developed with peripheral connectivity in mind. It natively maps SCSI (as SCSI FCP), making it an ideal technology for high speed I/O connectivity. Native Fibre Channel Arbitrated Loop (FC-AL) disk drives will allow storage applications to take full advantage of Fibre Channel's gigabaud bandwidth, passing SCSI data directly onto the channel with access to multiple servers or nodes.

FC-AL supports 127-node addressability and 10 KM cabling ranges between nodes. The peak transfer rate of a Fibre Channel port is 1.062 GB/s, or 100 Mbytes/second, which is the link rate of the full-speed interface. A Fibre Channel adapter can burst a 2048-byte frame at the link rate.

IBM discriminates SSA from Fibre Channel by pointing to SSA's greater cost efficiency and comparable performance. This view is echoed by SSA advocates, who have raised several points intended to sell their preferred serial interface.

One often voiced concern is that the performance of a Fibre Channel-Arbitrated Loop may fall short of expectations because of Arbitration. FC-AL developers promise bandwidth of 200 MB/s, while current implementations of SSA provide bandwidth of 80 MB/s. This fact would appear to favor FC-AL.

However, in actual field implementations, SSA advocates hypothesize, data transfer rates for FC-AL are likely to be considerably slower than 200 MB/s because of the technology's loop arbitration scheme. Arbitration effectively cancels out any advantage arising from its larger theoretical bandwidth.

This concern has been voiced since the earliest days of Fibre Channel development. It was even on the minds of International Data Corporation analysts in the mid-1990s, "One must fully consider the impact of various overheads and arbitration schemes. As such, SSA implementations will frequently outperform FC-AL, depending upon workload and configurations."[11]

Fibre Channel-Arbitrated Loop, as the name suggests, is an all-or-nothing or first-come, first-served technology. The arbitrated loop is

"owned" by the end station initiating a transfer until the transfer is complete. Others must wait in a queue until the communication is complete. By contrast, SSA provides four, full-duplex, frame-multiplexed, channels. In the field, SSA advocates charge, the greater the number of disks in an SSA loop, the better the array actually performs. In a simple, unswitched FC-AL implementation, the more disks in a loop, the greater the possibility of an I/O bottleneck.

The answer to this concern, according to Fibre Channel enthusiasts, is twofold. The larger frame sizes supported by Fibre Channel and the greater bandwidth it offers makes loop arbitration less of an issue. If more data can be moved faster, enabling shorter communications sessions, then greater loop availability should be the result.

An alternative is to use a switched fabric, rather than a straight loop or hub-attached loop topology. Switching between multiple loops is a straightforward means for increasing the number of devices that can be included in a Fibre Channel network—well beyond the specified loop limits.

An SSA connection consists of two ports conducting four concurrent conversations at 20 MB/s, thereby generating a total bandwidth of 80 MB/s. The topology is a series of point-to-point links, with ports connected between nodes by three-way routers. Point-to-point duplex connections establish the means for every SSA node to communicate with every other SSA node in either direction on the loop. This topology accounts for the high reliability of SSA systems. By contrast, critics say, a FC-AL topology provides opportunities for single points of failure.

Fibre Channel backers respond that FC-AL can be made fault tolerant by cabling two fully independent, redundant loops. This cabling scheme provides two independent paths for data with fully redundant hardware. Most disk drives and disk arrays targeted for high availability environments have dual ports specifically for the purpose. Alternatively, cabling an arbitrated loop through a hub or concentrator will isolate/protect the rest of the nodes on the loop from the failure of an individual node.

Fibre Channel nodes can be directly connected to one another in a point-to-point topology or can be cabled through a concentrator/hub or switch. Because each Fibre Channel node acts as a repeater for every other node on a loop, one down or disconnected node can take the entire loop down. Concentrators, with their ability to automatically bypass a node that has gone out of service, are an essential availability tool in many Fibre Channel networks.

For disk subsystems and RAID subsystems connected to a arbitrated loop, the Fibre Channel Association, an industry group backing the tech-

nology, is strongly recommending that each device or node within the subsystem have a port bypass circuit associated with it so that any node may be bypassed and allow for "hot swapping" of a device. With the use of the PBC, if a device failure occurs, the failed device can be bypassed automatically, eliminating any disruption of data transfers or data integrity.

Many have observed that Fibre Channel is a relatively unique application of networking technology to peripherals and thus presents a new challenge for IT professionals concerned with storage. While connecting peripherals through a concentrator is a somewhat foreign concept, with the networking of peripherals comes the need for protecting the availability of networked peripherals. This can require redundancy or the use of concentrators in peripheral networks.

THINKING OUTSIDE THE BOX

Because of their roles as both device interfaces and peripheral networking technologies, Fibre Channel, SSA, and SCSI force server administrators to "think outside the box"—to develop a view of storage that goes beyond the disk drive installed on a ribbon cable in the cabinet of a general purpose server. According to many analysts, while captive storage will likely remain a part of the corporate computing landscape for the foreseeable future, new mass storage arrays and storage area networks will become increasingly prevalent as companies seek to build a more manageable enterprise storage architecture. The next chapter examines array technologies in greater detail.

ENDNOTES

1. Magnetic tape was also an early form of sequential data storage and has persisted to the present as a preferred medium for archival and backup data storage.
2. Quantum Corporation, *Storage Basics*, Milpitas, CA, 1998.
3. "IBM Introduces World's Smallest Hard Disk Drive," IBM News Release, September 9, 1998.
4. "Data Storage Breakthrough: Seagate Demonstrates World's Highest Disc Drive Areal Density," Seagate Technology Press Release, February 3, 1999.
5. "MR Heads: The Next Step in Capacity and Performance," Seagate Technology White Paper, 1998.

6. "Magnetoresistive Head Technology: A Quantum White Paper," Quantum Corporation, 1998.
7. This phenomenon was discovered by Lord Kelvin in 1857 and today is called the anisotropic magnetoresistance (AMR) effect.
8. "MR Heads: The Next Step in Capacity and Performance," Seagate Technology White Paper, op. cit.
9. "Ultra SCSI White Paper," Quantum Corporation, Milpitas, CA, 1996.
10. "IBM Reaches One Petabyte Milestone in Serial Storage Architecture Shipments," IBM Press Release, San Jose, CA, February 25, 1997.
11. "The Case for Serial Interfaces," International Data Corporation, Framingham, MA, September 1995.