Chapter 5

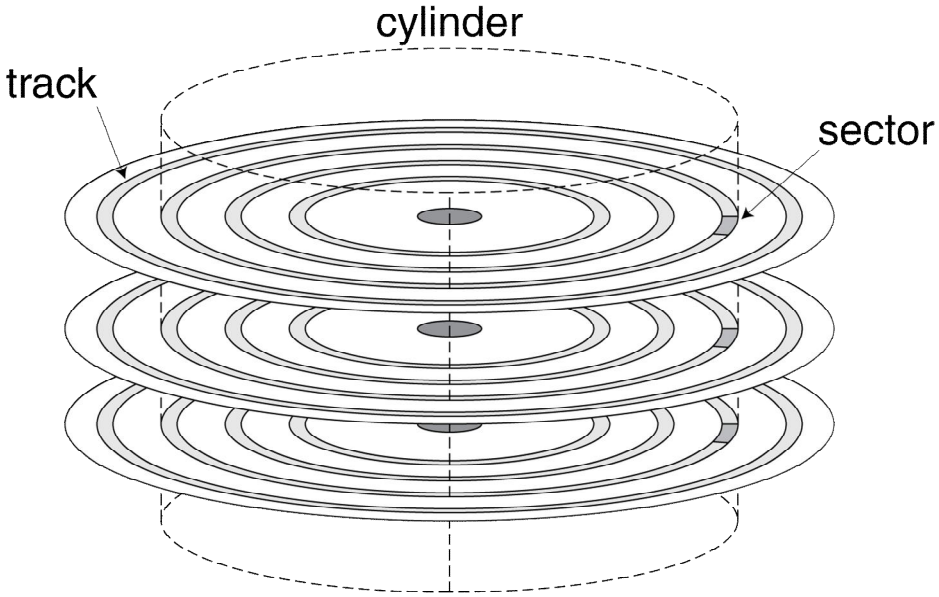
HDD Working

Memory Technology

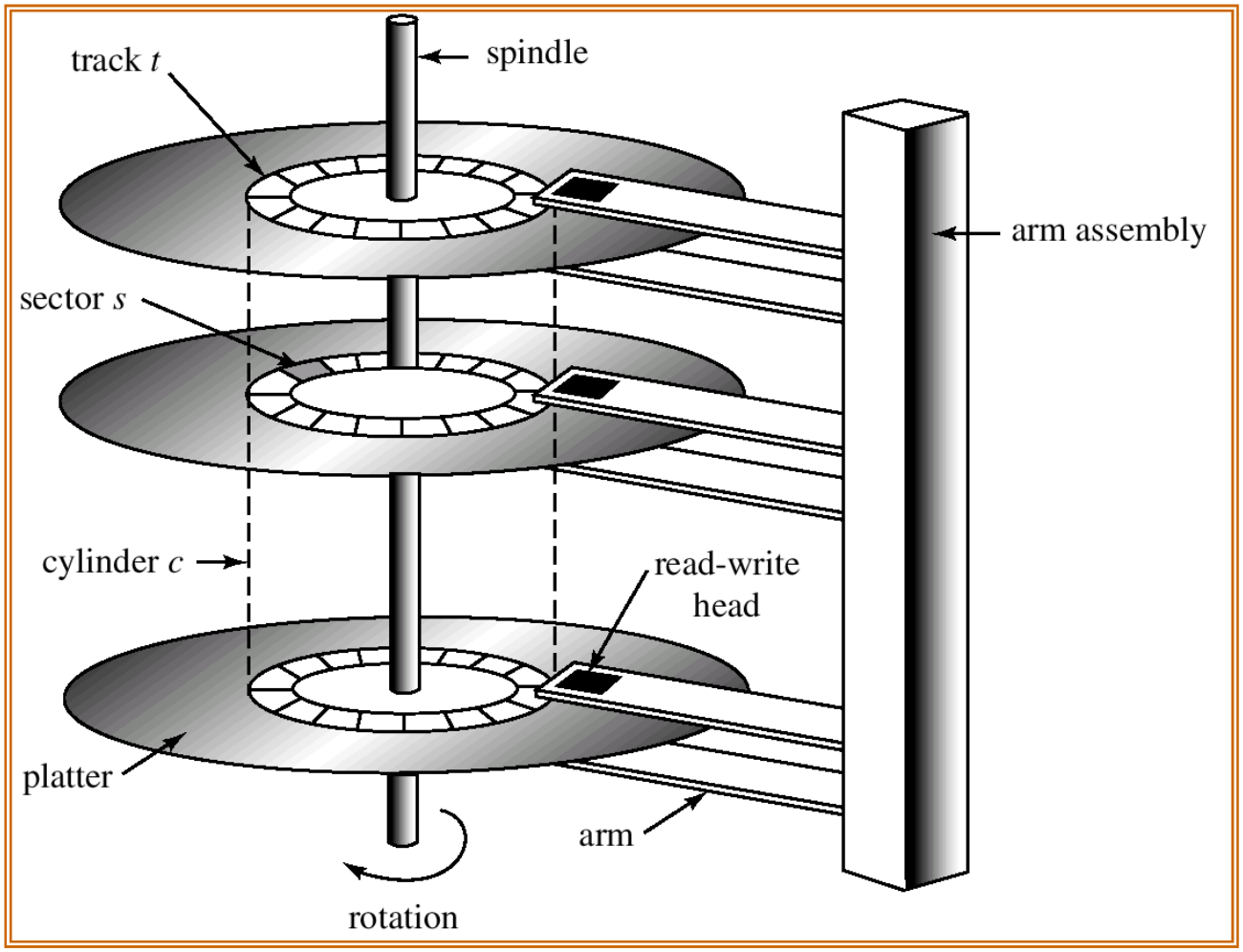
- Static RAM (SRAM)
  - 0.5ns – 2.5ns, $2000 – $5000 per GB
- Dynamic RAM (DRAM)
  - 50ns – 70ns, $20 – $75 per GB
- Magnetic disk
  - 5ms – 20ms, $0.20 – $2 per GB
- Ideal memory should have:
  - Access time of SRAM
  - Capacity and cost/GB of disk

Disk Storage

- Nonvolatile, rotating magnetic storage



Moving Head Disk Mechanism



Disk Sectors and Access

- Each sector records
  - Sector ID
  - Data (512 bytes, 4096 bytes proposed)
  - Error correcting code (ECC)
    - Used to hide defects and recording errors
  - Synchronization fields and gaps
- Access to a sector involves
  - Queuing delay if other accesses are pending
  - Seek: move the heads
  - Rotational latency
  - Data transfer
  - Controller overhead

Disk Access Example

- Given
  - 512B sector, 15,000rpm, 4ms average seek time, 100MB/s transfer rate, 0.2ms controller overhead, idle disk
- Average read time
  - 4ms seek time
    + ½ / (15,000/60) = 2ms rotational latency
    + 512 / 100MB/s = 0.005ms transfer time
    + 0.2ms controller delay
    = 6.2ms
- If actual average seek time (often less than specified seek time) is 1ms
  - Average read time = 3.2ms

Note that reliability and availability are actually quantifiable measures, rather than just synonyms for dependability.

What is the cause of failures? Figure 6.3 summarizes many papers that have collected data on reasons for computer systems and telecommunications systems to fail. Clearly, human operators are a significant source of failures.

Operator	Software	Hardware	System	Year data collected
42%	25%	18%	Datacenter (Tandem)	1985
15%	55%	14%	Datacenter (Tandem)	1989
18%	44%	39%	Datacenter (DEC VAX)	1985
50%	20%	30%	Datacenter (DEC VAX)	1993
50%	14%	19%	U.S. public telephone network	1996
54%	7%	30%	U.S. public telephone network	2000
60%	25%	15%	Internet services	2002

FIGURE 6.3 Summary of studies of reasons for failures. Although it is difficult to collect data to determine whether operators are the cause of errors, since operators often record the reasons for failures, these studies did capture that data. There were often other categories, such as environmental reasons for outages, but they were generally small. The top two rows come from a classic paper by Jim Gray [1990], which is still widely quoted almost 20 years after the data was collected. The next two rows are from a paper by Murphy and Gent, who studied causes of outages in VAX systems over time ["Measuring system and software reliability using an automated data collection process," *Quality and Reliability Engineering International* 11:5, September–October 1995, 341–53]. The fifth and sixth rows are studies of FCC failure data about the U.S. public switched telephone network by Kuhn ["Sources of failure in the public switched telephone network," *IEEE Computer* 30:4, April 1997, 31–36] and by Patty Enriquez. The study of three Internet services is from Oppenheimer, Ganapath, and Patterson [2003].

To increase MTTF, you can improve the quality of the components or design systems to continue operation in the presence of components that have failed. Hence, failure needs to be defined with respect to a context. A failure in a component may not lead to a failure of the system. To make this distinction clear, the term *fault* is used to mean failure of a component. Here are three ways to improve MTTF:

1. *Fault avoidance:* Preventing fault occurrence by construction.
2. *Fault tolerance:* Using redundancy to allow the service to comply with the service specification despite faults occurring, which applies primarily to hardware faults. Section 6.9 describes the RAID approaches to making storage dependable via fault tolerance.
3. *Fault forecasting:* Predicting the presence and creation of faults, which applies to hardware and software faults, allowing the component to be replaced before it fails.

Shrinking MTTR can help availability as much as increasing MTTF. For example, tools for fault detection, diagnosis, and repair can help reduce the time to repair faults by people, software, and hardware.

Check Yourself

Which of the following are true about dependability?

1. If a system is up, then all its components are accomplishing their expected service.
2. Availability is a quantitative measure of the percentage of time a system is accomplishing its expected service.
3. Reliability is a quantitative measure of continuous service accomplishment by a system.
4. The major source of outages today is software.

6.3 Disk Storage

As mentioned in Chapter 1, magnetic disks rely on a rotating platter coated with a magnetic surface and use a moveable read/write head to access the disk. Disk storage is **nonvolatile**—the data remains even when power is removed. A magnetic disk consists of a collection of platters (1–4), each of which has two recordable disk surfaces. The stack of platters is rotated at 5400 to 15,000 RPM and has a diameter from 1-inch to just over 3.5 inches. Each disk surface is divided into concentric circles, called **tracks**. There are typically 10,000 to 50,000 tracks per surface. Each track is in turn divided into **sectors** that contain the information; each track may have 100 to 500 sectors. Sectors are typically 512 bytes in size, although there is an initiative to increase the sector size to 4096 bytes. The sequence recorded on the magnetic media is a sector number, a gap, the information for that sector including error correction code (see **Appendix C**, page C-66), a gap, the sector number of the next sector, and so on.

Originally, all tracks had the same number of sectors and hence the same number of bits. With the introduction of zone bit recording (ZBR) in the early 1990s, disk drives changed to a varying number of sectors (and hence bits) per track, instead keeping the spacing between bits constant. ZBR increases the number of bits on the outer tracks and thus increases the drive capacity.

As we saw in Chapter 1, to read and write information the read/write heads must be moved so that they are over the correct location. The disk heads for each surface are connected together and move in conjunction, so that every head is over the same track of every surface. The term *cylinder* is used to refer to all the tracks under the heads at a given point on all surfaces.

To access data, the operating system must direct the disk through a three-stage process. The first step is to position the head over the proper track. This operation is called a **seek**, and the time to move the head to the desired track is called the *seek time.*

nonvolatile Storage device where data retains its value even when power is removed.

track One of thousands of concentric circles that makes up the surface of a magnetic disk.

sector One of the segments that make up a track on a magnetic disk; a sector is the smallest amount of information that is read or written on a disk.

seek The process of positioning a read/write head over the proper track on a disk.

Disk manufacturers report minimum seek time, maximum seek time, and average seek time in their manuals. The first two are easy to measure, but the average is open to wide interpretation because it depends on the seek distance. The industry has decided to calculate average seek time as the sum of the time for all possible seeks divided by the number of possible seeks. Average seek times are usually advertised as 3 ms to 13 ms, but, depending on the application and scheduling of disk requests, the actual average seek time may be only 25% to 33% of the advertised number because of locality of disk references. This locality arises both because of successive accesses to the same file and because the operating system tries to schedule such accesses together.

rotational latency Also called **rotational delay**. The time required for the desired sector of a disk to rotate under the read/write head; usually assumed to be half the rotation time.

Once the head has reached the correct track, we must wait for the desired sector to rotate under the read/write head. This time is called the **rotational latency** or **rotational delay**. The average latency to the desired information is halfway around the disk. Because the disks rotate at 5400 RPM to 15,000 RPM, the average rotational latency is between

$$\text{Average rotational latency} = \frac{0.5 \text{ rotation}}{5400 \text{ RPM}} = \frac{0.5 \text{ rotation}}{5400 \text{ RPM}/\left(60 \frac{\text{seconds}}{\text{minute}}\right)}$$

$$= 0.0056 \text{ seconds} = 5.6 \text{ ms}$$

and

$$\text{Average rotational latency} = \frac{0.5 \text{ rotation}}{15{,}000 \text{ RPM}} = \frac{0.5 \text{ rotation}}{15{,}000 \text{ RPM}/\left(60 \frac{\text{seconds}}{\text{minute}}\right)}$$

$$= 0.0020 \text{ seconds} = 2.0 \text{ ms}$$

The last component of a disk access, *transfer time*, is the time to transfer a block of bits. The transfer time is a function of the sector size, the rotation speed, and the recording density of a track. Transfer rates in 2008 were between 70 and 125 MB/sec. The one complication is that most disk controllers have a built-in cache that stores sectors as they are passed over; transfer rates from the cache are typically higher and may be up to 375 MB/sec (3 Gbit/sec) in 2008. Today, most disk transfers are multiple sectors in length.

A *disk controller* usually handles the detailed control of the disk and the transfer between the disk and the memory. The controller adds the final component of disk access time, *controller time*, which is the overhead the controller imposes in performing an I/O access. The average time to perform an I/O operation will consist of these four times plus any wait time incurred because other processes are using the disk.

Disk Read Time

EXAMPLE

What is the average time to read or write a 512-byte sector for a typical disk rotating at 15,000 RPM? The advertised average seek time is 4 ms, the transfer rate is 100 MB/sec, and the controller overhead is 0.2 ms. Assume that the disk is idle so that there is no waiting time.

ANSWER

Average disk access time is equal to average seek time + average rotational delay + transfer time + controller overhead. Using the advertised average seek time, the answer is

$$4.0 \text{ ms} + \frac{0.5 \text{ rotation}}{15{,}000 \text{ RPM}} + \frac{0.5 \text{ KB}}{100 \text{ MB/sec}} + 0.2 \text{ ms} = 4.0 + 2.0 + 0.005 + 0.2 = 6.2 \text{ ms}$$

If the measured average seek time is 25% of the advertised average time, the answer is

$$1.0 \text{ ms} + 2.0 \text{ ms} + 0.005 \text{ ms} + 0.2 \text{ ms} = 3.2 \text{ ms}$$

Notice that when we consider measured average seek time, as opposed to advertised average seek time, the rotational latency can be the largest component of the access time.

Disk densities have continued to increase for more than 50 years. The impact of this compounded improvement in density and the reduction in physical size of a disk drive has been amazing, as Figure 6.4 shows. The aims of different disk designers have led to a wide variety of drives being available at any particular time. Figure 6.5 shows the characteristics of four magnetic disks. In 2008, these disks from a single manufacturer cost between $0.30 and $5.00 per gigabyte. In the broader market, prices generally range between $0.20 and $2.00 per gigabyte, depending on size, interface, and performance.

While disks will remain viable for the foreseeable future, the conventional wisdom about where block numbers are found has not. The assumptions of the sector-track-cylinder model are that nearby blocks are on the same track, blocks in the same cylinder take less time to access since there is no seek time, and some tracks are closer than others. The reason for the breakdown was the raising of the level of the interfaces. Higher-level intelligent interfaces like **ATA** and **SCSI** required a microprocessor inside a disk, which lead to performance optimizations.

Advanced Technology Attachment (ATA) A command set used as a standard for I/O devices that is popular in the PC.

Small Computer Systems Interface (SCSI) A command set used as a standard for I/O devices.

To speed-up sequential transfers, these higher-level interfaces organize disks more like tapes than like random access devices. The logical blocks are ordered in serpentine fashion across a single surface, trying to capture all the sectors that are recorded at the same bit density. Hence, sequential blocks may be on different tracks. We will see an example in Figure 6.19 of the pitfall of assuming the conventional sector-track-cylinder model.

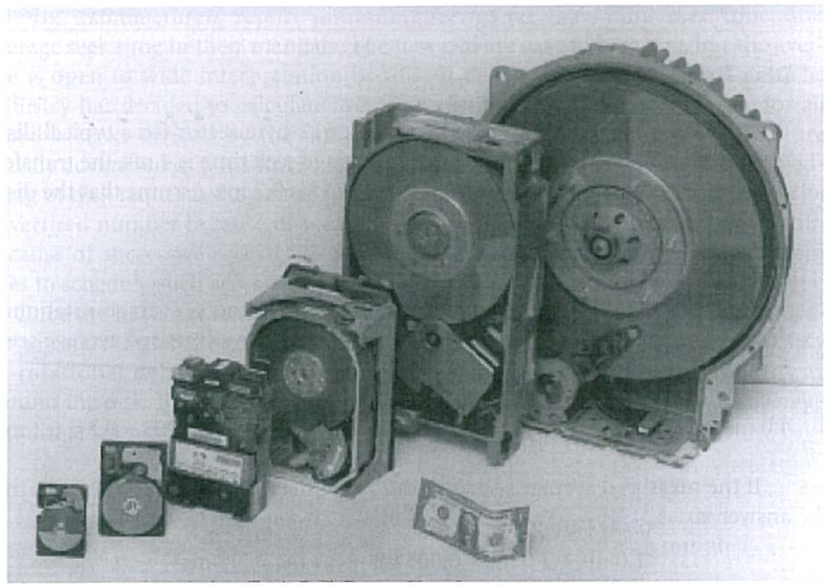


FIGURE 6.4 Six magnetic disks, varying in diameter from 14 inches down to 1.8 inches. The pictured disks were introduced over more than 15 years ago and hence are not intended to be representative of the best capacity of modern disks of these diameters. This photograph does, however, accurately portray their relative physical sizes. The widest disk is the DEC R81, containing four 14-inch diameter platters and storing 456 MB. It was manufactured in 1985. The 8-inch diameter disk comes from Fujitsu, and this 1984 disk stores 130 MB on six platters. The Micropolis RD53 has five 5.25-inch platters and stores 85 MB. The IBM 0361 also has five platters, but these are just 3.5 inches in diameter. This 1988 disk holds 320 MB. In 2008, the most dense 3.5-inch disk had 2 platters and held 1 TB in the same space, yielding an increase in density of about 3000 times! The Conner CP 2045 has two 2.5-inch platters containing 40 MB and was made in 1990. The smallest disk in this photograph is the Integral 1820. This single 1.8-inch platter contains 20 MB and was made in 1992.

Elaboration: These high-level interfaces let disk controllers add caches, which allow for fast access to data that was recently read between transfers requested by the processor. They use write-through and do not update on a write miss, and often also include prefetch algorithms to try to anticipate demand. Controllers also use a command queue that allow the disk to decide in what order to perform the commands to maximize performance while maintaining correct behavior. Of course, such capabilities complicate the measurement of disk performance and increase the importance of workload choice when comparing disks.

Characteristics	Seagate ST33000655SS	Seagate ST31000340NS	Seagate ST973451SS	Seagate ST9160821AS
Disk diameter (inches)	3.50	3.50	2.50	2.50
Formatted data capacity (GB)	147	1000	73	160
Number of disk surfaces (heads)	2	4	2	2
Rotation speed (RPM)	15,000	7200	15,000	5400
Internal disk cache size (MB)	16	32	16	8
External interface, bandwidth (MB/sec)	SAS, 375	SATA, 375	SAS, 375	SATA, 150
Sustained transfer rate (MB/sec)	73–125	105	79–112	44
Minimum seek (read/write) (ms)	0.2/0.4	0.8/1.0	0.2/0.4	1.5/2.0
Average seek read/write (ms)	3.5/4.0	8.5/9.5	2.9/3.3	12.5/13.0
Mean time to failure (MTTF) (hours)	1,400,000 @ 25°C	1,200,000 @ 25°C	1,600,000 @ 25°C	—
Annual failure rate (AFR) (percent)	0.62%	0.73%	0.55%	—
Contact start-stop cycles	—	50,000	—	>600,000
Warranty (years)	5	5	5	5
Nonrecoverable read errors per bits read	<1 sector per 10^{16}	<1 sector per 10^{15}	<1 sector per 10^{16}	<1 sector per 10^{14}
Temperature, shock (operating)	5°–55°C, 60 G	5°–55°C, 63 G	5°–55°C, 60 G	0°–60°C, 350 G
Size: dimensions (in.), weight (pounds)	1.0" × 4.0" × 5.8", 1.5 lbs	1.0" × 4.0" × 5.8", 1.4 lbs	0.6" × 2.8" × 3.9", 0.5 lbs	0.4" × 2.8" × 3.9", 0.2 lbs
Power: operating/idle/standby (watts)	15/11/—	11/8/1	8/5.8/—	1.9/0.6/0.2
GB/cu. in., GB/watt	6 GB/cu.in., 10 GB/W	43 GB/cu.in., 91 GB/W	11 GB/cu.in., 9 GB/W	37 GB/cu.in., 84 GB/W
Price in 2008, $/GB	~ $250, ~ $1.70/GB	~ $275, ~ $0.30/GB	~ $350, ~ $5.00/GB	~ $100, ~ $0.60/GB

FIGURE 6.5 Characteristics of four magnetic disks by a single manufacturer in 2008. The three leftmost drives are for servers and desktops while the rightmost drive is for laptops. Note that the third drive is only 2.5 inches in diameter, but it is a high performance drive with the highest reliability and fastest seek time. The disks shown here are either serial versions of the interface to SCSI (SAS), a standard I/O bus for many systems, or serial version of ATA (SATA), a standard I/O bus for PCs. The transfer rates from the caches is 3–5 times faster than the transfer rate from the disk surface. The much lower cost per gigabyte of the SATA 3.5-inch drive is primarily due to the hyper-competitive PC market, although there are differences in performance in I/Os per second due to faster rotation and faster seek times for SAS. The service life for these disks is five years. Note that the quoted MTTF assumes nominal power and temperature. Disk lifetimes can be much shorter if temperature and vibration are not controlled. See the link to Seagate at *www.seagate.com* for more information on these drives.

Check Yourself

Which of the following are true about disk drives?

1. 3.5-inch disks perform more IOs per second than 2.5-inch disks.
2. 2.5-inch disks offer the highest gigabytes per watt.
3. It takes hours to read the contents of a high capacity disk sequentially.
4. It takes months to read the contents of a high capacity disk using random 512-byte sectors.