Review of

Computer Organization

Operating System (CSC1036 & INF1036)

Dr. Hasin A. Ahmed
Assistant Professor
Department of Computer Science
Gauhati University

Primary Memory

PRIMARY MEMORY

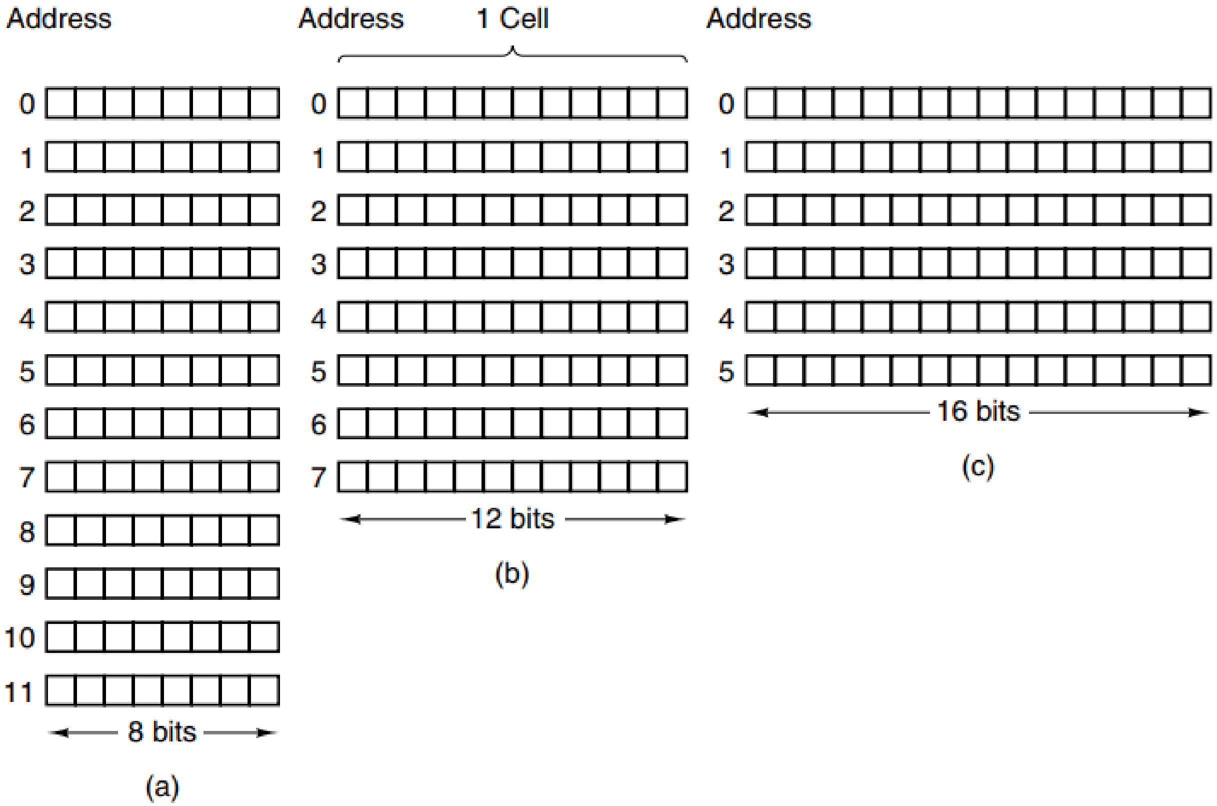
- The memory is that part of the computer where programs and data are stored
- Some use the term store or storage rather than memory
- Although more and more, the term “storage” is used to refer to disk storage
- The basic unit of memory is the binary digit, called a bit. A bit may contain a 0 or a 1

Name	Equal To	Size(In Bytes)
Bit	1 Bit	1/8
Nibble	4 Bits	1/2 (rare)
Byte	8 Bits	1
Kilobyte	1024 Bytes	1024
Megabyte	1, 024 Kilobytes	1, 048, 576
Gigabyte	1, 024 Megabytes	1, 073, 741, 824
Terrabyte	1, 024 Gigabytes	1, 099, 511, 627, 776
Petabyte	1, 024 Terabytes	1, 125, 899, 906, 842, 624
Exabyte	1, 024 Petabytes	1, 152, 921, 504, 606, 846, 976
Zettabyte	1, 024 Exabytes	1, 180, 591, 620, 717, 411, 303, 424
Yottabyte	1, 024 Zettabytes	1, 208, 925, 819, 614, 629, 174, 706, 176

MEMORY ADDRESSES

- Memories consist of a number of cells (or locations), each of which can store a piece of information.
- Each cell has a number, called its address, by which programs can refer to it.
- If a memory has n cells, they will have addresses 0 to n − 1.
- All cells in a memory contain the same number of bits.

96 BIT MEMORY



MEMORY ADDRESSES

- Computers that use the binary number system express memory addresses as binary numbers.
- If an address has m bits, the maximum number of cells addressable is 2^m.
- For example, an address used to reference the memory of Fig. (a) needs at least 4 bits
- A 3-bit address is sufficient for Fig. (b) and (c)

MEMORY CELLS

- The significance of the cell is that it is the smallest addressable unit
- In recent years, nearly all computer manufacturers have standardized on an 8-bit cell, which is called a byte.
- The term octet is also used.

MEMORY WORDS

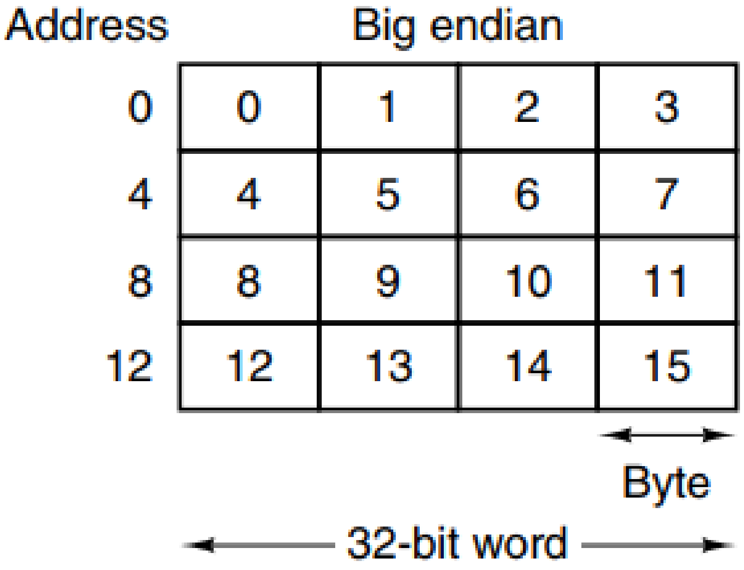
- Bytes are grouped into words.
- A computer with a 32-bit word has 4 bytes/word, whereas a computer with a 64-bit word has 8 bytes/word.
- The significance of a word is that most instructions operate on entire words, for example, adding two words together.

MEMORY WORDS

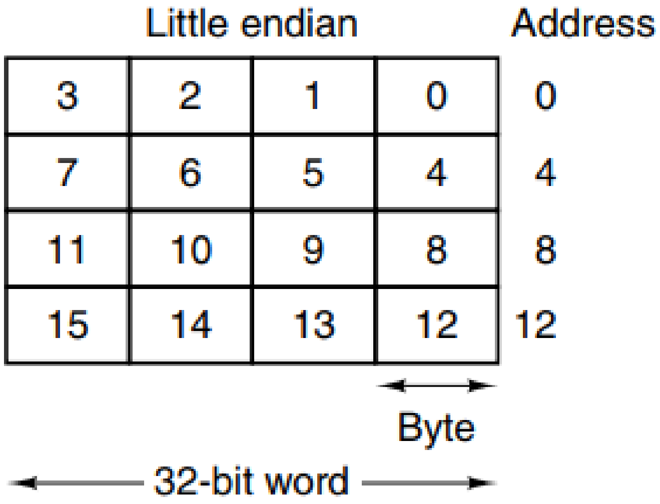
- Thus a 32-bit machine will have 32-bit registers and instructions for manipulating 32-bit words, whereas a 64-bit machine will have 64-bit registers and instructions for moving, adding, subtracting, and manipulating 64-bit words

BYTE ORDERING

- Figure depicts part of the memory of a 32-bit computer whose bytes are numbered from left to right, such as the SPARC or the big IBM mainframes.



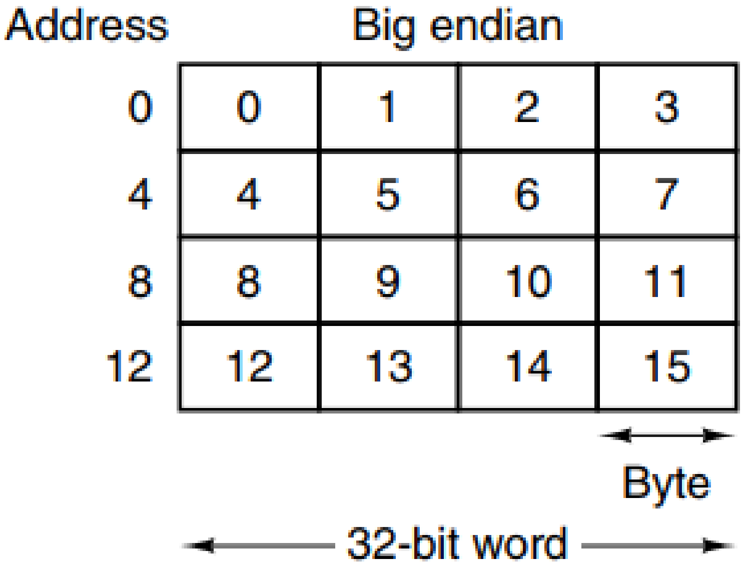
BYTE ORDERING



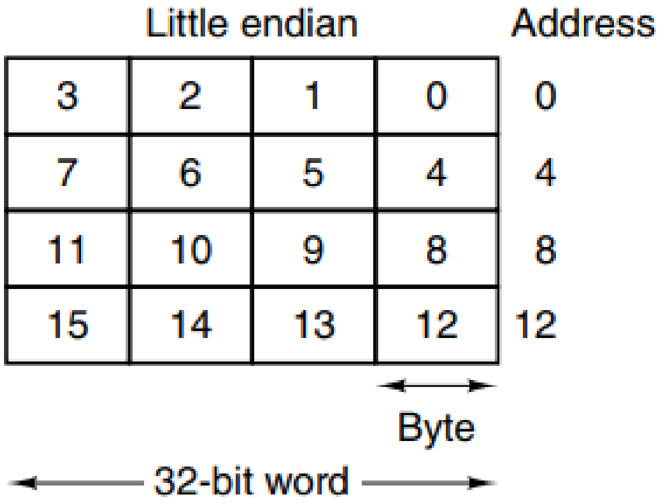
- Figure gives the analogous representation of a 32-bit computer using right-to-left numbering, such as the Intel family.

BIG ENDIAN SYSTEM

- The system, where the “big” (i.e., high-order or most significant) byte is stored first, at the lowest storage address



LITTLE ENDIAN SYSTEM



- The system, where the "little" (i.e., low-order or least significant) byte is stored first, at the lowest storage address

CACHE MEMORY

- As it becomes possible to put more and more circuits on a chip, CPU designers are using these new facilities for pipelining and superscalar operation, making CPUs go even faster
- Memory designers have usually used new technology to increase the capacity of their chips, not the speed

CACHE MEMORY

- After the CPU issues a memory request, it will not get the word it needs for many CPU cycles
- The problem is not technology, but economics
- Memory have to be located on the CPU chip
- Putting a large memory on the CPU chip makes it bigger expensive

CACHE MEMORY

- Even if cost were not an issue, there are limits to how big a CPU chip can be made
- Cache memory is the solution
- The basic idea behind a cache is simple: the most heavily used memory words are kept in the cache

CACHE MEMORY

- When the CPU needs a word, it first looks in the cache.
- Only if the word is not there does it go to main memory.
- If a substantial fraction of the words are in the cache, the average access time can be greatly reduced.

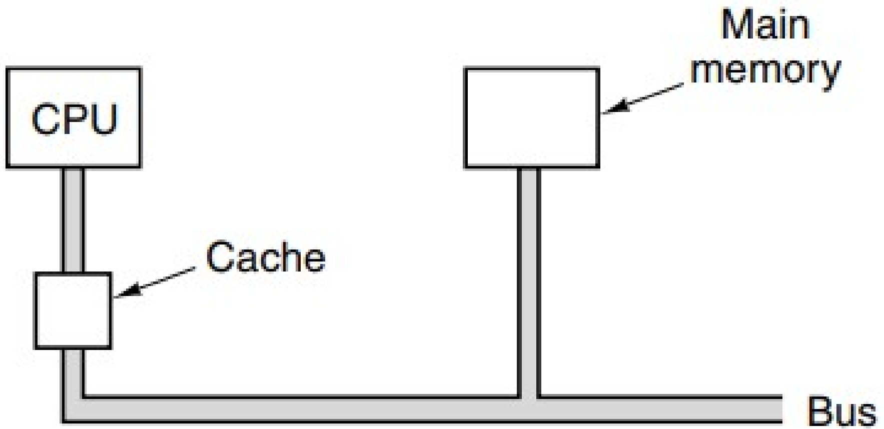
CACHE MEMORY

- Success or failure thus depends on what fraction of the words are in the cache.
- For years, people have known that programs do not access their memories completely at random.
- If a given memory reference is to address A, it is likely that the next memory reference will be in the general vicinity of A

CACHE MEMORY

- The observation that the memory references made in any short time interval tend to use only a small fraction of the total memory is called the **locality of reference** principle and forms the basis for all caching systems.

CACHE MEMORY

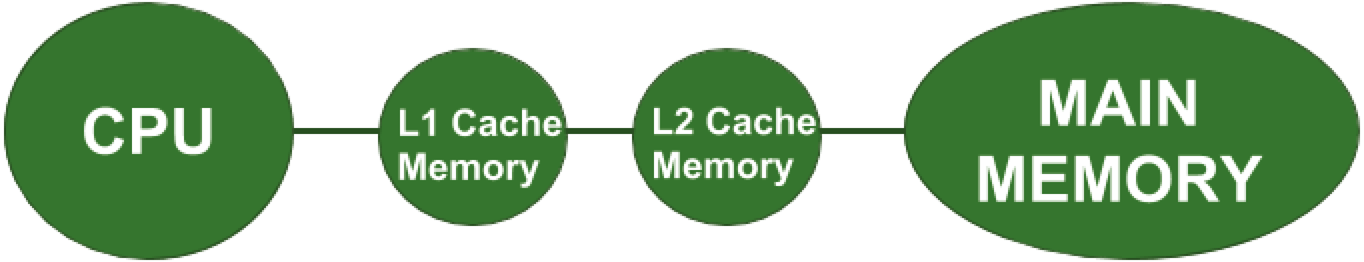


CACHE MEMORY

We can formalize mean access time introducing c, the cache access time, m, the main memory access time, and h, the hit ratio, which is the fraction of all references that can be satisfied out of the cache as

Effective Access Time = h*c+(1-h)*m

CACHE MEMORY



Tavg = H1 * C1 + (1 – H1) * (H2 * C2 +(1 – H2) *M)

Where

H1 is the Hit rate in the L1 caches.

H2 is the Hit rate in the L2 cache.

C1 is the Time to access information in the L1 caches.

C2 is the Miss penalty to transfer information from the L2 cache to an L1 cache.

M is the Miss penalty to transfer information from the main memory to the L2 cache.

CACHE MEMORY

Find the Average memory access time for a processor with a 2 ns clock cycle time, a miss rate of 0.04 misses per instruction, a missed penalty of 25 clock cycles, and a cache access time (including hit detection) of 1 clock cycle. Also, assume that the read and write miss penalties are the same and ignore other write stalls.

Solution:

Average Memory access time(AMAT)= Hit Time + Miss Rate * Miss Penalty.

Hit Time = 1 clock cycle (Hit time = Hit rate * access time) but here Hit time is directly given so,

Miss rate = 0.04

Miss Penalty= 25 clock cycle (this is the time taken by the above level of memory after the hit)

so, AMAT= 1 + 0.04 * 25

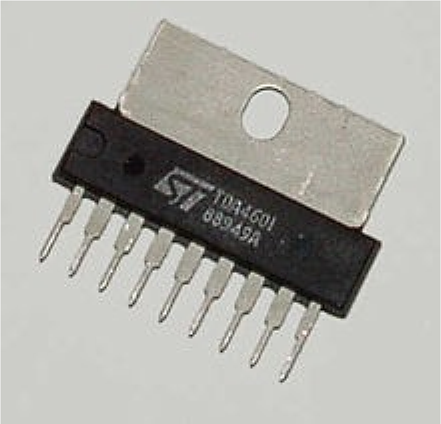
AMAT= 2 clock cycle

according to question 1 clock cycle = 2 ns

AMAT = 4ns

SIMM and DIMM

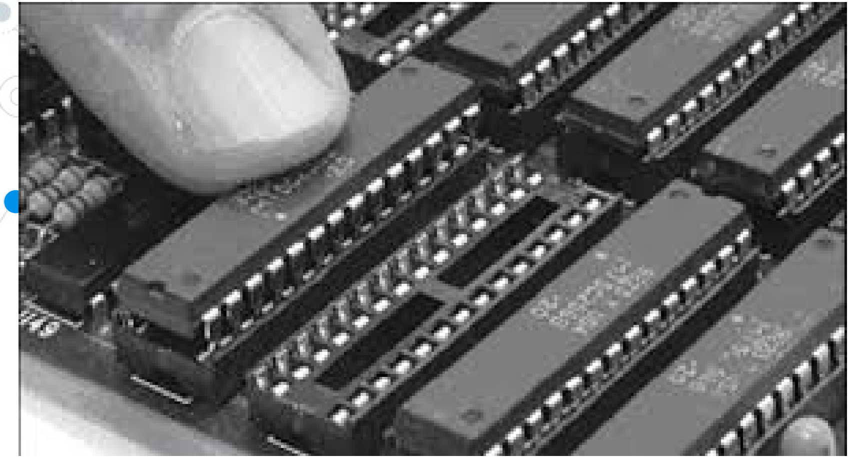
- Single in-line package (SIP)
- Dual in-line package (DIP)

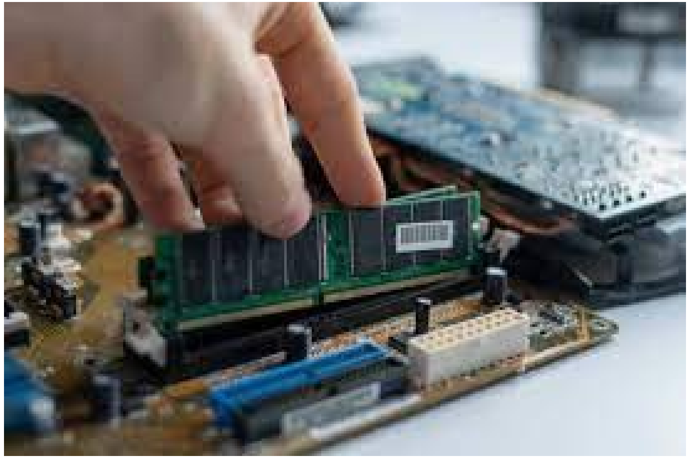


SIMM and DIMM

- Until the early 1990s, memory was manufactured, bought, and installed as single chips.
- Chip densities went from 1K bits to 1M bits and beyond, but each chip was sold as a separate unit.
- Early PCs often had empty sockets into which additional memory chips could be plugged, if and when the purchaser needed them

SIMM and DIMM





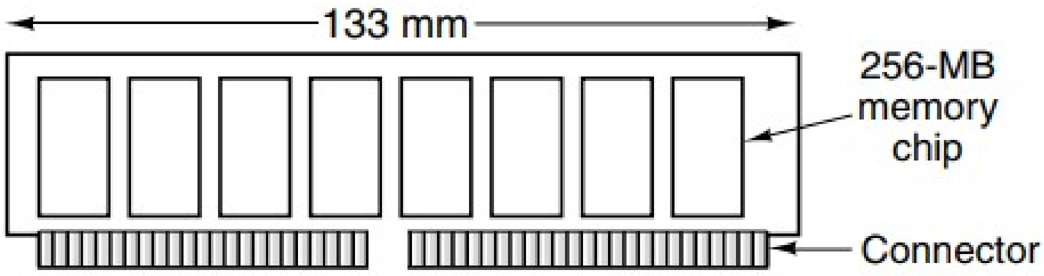
SIMM and DIMM

- Since the early 1990s, a different arrangement has been used.
- A group of chips, typically 8 or 16, is mounted on a printed circuit board and sold as a unit.
- This unit is called a SIMM (Single Inline Memory Module) or a DIMM (Dual Inline Memory Module), depending on whether it has a row of connectors on one side or both sides of the board.

SIMM and DIMM

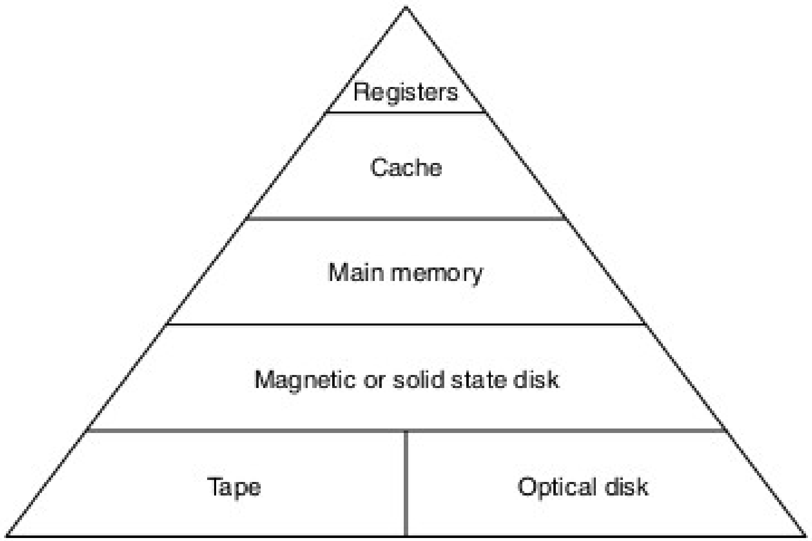
- SIMMs have one edge connector with 72 contacts and transfer 32 bits per clock cycle (DRAM). They are rarely used these days.
- DIMMs usually have edge connectors with 120 contacts on each side of the board (DDR3), for a total of 240 contacts, and transfer 64 bits per clock cycle.

SIMM and DIMM



Top view of a DIMM holding 4 GB with eight chips of 256 MB on each side. The other side looks the same.

Memory Hierarchy



Memory Hierarchy

- As we move down the hierarchy, three key parameters increase
- First, the access time gets bigger
- Second, the storage capacity increases as we go downward
- Third, the number of bits you get per dollar spent increases down the hierarchy

Magnetic Hard Disk

- A magnetic disk consists of one or more aluminum platters with a magnetizable coating.
- They are typically 3 to 9 cm in diameter
- A disk head containing an induction coil floats just over the surface
- When a positive or negative current passes through the head, it magnetizes the surface
- When the head passes over a magnetized area, a positive or negative current is induced in the head

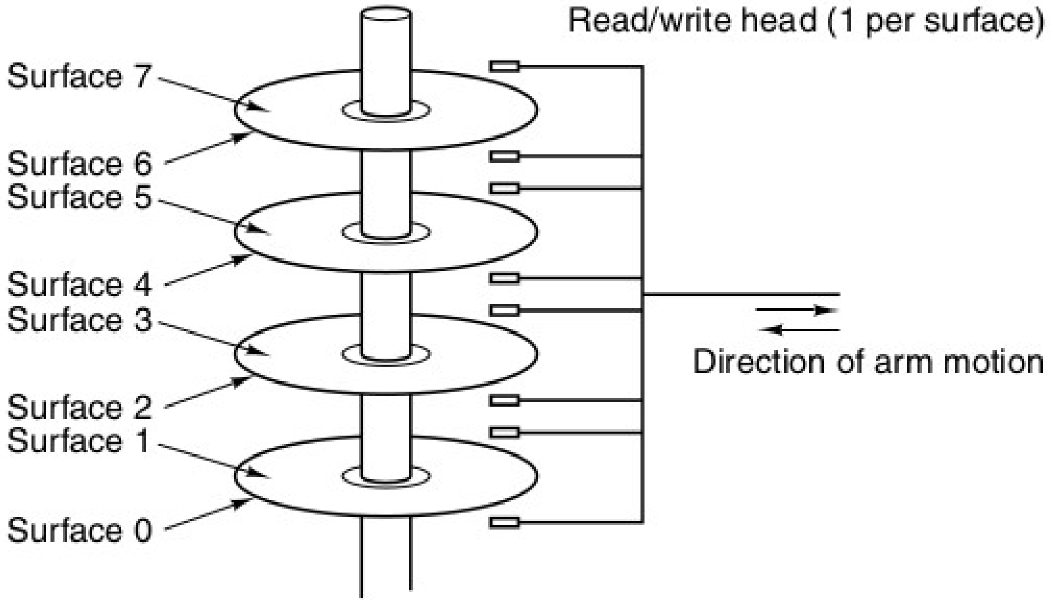
Magnetic Hard Disk

- The circular sequence of bits written as the disk makes a complete rotation is called a track.
- Each track is divided up into some number of fixed-length sectors, typically containing 512 data bytes, preceded by a preamble that allows the head to be synchronized before reading or writing.
- Following the data is an Error-Correcting Code (ECC), either a Hamming code or, more commonly, a code that can correct multiple errors called a Reed-Solomon code.
- Between consecutive sectors is a small intersector gap.

Magnetic Hard Disk

- All disks have movable arms that are capable of moving in and out to different radial distances from the spindle about which the platter rotates.
- At each radial distance, a different track can be read or written.
- The tracks are thus a series of concentric circles about the spindle.
- In order to achieve high surface and air quality, most disks are sealed at the factory to prevent dust from getting in.
- Such drives were originally called Winchester disks against Bernoulli disk and disk pack

Magnetic Hard Disk



Magnetic Hard Disk

- Named hard disks to differentiate them from the long-vanished floppy disks
- Most disks consist of multiple platters stacked vertically
- Each surface has its own arm and head.
- All the arms are ganged together so they move to different radial positions all at once.
- The set of tracks at a given radial position is called a cylinder.
- Current PC and server disks typically have 1 to 12 platters
- High-end disks can store 1 TB on a single platter

Magnetic Hard Disk

- To read or write a sector, first the arm must be moved to the right radial position. This action is called a seek.
- Average seek times (between random tracks) range in the 5- to 10-msec
- Once the head is positioned radially, there is a delay, called the rotational latency, until the desired sector rotates under the head.
- Most disks rotate at 5400 RPM, 7200 RPM, or 10,800 RPM, so the average delay (half a rotation) is 3 to 6 msec.
- Transfer time depends on the linear density and rotation speed.
- With typical internal transfer rate of 150 MB/sec, a 512-byte sector takes about 3.5 μ sec

Magnetic Hard Disk

- Since all magnetic disks rotate at a constant angular velocity, no matter where the heads are, this observation creates a problem.
- In older drives, manufacturers used the maximum possible linear density on the innermost track, and successively lowered linear bit densities on tracks further out.
- Nowadays, a different strategy is used. Cylinders are divided into zones (typically 10 to 30 per drive), and the number of sectors per track is increased in each zone moving outward from the innermost track.

Magnetic Hard Disk

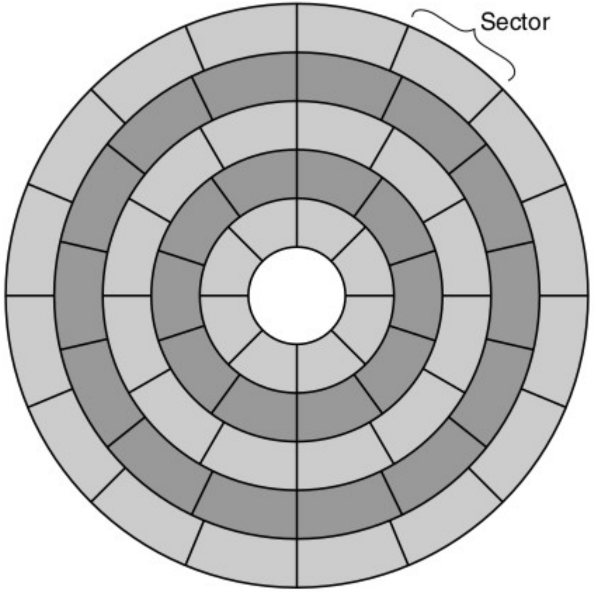


Figure 2-21. A disk with five zones. Each zone has many tracks.

Magnetic Hard Disk

- Associated with each drive is a disk controller, a chip that controls the drive.
- Some controllers contain a full CPU.
- The controller’s tasks include accepting commands from the software, such as READ , WRITE , and FORMAT (writing all the preambles), controlling the arm motion, detecting and correcting errors
- Some controllers also handle buffering of multiple sectors, caching sectors read for potential future use, and remapping bad sectors

CD-ROM

- A CD is prepared by using a high-power infrared laser to burn 0.8-micron diameter holes in a coated glass master disk.
- From this master, a mold is made, with bumps where the laser holes were.
- Into this mold, molten polycarbonate is injected to form a CD with the same pattern of holes as the glass master.
- Then a thin layer of reflective aluminum is deposited on the polycarbonate, topped by a protective lacquer and finally a label.
- The depressions in the polycarbonate substrate are called pits; the unburned areas between the pits are called lands.

CD-ROM

- When a CD is played back, a low-power laser diode shines infrared light with a wavelength of 0.78 micron on the pits and lands as they stream by.
- The laser is on the polycarbonate side, so the pits disburse the light.
- As a result, return less light to the player's photodetector than light bouncing off a land.
- Although it might seem simplest to use a pit to record a 0 and a land to record a 1, it is more reliable to use a pit/land or land/pit transition for a 1 and its absence as a 0, so this scheme is used.

CD-ROM

- **Constant Linear Velocity:** it is necessary for the pits and lands to stream by at a constant linear velocity.
- Consequently, the rotation rate of the CD must be continuously reduced as the reading head moves from the inside of the CD to the outside

DVD (Digital Versatile Disk)

- DVDs use the same general design as Cds polycarbonate disks containing pits and lands that are illuminated by a laser diode and read by a photodetector
- What is new is the use of
  - Smaller pits (0.4 microns versus 0.8 microns for Cds).
  - A tighter spiral (0.74 microns between tracks versus 1.6 microns for CDs).
  - A red laser (at 0.65 microns versus 0.78 microns for CDs).

DVD (Digital Versatile Disk)

- Together, these improvements raise the capacity sevenfold, to 4.7 GB.
- A 1x DVD drive operates at 1.4 MB/sec (versus 150 KB/sec for CDs).
- Four formats have been defined:
  - Single-sided, single-layer (4.7 GB)
  - Single-sided, dual-layer (8.5 GB)
  - Double-sided, single-layer (9.4 GB)
  - Double-sided, dual-layer (17 GB)

Blu-ray

- The successor to DVD is Blu-ray, so called because it uses a blue laser instead of the red one used by DVDs.
- A blue laser has a shorter wavelength than a red one, which allows it to focus more accurately and thus support smaller pits and lands.
- Single-sided Blu-ray disks hold about 25 GB of data; double-sided ones hold about 50 GB.
- The data rate is about 4.5 MB/sec, which is good for an optical disk, but still insignificant compared to magnetic disks

thank you

? ? ?