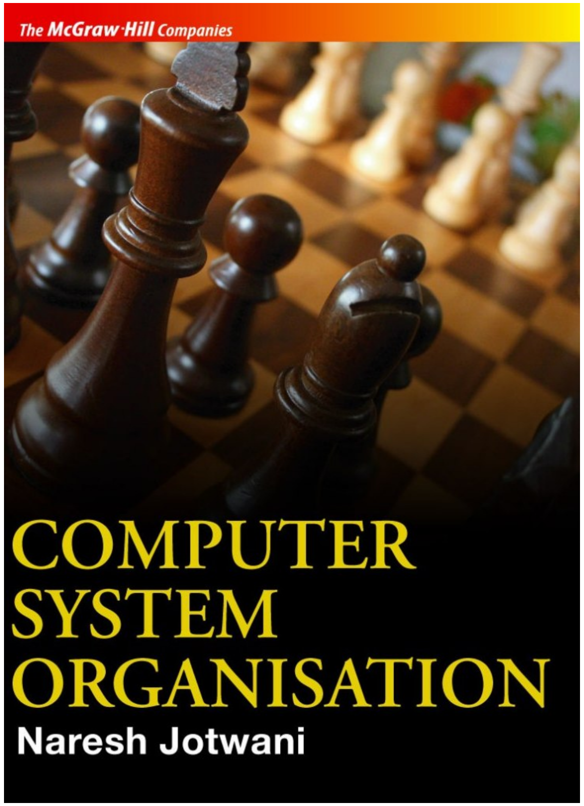
McGraw-Hill Education

COMPUTER SYSTEM ORGANISATION

Naresh Jotwani

PowerPoint Slides



CHAPTER 12

SECONDARY STORAGE AND OTHER DEVICES

Introduction

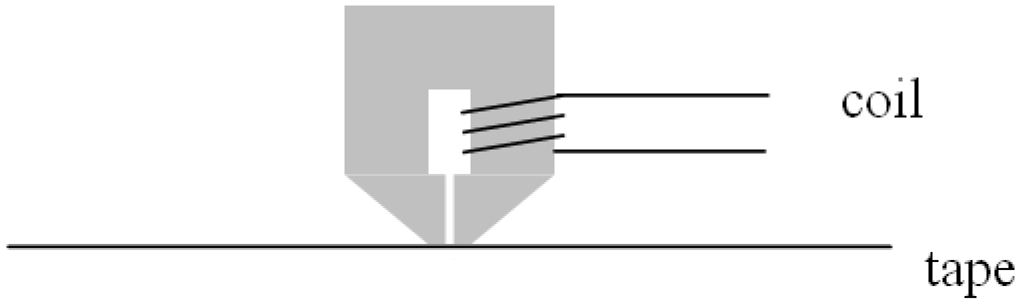
- For the last several decades, *magnetic disks* have been used to provide secondary storage.

- For longer term storage – e.g. *archival storage*, *backup*, and *distribution* – data must be stored on *removable media*, such as *optical disks*.

- Other useful devices typically available on a computer system are – *clock*, *printer*, *speaker* & *microphone* pair, and a network connection through wired and/or wireless LAN.

Magnetic Disks

- Magnetic media have been used for non-volatile storage since the 1950s.
- The *magnetic tape* in a tape recorder stores a piece of music because the pattern of magnetization along the tape mimics the sound of music.
- Tape is magnetized during the *record* operation; *playback* means sensing the magnetization pattern and converting it into sound waves.
- A tape recorder is equipped with two *heads* – one for recording and one for playback. Respective operation is performed when the moving tape passes under the appropriate head.



- The record head is a tiny electromagnet coming in contact with the moving tape, as shown above.

- Current through the coil generates a magnetic field corresponding to the sound signal being recorded, and this field magnetizes the moving tape.

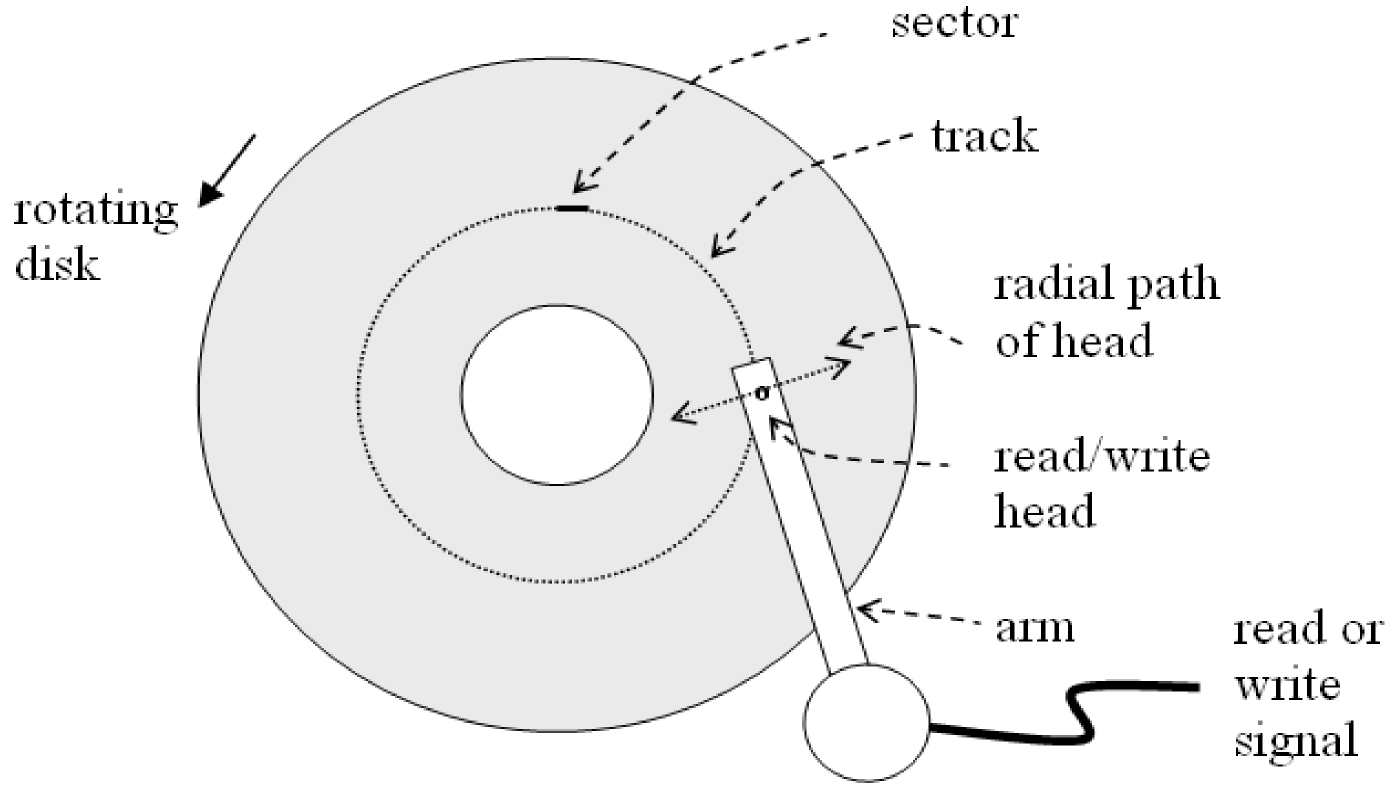
- During record and playback, the two respective principles of physics at work are:
  (a) current in a coil generates a magnetic field, and
  (b) a changing magnetic field induces a voltage across a conducting coil placed in the field.

- A magnetic disk makes use of these same basic principles. But the design of the disk is very different from that of a tape.

- The magnetic disk is a *digital* storage medium – i.e. it stores only ‘1’s and ‘0’s, using two directions of magnetization to represent binary values.

- Magnetic storage medium is coated on a rigid disk, which rotates around its spindle at a high speed.

- The combined *read/write* head of the magnetic disk is located on an arm which can move radially in and out over the surface of the disk.



- Figure shows top view of a magnetic disk. *Disk arm* on the right can be moved in and out; *read/write* head is located on this arm.

- For read or write operation, the *read/write* head is brought to the required radial position with accuracy of the order of a micron (10^{-3} mm).

- As the disk rotates, one *track* on the disk surface moves under the head in one revolution of the disk. Data is stored longitudinally along each such track.

- Each track is divided length-wise into many *sectors*, unit of data transfer between disk and main memory. One sector is of the order of 512 to 4096 bytes.

  This arrangement yields major benefits, e.g.:

  (i) Error detection and correction techniques are more effective on sector-size blocks of data, and

  (ii) Better utilization of disk space can be achieved by dividing tracks into the smaller units of sectors

Example:

- A magnetic disk has radial track density of 10000 tracks/cm, i.e. one track per micron.
- Disk has radial operating distance of 2.4 cm. The total number of tracks on one surface is therefore 10000 x 2.4 = 24000.
- Length of a track increases as we move radially out. Assume the <u>average</u> length of a track is 15 cm, and one sector occupies 1 mm of length along the track. So the total number of sectors on the disk = 24000 x 150 = 3600000.
- If each sector stores 512 bytes = 4096 bits of data, then the bit density of data stored along a sector is 4096/0.1 = 40960 bits/cm.
- Storage capacity of one surface is = 3600000 x 512 = 1800 megabytes. If the disk has six such surfaces stacked together on a common spindle, the storage capacity is 10800 megabytes = 10.8 gigabytes.

- *Latency* of a magnetic disk is the time taken for the required data sector to become available under the *read/write* head on the disk arm.

- This latency has two components:

  (i) *Seek time* - time taken for the head move radially to the required track.

  In the range of 1 to 10 ms, depending on the radial distance the arm has to be moved from its previous position.

  (ii) *Rotational delay* - time taken for the required sector on the track to reach a position under the head, so that it can be read or written.

  On average, this is half the time taken by the disk to make one revolution, since the required sector is equally likely to be at any position along the circular track.

- *Total latency = seek time + rotational delay*.

- Both components of the total latency are random in nature, since they depend on the position of the disk arm prior to executing a read or write command.

- Assume that the disk of the previous example is rotating at an angular speed of 6000 rpm, i.e., 100 revolutions per second.

- i.e. one complete track passes under the head in 10 ms, and contains on average 150 sectors.

- Since each sector has 0.5 kilobytes of data, 150 x 0.5 kilobytes of data pass under the head in 10 ms. So the rate at which data is read from or written into the disk is 150 x 0.5 / (10^{-2}) kilobytes/second = 7.5 megabytes/second -- after the required sector reaches *read/write* head on the disk arm.

- Time of one revolution is 10 ms, and therefore rotational delay is between 0 and 10 ms.

- Assuming that seek time is in the range of 2 to 10 ms, we can say that the total latency is in the range from 2 ms to 20 ms.

- Therefore the average latency is (2+20)/2 = 11 ms.

- Magnetic disk is known as a *direct access* storage device -- because any required data sector on it can be accessed directly, without having to access any intervening sectors.

- In other words, the average latency figure applies to any data sector on the disk, regardless of its location, or the prior position of the read/write head.

- For a read or write operation, the program sends to the disk controller the track number, sector number, and number of sectors to read/ write.

- Tracks and sectors on the disk are formed by patterns of magnetization along the track and between tracks, while *formatting* the disk.

- Sensor electronics associated with the read/write head is capable of locating the required track and sector by its track number and sector number.

- Some fraction of space along each track is used for recording sector numbers, inter-sector gaps etc.

- Usable data in each sector excludes such overhead, and it also excludes additional bits stored with each sector for error detection and correction.

- In a very small package, magnetic disks store many gigabytes of data, and make it available to programs with small latency and high transfer rate.

- Large websites today utilize hundreds of disk drives and high-speed interconnections.

- But there are certain important requirements for which the magnetic disk is not suited:
  - Magnetic disk is a *non-removable storage medium*.
  - Therefore it is not useful for *long term backup*, *archival storage*, and *distribution* of data, multi-media content and software.

- Today *optical disk* and *flash memory* provide far more suitable media for such requirements.

- **Optical disk** makes use of a beam of laser light.

  ‘1’ and ‘0’ are distinguished by the way in which the beam is reflected back by a shiny metallic coating on a transparent polycarbonate disk.

  – The beam is focused on an area of disk which is much less than a micron (10^{-3} mm) across.

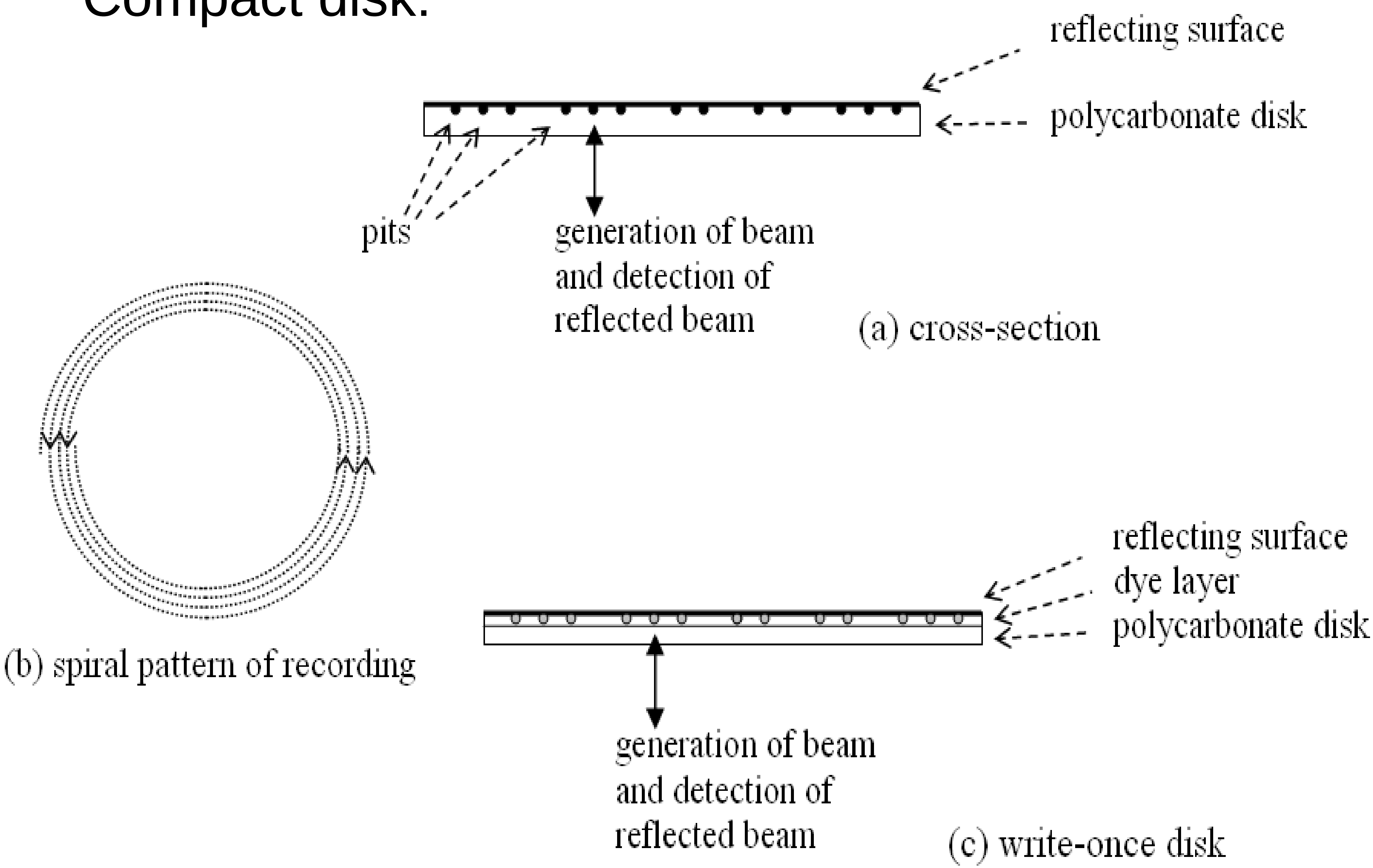
- A *compact disk* (CD) has tiny ‘pits’ on the inside of its reflecting surface.

  – Pits and *lands* (flat spots without pits) reflect the laser beam in different manner.

  – As the disk rotates, the serial pattern of pits and lands moves over the electronic circuit which generates the beam and detects its reflection. [See part (a) of the next figure.]

  – The circuit generates the electronic ‘1’s and ‘0’s which are sent to the computer system.

Compact disk:



- Part (b) of the figure shows the *spiral* pattern of recorded data on the surface of an optical disk. The pattern is that of a single long spiral running over the entire recording surface.

- *Seek times* of the CD drive are much longer than that of magnetic disks.

- The disk rotates about its axis at a variable angular speed – such that the linear speed of the pits and lands over the read head remains constant, and thus also the rate at which ‘1’s and ‘0’s are read.

- Such CDs can be ‘stamped out’ from a master disk at very little cost per piece. Therefore the *read only* CD is widely used for distribution of music, software, directories, and so on.

- Use of laser beams allow very high data storage density to be achieved on optical disks.

- An optical disk has outer diameter of 120 mm. Assuming 8000 square mm of storage area, and storage density of 1 bit per square micron, the total number of bits recordable is 8000×10^6 = 8000 megabits.

- This number includes data bits as well as error correction bits, header bits, etc. Assuming that 50% of all bits are data bits, the data storage capacity of such a disk is 4000 megabits, or 500 megabytes.

- With these numbers, the total length of the spiral of pits and lands on the disk can be roughly estimated as (8000×10^6 bits) x (10^{-6} meters/bit) = **8 km**.

- The speed of data transfer to/from device depends both on storage density and rotation speed.

- For mechanical reasons, a removable rotating storage device must rotate much more slowly than a non-removable device. The speed of a non-removable magnetic disk may be 7200 rpm, but that of a CD is limited to a few hundred rpm.

- After CDs achieved widespread use, the need was felt to produce *writable* versions of optical disks.

- For this, the laser beam in the optical disk drive must alter the reflectivity of a tiny spot on the disk – this spot taking the place of the ‘pit’ seen above.

- In *write-once* optical disks, this is achieved by using a layer of a chemical dye to control the reflectivity of a spot on the reflector surface, as shown in part (c) above.

- When initially coated on the layer, the dye is opaque to light, but a laser beam with slightly higher power can break up dye molecules and make a tiny spot trans-parent rather than opaque.

- Pattern of opaque and transparent spots corresponds to the mechanical pits and lands seen in part (a).

- For reading the disk, the laser beam is given less power, so that it does not break up the dye molecules.

- Dye-based disk allows only one write operation on the disk, because once-broken dye molecules at a spot cannot be regenerated.

- To allow multiple write operations on an optical disk, a different technique is used.
  - Spots on a layer of a specially designed metal alloy can be switched between crystalline and amorphous state, by the application of appropriate amount of heat through a laser beam.
  - Optical properties of the material are different in its crystalline and amorphous states.
  - A layer of such material on the underside of the reflector on the disk thus allows ‘1’s and ‘0’s to be written in the form of tiny spots with different effective reflectivity.

- Improvements in technology have led to the development of *digital video disk* (DVD) storage format, with much higher storage density and therefore much higher storage capacity per disk.

- The speed at which data transfer takes place between the disk and main memory has increased over the years, to keep up with higher demands such as those of video applications.

- On drives, these speeds are indicated by factors such as 32x, which means that the drive transfers data 32 times faster than the original CD rate (1x) of approximately 150 kilobytes per second.

Flash Memories

- Semiconductor memory based on transistors switching ‘on’ and ‘off’ is *volatile*, since transistors cannot operate without power supply.

- But there has always been need for *non-volatile* semiconductor memory and *read-only* memory.
  – Non-volatile semiconductor memory allows for long-term storage of data, with the advantage over disks of not requiring any moving parts.
  – Read-only memory (ROM) protects programs and data against accidental deletion and corruption.

- The earliest ROMs could only be ‘programmed’ at the time of their manufacture – since that involved making physical changes in the array-like circuits of memories.

- This is a handicap, because computer companies need to program such memories at the final stage of assembly of computer systems.

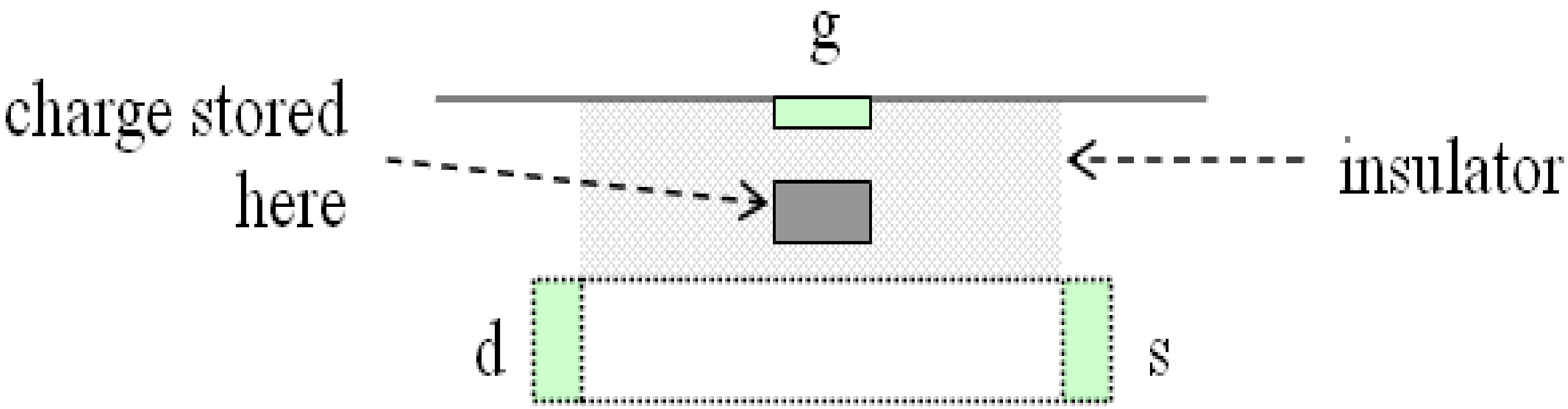
- *Programmable* read-only memories (PROMs) allowed computer companies to program the memories as needed.
  - PROMs are non-volatile and, after one round of programming, they became read-only.
  - EPROMs can be *erased* by using UV light, to become ready for another round of programming
  - *Electrically erasable* PROMs (EEPROMs) can be erased by applying higher voltages to the memory circuits.
  - EEPROMs satisfy the requirements of non-volatility as well as protection against accidental deletion or corruption.

- Development of flash memories is a follow-up on EEPROM technology, with two important differences:

  (i) Far greater density and capacity of storage is now possible, because of VLSI technology

  (ii) Flash memory can also provide secondary storage, rather than a part of main memory

- Flash memories are designed with the block-oriented format more suitable for secondary storage, and not the byte-addressable format used with main memories.



- Figure shows the principle of operation of a semiconductor single-bit cell of an EEPROM and flash memory:
  - The darker rectangle in the centre is a kind of 'metallic island', which is surrounded by insulating material on all sides.
  - Three terminals – labeled 's', 'd' and 'g' in the figure – are available for the application of appropriate voltages to the device.

- With the right combination of voltages, electric charge can be deposited or removed from the island.
  - In the absence of any voltage on the three terminals, charge remains on the island <u>almost indefinitely</u>, since it is surrounded on all sides by insulating material.
  - Presence or absence of charge indicates ‘1’ or ‘0’, and thus we have non-volatile storage of one bit.

- For sensing (i.e. reading) the state of the cell, a lower voltage is applied to the terminals.
  - This voltage does not alter the charge present on the island, but its presence or absence can be sensed across the two terminals ‘s’ and ‘d’.

- For creating larger amounts of storage, a number of such cells must be connected together.

- Such memories can be made compact by allowing *erase* and *write* operations to take place only on a block of bytes, rather than on individual bytes.

- For their use as secondary storage devices, this restriction on flash memories does not create any difficulty, since secondary storage operations take place a block at a time in any case.

Communication Links and Devices

- Computers should be able to *communicate* – i.e. send and receive data from one another – without the need to physically carry storage media.

- Communication between two computers requires the following:

1. A *communication link* to carry an electrical, electro-magnetic, or optical signal from one computer to the other,
2. Devices – also known as *adaptors* – which translate the user's instructions [to send / receive data] into appropriate operations to be performed over the communication link
3. *Software* which carries out the functions specific to the task of communication – for example, error detection and correction.

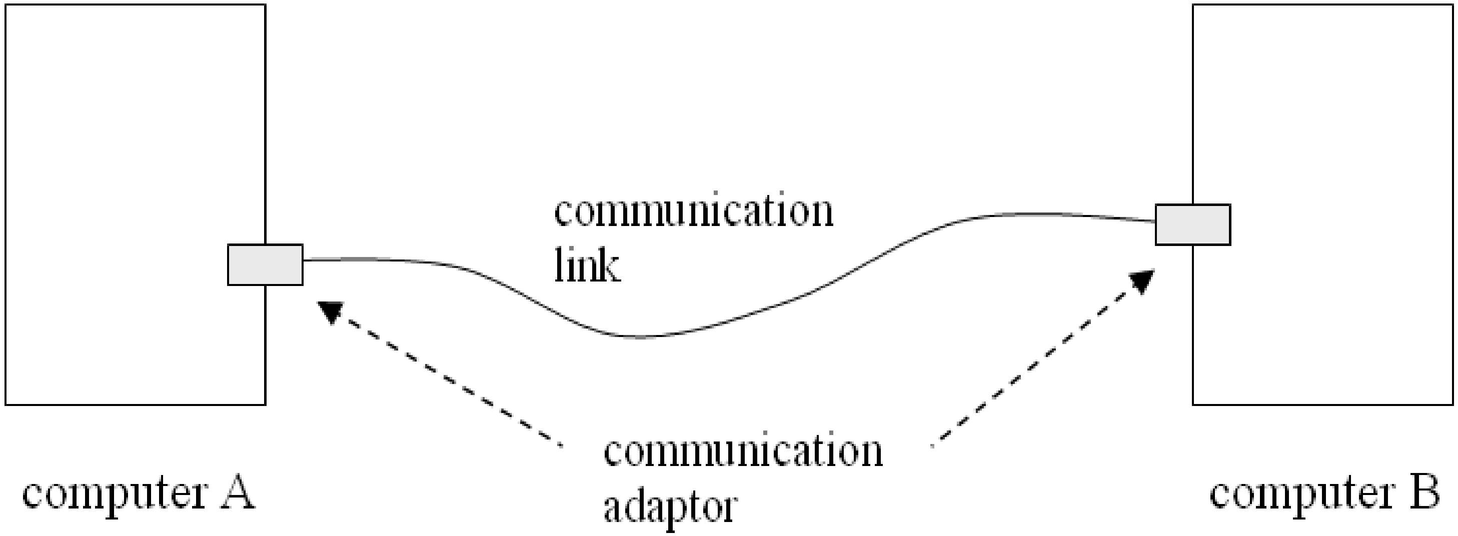


Figure shows a schematic diagram of such an arrangement between two computers.

Send is analogous to an output operation, and *receive* is analogous to an input operation.

- A communication link capable of carrying data in one direction, but <u>not</u> in the other direction, is known as a *simplex* link.
- If the link can carry data in either direction, but only in one direction at a time, it is a *half-duplex* link.
- A *full-duplex* link is one which can carry data in both directions simultaneously.
- Process of converting binary data into signals for a communication link is known as *modulation*; and the converse process of converting signals on the link back into binary data is known as *demodulation*.
-
  - The device which connects computers to telephone lines is known as *<u>mo</u>dulator-<u>dem</u>odulator*, or *modem*. Such modems can send / receive data at the speed of 56 kilobits per second (kbps).

- A modem can be used to connect a home PC to other computers, using the simple telephone line to connect to the Internet.

- The many inter-related factors driving the growth in computer networks may be listed as:

  - higher processing speeds and memory capacities of computers,
  - steadily decreasing costs,
  - increasing numbers of computers in use around the world, and
  - the vast range of applications for which computers are used.

- Today speeds as high as several megabits per second are possible over communication links.
- Computers and different types of communication links combine to form *computer networks*. The Internet can be thought of a network of networks.
- Based on the technology used and physical reach of the network, we may have a *local area network* (LAN), a *wide area network*, a *wireless network*, and so on.
- A typical computer system today has provision for several common types of communication links:
  - a modem for use with a telephone,
  - a LAN & wireless LAN connection, and
  - the general-purpose serial link known as the *universal serial bus* (USB).

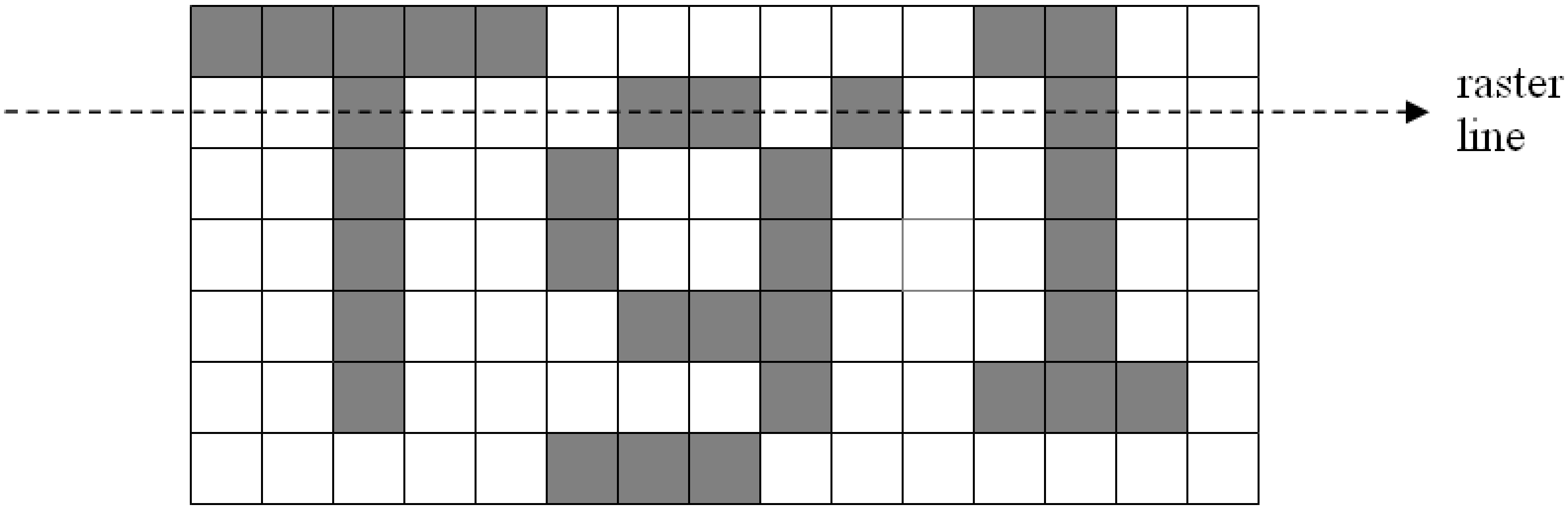
- Typical functions performed by a communication adaptor are:
  - (i) Buffering of data – to handle speed mismatches between the computer and the link, without loss of data,
  - (ii) Modulation and demodulation (described above),
  - (iii) Error detection and correction – Communication links are prone to data transmission errors; hence some error detection and/or correction is provided in adaptors.
  - (iv) Book-keeping – i.e. keeping track of data which has or has not been correctly sent and received.
  - (v) Link control – functions specific to the communication link must also be performed by the controller

- A communication adaptor must also perform data transfers to and from main memory, a byte or a word at a time, using either programmed I/O or direct memory access.

Printers

- Used to transfer electronic data from computer onto paper, by creating tiny dots of ink on paper according to the required characters and/or graphics.

- Figure illustrates how the three characters '`Tgl`' can be created using a *dot matrix* of 7 rows and 5 columns for each character.



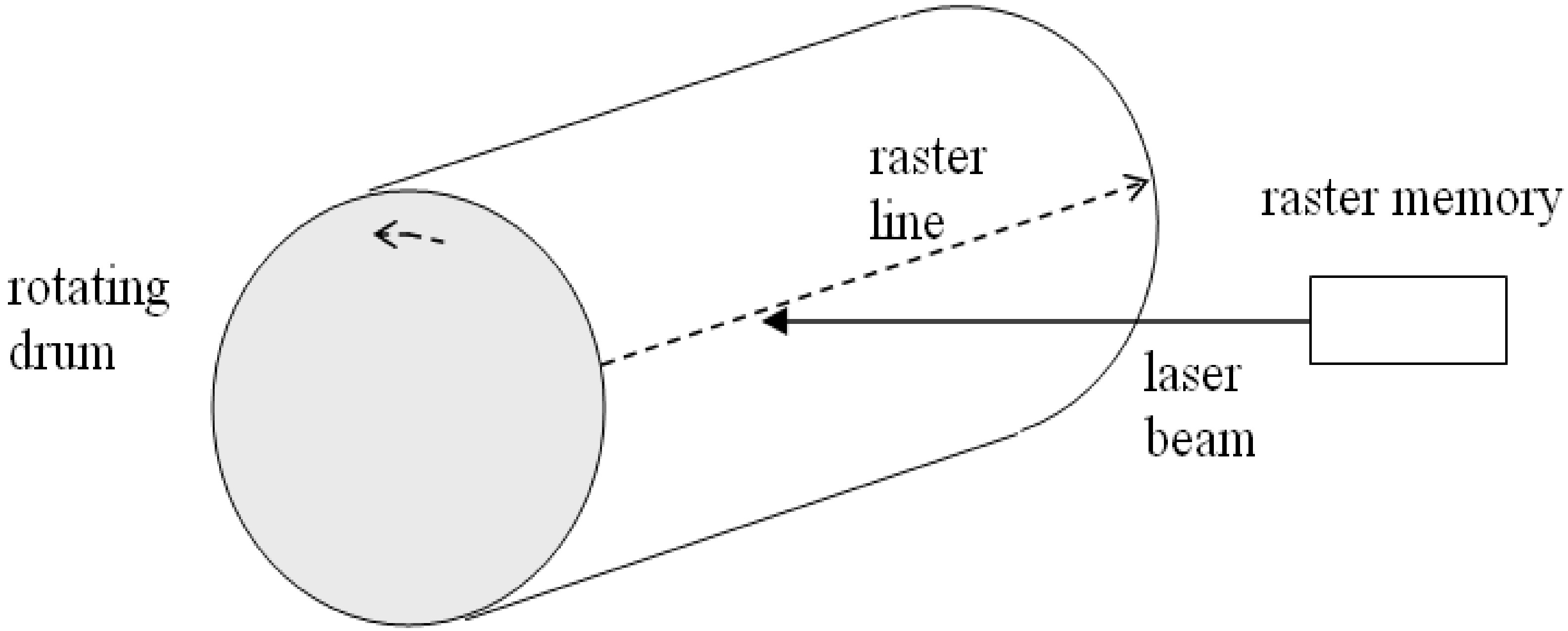
Laser printer

- Prints a page after creating in its internal memory a dot-by-dot image of the entire page. The dots making up one row of the page image make up one *raster line* (or *scan line*).

- This internal memory of the printer is known as *raster memory*. A laser beam transfers page image from raster memory to a rotating photosensitive drum.

- The laser beam determines – at each dot position on the drum – whether or not there is an electric charge at the dot position.

- The presence or absence of electric charge at each dot position determines whether or not the dot will be inked by the ‘toner’.



- The way in which the laser beam converts the binary data of the raster image into dots of charge on the drum is illustrated here.
- In effect, the laser beam ‘writes’ the page on the drum – one *raster line* at a time – in the form of dots of electric charge.
- *‘Toner’* is a very fine powder of plastic and ink.

- As the drum rotates, toner particles are attracted to the charged dots on the drum. Therefore dots of toner particles on the drum correspond to the image to be printed.
- Image is transferred to paper using physical contact between drum and paper. With heat and pressure, toner particles are made to fuse with the paper, producing the final printed page.

Dot-matrix and thermal printers

- Typewriters are *impact* devices. A small *hammer*, with its head shaped in mirror-image of character, hits the paper at the required spot.
- Inked ribbon intervenes between striking hammer and paper; force of impact transfers ink from ribbon to paper, to print character on paper.

- In a dot-matrix printer, the impact is made by tiny dot-shaped hammers. Imagine tiny dot-shaped hammers – actually pins of stiff wire – vertically arranged in the shape of one column of a dot matrix.

- The pins ride on the *print head*. As the head moves left to right on its carriage, the correct pins for each column hit the inked ribbon with the paper behind it, producing the image required.

- A thermal printer uses tiny spots of heat to produce patterns of dots on a special 'thermal' paper.

- Dot-matrix and thermal printers can accept *continuous stationery*, whereas a laser printer prints only on cut sheets, e.g. A4 sized sheets.

- An impact printer can produce multiple copies using multi-part stationery, which has intervening sheets of carbon paper. *Form length* of the stationery can be defined by the user.
  - Receipts from point-of-sale terminals and credit card transactions are often printed on continuous rolls of thermal paper, with a knife-edge to easily tear off a printed slip.
- *Ink-jet printers*
- They combine to some extent the advantages of print quality and lower cost.
  - Mechanism of printing is a ‘print head’ which ejects microscopic droplets of liquid ink towards the paper, without coming into contact with it.
  - Ink droplets are produced by applying heat or pressure to a chamber of ink behind the print-head.

Clock

- OS and running programs perform several crucial functions which need the functionality of a clock, e.g:
  - Time of day is to be shown in a corner of the display screen,
  - A program is to be run for fifty milliseconds at a time, and
  - A program needs to auto-save a file every ten minutes.

- Computer systems are provided with a real-time clock, basically an I/O device which performs clock and timing functions. Examples of its functions are:

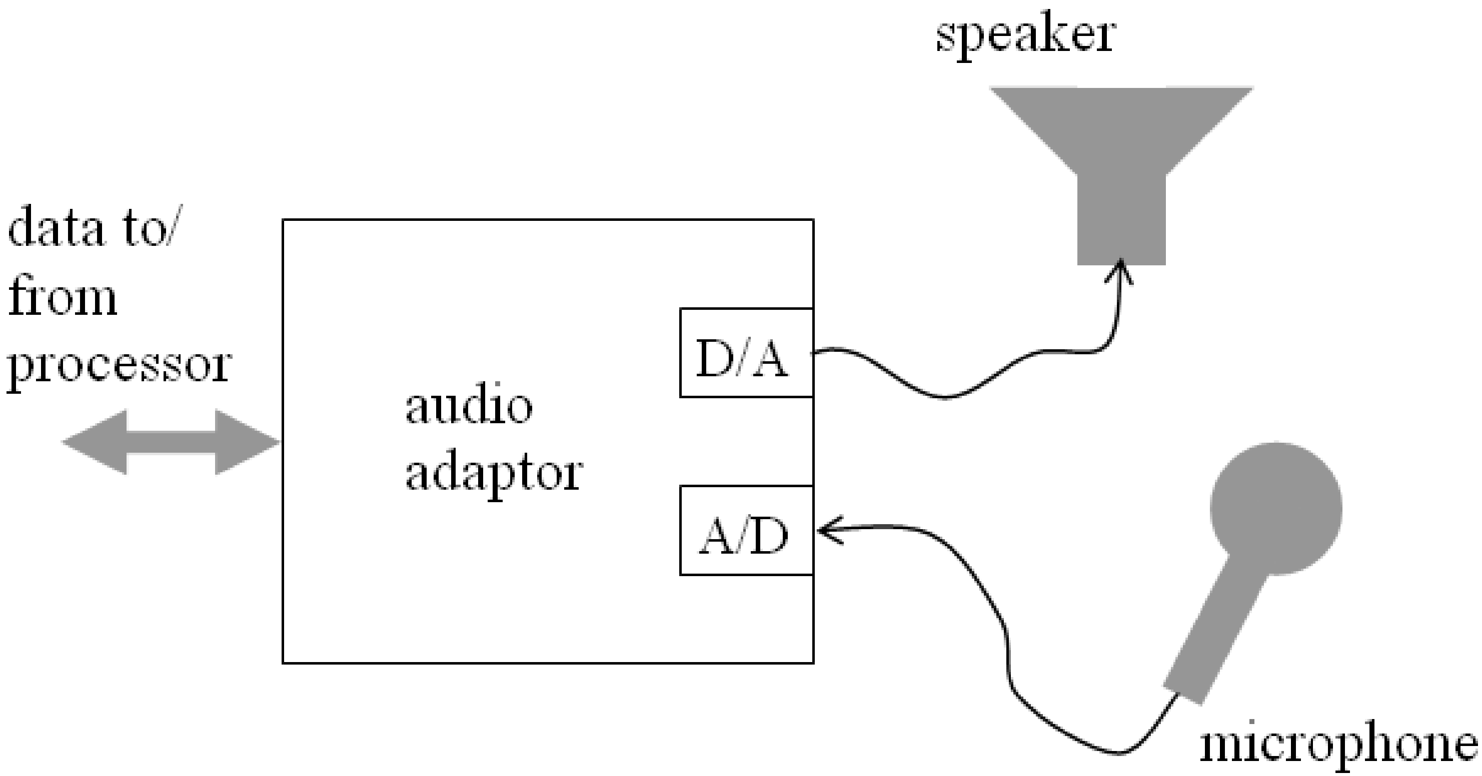
  setting time and date, reading time and date, generating an interrupt after a certain time, generating periodic timer interrupts, etc.

- Within a computer system, a small battery is provided with the real-time clock, to retain information in the absence of power.

Speaker and microphone:

- A microphone converts sound waves into an analog electronic signal which mimics closely the variations of the sound wave.
- A speaker converts an analog electronic signal into corresponding sound waves.
- Internally, the computer represents sound in *digital* form. In any such digital system, audio information is stored in binary.
- Depending on the quality of sound required, one second of sound may require storage varying from a few thousand to several tens of thousands of bits.

Digital audio systems are also designed using processor and memory; techniques used for audio input & output are basically the same as those used in computers.

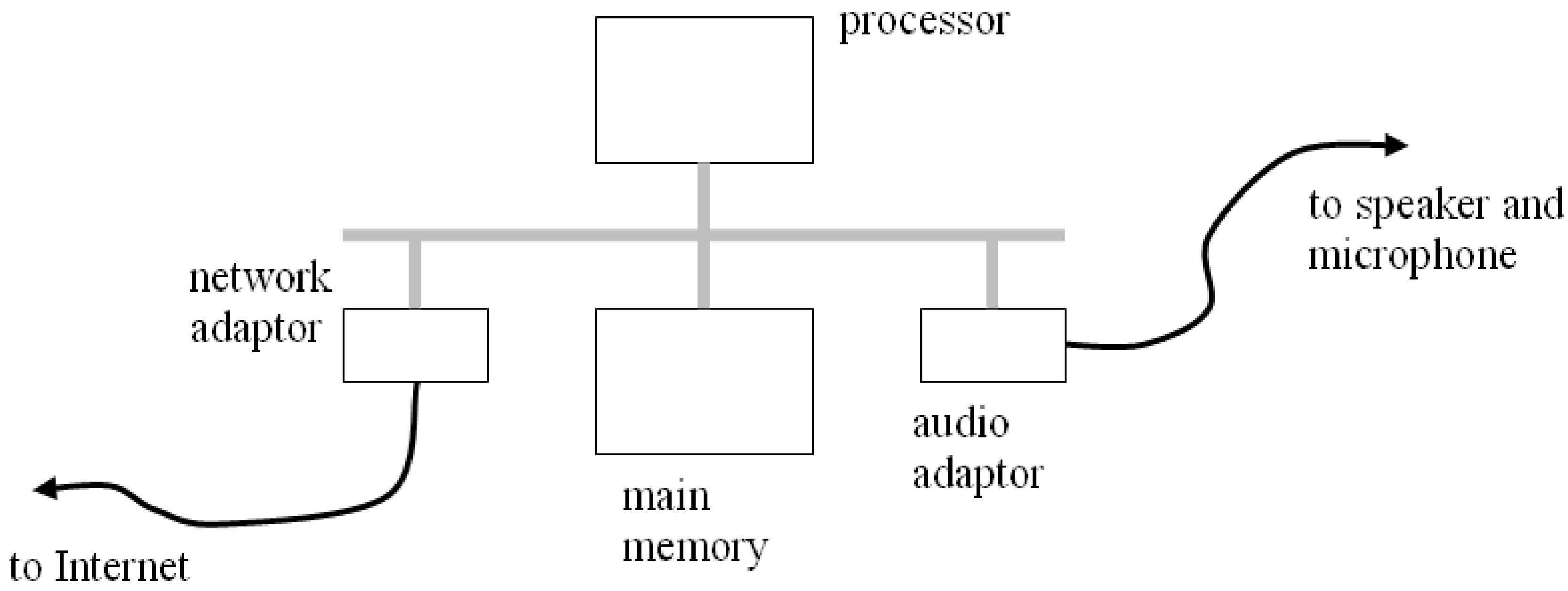


- Audio input & output requires a specialized device controller – or *adaptor* – on the system. Binary audio data is transferred between memory and the adaptor over the system data bus.

- Buffer memory is provided on the adaptor to ensure that the process of recording and playback proceeds at its required rate.

- For playback, digital audio signal is converted into analog form using an electronic circuit known as *digital to analog converter* (shown as D/A in the figure). An amplifier (not shown in the figure) regulates the speaker volume.

- For recording, analog signal from the microphone – or another source of sound – is converted to digital form using an *analog to digital converter* (shown as A/D).



Sending and receiving audio across the Internet:

- Audio bit-stream is received from the network into main memory, and then sent to audio adaptor, where it is converted to analog form; analog signal goes to the speaker.
- In the other direction, analog input from the microphone is digitized in audio adaptor, transferred to main memory, and sent out over the network.
- If any one element in the system cannot keep up with the required data rate, the quality of voice suffers.

Summary

We studied principles of operation of:

- Magnetic disk
- Optical disk (various types)
- Flash memory
- Communication links and devices
- Printers
- Clock, speaker and microphone