UNIT-IV –PART-II

Memory Organization

Memory Hierarchy, Main Memory, Auxiliary memory, Associate memory, Cache Memory, Mapping Techniques, Replacement Algorithms, Write Policies, Memory interleaving.

Memory Hierarchy

Introduction

• The memory unit is an essential component in any digital computer since it is needed for storing programs and data.

• A very small computer with a limited application may be able to fulfill its intended task without the need of additional storage capacity.

• Most general-purpose computers would run more efficiently if they were equipped with additional storage beyond the capacity of the main memory. • There is just not enough space in one memory unit to accommodate all the programs used in a typical computer.

• Moreover, most computer users accumulate and continue to accumulate large amounts of data processing software.

• Not all accumulated information is needed by the processor at the same time. • Therefore, it is more economical to use low-cost storage devices to serve as a backup for storing the information that is not

currently used by the CPU.

Main Memory and Auxiliary Memory

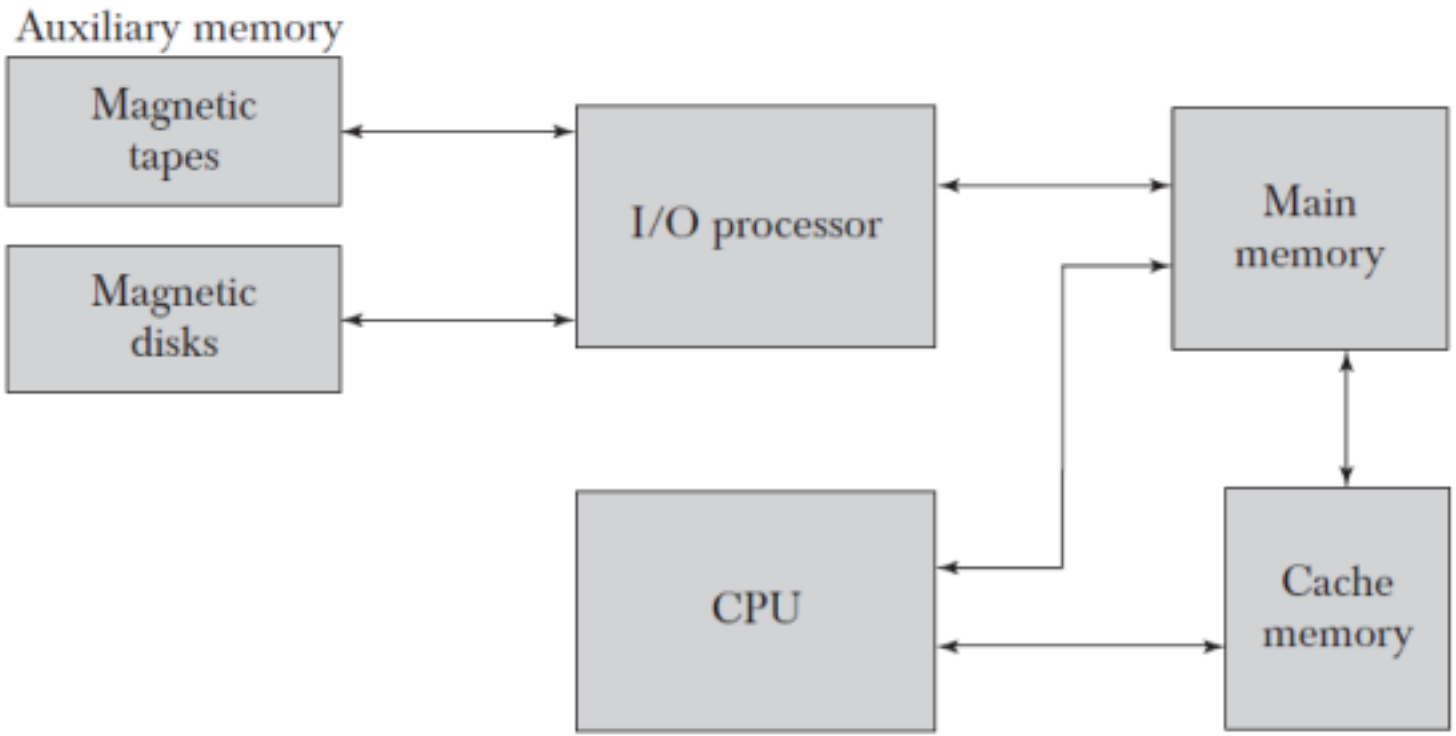
- The memory unit that communicates directly with the CPU is called the **main memory.**
- Devices that provide backup storage are called **auxiliary memory.** • The most common auxiliary memory devices used in computer systems are magnetic disks and tapes.
- They are used for storing system programs, large data files, and other backup information.
- Only programs and data currently needed by the processor reside in

main memory.

- All other information is stored in auxiliary memory and transferred to main memory when needed.

Memory hierarchy in a computer system

The total memory capacity of a computer can be visualized as being a hierarchy of



components.

Memory hierarchy in a computer system

• The memory hierarchy system consists of all storage devices employed in a computer system from the slow but high-capacity auxiliary memory

to a relatively faster main memory, to an even smaller and faster cache memory accessible to the high-speed processing logic.

• While the I/O processor manages data transfers between auxiliary memory and main memory, the cache organization is concerned with the transfer of information between main memory and CPU.

• Thus each is involved with a different level in the memory hierarchy system.

Memory hierarchy in a computer system

• At the bottom of the hierarchy are the relatively slow magnetic tapes used to store removable files.

• Next are the magnetic disks used as backup storage.

• The main memory occupies a central position by being able to communicate directly with the CPU and with auxiliary memory devices through an I/O processor.

• When programs not residing in main memory are needed by the CPU, they are brought in from auxiliary memory.

• Programs not currently needed in main memory are transferred into auxiliary memory to provide space for currently used programs and data.

Cache Memory

• A special very-high-speed memory called a cache is sometimes used to increase the speed of processing by making current programs and data available to the CPU at a rapid rate.

• The cache memory is employed in computer systems to compensate for the speed differential between main memory access time and processor logic.

• CPU logic is usually faster than main memory access time, with the result that processing speed is limited primarily by the speed of main memory.

Cache Memory

• A technique used to compensate for the mismatch in operating speeds is to employ an extremely fast, small cache between the CPU and main memory whose access time is close to processor logic clock cycle time.

• The cache is used for storing segments of programs currently being

executed in the CPU and temporary data frequently needed in the present calculations.

• By making programs and data available at a rapid rate, it is possible to increase the performance rate of the computer.

Reason: for having two or three levels of memory hierarchy

• The reason for having two or three levels of memory hierarchy is economics. •

As the storage capacity of the memory increases, the cost per bit for storing binary information decreases and the access time of the memory becomes longer.

• The auxiliary memory has a large storage capacity, is relatively inexpensive, but has low access speed compared to main memory.

• The cache memory is very small, relatively expensive, and has very high

access speed.

- Thus as the memory access speed increases, so does its relative cost.

• The overall goal of using a memory hierarchy is to obtain the highest possible average access speed while minimizing the total cost of the entire memory system.

Direct Memory Access

- Auxiliary and cache memories are used for different purposes.
- The cache holds those parts of the program and data that are most heavily used, while the auxiliary memory holds those parts that are not presently used by the CPU.

• Moreover, the CPU has direct access to both cache and main memory but not to auxiliary memory.

• **The transfer from auxiliary to main memory is usually done by means of direct memory access of large blocks of data.**

Access Time

• The typical access time ratio between cache and main memory is about 1 to 7.

• For example, a typical cache memory may have an access time of 100 ns, while main memory access time may be 700 ns.

• Auxiliary memory average access time is usually 1000 times that of main memory.

• Block size in auxiliary memory typically ranges from 256 to 2048 words, while cache block size is typically from 1 to 16 words.

Multiprogramming

- Many operating systems are designed to enable the CPU to process a number of independent programs concurrently.
- **This concept, called multiprogramming, refers to the existence of two or more programs in different parts of the memory hierarchy at the same time.**
- In this way it is possible to keep all parts of the computer busy by working with several programs in sequence.
- For example, suppose that a program is being executed in the CPU and

an I/O transfer is required. The CPU initiates the I/O processor to start executing the transfer. This leaves the CPU free to execute another program.

- In a multiprogramming system, when one program is waiting for input or output transfer, there is another program ready to utilize the CPU.

Multiprogramming

- With multiprogramming the need arises for running partial programs, for varying the amount of main memory in use by a given program, and for moving programs around the memory hierarchy.
- Computer programs are sometimes too long to be accommodated in the total space available in main memory.

- Moreover, a computer system uses many programs and all the programs cannot reside in main memory at all times.
- A program with its data normally resides in auxiliary memory.
- When the program or a segment of the program is to be executed, it is transferred to main memory to be executed by the CPU.

Multiprogramming

- Thus one may think of auxiliary memory as containing the totality of information stored in a computer system.
- It is the task of the operating system to maintain in main memory a portion of this information that is currently active.
- The part of the computer system that supervises the flow of

information between auxiliary memory and main memory is called the **memory management system.**

Case Study

Where the OS fits in the computer system?

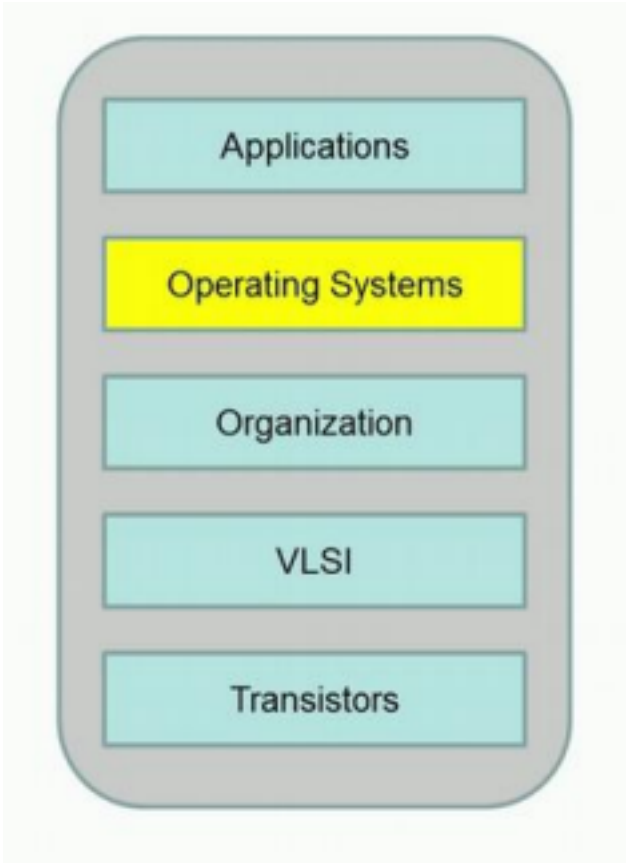
Layers in a Computing System

- **Hardware Abstraction**

  turns hardware into something that applications can use

- **Resource Management**

  manage system's resources



The OS is used for following two purposes:

“The Operating System essentially would manage both the Applications that execute on the Computer as well as it would manage how the Resources are utilized in the system”

A Sample C Program

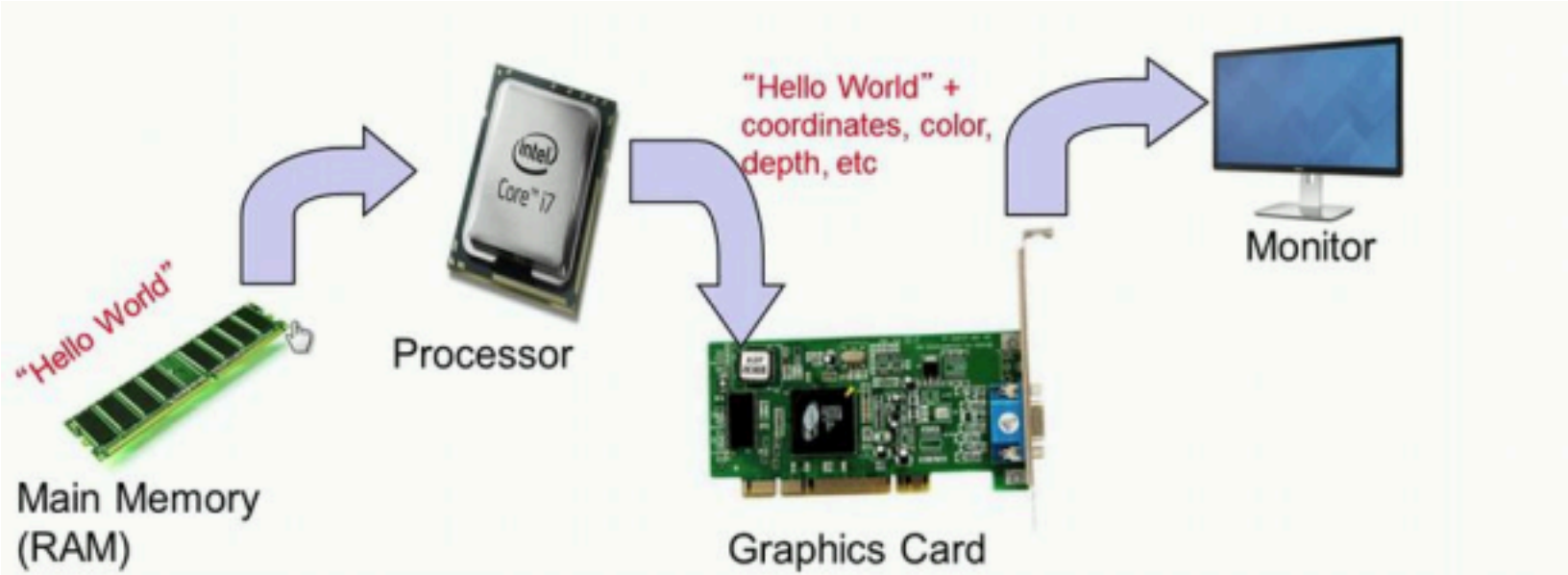
```c
#include <stdio.h>

int main(){
  char str[] = "Hello World\n";
  printf("%s", str);
}
```

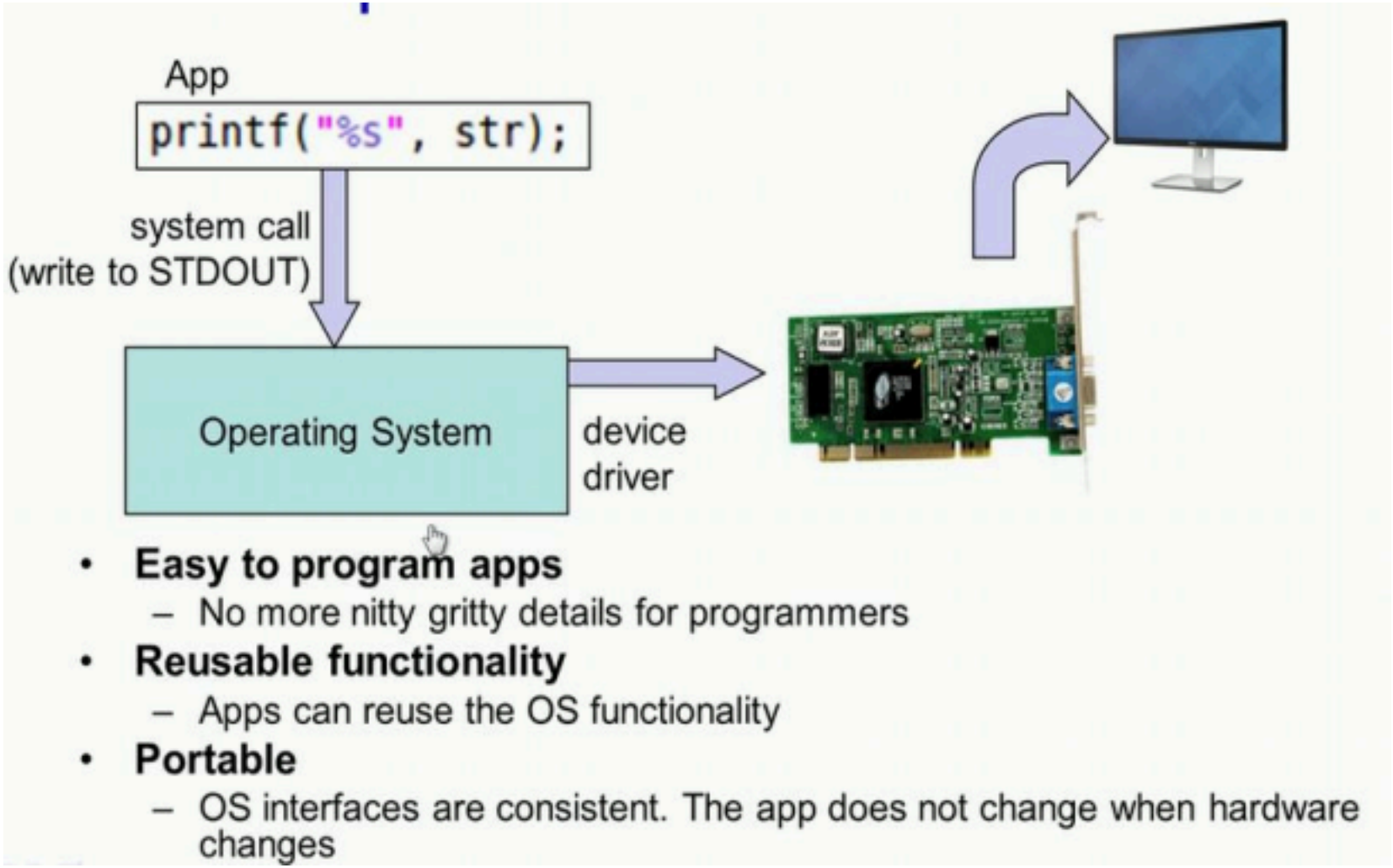
string “Hello World” is stored in memory and it is pointed to by this pointer str[ ]. Now, printf is passed this pointer str and would result in the string being printed on to the monitor.

How exactly is the string displayed on to the monitor?

OS Makes this simple

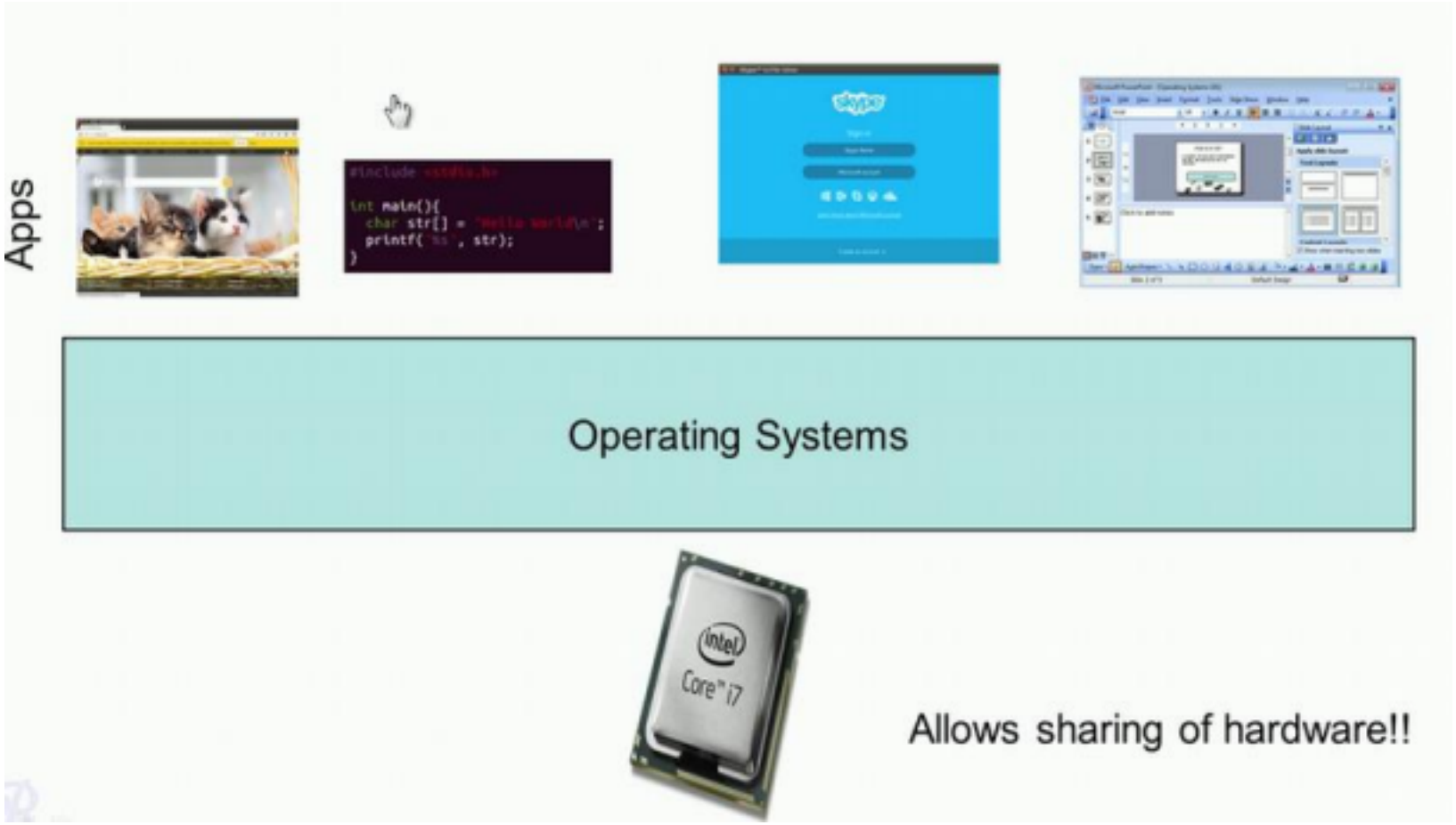


1. 1 Hardware Abstraction

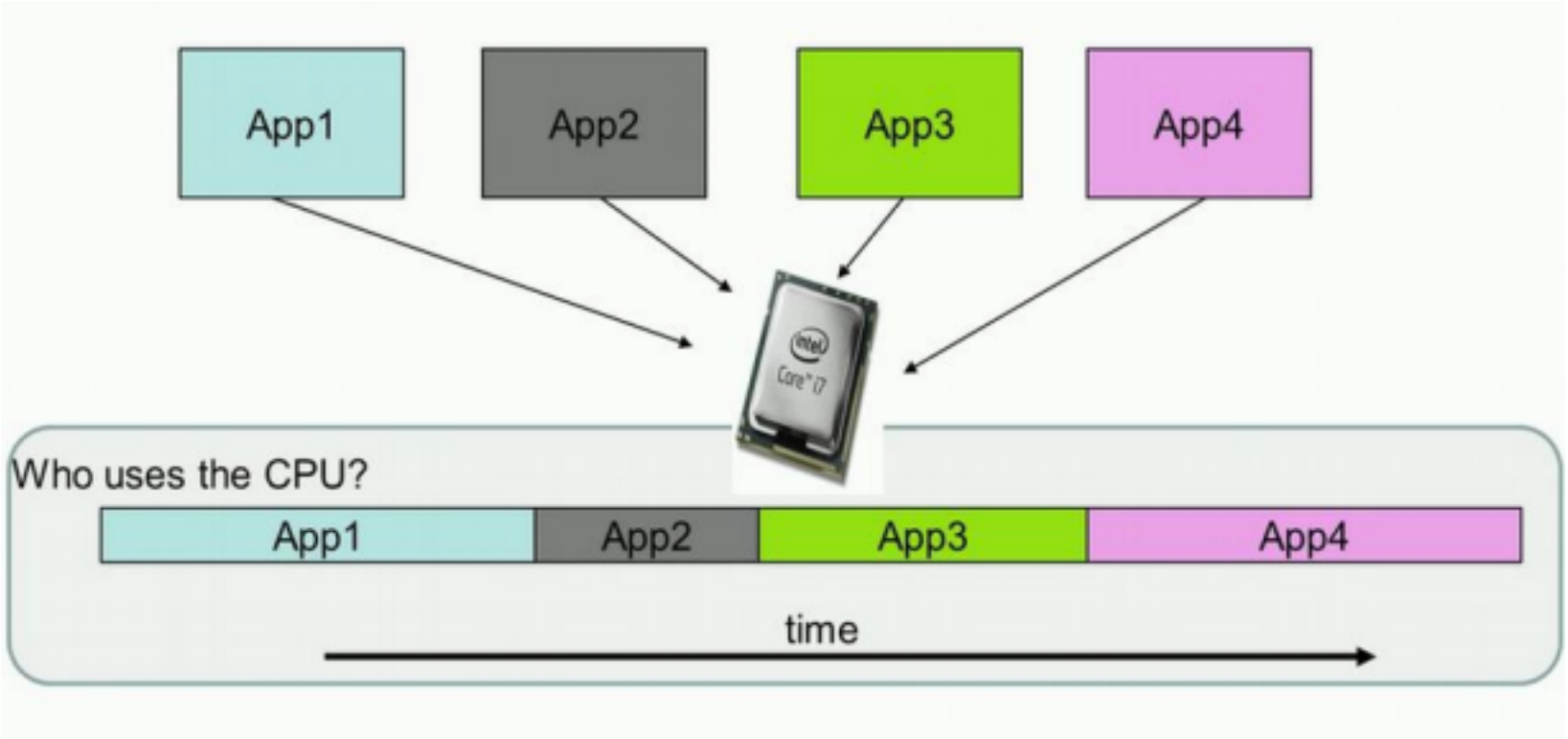


- **Easy to program apps**
  - No more nitty gritty details for programmers
- **Reusable functionality**
  - Apps can reuse the OS functionality
- **Portable**
  - OS interfaces are consistent. The app does not change when hardware changes

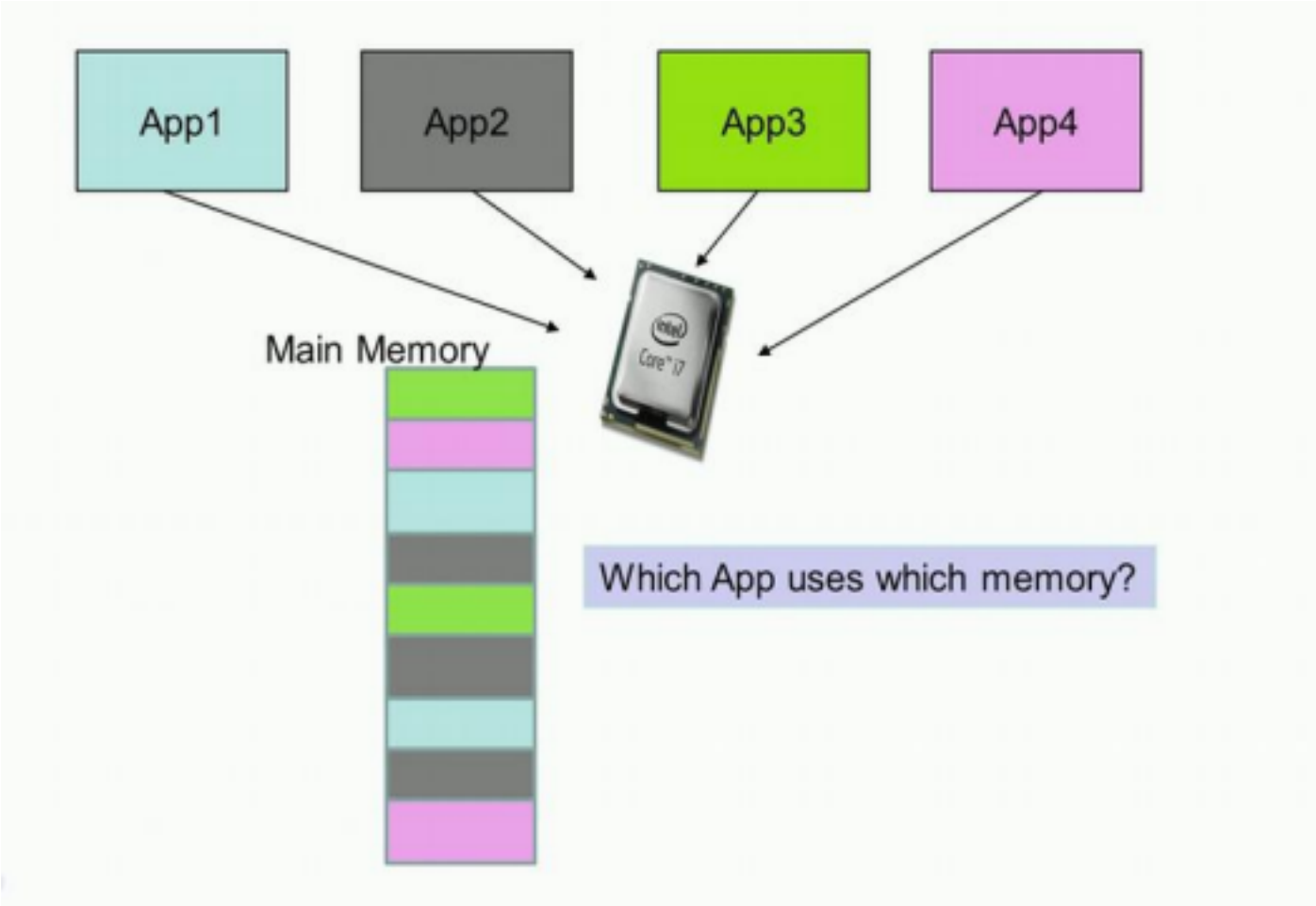
2. Resource Management



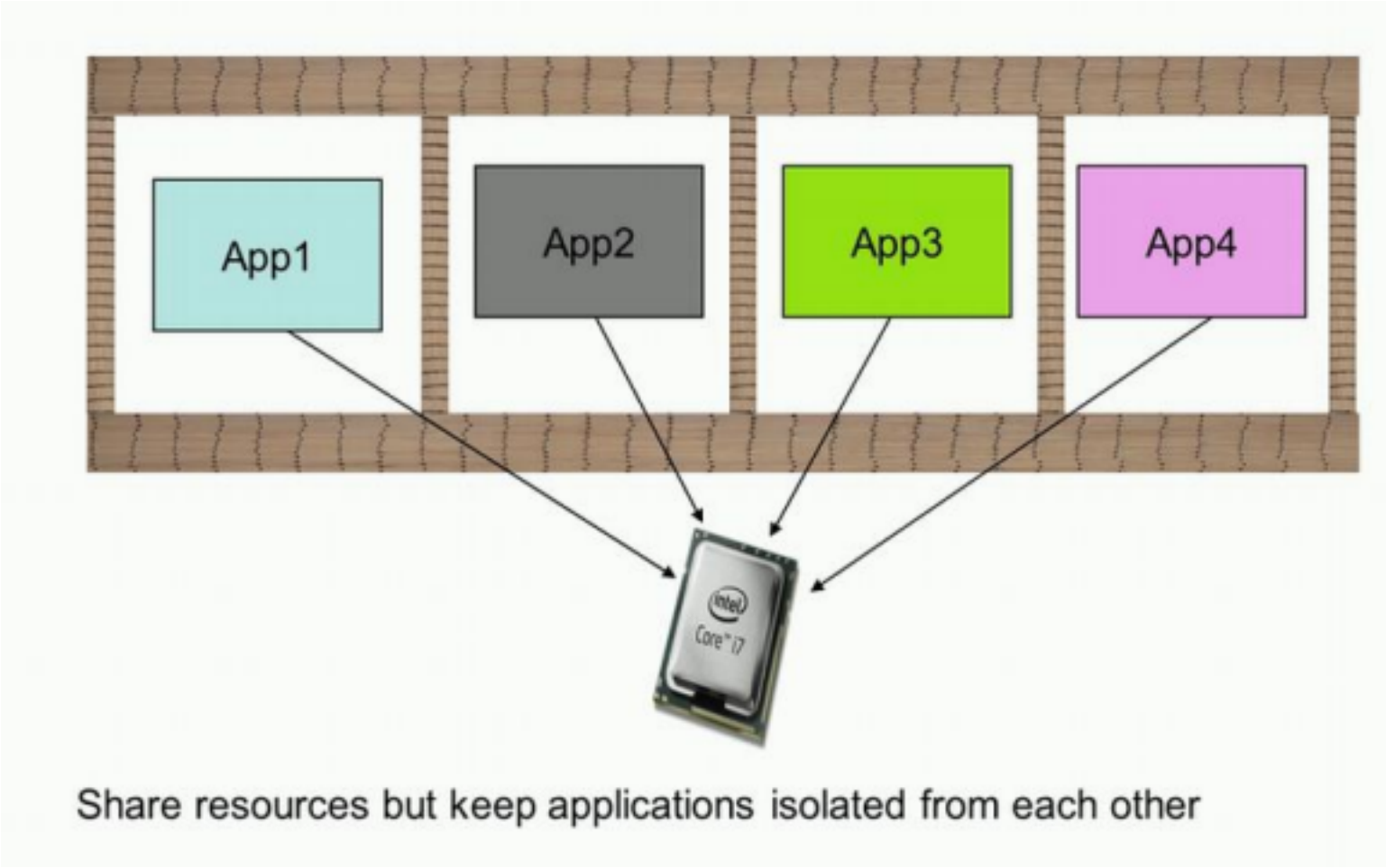
Sharing the CPU: Scheduling



Sharing Memory



Secure Sharing of Resources


App1
App2
App3
App4
Share resources but keep applications isolated from each other

Main Memory

Main Memory

• The main memory is the central storage unit in a computer system.

• It is a relatively large and fast memory used to store programs and data during the computer operation.

• The principal technology used for the main memory is based on semiconductor integrated circuits.

Random-Access Memory (RAM)

• The main memory is the central storage unit in a computer system.

• It is a relatively large and fast memory used to store programs and data

during the computer operation.

- The principal technology used for the main memory is based on semiconductor integrated circuits.
- Integrated circuit RAM chips are available in two possible operating modes, static and dynamic.

Static RAM

- The static RAM consists essentially of internal flip-flops that store the binary information.
- The stored information remains valid as long as power is applied to the unit.

Dynamic RAM

- The dynamic RAM stores the binary information in the form of electric charges that are applied to capacitors.
- The capacitors are provided inside the chip by MOS transistors.
- The stored charge on the capacitors tend to discharge with time and the capacitors must be periodically recharged by refreshing the dynamic memory.
- Refreshing is done by cycling through the words every few milliseconds to restore the decaying charge.

Static RAM Vs Dynamic RAM

- The dynamic RAM offers reduced power consumption and larger

storage capacity in a single memory chip.

- The static RAM is easier to use and has shorter read and write cycles.
- **One of the major applications of the static RAM is in implementing the cache memories.**
- **The dynamic RAMs are used for implementing the main memory.**
- Most of the desktop personnel computer systems are dynamic RAMs with improved performance characteristics such as multibank DRAM, extended dataout DRAM, synchronous DRAM, and Direct RAM bus DRAM.

Read-Only Memory (ROM)

- Most of the main memory in a general-purpose computer is made up of

RAM integrated circuit chips, but a portion of the memory may be constructed with ROM chips.

• Originally, RAM was used to refer to a random access memory, but now it is used to designate a read/write memory to distinguish it from a read-only memory, although ROM is also random access.

• RAM is used for storing the bulk of the programs and data that are subject to change.

• ROM is used for storing programs that are permanently resident in the computer and for tables of constants that do not change in value once the production of the computer is completed.

Bootstrap Loader

- Among other things, the ROM portion of main memory is needed for storing an initial program called a bootstrap loader.
- The bootstrap loader is a program whose function is to start the computer software operating when power is turned on.
- Since RAM is volatile, its contents are destroyed when power is turned off.
- The contents of ROM remain unchanged after power is turned off and on again.

Computer Startup

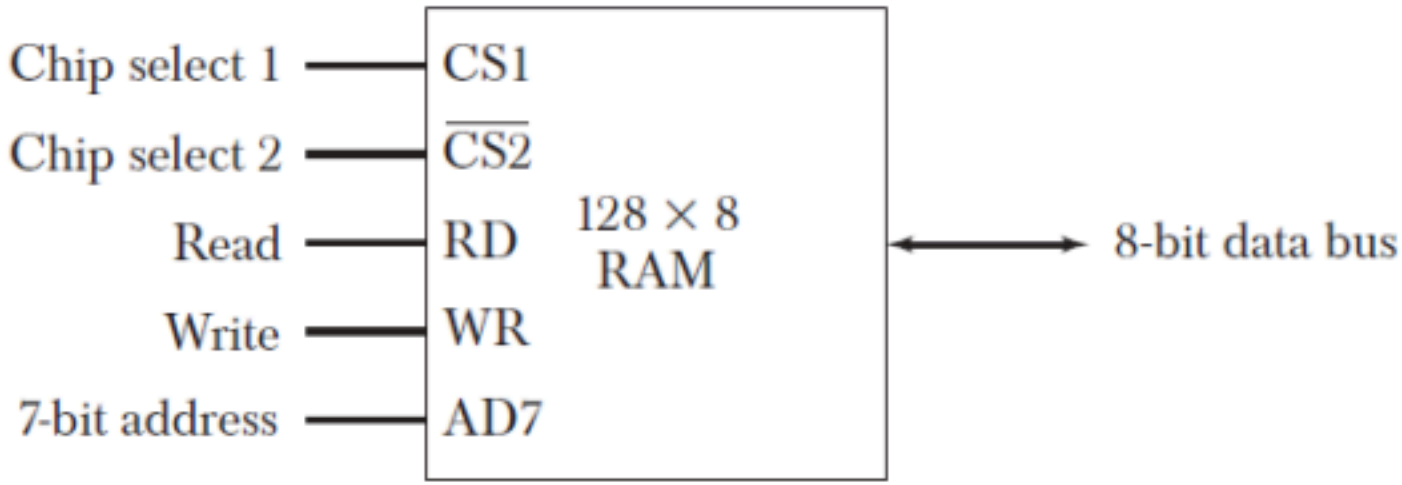
- The startup of a computer consists of turning the power on and starting the execution of an initial program.
- Thus when power is turned on, the hardware of the computer sets the program counter to the first address of the bootstrap loader.
- **The bootstrap program loads a portion of the operating system from disk to main memory and control is then transferred to the operating system, which prepares the computer for general use.**

RAM and ROM chips

RAM and ROM Chips

- RAM and ROM chips are available in a variety of sizes.
- If the memory needed for the computer is larger than the capacity of one chip, it is necessary to combine a number of chips to form the required memory size.
- To demonstrate the chip interconnection, we will show an example of a 1024 x 8 memory constructed with 128 x 8 RAM chips and 512 x 8 ROM chips.

Typical RAM Chip



(a) Block diagram

CS1	$\overline{CS2}$	RD	WR	Memory function	State of data bus
0	0	×	×	Inhibit	High-impedance
0	1	×	×	Inhibit	High-impedance
1	0	0	0	Inhibit	High-impedance
1	0	0	1	Write	Input data to RAM
1	0	1	×	Read	Output data from RAM
1	1	×	×	Inhibit	High-impedance

(b) Function table

Typical RAM Chip

- A RAM chip is better suited for communication with the CPU if it has one or more control inputs that select the chip only when needed. • Another common feature is a bidirectional data bus that allows the transfer of data either from memory to CPU during a read operation, or from CPU to memory during a write operation.
- A bidirectional bus can be constructed with three-state buffers. • A three-state buffer output can be placed in one of three possible states: a signal equivalent to logic 1, a signal equivalent to logic 0, or a high-impedance state.
- The logic 1 and 0 are normal digital signals.

- The high-impedance state behaves like an open circuit, which means that the output does not carry a signal and has no logic significance.

Typical RAM Chip

- The block diagram of a RAM chip is shown in Fig. (a)
- The capacity of the memory is 128 words of eight bits (one byte) per word.
- This requires a 7-bit address and an 8-bit bidirectional data bus. • The read and write inputs specify the memory operation and the two chips select (CS) control inputs are for enabling the chip only when it is selected by the microprocessor.
- The availability of more than one control input to select the chip facilitates the decoding of the address lines when multiple chips are used in the microcomputer.

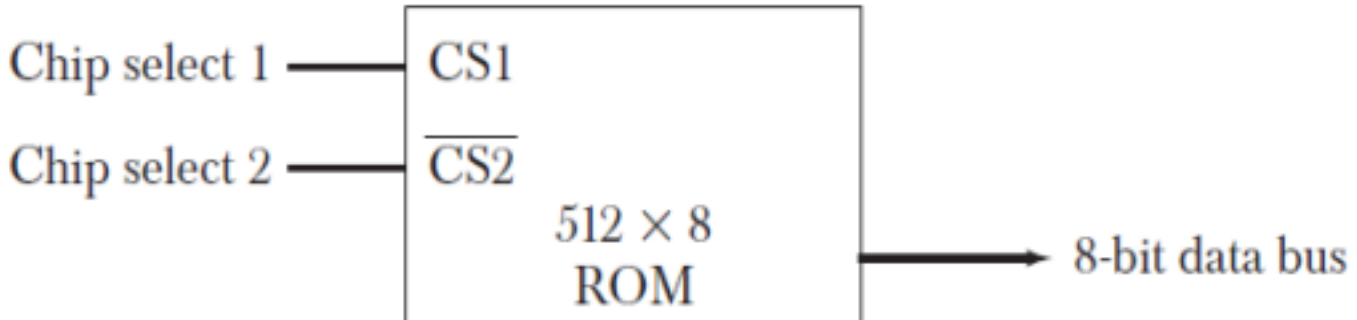
- The read and write inputs are sometimes combined into one line labeled R/W.
- When the chip is selected, the two binary states in this line specify the two operations of read or write.

Typical RAM Chip

- The function table listed in Fig.(b) specifies the operation of the RAM chip. • The unit is in operation only when *CS1 =1* and ����2 =0.
- The bar on top of the second select variable indicates that this input is enabled when it is equal to 0.
- If the chip select inputs are not enabled, or if they are enabled but the read or write inputs are not enabled, the memory is inhibited and its data bus is in a high-impedance state.

- When *CS1 =1* and ����2 =0, the memory can be placed in a write or read mode.
- When the WR input is enabled, the memory stores a byte from the data bus into a location specified by the address input lines.
- When the RD input is enabled, the content of the selected byte is placed into the data bus.
- The RD and WR signals control the memory operation as well as the bus buffers associated with the bidirectional data bus.

Typical ROM Chip



A ROM chip is organized externally in a similar manner.

However, since a ROM can only read, the data

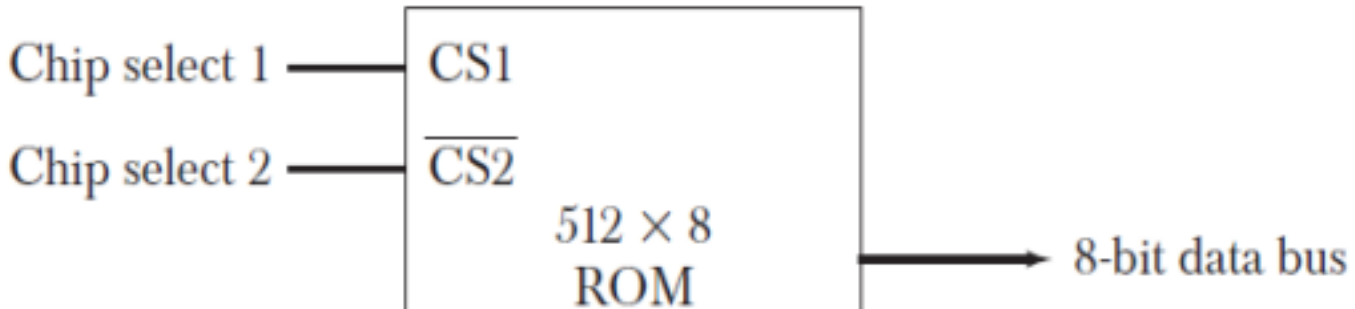
bus can only be in an output mode.

The block diagram of a ROM chip is shown in Fig.

For the same-size chip, it is possible to have more bits of ROM than of RAM, because the internal binary cells in ROM occupy less space than in RAM.

For this reason, the diagram specifies a 512-byte ROM, while the RAM has only 128 bytes.

Typical ROM Chip



The nine address lines in the ROM chip
specify any one of the 512 bytes stored in it.

The two chip select inputs must

be *CS1 =1*
and ����2 =0 for the unit to operate.

Otherwise, the data bus is in a high-impedance state.

There is no need for a read or write control because the unit can only read.

Thus when the chip is enabled by the two select inputs, the byte selected by the address lines appears on the data bus.

Memory Address Map & Memory Connection to CPU

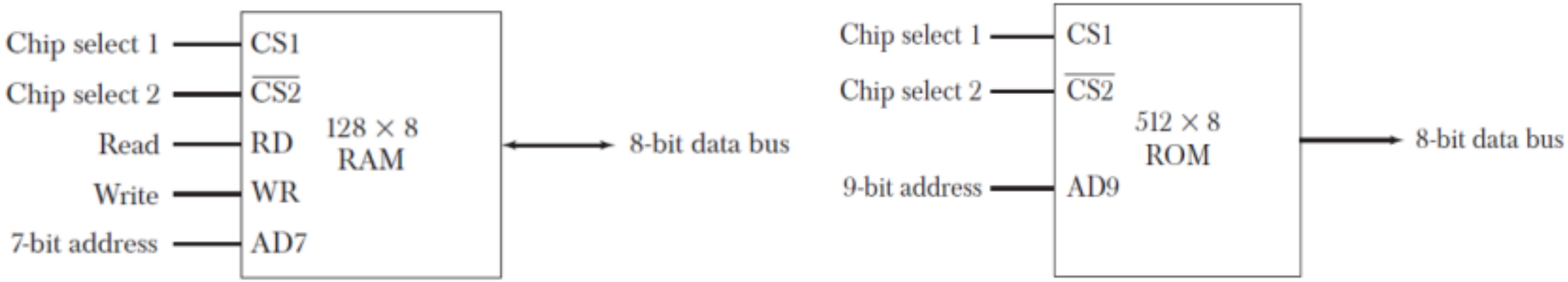
Memory Address Map

- The designer of a computer system must calculate the amount of memory required for the particular application and assign it to either RAM or ROM.
- The interconnection between memory and processor is then established from knowledge of the size of memory needed and the type of RAM and ROM chips available.
- The addressing of memory can be established by means of a table that specifies the memory address assigned to each chip.
- The table, called a memory address map, is a pictorial representation of assigned address space for each chip in the system.

Example

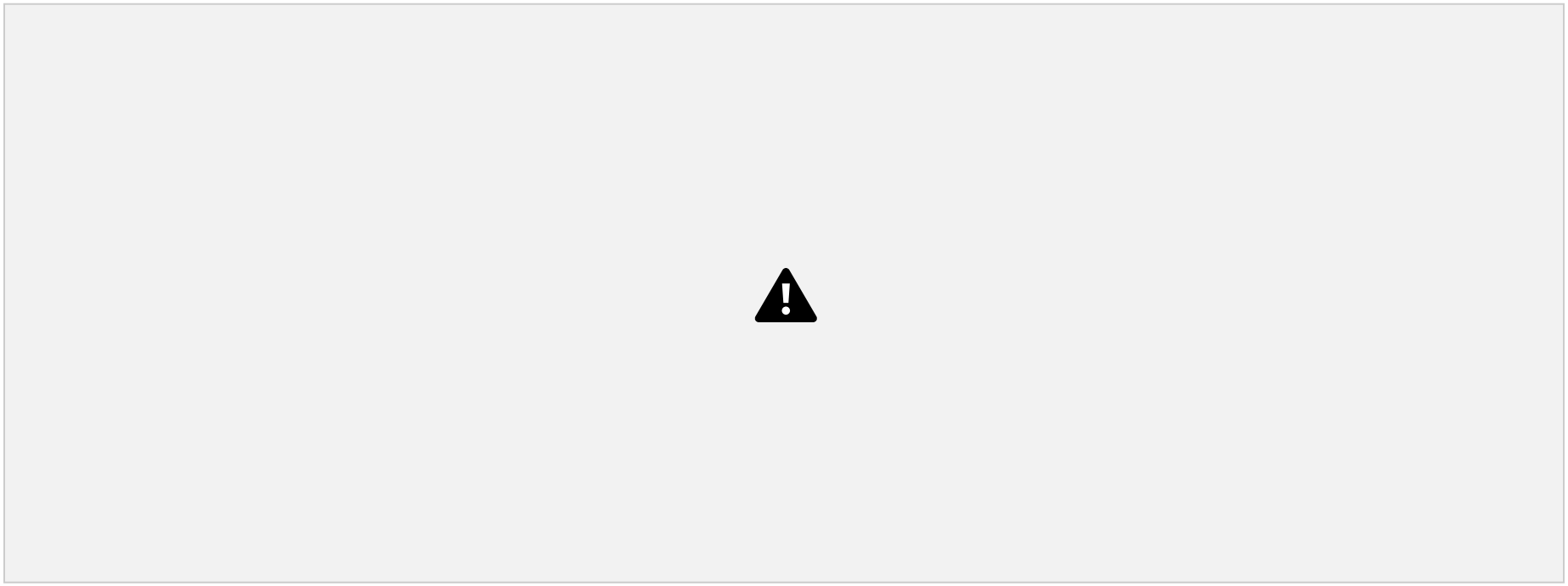
To demonstrate with a particular example, assume that a computer system needs 512 bytes of RAM and 512 bytes of ROM.

The RAM and ROM chips to be used are specified as follows:



Memory Address Map for Microprocomputer

The memory address map for this configuration is shown in Table



(Refer Notes)

Memory connection to the CPU

RAM and ROM chips are connected to a CPU through the data and address buses.

The low-order lines in the address bus select the byte within the chips and other lines in the address bus select a particular chip through its chip select inputs.

The connection of memory chips to the CPU is shown in Fig.

This configuration gives a memory capacity of 512 bytes of RAM and 512 bytes of ROM.

It implements the memory map of Table.

Each RAM receives the seven low-order bits of the address bus to select one of 128 possible bytes.

The particular RAM chip selected is determined from lines 8 and 9 in the address bus. This is done through a 2 x4 decoder whose outputs go to the CS1 inputs in each RAM chip.

Thus, when address lines 8 and 9 are equal to 00, the first RAM chip is selected. When 01, the second RAM chip is selected, and so on. The RD and WR outputs from the microprocessor are applied to the inputs of each RAM chip.

Memory connection to the CPU The

selection between RAM and ROM is achieved through bus line 10.

The RAMs are selected when the bit in this line is 0, and the ROM when the bit is 1.

The other chip select input in the ROM is connected to the RD control line for the ROM chip to be enabled only during a read operation.

Address bus lines 1 to 9 are applied to the input address of ROM without going through the decoder.

This assigns addresses 0 to 511 to RAM and 512 to 1023 to ROM.

The data bus of the ROM has only an output capability, whereas the data bus connected to the RAMs can transfer information in both directions.

Auxiliary Memory

Auxiliary Memory

- The most common auxiliary memory devices used in computer systems are magnetic disks and tapes.
- Other components used, but not as frequently, are magnetic drums, magnetic bubble memory, and optical disks.

• The important characteristics of any device are its access mode, access time, transfer rate, capacity, and cost.

Access Time

• The average time required to reach a storage location in memory and obtain its contents is called the access time.

• In electromechanical devices with moving parts such as disks and tapes, the **access time** consists of a **seek time** required to position the read-write head to a location and a **transfer time** required to transfer data to or from the device.

• Because the seek time is usually much longer than the transfer time, auxiliary storage is organized in **records or blocks**.

• A record is a specified number of **characters or words.**

• Reading or writing is always done on entire records.

• The **transfer rate** is the number of characters or words that the device can transfer per second, after it has been positioned at the beginning of the record.

Magnetic drums and disks

• Magnetic drums and disks are quite similar in operation.

• Both consist of high-speed rotating surfaces coated with a magnetic recording medium.

• The rotating surface of the drum is a cylinder and that of the disk, a round flat plate.

• The recording surface rotates at uniform speed and is not started or stopped during access operations.

• Bits are recorded as magnetic spots on the surface as it passes a stationary mechanism called a write head.

• Stored bits are detected by a change in magnetic field produced by a recorded spot

on the surface as it passes through a read head.

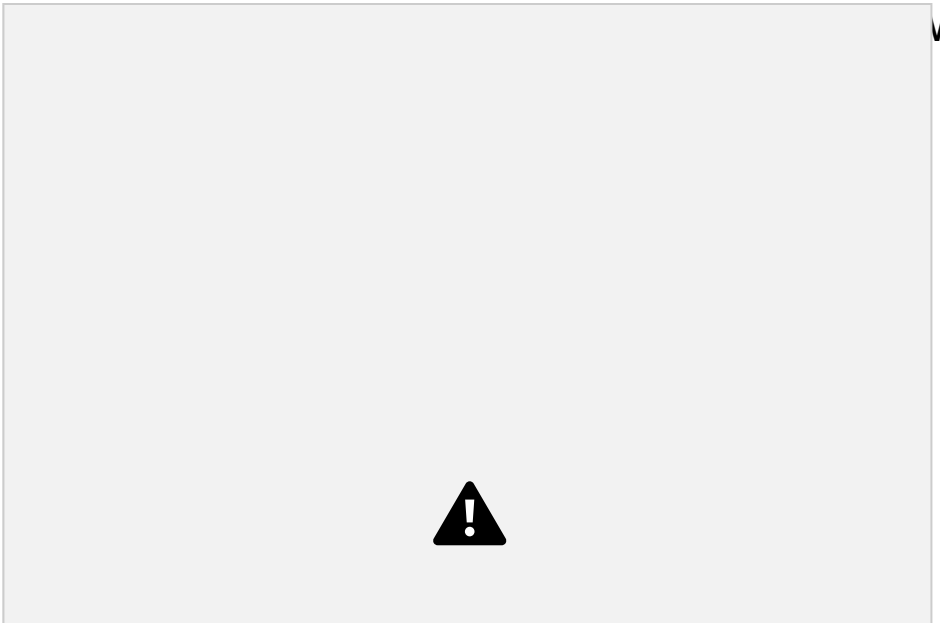
• The amount of surface available for recording in a disk is greater than in a drum of equal physical size.

• Therefore, more information can be stored on a disk than on a drum of comparable size.

• For this reason, disks have replaced drums in more recent computers.

Magnetic Disks

A magnetic disk is a circular plate constructed of metal or plastic coated with magnetized material.



Often both sides of the disk are used and several disks may be stacked on one spindle with read/write heads available on each surface.

All disks rotate together at high speed and are not stopped or started for access purposes.

Bits are stored in the magnetized surface in spots along concentric

circles called tracks.

The tracks are commonly divided into sections called sectors.

In most systems, the minimum quantity of information which can be transferred is a sector.

Magnetic Disks

- Some units use a single read/write head for each disk surface.
- In this type of unit, the track address bits are used by a mechanical assembly to move the head into the specified track position before reading or writing.
- In other disk systems, separate read/write heads are provided for each track in each surface.
- The address bits can then select a particular track electronically through a

decoder circuit.

- This type of unit is more expensive and is found only in very large computer systems.

Magnetic Disks

- Permanent timing tracks are used in disks to synchronize the bits and recognize the sectors.
- A disk system is addressed by address bits that specify the disk number, the disk surface, the sector number and the track within the sector.
- After the read/write heads are positioned in the specified track, the system has to wait until the rotating disk reaches the specified sector under the read/write head.
- Information transfer is very fast once the beginning of a sector has been reached.

- Disks may have multiple heads and simultaneous transfer of bits from several tracks at the same time.

Magnetic Disks

- A track in a given sector near the circumference is longer than a track near the center of the disk.
- If bits are recorded with equal density, some tracks will contain more recorded bits than others.
- To make all the records in a sector of equal length, some disks use a variable recording density with higher density on tracks near the center than on tracks near the circumference.
- This equalizes the number of bits on all tracks of a given sector.

Magnetic Disks

- Disks that are permanently attached to the unit assembly and cannot be removed by the occasional user are called hard disks.
- A disk drive with removable disks is called a floppy disk.
- The disks used with a floppy disk drive are small removable disks made of plastic coated with magnetic recording material.
- There are two sizes commonly used, with diameters of 5.25 and 3.5 inches.
- The 3.5-inch disks are smaller and can store more data than can the 5.25-inch disks.
- Floppy disks are extensively used in personal computers as a medium for distributing software to computer users.

Magnetic Tape

• A magnetic tape transport consists of the electrical, mechanical, and electronic components to provide the parts and control mechanism for a magnetic-tape unit. • The tape itself is a strip of plastic coated with a magnetic recording medium. • Bits are recorded as magnetic spots on the tape along several tracks. • Usually, seven or nine bits are recorded simultaneously to form a character together with a parity bit.

- Read/write heads are mounted one in each track so that data can be recorded and read as a sequence of characters.
- Magnetic tape units can be stopped, started to move forward or in reverse, or can be rewound.
- However, they cannot be started or stopped fast enough between individual characters.
- For this reason, information is recorded in blocks referred to as records.

Magnetic Tape

- Gaps of unrecorded tape are inserted between records where the tape can be stopped.
- The tape starts moving while in a gap and attains its constant speed by the time it reaches the next record.
- Each record on tape has an identification bit pattern at the beginning and end.
- By reading the bit pattern at the beginning, the tape control identifies the record number.
- By reading the bit pattern at the end of the record, the control recognizes the beginning of a gap.
- A tape unit is addressed by specifying the record number and the number of characters in the record.

• Records may be of fixed or variable length.

Associative Memory

Introduction

• Many data-processing applications require the search of items in a table stored in memory.

• An assembler program searches the symbol address table in order to extract the symbol's binary equivalent.

• An account number may be searched in a file to determine the holder's name and account status.

• **The established way to search a table is to store all items where they can be addressed in sequence.**

- **The search procedure is a strategy for choosing a sequence of addresses, reading the content of memory at each address, and comparing the information read with the item being searched until a match occurs.**
- **The number of accesses to memory depends on the location of the item and the efficiency of the search algorithm.**
- Many search algorithms have been developed to minimize the number of accesses while searching for an item in a random or sequential access memory.

Associative Memory/Content Addressable Memory (CAM)

- The time required to find an item stored in memory can be reduced considerably if stored data can be identified for access by the content of the data itself rather than by an address.

- A memory unit accessed by content is called an associative memory or content addressable memory (CAM).
- This type of memory is accessed simultaneously and in parallel on the basis of data content rather than by specific address or location.

Associative Memory/Content Addressable Memory (CAM)

- When a word is written in an associative memory, no address is given.
- The memory is capable of finding an empty unused location to store the word.
- When a word is to be read from an associative memory, the content of the word, or part of the word, is specified.
- The memory locates all words which match the specified content and

marks them for reading.

- Because of its organization, the associative memory is uniquely suited to do parallel searches by data association.
- Moreover, searches can be done on an entire word or on a specific field within a word.

Associative Memory/Content Addressable Memory (CAM)

- An associative memory is more expensive than a random access memory because each cell must have storage capability as well as logic circuits for matching its content with an external argument.
- For this reason, associative memories are used in applications where the search time is very critical and must be very short.

Hardware Organization

It consists of a memory array and logic for *m* words with *n* bits per word.

The argument register *A* and key register *K* each have *n* bits, one for each bit of a word.

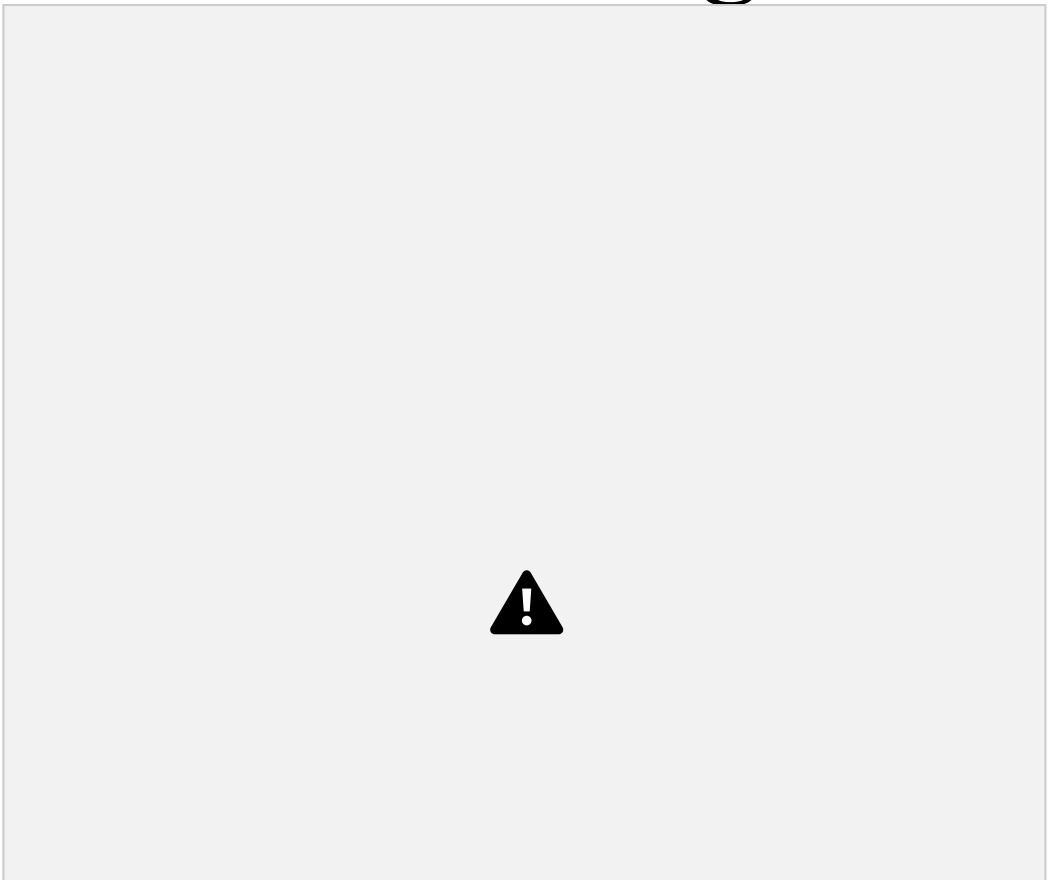
The match register *M* has *m* bits, one for each memory word. Each word in memory is compared in parallel with the content of the argument register.

The words that match the bits of the argument register set a corresponding bit in the match register.

After the matching process, those bits in the match register that have been set indicate the fact that their corresponding words have been matched.

Reading is accomplished by a sequential access to memory for those words whose corresponding bits in the match register have been set.

Hardware Organization



The key register provides a mask for choosing a particular field or key in the argument word.

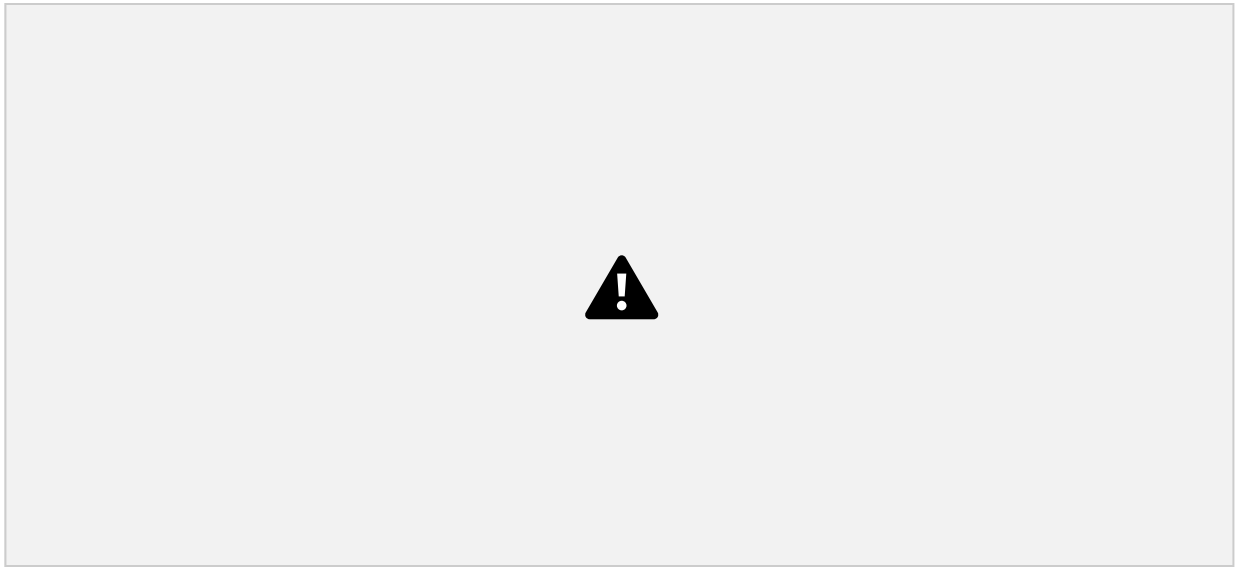
The entire argument is compared with each memory word if the key register contains all 1’s.

Otherwise, only those bits in the argument that have 1’s in their corresponding position of the key register are compared.

Thus the key provides a mask or identifying piece of information which specifies how the reference to memory is made.

To illustrate with a numerical

example, suppose that the argument register *A* and the key register *K* have the bit configuration shown below.
Only the three leftmost bits of *A* are compared with memory words because *K* has 1's in these positions.



Word 2 matches the unmasked argument field because the three leftmost bits of the argument and the word are equal.

Cache Memory

Locality of Reference

- Analysis of a large number of typical programs has shown that the references to memory at any given interval of time tend to be confined within a few localized areas in memory.
- This phenomenon is known as the property of **locality of reference.**

Locality of Reference

- The reason for this property may be understood considering that a typical computer program flows in a straight-line fashion with program loops and subroutine calls encountered frequently.

- When a program loop is executed, the CPU repeatedly refers to the set of instructions in memory that constitute the loop.
- Every time a given subroutine is called, its set of instructions are fetched from memory. **Thus loops and subroutines tend to localize the references to memory for fetching instructions.**

Locality of Reference

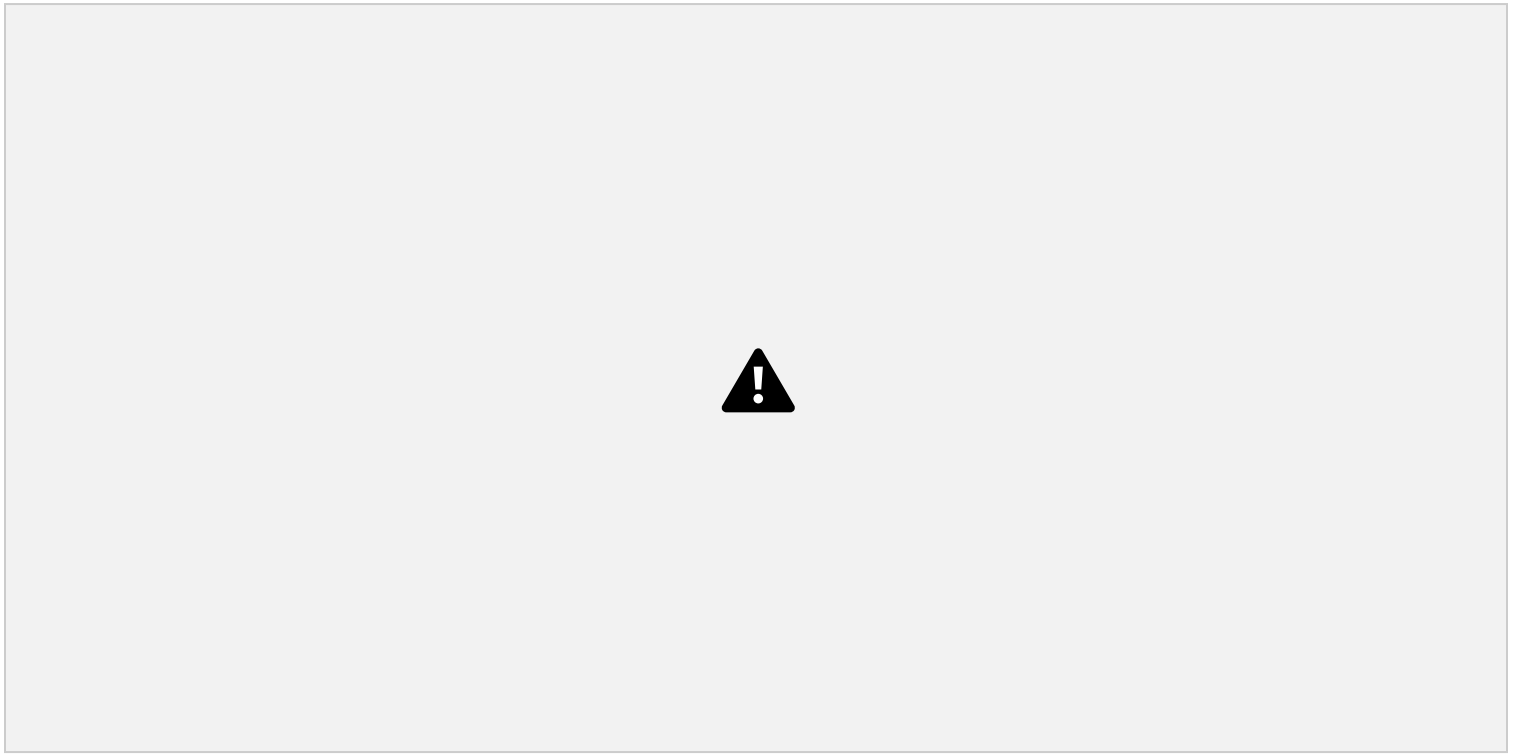
- To a lesser degree, memory references to data also tend to be localized.
- Table-lookup procedures repeatedly refer to that portion in memory where the table is stored.

- Iterative procedures refer to common memory locations and array of numbers are confined within a local portion of memory.
- **The result of all these observations is the locality of reference property, which states that over a short interval of time, the addresses generated by a typical program refer to a few localized areas of memory repeatedly, while the remainder of memory is accessed relatively infrequently.**

Cache Memory

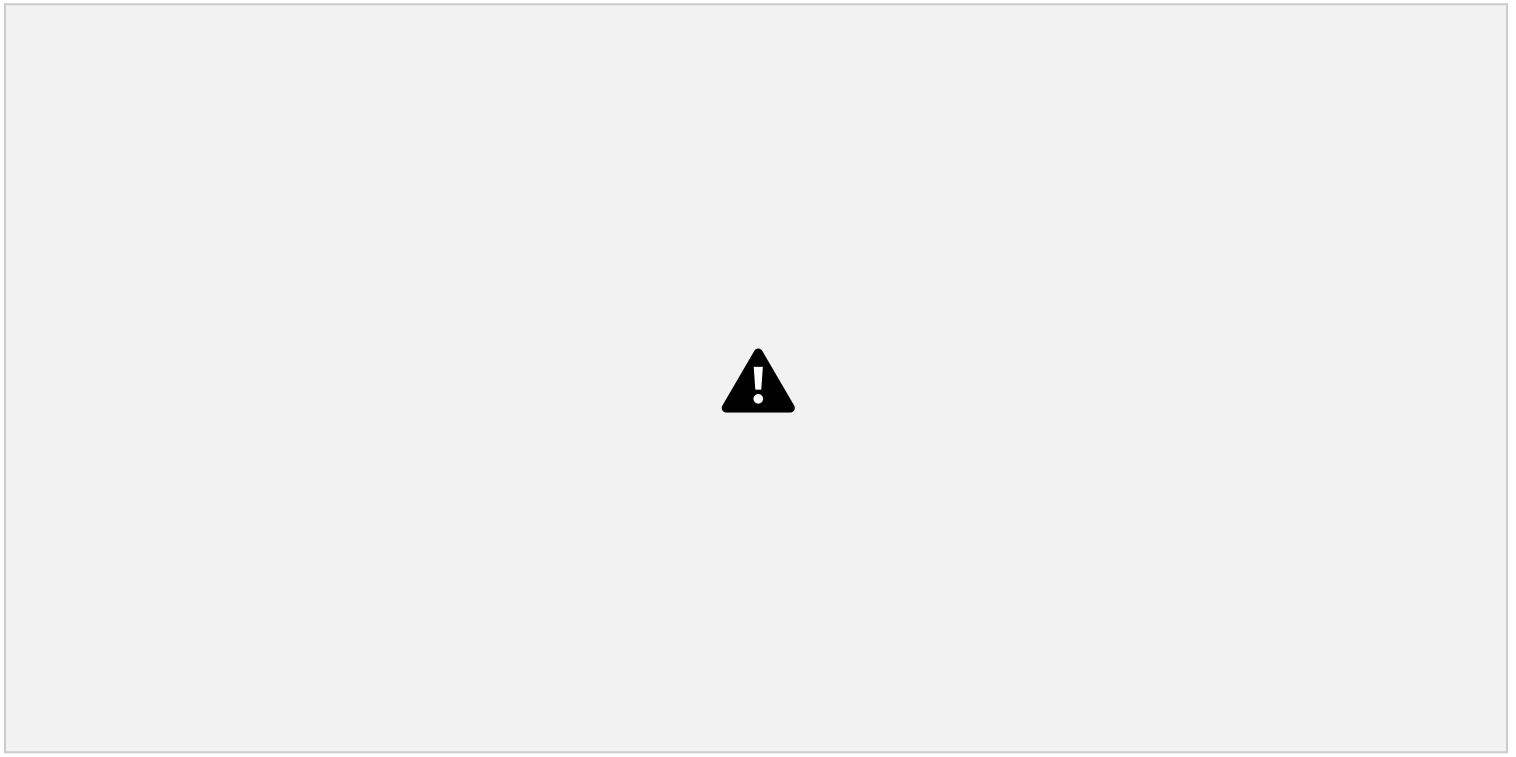
- If the active portions of the program and data are placed in a fast small memory, the average memory access time can be reduced, thus reducing the total execution time of the program.

• Such a fast small memory is referred to as a **cache memory.** • It is placed between the CPU and main memory



Cache Memory

- The cache memory access time is less than the access time of main memory by a factor of 5 to 10.
- The cache is the fastest component in the memory hierarchy and approaches the speed of CPU components



Cache Organization

- The fundamental idea of cache organization is that by keeping the

most frequently accessed instructions and data in the fast cache memory, the average memory access time will approach the access time of the cache.

- Although the cache is only a small fraction of the size of main memory, a large fraction of memory requests will be found in the fast cache memory because of the locality of reference property of programs.

Operation of Cache

The basic operation of the cache is as follows.

• When the CPU needs to access memory, the cache is examined. • If the word is found in the cache, it is read from the fast memory. • If the word addressed by the CPU is not found in the cache, the main memory

is accessed to read the word.

- A block of words containing the one just accessed is then transferred from main memory to cache memory. The block size may vary from one word (the one just accessed) to about 16 words adjacent to the one just accessed.
- In this manner, some data are transferred to cache so that future references to memory find the required words in the fast cache memory.

Hit Ratio

- The performance of cache memory is frequently measured in terms of a quantity called hit ratio.
- When the CPU refers to memory and finds the word in cache, it is said to

produce a hit.

- If the word is not found in cache, it is in main memory and it counts as a miss. • The ratio of the number of hits divided by the total CPU references to memory (hits plus misses) is the hit ratio.
- The hit ratio is best measured experimentally by running representative programs in the computer and measuring the number of hits and misses during a given interval of time.
- Hit ratios of 0.9 and higher have been reported.
- This high ratio verifies the validity of the locality of reference property.

Memory Access Time

- The average memory access time of a computer system can be improved considerably by use of a cache.

- If the hit ratio is high enough so that most of the time the CPU accesses the cache instead of main memory, the average access time is closer to the access time of the fast cache memory.
- For example, a computer with cache access time of 100 ns, a main memory access time of 1000 ns, and a hit ratio of 0.9 produces an average access time of 200 ns.
- This is a considerable improvement over a similar computer without a cache memory, whose access time is 1000 ns.
- The basic characteristic of cache memory is its fast access time. •

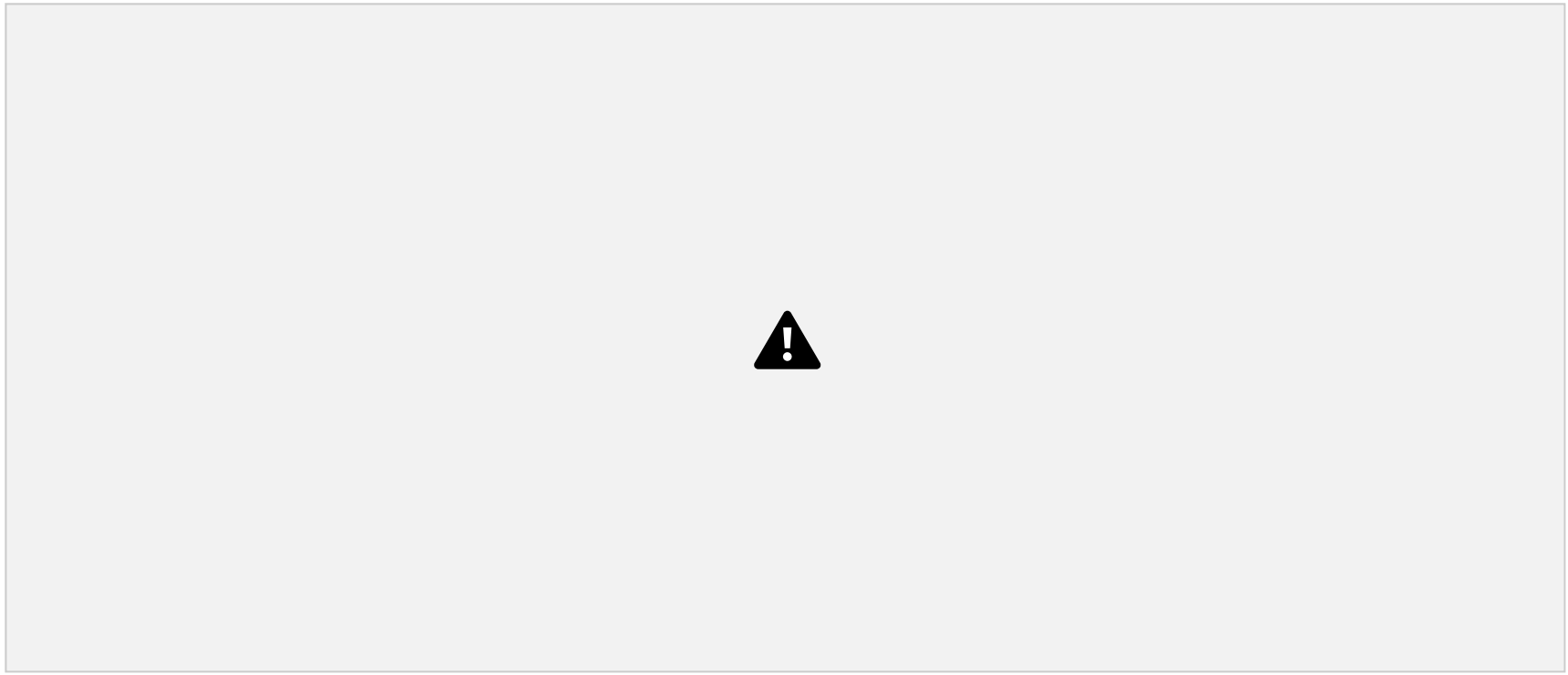
Therefore, very little or no time must be wasted when searching for words in the cache.

Mapping

- The transformation of data from main memory to cache memory is referred to as a mapping process.
- Three types of mapping procedures are of practical interest when considering the organization of cache memory:

1. Associative mapping
2. Direct mapping
3. Set-associative mapping

To help in the discussion of these three mapping procedures we will use a specific example of a memory organization as shown in Fig.



Description

- The main memory can store 32K words of 12 bits each.

- The cache is capable of storing 512 of these words at any given time. For every word stored in cache, there is a duplicate copy in main memory.
- The CPU communicates with both memories.
- It first sends a 15-bit address to cache. If there is a hit, the CPU accepts the 12-bit data from cache.
- If there is a miss, the CPU reads the word from main memory and the word is then transferred to cache.