High Performance Computer Architecture
Prof. Ajit Pal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

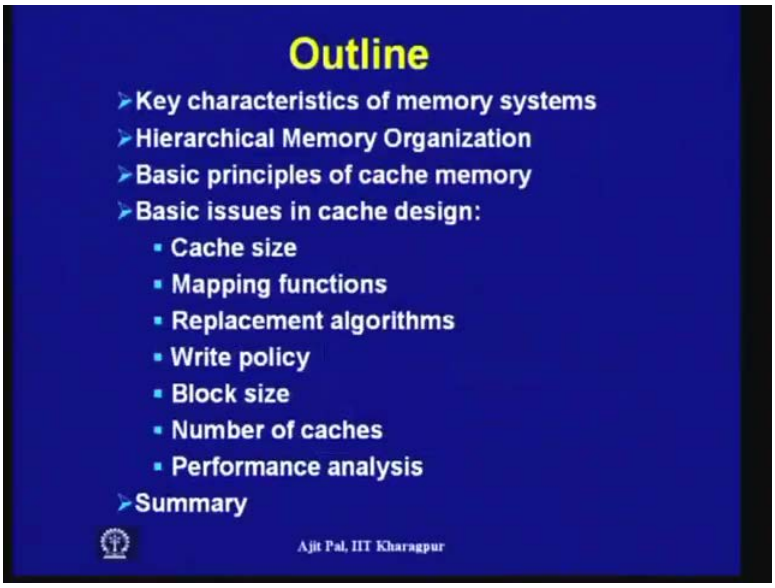
Lecture - 21
Hierarchical Memory Organization

Hello and welcome to today's lecture on Hierarchical Memory Organization. So far, we have focused on different techniques for improving the performance of the professor, but in a computer system, it is not only the processor, they are other components, which are present and the performance is dependent not only on the processor, but those but also on those components. And memory is one such very important component of a computer system.

(Refer Slide Time: 01:27)

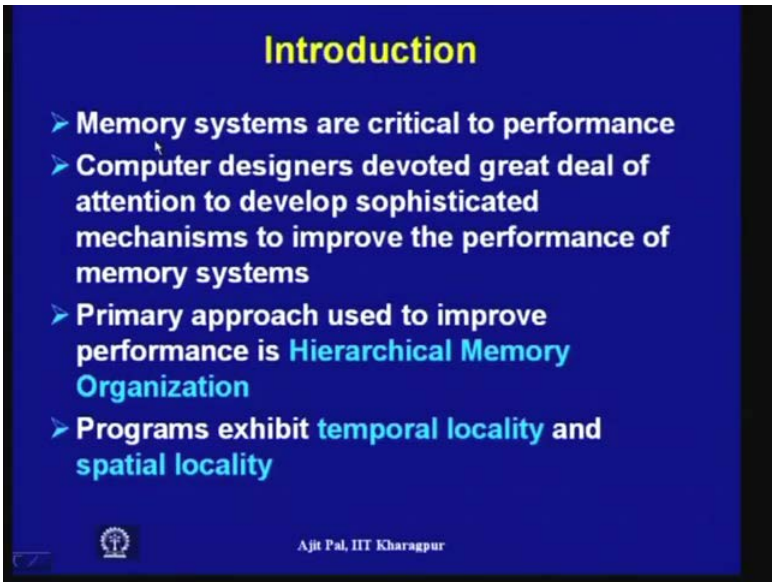
So starting today, we shall focus on hierarchical memory organization, which is the technique for improving the performance of the memory system.

(Refer Slide Time: 01:28)



So, I shall discuss about key characteristics of memory systems, which will provide you necessary background, to discuss about to know about hierarchical memory organization or other understand hierarchical memory organization. Then, I shall particularly today start our discussion on cache memory, which is present nowadays, present in almost all processors, then discuss about various this basic issues in cache memory design, like cache size, mapping function, replacement algorithms, write policy, block size, number of caches, performance analysis and so on. So, these are the things, I shall discuss maybe, it may not be able to cover in one lecture, but we may cover in two lectures.

(Refer Slide Time: 02:32)

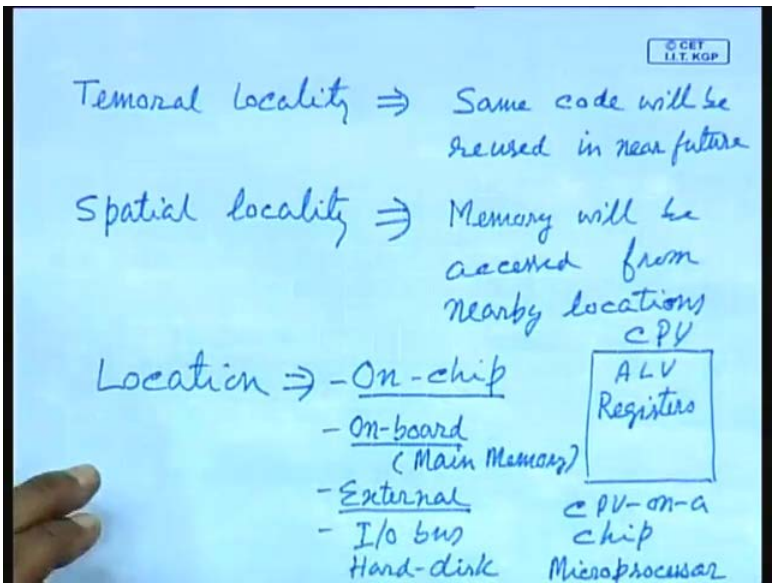


As I maintained in the beginning, memory systems are critical to performance, why it is so, the reason for that is as we have seen, the program and data are present in the memory, the CPU has to fetch it from the memory, for the purpose of computation. So, instructions have to be fetched from memory, data to be fetched from memory, so if the memory is very slow, cannot provide the instruction and data at the same rate, at which the processor is computing, then the execution of instructions cannot proceed.

And that is the reason, why memory system are critical to performance of the overall computer system and is an consequence computer designers have devoted, a great deal of attention to develop sophisticated mechanisms, to improve the performance of memory system. So, many sophisticated techniques have been developed, which I shall discuss in a couple of lectures, to improve the performance of memory systems and the most primary approach, used to improve performances is hierarchical memory organization.

So, this is the key technique, which has been used to improve the performance of the memory system and that is the reason why, we shall discuss about it in more details. Now, we may be asking, what is the basic idea behind this hierarchical memory organization, actually it is best on the observation that programs exhibit, temporal locality and spatial locality, what do you mean by temporal locality by temporal locality we mean that a piece of program, which is being used now, will be used in your or future.

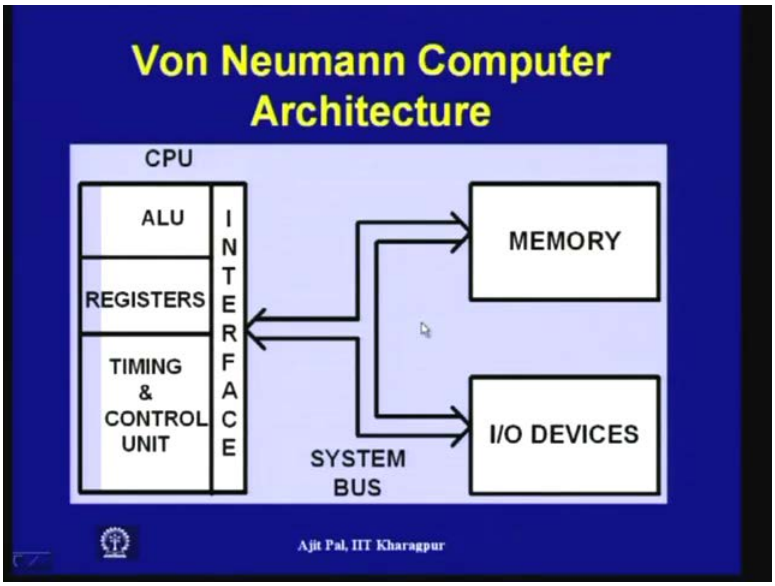
(Refer Slide Time: 04:52)



That means, that temporal locality is concerned with time that means, same code will be reused in near future, near future means, recently it will be used, then comes the spatial locality. What does it mean, spatial locality means that, as we know the programs are codes are stored in a memory, now spatial locality says that, if we are executing a code, on a part of a I mean from a particular memory location, the adjust den locations will be accessed in their future.

That means, the memory will be accessed from, nearby locations, nearby memory locations, where there it is instruction on data and particularly it is for instruction, it is very common. So, these 2 properties can be exploited to build a hierarchical memory organization and we shall see, how it can be done.

(Refer Slide Time: 06:43)



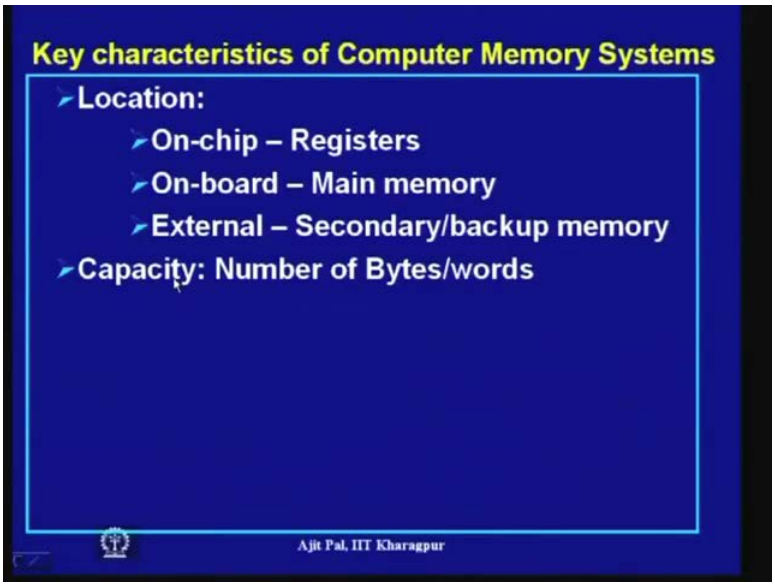
Now, before we going to the details of hierarchical memory organization let us start with the simplest computer system based on Von Neumann architecture that was the basis for stored program computer, as you have seen we have got a central processing unit. Comprising arithmetic and logic unit registers timing and control unit and there is a interface through, which it is connected to the memory and I O devices.

And this particular external box through which is connected is known as system box, as you know address at this lines data lines and many control lines are used, for the purpose of accessing memory I O devices. So, you see in general we assume that, your program

and data is outside the C P U and it is inside the memory, which had to be accessed through a system box and obviously, it will slower.

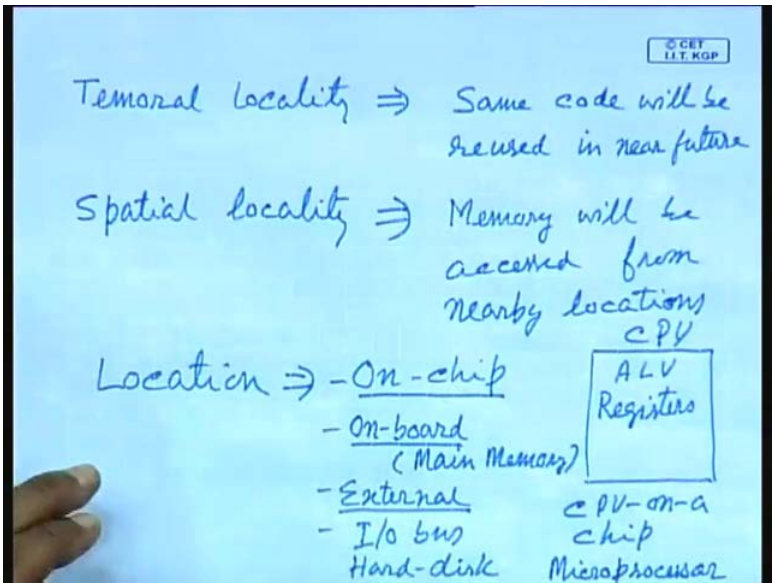
Of course, this is the Von Neumann computer architecture and subsequently as we have incorporated different kinds of memory the architecture had changed an as we shall see, how it had changed, particularly to tech here of hierarchical memory organization.

(Refer Slide Time: 08:08)



So, let us have look at the various characteristics of key characteristics of computer memory systems, first is start with the location, location means where it is located, where the memory is located.

(Refer Slide Time: 08:28)

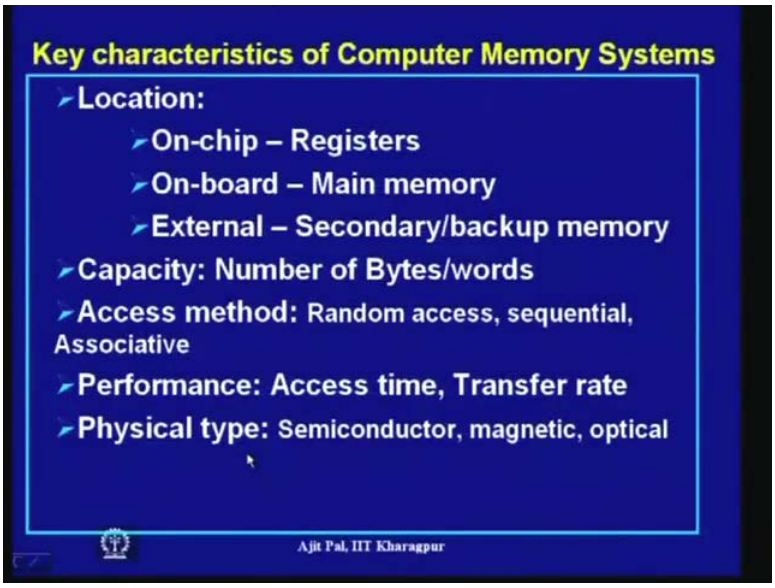


It can be say on chip, on chip means you know that nowadays, we know that C P U is implemented on a single chip that means, the C P U on a chip is commonly referred to as microprocessor. And microprocessors are used as central processing unit, in present day computer systems, now since it is realized on chip, some memory is president on chip and particularly as you have seemed, the registers apart from A L U arithmetic logic unit, registers which is a kind of memory is stored on chip.

So, we can say that the first type of location, where memory can be stored is on chip, so one is on chip, now whenever you built a system, obviously, you cannot build a system just using the C P U, you will require main memory and other types of memories. So, that will present, one possibility is that on board means, he will make a printed circuit board, on which you will put your C P U central processing unit and main memory and other types memories.

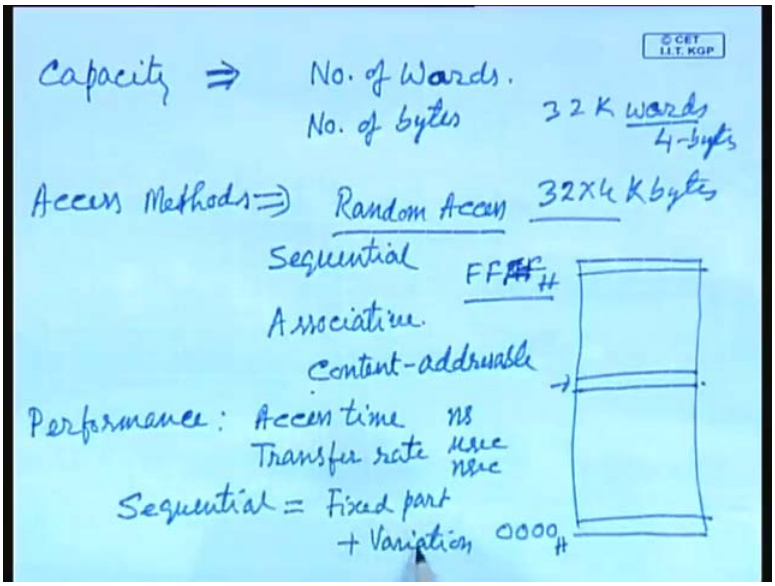
So, that is why second type of memory can be on board, on board particularly main memory, can be on board, third type of memory, which can be an external. External means, it is not an chip, it is not on the board, but outside means, you have to connect it usually through, what is known as I O bus for example, hard disk and another type of memories our external and usually connect through I O bus. So, we can say broadly there are 3 possible locations, one is on chip, another is on board, another is external and obviously, the time required to access memory will be depended, where they are located.

(Refer Slide Time: 11:15)



Second important characteristics is capacity, capacity of a memory is specified, usually in terms of bytes, there are possible, there are several approaches.

(Refer Slide Time: 11:32)



Capacity the preferred unique for capacity can be number of words that is present in particular memory system, that means, 8 bit in processor the word size can be 8 bit, for 16 bit processor, the word size can be 16 bit, for 32 bit processor, it can be 32 bit. So, best on the word size can be considered as the capacity, I mean the in terms of the

number of words the capacity is specified, that means, say 32 kilo words can be the capacity.

However, for the sake of generalization, that means, irrespective of the word size you can specify the capacity, for example, just by the capacity, you may not, you will not able to, identify the exact size, unless the word size is specified. So, just from these that 32 K, you cannot say, how many bits or bytes are present that is why, nowadays the common approach is to used number of bytes as the units of capacity.

And you will see, subsequently we shall be always using, the number of bytes as the capacity of a memory. So, if it is a say word size is 4, I mean 4 bytes then, essentially you have got 32 into 4 K bytes, so you have to specify in terms of bytes and in that case capacity as to represented in this way. So, this is how the capacity is represented then comes the access method, there are various types of access methods, which can be used for different types of memory.

The most common ones, access methods number 1 is random access, what you really mean by random access, we can mean, that wherever the byte or word is present can be accessed.

Accessed at the same time, you will required for accessing irrespective of it is location, so for example, you have got memory, which is really early organized, starting with location 0 0 0 0 2 1 1, I mean 0 0 0 0 0 may be in H and this can be say FFFF in H. So, this is the you have got 64 kilo bytes of memory, now irrespective of whether, it is where it is locating, where it is located, whether it is 0 location or this location, it can be accessed at the same time.

Same time will be required for accessing that is why, it is called random access, that means, some kind of and this location is fixed for particular word, I mean you can be, it is specified with the help of the address, that means, address specifies the data, that is present (Refer Time: 15:24), what data is present there. And it can be accessed, using this, I mean at the same time, another possibility is sequential access, in case of sequential access, memory is organized, in such a way that, you have to access it serially as it happens in your tape, as it happens in your C D room.

So, I will find that you have to access one after the other, starting from a particular point and then after you reach that point then, you will able to access, so in such cases, you will see that access time cannot be specified by a fixed time. So, in this case it is fixed time, but in case of sequential access that, where it is located that will also decide, how much time will be required to access it, so it is will be a variable time depending on it is location, that happens in case of tape and also in case of disk to some extent.

And then third type of access is known as associative, in case of random access, we have seen, for a given address it is the location is fixed, the data where it is present is fixed, now in case of associative memory, it can be decided based on content. A part of the content can be used, to access it and that is why it is also called content addressable, that means, the location cannot decided best on is not fixed, for a particular data location is not fixed.

It can be present anywhere and part of the content will decide, where it is located and later on when I shall be discussing cache memory you will see, this type of associative memory is used and we shall see how a part of the content can be used to access it later on. Then comes the performance, performance plays a very important role and obviously, you have to we have to specify performance, what do you really mean by performance, performance means C P U is trying to access, either data or instruction from the memory.

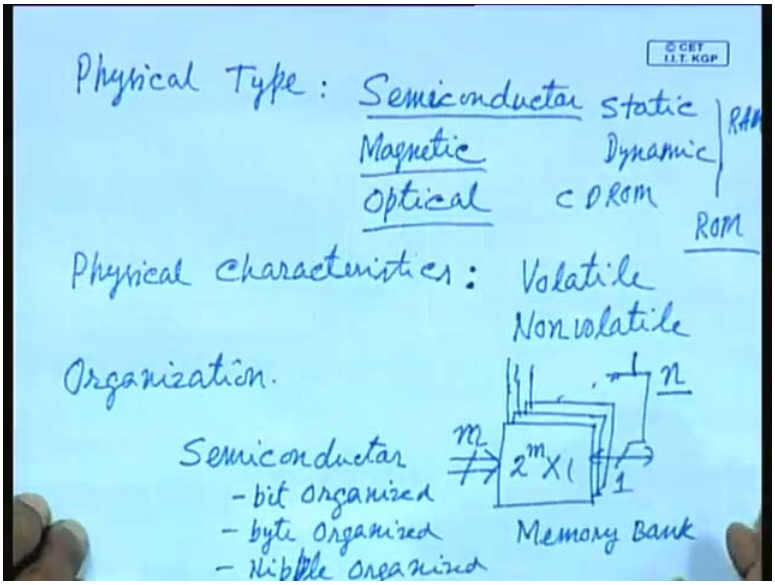
So, what is the C P U will do, C P U will generate some address, for that instruction on data and it as to wait for some more time to get the data, that means, memory system will take some time to provide the data to the processor. And that is why, they access time is very important, that means, time required from the beginning of providing the address and when the correct data is available on the data bars or through, which it is access.

Access time usually can be specified in terms of say the rate, at which it can be specified in terms of maybe nanosecond or microsecond or mini second, if depending on the type of memory that is being used. For example, for main memory system, it can be 10 to 100 of nanosecond and later on, we shall discuss cache memory, you will see that, for that the access time will be dependent is fraction of nanosecond. Because, it depends on the technology that is being used to implement that memory, so access time and also it can be specified in some other way or another way that your transfer rate.

The rate at which data transfers can take place, so if it is access time in nanosecond 1 by x-is time will be the transfer rate, rate the rate at which transfer can place. Now, as I maintained for random access memory, the transfer rate will be fixed irrespective of the location, on the other hand whenever it is sequential then access time, there will be 2 parts, one is fixed part, an another one is variable part that, variable part will be dependent on the exact location and the number of bytes that as been transferred.

So, this sequential access time is a is a that transfer rate is variable cannot be fixed, in case of sequential access, then comes the physical type.

(Refer Slide Time: 20:43)



Physical type means, how it has been realized, I mean the technology that is been used. As you know, nowadays we use semiconductor memory, for our main memory and cache memory. And particularly the semiconductor technologies is used to implement the processor and the same technology is being used to implement the main memory and cache memory and has a consequence, there is good electrical compatibility. And depending on the technology the that is realization, it can be two types as you know, one can be static gram and another can be dynamic.

And depending on that, the that access time or cycle time will be different and accordingly, we shall used it for different purposes. So, semiconductor memory is very common and that is used, because of they are speed of operation another possibilities is magnetic technique, for example, hard disk magnetic tape, the data is stored with a help

of magnetic technique. So, the that magnetic medium is used to store it and you have to you and in such cases you will find the access time will be longer and another technique that being use optical technique that is used in your C D rom.

So, these are the different types of physical type that can be use, for memory of computer system, then comes the physical characteristics, physical characteristics means, we have seen that we can use different types of memory and their characteristics can be different. And particularly it is usually classified into 2 types, one is known as volatile, what do you mean by volatile, that means, as long as power is present then information is available, information is not lost, information is retained.

And that is why, I mean is not volatile means, as long as for is their information is retained, but has you with draw the power information is lost, as the happens in your semiconductor memories particularly static and dynamic RAM. So volatile, I mean that is the one of the physical characteristics, but you will see in your computer memory system, you will always require 2 types of memories, you will require volatile as well as you will require nonvolatile by nonvolatile, I mean information will not be lost, if in if you with draw power, that means, when the power supply is withdrawn.

And in that case also the information will not be lost, that means, it can be retained for longer duration and it can be subsequently used and reused. So, nonvolatile is an another important property particularly optically disc magnetic, I mean, the using magnetic property, magnetic types of memory devices, optical type of memory devices belong to this category. And in semiconductor memory device is also, you can realize nonvolatile and that is known as read-only memory ROM. So, physical characteristics are can be broadly divided into 2 types, volatile and nonvolatile, then comes the organization.

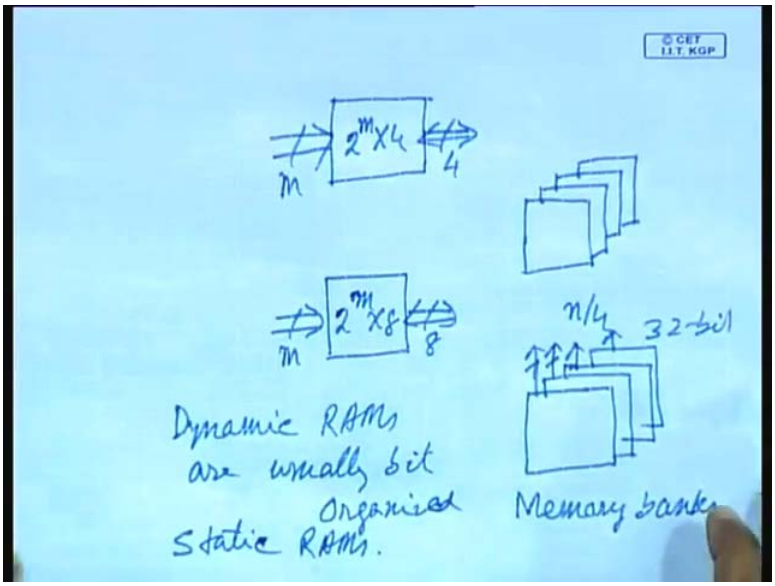
So, let me specify the organization in terms of the memory that is being used particularly semiconductor memory, we have seen that the memory usually is organized in terms of bytes or words, particularly the CPU will access in terms of words. But, it may be several bytes, now on the other hand your memory can be organized in a different way there are several alternatives, that means, it can be bit organized.

What do you mean by bit organized that means, you provide some address say m bit addresses is provided and inside this memory, it is available in terms of bits; that means, it produces in terms of bites, it maybe usually by directional, that means, it is 1. So, you

have got inside the chip 2 to the power m memory locations and each counting in one bit, so it is called bit organized. Now, whenever you have to realize a n bit memory system, where n is the number of bits for the word you have access n bits in parallel, how will you do, you will be using several such memories number of such memories.

So, n such memory elements will be required to implements, so this will provide first to one bit basic provide another bit this will provide bit in this way, you will get n bits and this is called memory bank. So, if it is bit organized to realize n bits, you will require n chips, to have n bits simultaneously, now it is not always necessary that, the memory devices will be always bit organized, it can be byte organized or it can be nimble organized, what is the difference between bit organized and nimble organized.

(Refer Slide Time: 27:59)



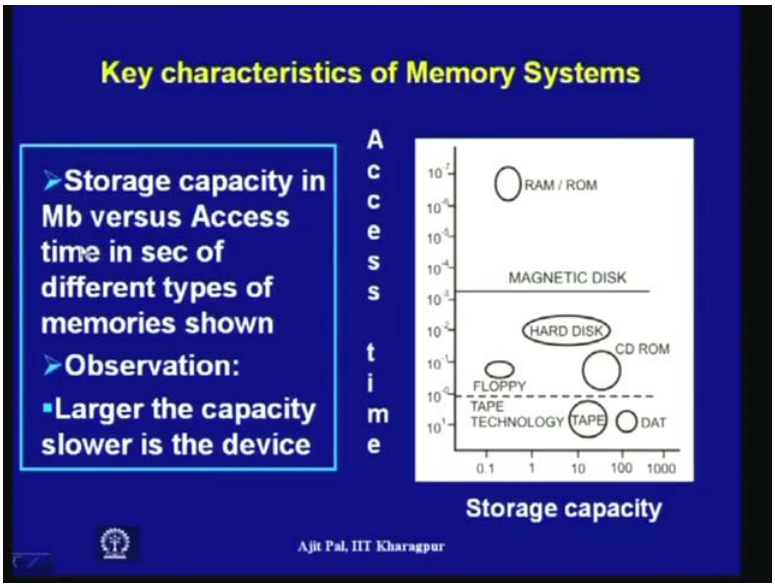
In case of nibble organized, you will have, you can access 4 bits simultaneously, that means, it will have m addresses is present and inside it is organized as 2 to the power m in to 4 and has a consequence of you have to realize in bits of memory, you will require n by 4, such memory chips, that means, the number of devices that, you require in your computer systems will be n by 4 in parallel. So, here also, you will require a memory bank, but number of devices that will be required will be lesser, instead in compared to bit organized.

Similarly it is byte organized, you will get 8 bits at a time and if the number of at this lines again m inside the chip, it is organized as 2 the power m into 8. So, you can, you

will also require memory bank, whenever it these byte organized and particularly, for example, if the word sizes is 4, the 32 4 bytes or 32 bits, you will require 4, such memory chips, to have 32 bits. So, it will give 8 bit, it will give 8 bit and this will also give 8 bits, so this way, you will get 32 bit from 8, I mean 4 memory chips.

So, this is how memory banks are realized and in all modern computer systems, you will find this is how they memories organized particularly will see that, if it is dynamic Ram. Dynamic rams are usually bit organized on the other hand, static Rams may be in the nibble organized or byte organized, so whenever, you design a memory system, you have to take into consideration, these characteristics in realizing the memory system.

(Refer Slide Time: 30:33)



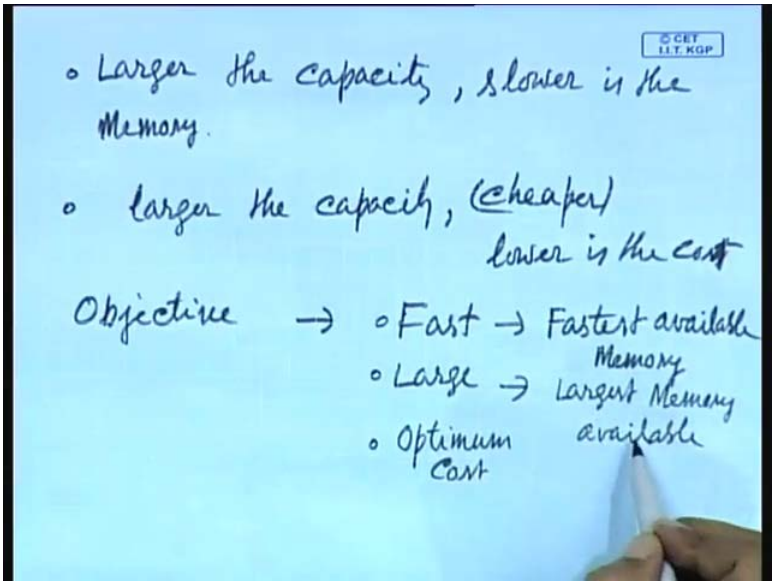
Now, let us have a look at the key characteristics of memory systems, number 1 is storage capacity in megabytes versus access time, in second of different types of memories. So, a computer system will usually required different types of memories, now what are they are important characteristics, relative characteristics, here we find that on the x-axis, we have the storage capacity of different types of memories that is been used and on the y-axis you have got the access time.

So, access time is on the y-axis and on the x-axis you have got the storage capacity, we find that when the access time is very small 10 to the power minus 7 second, then your capacity is very small. So, when the access time is I mean a small or other the when it is speed is high, then capacity is small, on the other hand, we can see as the access time is

larger capacities is more hard disk will have a larger capacity and magnetic device magnetic tapes and other things will have still larger capacity, but their access time is very large.

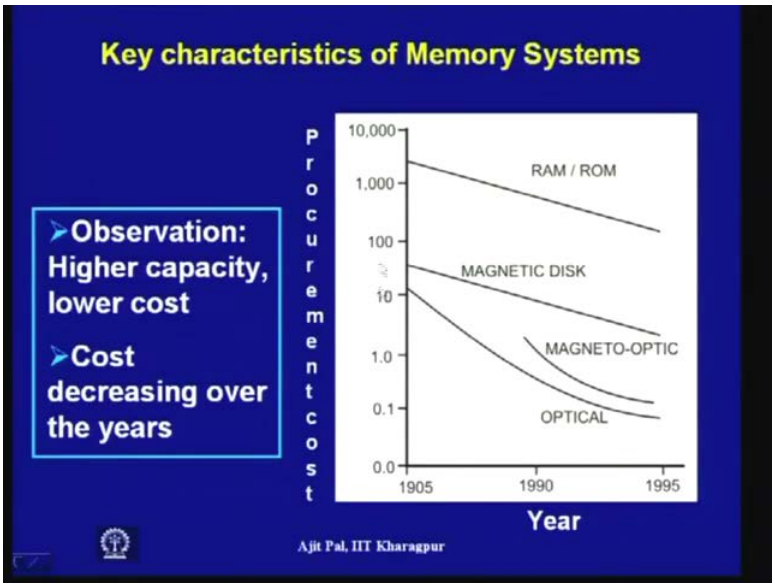
So, we find that larger the capacity slower is the device, this is a very key observation, important of the observation, which will be used in implementing the hierarchical memory organization.

(Refer Slide Time: 32:23)



That means, the larger the capacity slower is the memory, but that particular type of memory. So, these way you will play a very crucial role in deciding or in implementation hierarchical memory organization.

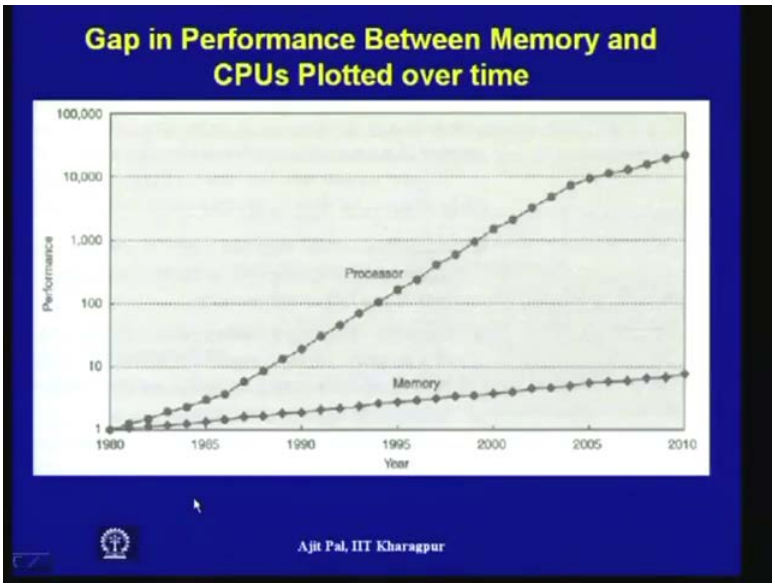
(Refer Slide Time: 32:51)



Now, second observation is that, cost we find that on the x-axis, we have the number of years and on the y-axis have the procurement cost, we find that the Ram and Rom are costlier although their cost is gradually decreasing over the years because of the advancement of technology process technology is improving. So, it is getting cheaper and cheaper, but if you look at their relative positions, we find that Ram and Rom that is your semiconductor memory devices are costlier than magnetic disks and magnetic disks are costlier than optical disks or magneto optic disks.

Although their cost is decreasing over the years, but these relative position has remained the same, that means, observation is higher capacity lower cost. So, here first observation was larger the capacity slow slower is the memory, another we can write larger capacity, it is cheaper, that means, lower you can write cheaper, you can specify lower is the cost. So, these 2 plays a very important role in implementing hierarchical memory organization.

(Refer Slide Time: 34:38)

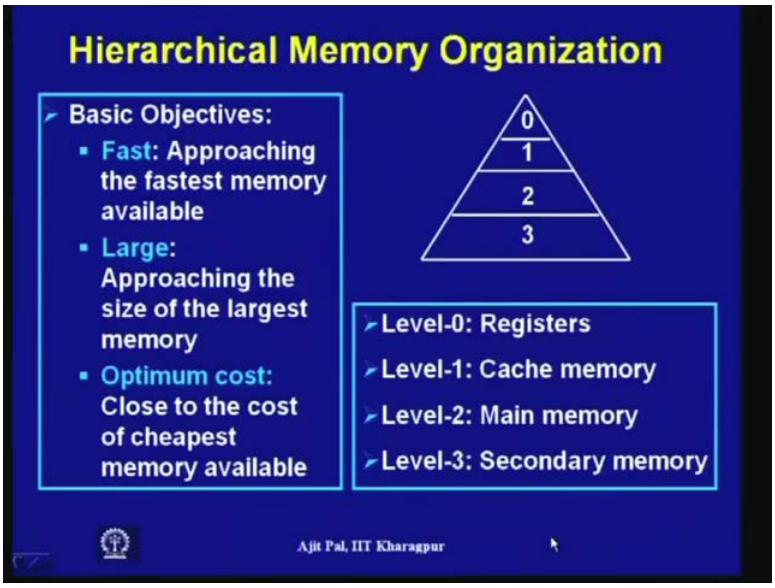


Another important observation is you can see, how the performance is changing over the years, for processor and memory system, here actual the memory by memory, we mean that dynamic Ram. So, dynamic Ram has been used, as the representative memory here and here is your professor, you can see the performance of the professor is improving, rapidly, compared to the memory systems, it is not that the memory the performance of memory system is not improving.

But, the improvement is very slow the rate at the performance of memory the devices is improving is very slow, compared to the rate, at which the performance of the processor is improving. Here you can see the processor performance is improving at much higher rate maybe 7 percent per year, on the other hand for memory it may be point seven zero percent per year. So, the rate of improvement is much lower for memory, what does it really mean, that means, has the here is progressing, may be in the year 1980, the performance of processor and the performance of the dynamic Ram was same.

But, it has changed over the here and you can see, this gap is gradually becoming wider and wider, so question naturally arises, how to bridge the gap, because the processor has to access instructions and data from memory. And if the gap is increasingly the processor is becoming faster and faster memory is not becoming faster at the same rate, the gap is widening. So, we have 2 bridge this widening gap by using suitable technique and hierarchical memory organization is the technique that can be used.

(Refer Slide Time: 36:50)



So, we have now, we have given you necessary background, for hierarchical memory organization, now let us see, what is the basic objective of hierarchical memory organization, first is fast. So, whenever you want to have a memory system, your objective, first objective is you want it fast, that means, obviously, you wanted compatible with the processor, whenever it is compatible with the processor, than the day memory the processor will not incur any you know, wet cycle or drilling.

So, the rate, at which the processor needs it, it will get it from the memory, if the speed is compatible, so that is that means, you want as fast as the CPU, that is your first objective, second objective is large. So, first means obviously, we have seen, you have got different types of memories, so fastest available memory, then second objective is it has to be large, obviously, this requirement is arising from, the programmers demand as you know the size of the code is increasing over the years, the problems are becoming more and more complex.

So, you will require more memory to store your program and obviously, you will require large memory and what is the basic objective, objective is it should be as large as the largest memory available. So, whenever we say fast, fast means that the this speed should be as close, as possible to the fastest memory, as we know the semiconductor memory devices are the fastest. So, this speed should be closer to the semiconductor

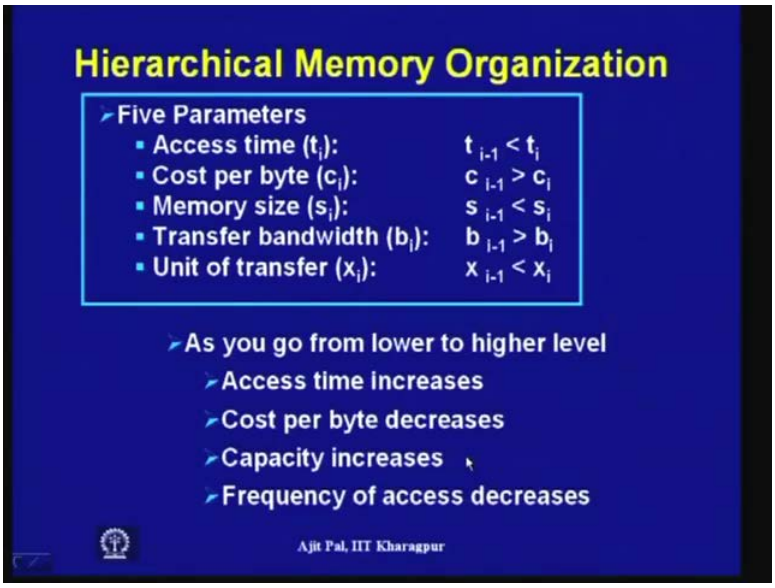
memory devices on the other hand, whenever we say large, we know that your secondary memory or backup memory, magnetic tape has the largest capacity.

So, capacity should be close to the a magnetic tape, but can we get these 2 together; that means, it should be fast it should be large. So, these 2 you want to gather, but unfortunately as we have seen, fastest memory is the costliest and smallest in size and largest memory is slower and cheaper. And last requirement is it has to be, we have to get it at an optimum cost, so we have to use, if we say suppose, you decide that, you will use only semiconductor memory, which is very fast and there very large size, then cost will be very high.

So, that will not serve your purpose, on the other hand, if it is as large, if you use decide to use the I mean largest memory, than it will be slow, it will not serve the purpose, that means, what we have to do. We have to devise a mechanism, such that speed will be closer to the fastest memory, Size will be closer to the largest memory cost will be optimum, optimum means, it will not be very high. So, that is the basic objective of hierarchical memory organization and with this basic objective memory has been organized in a hierarchical manner.

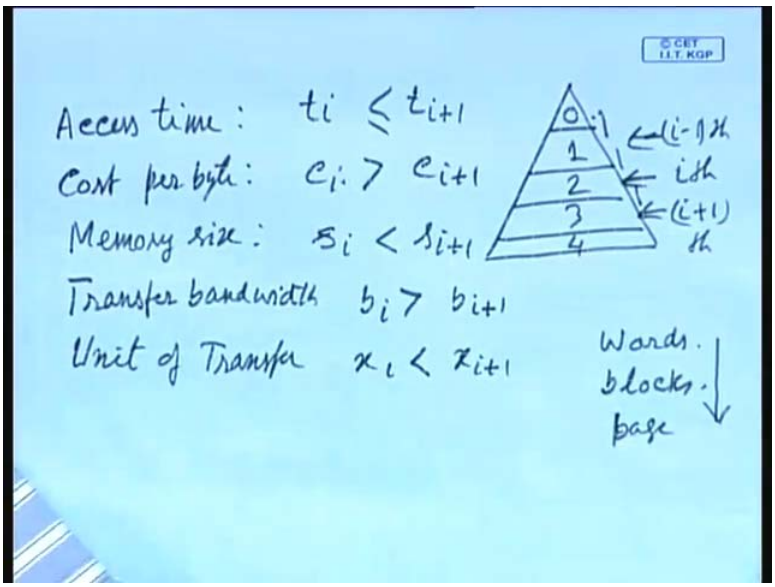
And this is how it is being done, say we have got, we shall be using different types of memories in the level 0, we shall be having registers level 1, we shall be using cache memory and in level 2, we shall be using main memory and level 3 secondary memory. So, as we can see, we shall be I have shown here, here 4 different levels of memory.

(Refer Slide Time: 41:55)



And now they are the 4 different memories can be, since they are organized in a hierarchical manner, you see how they are parameters will be changing different parameters will be different, for different types of memories, number 1 is access time.

(Refer Slide Time: 42:23)



Say here, we have organized in a hierarchical manner, so this is the 0 at level, this is the first level, this is the second level, this is the third level, you may have 4th level. Now, if we consider, it as the i th level, this is i minus 1 th level and this is i plus 1 th level. So, we can say, it this way now, what is the how the access time varies, whenever you

organize in this hierarchical manner. So, access time will vary in this way that time t i that is access time for the highly, it level we have seen that, then this is the fastest and this is the slowest, so access time, this will be, this is t i plus 1.

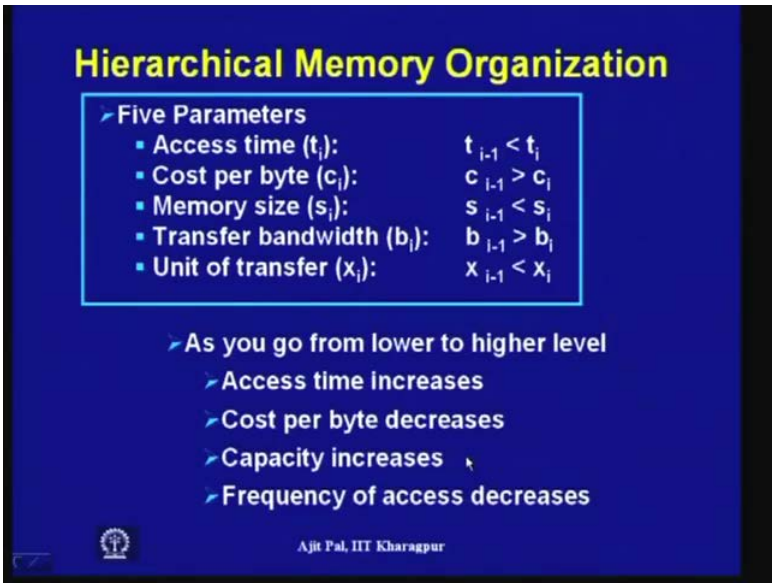
So, this will be more that means, that accessed time will be more has we go down words access time is more, for your larger memory size, so we can state it in this way, that means, as we go towards hierarchical memory the access time will become more. Second is your cost per byte say C i, C i minus C this is C i and this is C i plus 1. So, cost per byte, of which one will be more or which one will be less, how the cost will be changing we know that has we go down the cost is lower; that means, this cost for a lower level will be more, for byte then the cost for higher level.

So, C i is more than C i plus 1 or what about the memory size, memory sizes a size and s i plus 1. So, you have got say 2 different sizes, as we go downwards the sizes is increasing, that means, s i plus 1 will be larger than s i. So, memory size will be larger for i plus 1 at level compared to highest level, that means, it will be removed then comes the transfer bandwidth b i and b i plus 1. So, what do you really mean by bandwidth, bandwidth means, rate of transfer, their rate at which transfer will take place.

So, rate of transfer obviously, will be more between these 2 tables or between these 2 tables or between these 2 tables, that means, the rate of transfer will be, that within the hear, it will be more t compared to b i plus 1. Then comes the unit of transfer, unit of transfer is x i and x i plus 1, what you really mean by unit of transfer, you see we can transfer 1 width, 1 byte, 1 word. So, whenever you transfer, what is the minimum size that we transfer, as we all see the whenever it is between register and cache memory, it is usually in terms of words.

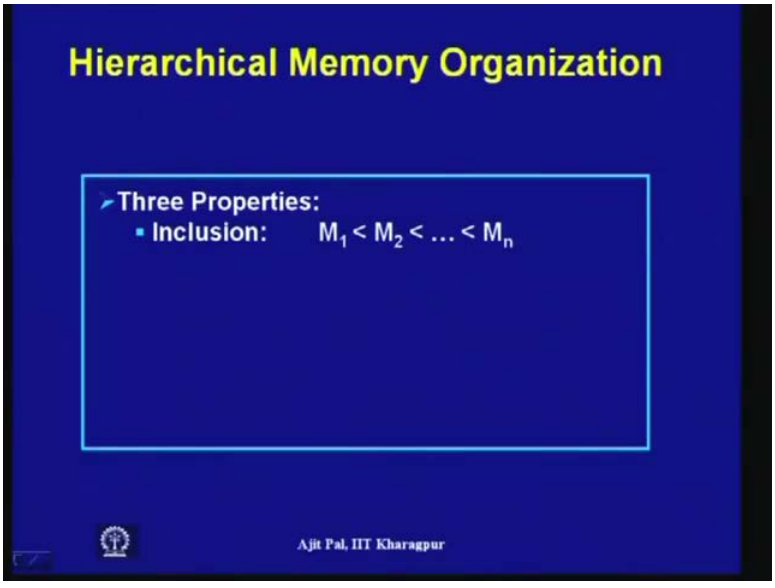
And whenever it is between cache memory and main memory, it will be in terms of blocks and whenever it is in terms of between main memory and secondary memory, it will be in terms of maybe in some other unit, maybe page. And you will see that their the sizes is gradually increasing and that is the reason why, this will be more; that means, units of transfer is more, as we go towards higher and higher levels.

(Refer Slide Time: 47:19)



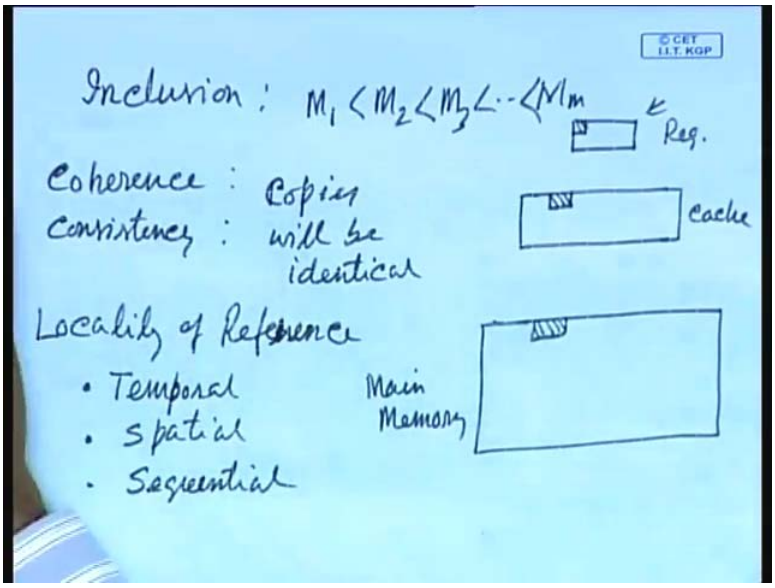
So, these parameters will be applicable, whenever is organized in a hierarchical manner, in addition to this, as we go from lower to higher level access time increases, costs per byte decreases, capacity increases, frequency of access decreases and these characteristics will be maintain.

(Refer Slide Time: 47:35)



Whenever you organized, whenever you realized hierarchical memory organization, now in addition to that 3 important properties are mentioned, what are the 3 properties.

(Refer Slide Time: 47:54)



Number 1 is inclusion, say hear you have got, say registers, let us assume this is your cache memory and here is your main memory. Now inclusion property says that, if a particular data is present here, it must be present here, it must also be present here, that means, they are convert the any data, which is present at lower level is not present in higher level, this is the inclusion property. Inclusion will guarantee that, any data which you get, in a lower level of memory, as to be present in higher level.

So, it can be specified as M i M 1 covers M 2 covers M 3, so if you have got M levels, this will be maintained. In other words that, a subset of data that is present in the main memory will be present the cache memory and a subset of the data present, which is present in the cash memory will be present in the register. So, this is called the inclusion property, which will be maintained, whenever you organized memory, in hierarchical manner.

Then second is coherence, now see here you have got one copy, a copy of a particular byte, another copy present here, another copy present here, so from here it was copied from here and from here it was copied from here. So, initially it was present in the main memory from where it was transferred to the cache memory and from cache memory, it is transferred to the registers. Now coherence property says that, coherence or consistence, it is also called consistency, it says that the copies will be consistent or identical.

What does it really mean, when you say all the copies of the same data present in different levels has to be identical, what does it really mean, it means that say if C P U modifies here, it is essential that it as to be modified here, it is also essential to be modified in the main memory. That means, you know that CP that registers are closed to the processor or rather registers are present inside the C P U, so whenever it executes a program, modification takes place in the registers.

So, as you do that, it will be necessary to transfer the same modification, implement the same modification at high higher levels, so they have to be identical, this is the second property to be maintained. Third property is locality of reference, I have already told you about, the locality of reference, one is your temporal locality, another is especial locality and third locality that is known as sequential locality. So, temporal locality as I told, if a if we use some data or this program instructions, at this moment in your future, you will be using it, that means, it will be re used in your future.

So, that is quite common because, if you are running a particular, I mean application right now, same application will be running in your future. So, we just like you know, if you are reading a book at a particular in stand semester, you will be reading the same book in your future. Second is your special locality, that means, some books some books, which are related to the book that are you are reading, say computer organization book other books related to that you will be reading in your future.

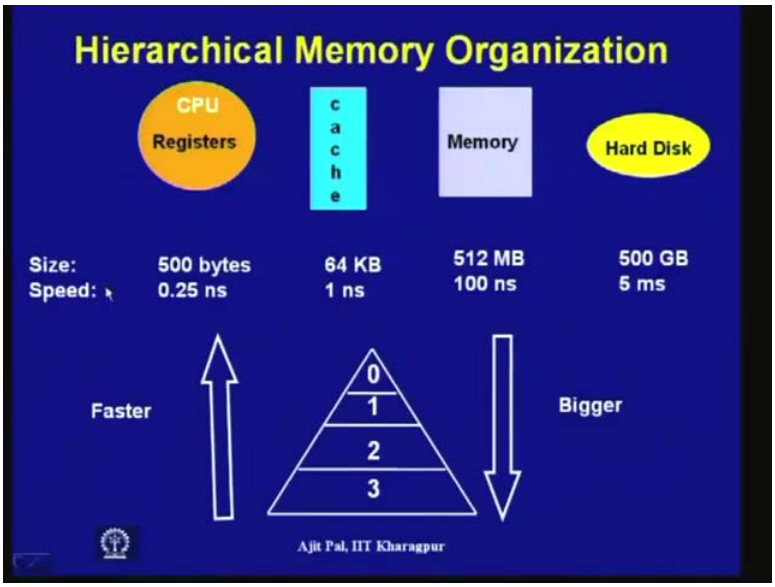
So, this is somewhat similar to some adjacent memory location, which are even not using at this movement, but it is likely to be used in your future, but as I have already told this is your special locality. Third locality, which is somewhat similar to special locality, but little different, which particularly arises, in the context of you know programs, execution of instructions.

So, we know that normally the what is a program, a program is nothing but, sequence of instructions, normally the sequence are stored in contiguous memory locations. So, whenever you are executing a program, it is very lively that you will be executing, will be fetching instructions from sequential memory locations, on the other hand that special locality is applicable to data.

And sequential locality applicable to instructions that is present in a memory, however, whenever there are branches, subroutine calls interrupt in such cases it will not be it will

not be sequential it will fetch from non sequential memory locations. But, most of the time, you will be doing fetching from sequential memory locations, so that means, this locality of reference is being is followed, in memory access by the processors and which will be exploited in implementing, hierarchical memory organization.

(Refer Slide Time: 54:58)



So, best on these, we can say here, typical example is given, you can see, you have got registers present, which is part of the C P U, cache memory and this is main memory and hard disk. And typical sizes are given, as you can see this size is gradually increasing 500 bytes maybe that is the size of the registers and cache memory size is 64 kilo byte main memory size is 512 mega byte. So, all in units are in terms of megabytes and hard disk capacity is 500 gega byte, so you can see that, it is roughly about 10 orders of magnitude.

So, the registers the cache memory is about 10 orders of magnitude than, the 10 to the power 3 compared to the register size. Similarly, main memory size is also about 3 orders of magnitude compared to the cache memory, similarly it is three orders of magnitude compared to hard disk size is compared to here also magnitude have then main memory. So, and so this is that means, size it is becoming bigger, as you go down, similarly you can look at a typical speed of present day memory devices.

The registers can be accessed with access time of 0.25 nanosecond, the cache memory can be accessed with the access time of 1 nanosecond the main memory can be accessed

with the access of 100 nanosecond, hard disk can be accessed with the access time of 5 mille second.

So, you can see I mean that is a typical value, but as I said that will be little bit variable, so this is how the hierarchical memory is organized. So, with this introduction in my next lectures, I shall discuss about, first the cache memory organization, that is the first level of hierarchical memory organization, that is used that is being used in all the contemporary computer systems.

Thank you.