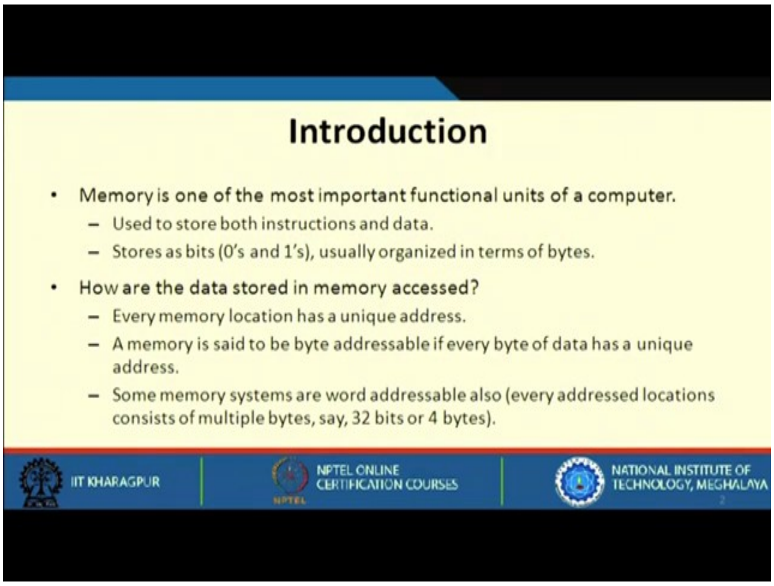
Computer Architecture and Organization
Prof. Kamalika Datta
Department of Computer Science and Engineering
National Institute of Technology, Meghalaya

Lecture - 23
Processor Memory Interaction

Welcome to week 5. In this week we shall be discussing on memory system design. We will look into the various technologies that are used to build the memory that we use in computers. And we will also look into how these memories are used to design and organize them in the system.

(Refer Slide Time: 00:49)



Introduction

- Memory is one of the most important functional units of a computer.
  - Used to store both instructions and data.
  - Stores as bits (0's and 1's), usually organized in terms of bytes.
- How are the data stored in memory accessed?
  - Every memory location has a unique address.
  - A memory is said to be byte addressable if every byte of data has a unique address.
  - Some memory systems are word addressable also (every addressed locations consists of multiple bytes, say, 32 bits or 4 bytes).

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NATIONAL INSTITUTE OF TECHNOLOGY, MEGHALAYA
2

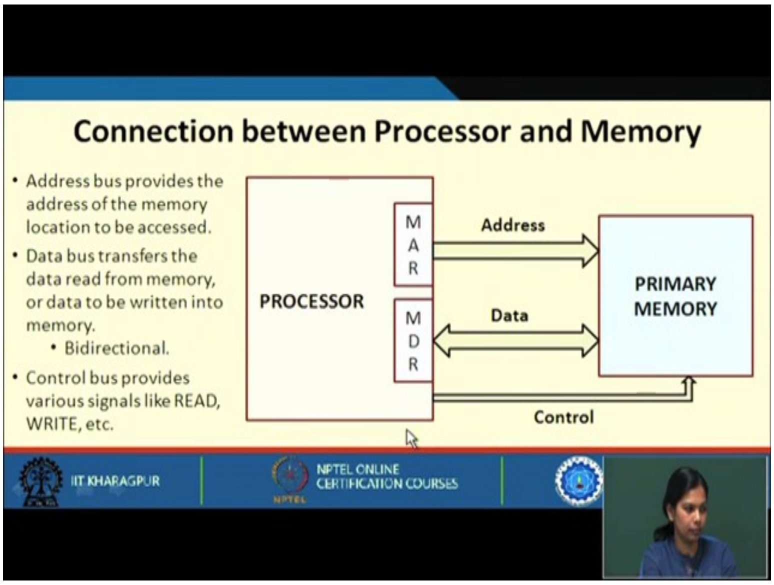
Memory is one of the most important functional units of a computer. We all know that. It is used to store both instructions and data. And it stores bits like 0's and 1's. So, as we have already seen, we encode an instruction with bits of 0's and 1's. So, in the memory location we say we store both instructions and data, those instructions and data are organized in bits of 0's and 1's. And they are usually organized in terms of bytes.

We will see here how are the data stored in the memory are accessed. We need to know the mechanism how we can access the data from the memory. We should also know how we store data into the memory. These are the two things we need to look into. Every memory location has a unique address.

And memory is byte addressable, that is, every byte i.e. a group of 8 bits, has a unique address. Some memory systems are word addressable. And by word addressable we mean that, each location consists of multiple bytes depending on the word size. If one word is 4 bytes or 32 bits, then the memory location will be changed as 0, 4, 8, and so on.

So, if it consists of 8 bytes or 64 bits then the word length is 64, then the memory address will be incremented by 8.

(Refer Slide Time: 03:10)



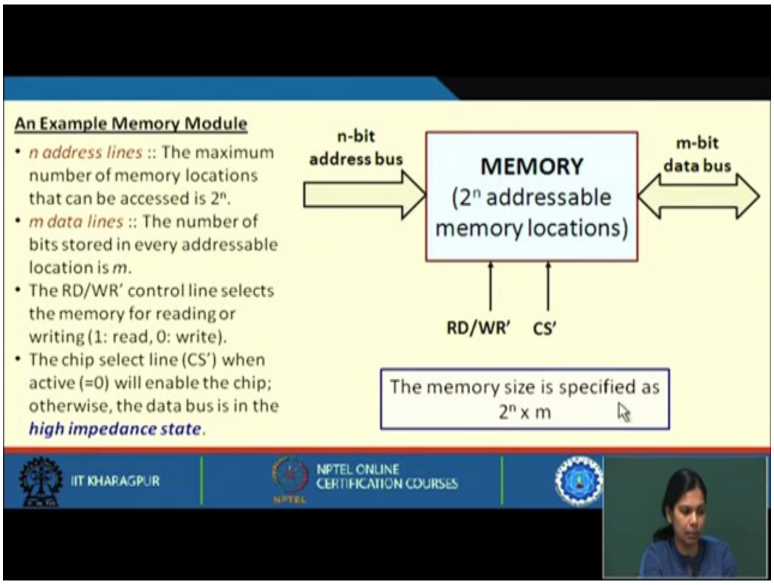
Connection between Processor and Memory

- Address bus provides the address of the memory location to be accessed.
- Data bus transfers the data read from memory, or data to be written into memory.
    - Bidirectional.
- Control bus provides various signals like READ, WRITE, etc.

Now, see the connection between processor and memory. As you know that in processor we have two important registers. One is memory address register (MAR), another is memory data register (MDR). MAR contains the address of an instruction or data that is to be read from the memory, or the address of data that is to be written into the memory. And that particular data which is to be read comes through this data bus. So, this is the data bus, and whatever address is here that address is hit and then from that address whatever data is present that data comes through this data bus, and it comes to MDR. Now you see that the data bus is bidirectional because, we can read the data from the memory. So, the data is coming from the memory to MDR and for write we have to write the data.

So, from the MDR it will go through this data bus into memory. And along with this we also require some control signals, like read, write etc.

(Refer Slide Time: 04:34)



An Example Memory Module

- *n address lines* :: The maximum number of memory locations that can be accessed is 2^n.
- *m data lines* :: The number of bits stored in every addressable location is *m*.
- The RD/WR' control line selects the memory for reading or writing (1: read, 0: write).
- The chip select line (CS') when active (=0) will enable the chip; otherwise, the data bus is in the ***high impedance state***.

n-bit address bus

MEMORY (2^n addressable memory locations)

m-bit data bus

RD/WR'

CS'

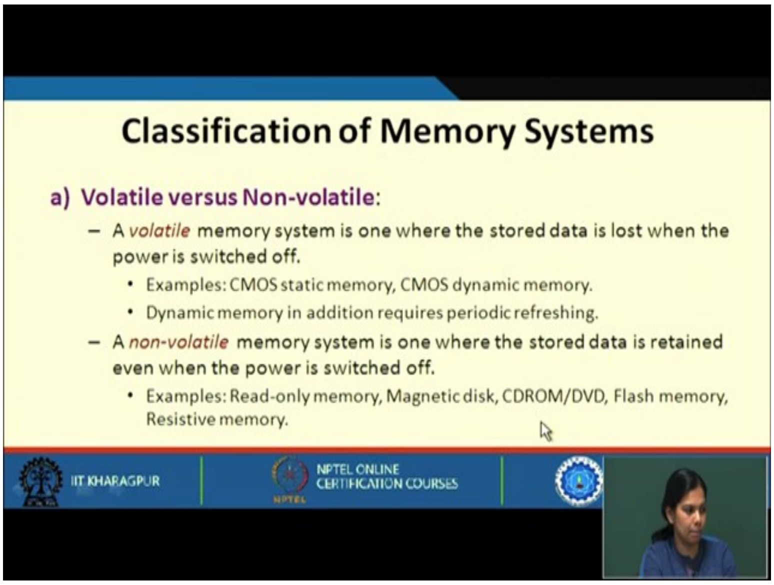
The memory size is specified as 2^n x m

So, if we have a n-bit address bus then, the memory addressable memory location will be 2^n. Like we already discussed if you have a 3-bit address bus then the total number of location will be 2^3. So, there will be 8 locations starting from 000, 001, we go on to 111. So n-bit address bus can have a maximum of 2^n addressable memory locations. And we can have a m-bit data bus. So, in that particular address the data that is present is m-bit, and m-bit data at a time can be transferred to memory. And we have other signals like read, write, and chip select.

We will be seeing that why chip select is required in course of time. So, the maximum number of memory location that can be accessed is 2^n. For m-bit data line the number of bits stored in every addressable location is m, and the read/write control signal selects the memory for reading or writing. So, for reading it is 1, for writing it is 0. As I said, chip select line; this is active low. So, it is active when it is 0. This will enable the chip when it is 0. Otherwise the data bus is in high impedance state. So, this memory module will not be selected in that case.

So, here we have n-bit address bus. We have 2^n addressable locations, and m-bit data bus. So, the total size of the memory is 2^n x m.

(Refer Slide Time: 06:38)



Classification of Memory Systems

a) Volatile versus Non-volatile:

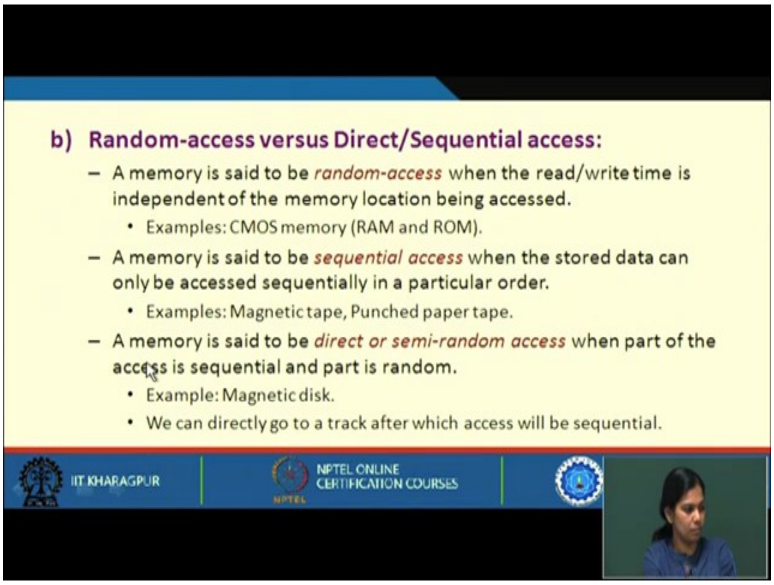
- A *volatile* memory system is one where the stored data is lost when the power is switched off.
  - Examples: CMOS static memory, CMOS dynamic memory.
  - Dynamic memory in addition requires periodic refreshing.
- A *non-volatile* memory system is one where the stored data is retained even when the power is switched off.
  - Examples: Read-only memory, Magnetic disk, CDROM/DVD, Flash memory, Resistive memory.

Now, classification of memory system; how we can classify a memory system? One way to classify memory system is volatile versus non-volatile, i.e. with respect to volatility. A volatile memory system is one where the stored data is lost when the power is switched off; that means, as long as the power is applied to it the data will remain, but as long as the power is taken off the data goes off; that means, it is volatile it goes off after the power is cut off. CMOS static memory and CMOS dynamic memory both these are volatile memories; that means, as long as power is supplied the data remains.

But in case of dynamic memory, even if we are supplying the power then also it requires periodic refresh. So, data cannot be retained for longer period of time. So, periodic refresh is necessary.

Now, what is non-volatile memory? A non-volatile memory system is one where the stored data is retained even when the power is switched off. So, where you will see such kind of non-volatile memory? We see that in read only memories where, once the data is there it retains. For example, magnetic disk, CDROM, DVD, flash memory, and some resistive memories. These are all non-volatile memories. So, even if the power is not supplied the data will remain.

(Refer Slide Time: 08:36)



b) Random-access versus Direct/Sequential access:

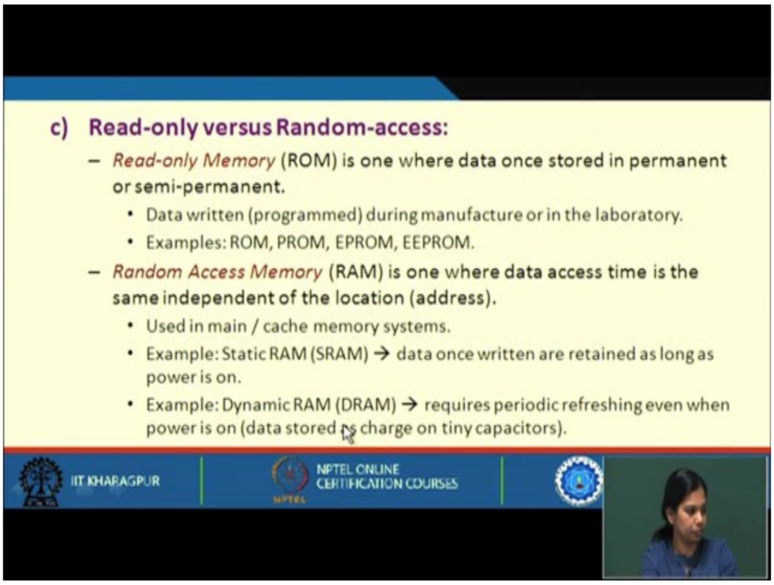
- A memory is said to be *random-access* when the read/write time is independent of the memory location being accessed.
  - Examples: CMOS memory (RAM and ROM).
- A memory is said to be *sequential access* when the stored data can only be accessed sequentially in a particular order.
  - Examples: Magnetic tape, Punched paper tape.
- A memory is said to be *direct or semi-random access* when part of the access is sequential and part is random.
  - Example: Magnetic disk.
  - We can directly go to a track after which access will be sequential.

Again we can differentiate a memory with respect to random access versus direct or sequential access. What do you mean by that random access? By random access we mean that, the read and write time is independent of the memory location being accessed. That means, you either hit location 0 or you hit the last location or the middle location, the access time is same.

So, whichever location you access the access time will be same irrespective of the location. The example is CMOS memory that is RAM and ROM, both are random access. And then what is sequential access? A memory is said to be sequential access, when the stored data can only be accessed sequentially in a particular order. Like, an example is magnetic tape. Here the data are accessed sequentially, one by one.

A memory is also said to be direct or semi-random access when, a part of the access is sequential, and a part is random; like your magnetic disk. Here we can directly go to a particular track, but after reaching that particular track we have to sequentially get the data one by one. This kind of memory is semi-random access, which is somewhat sequential, somewhat random.

(Refer Slide Time: 10:26)



c) **Read-only versus Random-access:**

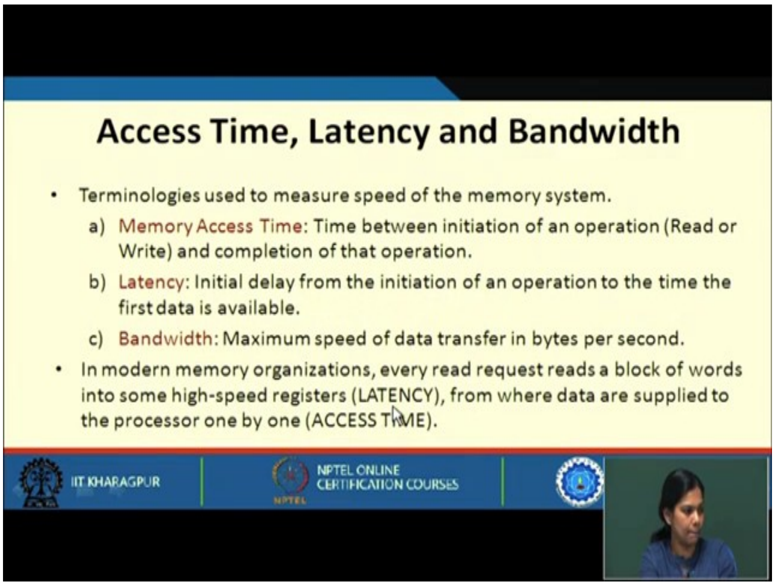
- *Read-only Memory* (ROM) is one where data once stored in permanent or semi-permanent.
  - Data written (programmed) during manufacture or in the laboratory.
  - Examples: ROM, PROM, EPROM, EEPROM.
- *Random Access Memory* (RAM) is one where data access time is the same independent of the location (address).
  - Used in main / cache memory systems.
  - Example: Static RAM (SRAM) → data once written are retained as long as power is on.
  - Example: Dynamic RAM (DRAM) → requires periodic refreshing even when power is on (data stored as charge on tiny capacitors).

Next let us see read only versus random access. What is read only memory? Read only memory is one where the data once stored, is permanent or semi permanent. What do you mean by permanent, what you mean by semi permanent? By permanent we mean that once we write into it, no changes can be made to it. And semi permanent means when we write, it remains, but if later we want to change it we can still do it. It remains permanent for a period of time. And again, if you want to change it we can change it and then it will again remain. So, the examples are PROM programmable read only memory, Erasable programmable read only memory, and electrically erasable programmable read only memory. So, these are all classes of read only memory, where the data are written during manufacturing, but can be changed later also.

So, ROM comes first, then PROM, then EPROM, and so on. Now random access memory is one where the data access time is same independent of the location. So, we access the first location or the last location the access time will be same. And where it is used? We will be talking extensively about your main memory and your cache memory. So, in both the memories such kind of memory, that is random access memory, are used. Some of the examples are static RAM.

Here once the data is written it retains as long as the power is supplied to it. And dynamic RAM is having the same feature of a RAM, but even if the power is supplied to it, it requires periodic refresh. So, periodic refresh is required, even if the power is supplied to it. And here the data is stored as charge on tiny capacitors. We will be looking into more details of static RAM and dynamic RAM in course of time.

(Refer Slide Time: 13:15)



Access Time, Latency and Bandwidth

- Terminologies used to measure speed of the memory system.
  a) Memory Access Time: Time between initiation of an operation (Read or Write) and completion of that operation.
  b) Latency: Initial delay from the initiation of an operation to the time the first data is available.
  c) Bandwidth: Maximum speed of data transfer in bytes per second.
- In modern memory organizations, every read request reads a block of words into some high-speed registers (LATENCY), from where data are supplied to the processor one by one (ACCESS TIME).

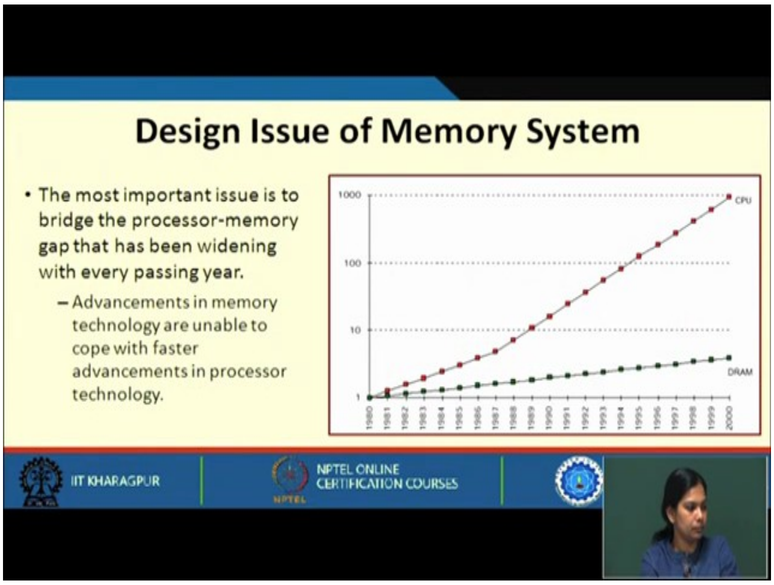
At this point of time, we need to know some of the terminologies that we will be using it very often. They are called access time, latency and bandwidth.

So, what is memory access time? By memory access time we mean that the time between initiation of an operation, either it can be read or write, and the completion of that operation. How much time it is required to access the particular data?

Next is latency. Latency is the initial delay from the initiation of an operation to the time the first data is available. Let me tell one thing at this point of time, that when we access a particular location in the memory, we do not just access or retrieve that particular data. We always transfer a block of data. That is why this latency is an important term because it will give the time required to access the first data. And then the subsequent data that are present can be accessed in a much faster rate.

So, latency is the initial delay from the initiation of an operation to the time the first data is available. And what is bandwidth? Bandwidth is the maximum speed of data transfer in bytes per second. In modern memory organization every read request reads a block of words into some high-speed register. That is when the first word is available. And from then the data are supplied to the processor one by one. So, the total access time will be depending on not only a single word, but a block transfer.

(Refer Slide Time: 15:45)



Design Issue of Memory System

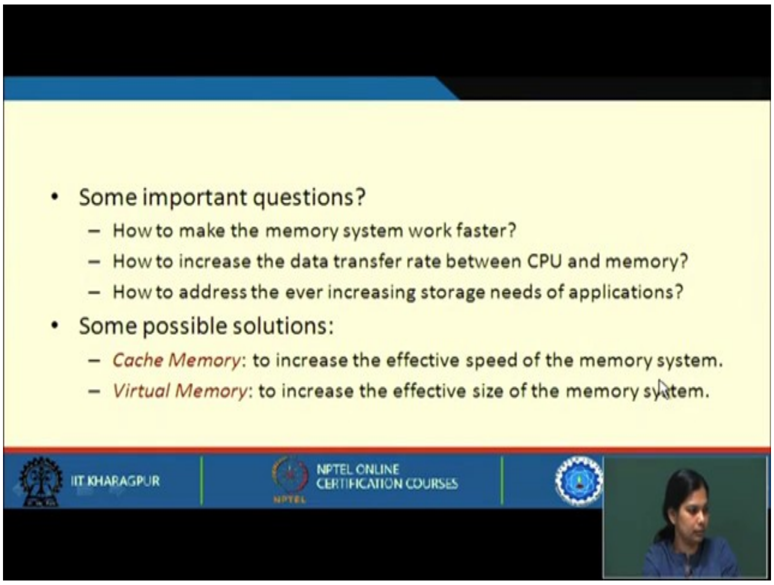
- The most important issue is to bridge the processor-memory gap that has been widening with every passing year.
  - Advancements in memory technology are unable to cope with faster advancements in processor technology.

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

Now, this graph I have already shown you earlier, while talking about evolution of computer systems, but now let us see the design issue of memory system. This red line shows the growth of processor in course of time and this green line shows the growth of memory technologies. Although you can see both are growing, but processor design is growing at a much higher pace, and memory design advancement is coming at a lower rate. But both are advancing. Technology is advancing in both, but this speed gap is steadily increasing. So, the most important issue is to bridge this processor-memory gap that has been widening with every passing year. Advancements in memory technologies are unable to cope with faster advancement in processor technology, but there are many techniques that are used to bridge this speed gap. At this point some important questions arise. How to make a memory system work faster? It has a limitation.

(Refer Slide Time: 17:32)

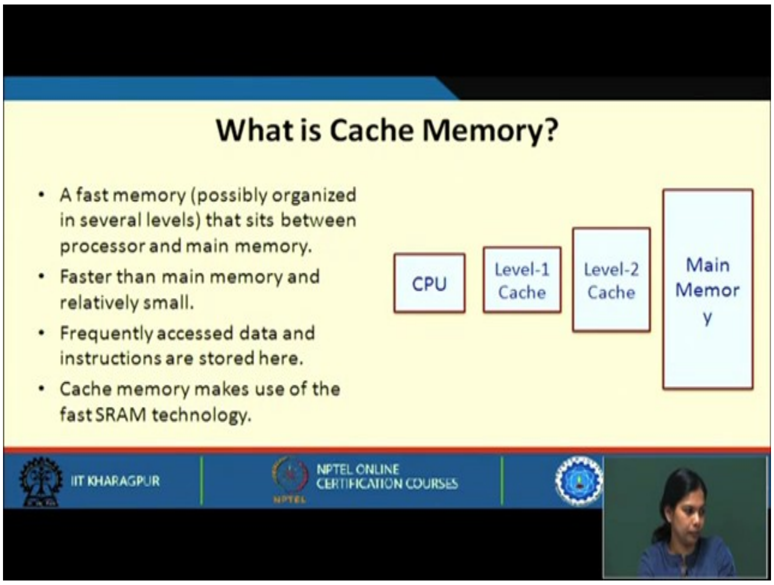


- Some important questions?
  - How to make the memory system work faster?
  - How to increase the data transfer rate between CPU and memory?
  - How to address the ever increasing storage needs of applications?
- Some possible solutions:
  - *Cache Memory*: to increase the effective speed of the memory system.
  - *Virtual Memory*: to increase the effective size of the memory system.

But how we can make it faster such that the processor and memory speed gap can be reduced, how to increase the data transfer rate between CPU and memory, the transfer of data; how it can be made faster, and how to address the ever increasing storage need of application? We need large memory as well. Not only we need faster, we also need larger memories. Because, there are various applications that require larger memory space.

So, we need to look into all the issues. First issue is how we can make this memory work faster. How we can have a larger memory and by all these thing how we can reduce the speed gap between processor and memory. Some possible solutions are cache memory and virtual memory. What a cache memory does, we will be looking into detail in later weeks. But what it does is, it increases the effective speed of memory system. And what virtual memory does? It increases the effective size of memory system. So, we will be looking into these in some detail in the later part of this week.

(Refer Slide Time: 19:19)



What is Cache Memory?

- A fast memory (possibly organized in several levels) that sits between processor and main memory.
- Faster than main memory and relatively small.
- Frequently accessed data and instructions are stored here.
- Cache memory makes use of the fast SRAM technology.

CPU

Level-1 Cache

Level-2 Cache

Main Memor y

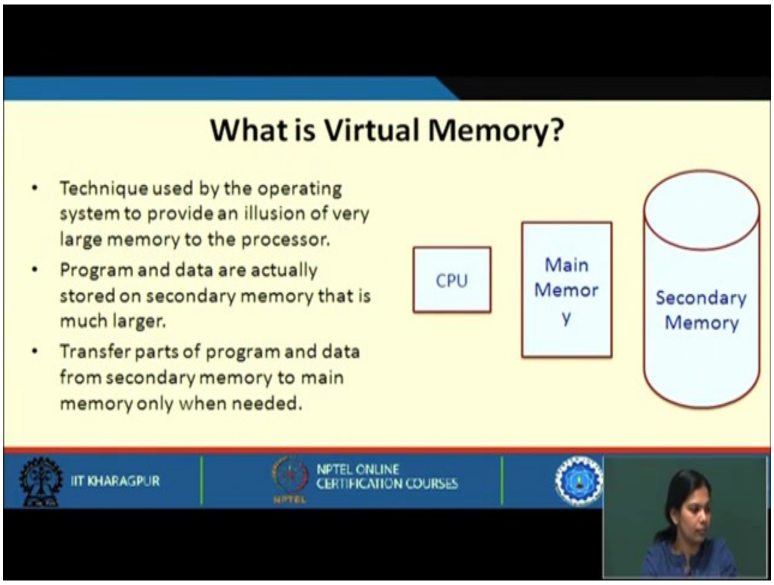
IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, very briefly what is cache memory? It is a fast memory that sits between CPU and main memory. And we can have many levels of cache memory. Why cache memory is in place? We will see this because of properties of computer programs called locality of reference. One is temporal locality of reference; other is spatial locality of reference.

So, we will see this in detail later. But for now let us understand that cache memory is a memory, which sits between CPU and main memory, and there can be many level of caches. But the cache memory cannot be very large. It is much smaller compared to main memory. So, frequently accessed data or instruction can only be brought here and executed. And what technology is used to build this cache memory? We use static RAM technology to build this cache memory.

(Refer Slide Time: 20:32)



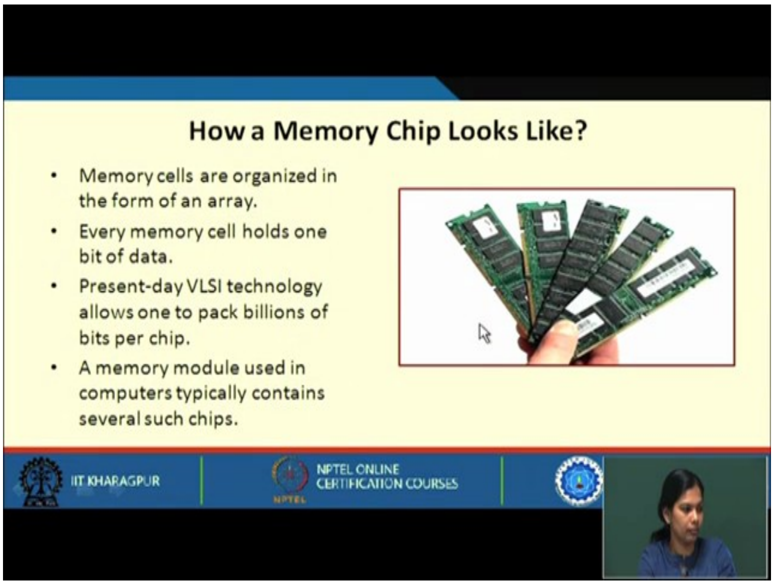
What is Virtual Memory?

- Technique used by the operating system to provide an illusion of very large memory to the processor.
- Program and data are actually stored on secondary memory that is much larger.
- Transfer parts of program and data from secondary memory to main memory only when needed.

Virtual memory is basically a concept that is used to give an illusion, that we have a very large memory space at our disposal. But actually we have space equal to main memory. But it gives an illusion to the programmer that you have a larger space to execute. So, the technique used by operating system to provide an illusion of a very large memory to the processor. Program and data are actually stored on secondary memory that is much larger. And data and instruction are brought into main memory as and when it is needed.

So, secondary memory is a concept where we say that we have a very large memory, but whenever we want to execute it we need to bring those data or instruction into main memory, and then it can be executed.

(Refer Slide Time: 21:44)

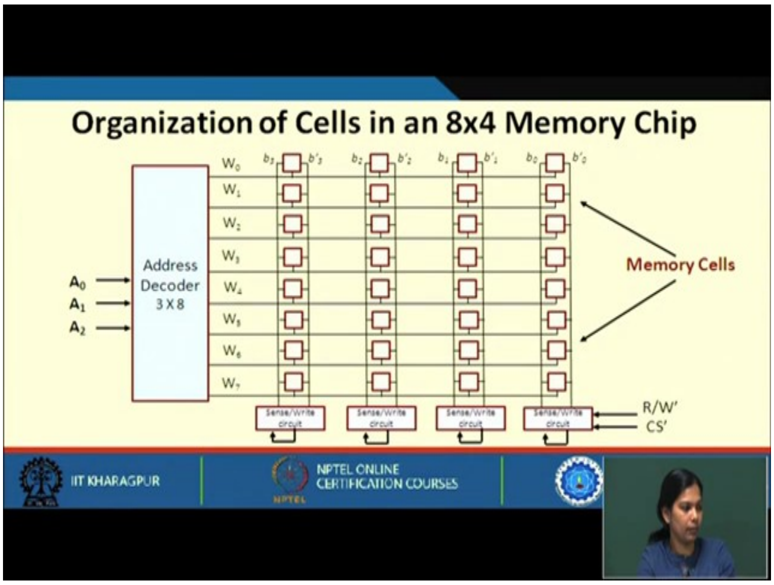


How a Memory Chip Looks Like?

- Memory cells are organized in the form of an array.
- Every memory cell holds one bit of data.
- Present-day VLSI technology allows one to pack billions of bits per chip.
- A memory module used in computers typically contains several such chips.

Now, let us see how a memory chip looks like. So, this is on a PCB (printed circuit board). So, these are memory. This is a separate memory modules that are placed. These memory cells are organized in the form of array. This may be a 4 GB memory and each having, say 1 GB, 1 GB, 1 GB like that. So, present-day VLSI technology allows one to pack billions of bits per chip. Memory modules used in computers typically contain several such chips. These chips are put into the memory slots that are present in the PC.

(Refer Slide Time: 22:47)



Organization of Cells in an 8x4 Memory Chip

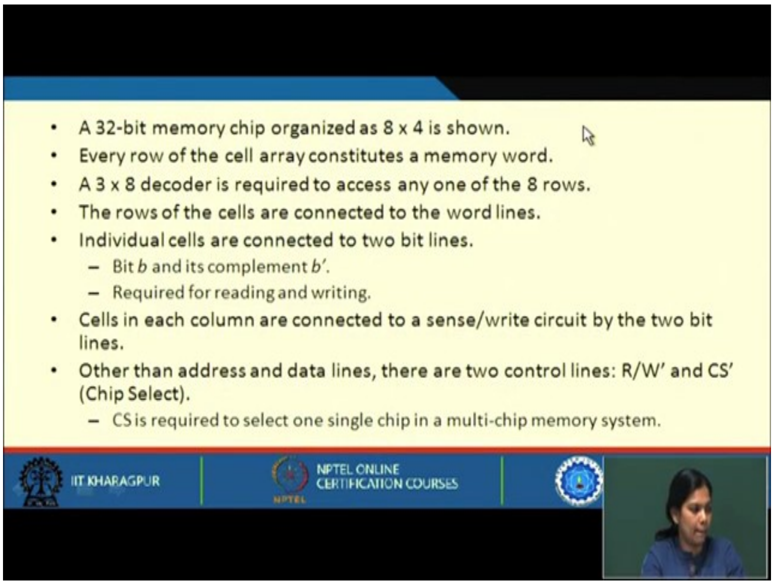
Now, let us see organization of cells in a 8 x 4 memory chip. So, this is a 8 x 4 memory chip. Let us see how it is organized. So, you have 8 rows; this is the first row, second row, third row, fourth row, and is the 8th row. And in each row, there are 4 bits that can be taken out. Let us consider this as a row. This row is connected with the word line. This is called word line: W0, W1, W2; these are word line. And individual cell is connected to 2 bit lines. One is b another is b-bar that is, complement of the other. And which is connected to the sensor write circuitry, and this sensor write circuitry is further connected to the data lines.

Here there are 8 rows. We have to select any one of the 8 rows. For that reason we require a 3 x 8 address decoder. So, these A0, A1, and A2 are applied to this address and then depending on this, say it is 0 0 0, the first word line will get selected. And then all the bits of the word line is transferred through this sensor write circuit to the data lines, if you want to read the data.

And suppose I want to write the data into the cells, then what will happen? The data present in these data lines that is coming from your MDR will get stored in these bits through this sensor write circuit. So, in this organization we can see that a 8 x 4 memory chip is there. So, the address is decoded with using a 3 x 8 decoder. And then each of the bit, each of the memory cells are connected to 2 bit lines. One is the complement of the other, which is connected to the sensor write circuitry and through the sensor write circuitry, is connected to the data lines. And we have to also supply these signals that are: either you want to read a data or want to write a data.

So, this is how the memory chip is organized. So, as I said a 32-bit memory chip is organized as 8 x 4 as shown in the previous figure.

(Refer Slide Time: 26:01)



- A 32-bit memory chip organized as 8 x 4 is shown.
- Every row of the cell array constitutes a memory word.
- A 3 x 8 decoder is required to access any one of the 8 rows.
- The rows of the cells are connected to the word lines.
- Individual cells are connected to two bit lines.
  - Bit *b* and its complement *b'*.
  - Required for reading and writing.
- Cells in each column are connected to a sense/write circuit by the two bit lines.
- Other than address and data lines, there are two control lines: R/W' and CS' (Chip Select).
  - CS is required to select one single chip in a multi-chip memory system.

Each row of the cell array constitutes a memory word. So, the entire row entire one word is the row. So, every row of the cell will constitute a memory word we need a 3 x 8 decoder to access any one of the 8 rows. And the rows of the cells are connected to the word lines. Individual cells are connected to 2 bit lines. One is b another is its complement, and it is required for reading and writing.

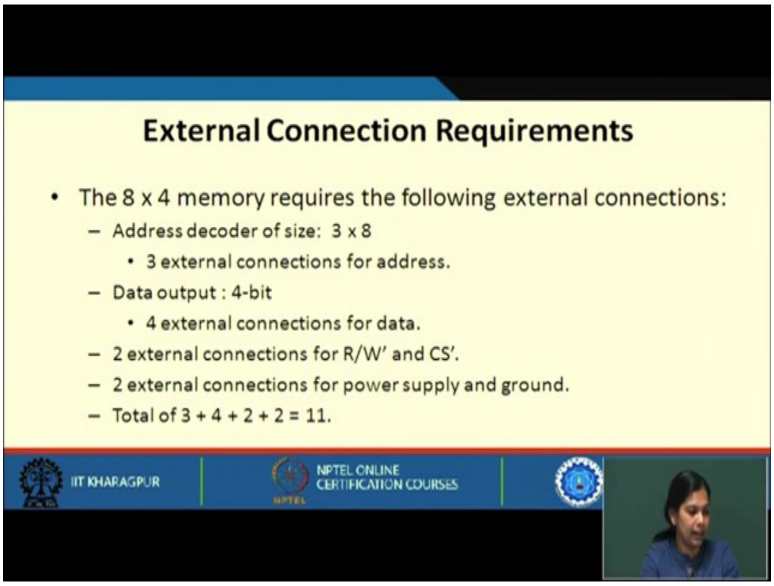
And cells in each column are connected to sensor write circuitry. So, this is one column, this is the next column, third column and fourth column. The cells in each column are connected to this sensor right circuitry. Other than the address and data lines there are 2 control lines, read/write and chip select. And why chip select is required? It is required to select one chip in a multi-chip memory system. We will be seeing this with examples later.

So, basically this read/write and chip select is connected, such that either it will specify that you have to read the content of any one of the words, or you have to write data into one of the words. Now in this diagram, how many external connections are required? What do you mean by external connection? Externally that is provided not within this memory chip. You can clearly make out that these address lines is A0, A1 and A2 are externally provided to this decoder, and it is decoded and a particular word is selected.

Now once you select a particular row, then all these bits will be transferred to the data line, through the sensor write circuit. So, it is connected to 4 bits of this. So, there will be

4 data lines through which this data will go. So, those are 4 more external signals that are required here. So, 3 for this address 4 for the data lines. And then you have 2 more signals, that is read/write and chip select, that should be also provided externally. Because the processor will tell either you have to read the data or you have to write a data. And there will be two more, that is, power supply and ground.

(Refer Slide Time: 29:06)

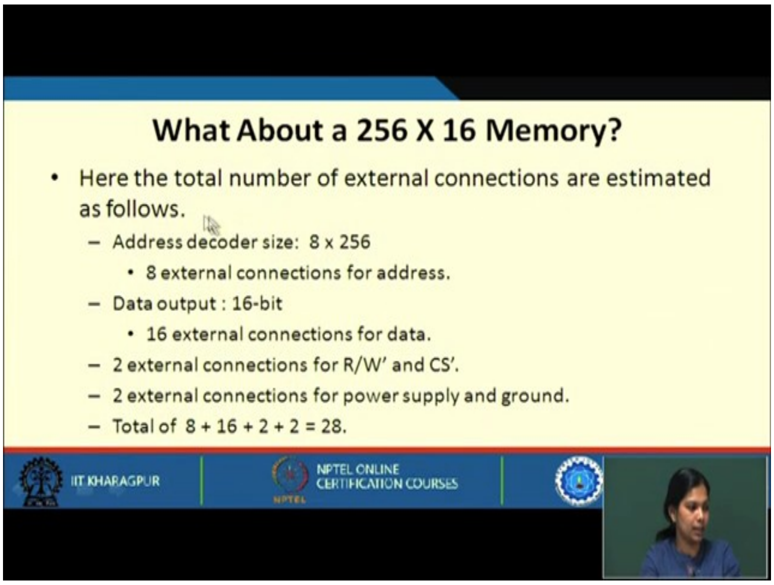


External Connection Requirements

- The 8 x 4 memory requires the following external connections:
  - Address decoder of size: 3 x 8
    - 3 external connections for address.
  - Data output : 4-bit
    - 4 external connections for data.
  - 2 external connections for R/W' and CS'.
  - 2 external connections for power supply and ground.
  - Total of 3 + 4 + 2 + 2 = 11.

So, external connection requirements that are there for this 8 x 4 memory, is 3 external connections for address, 4 external connections for data, 2 external connections for read/write and chip select, 2 external connections for power supply and ground. So, a total of 3 + 4 + 2 + 2 = 11 is required for this 8 x 4 memory chip.

(Refer Slide Time: 29:44)



What About a 256 X 16 Memory?

- Here the total number of external connections are estimated as follows.
  - Address decoder size: 8 x 256
    - 8 external connections for address.
  - Data output : 16-bit
    - 16 external connections for data.
  - 2 external connections for R/W' and CS'.
  - 2 external connections for power supply and ground.
  - Total of 8 + 16 + 2 + 2 = 28.

Now, let us see what about this 256 x 16 memory. There will be 256 rows. To select any one of the 256 rows, you require a 8 x 256 decoder. So, the address decoder size will be 8 x 256. So, 8 external connections for address will be required. Then the data output is 16. So, 16 external connections will be required to transfer the data, either to read the data or to write the data. Similarly, 2 external connections for read/write and chip select, and 2 for power supply and ground. So, a total of 28 external connections will be required.

So, we come to the end of this lecture where we briefly discussed about what is memory? How memory chips can be organized? And in the next few lectures, we will be seeing what kind of memory technologies are actually used to build this.

Thank you.