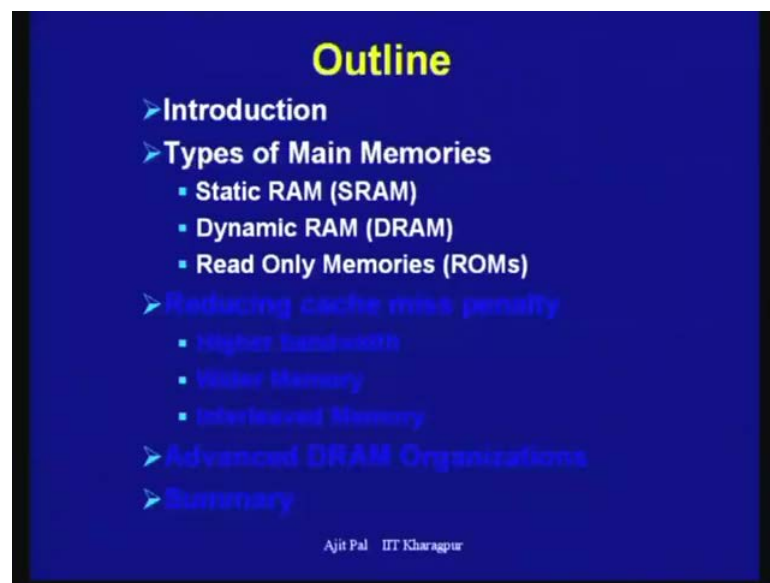


**High Performance Computer Architecture**  
**Prof. Ajit pal**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 27**  
**Main Memory Organization**

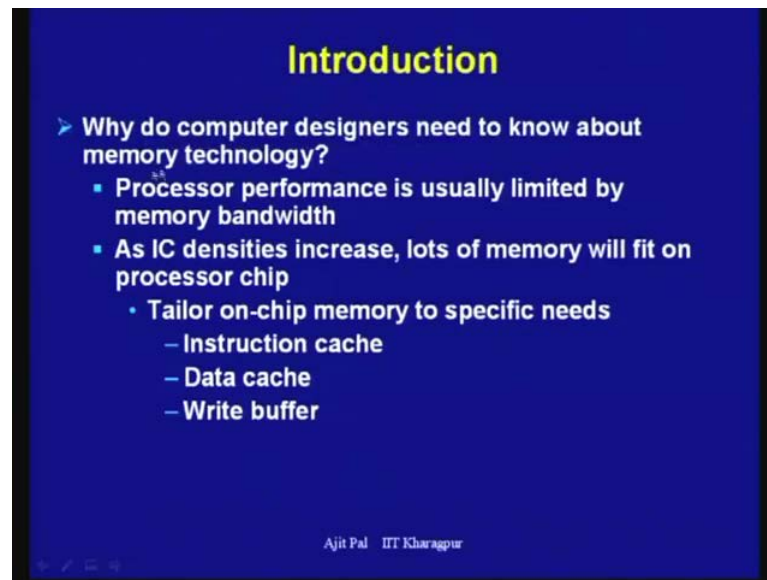
Hello and welcome to today's lecture, on main memory organization. In the last 3 lectures, we have discussed about various techniques, by which the performance of the cache memory can be improved, like its hit time can be reduced, miss rate can be reduced and miss penalty can be reduced. Now apart from cache memory, there is another memory which is also very important, which is known as prime memory or main memory. And today, we shall focus on main memory.

(Refer Slide Time: 01:37)



First, after giving a brief introduction, I shall discuss about the various types of main memory's that is used in computer systems, contemporary computer systems.

(Refer Slide Time: 01:51)



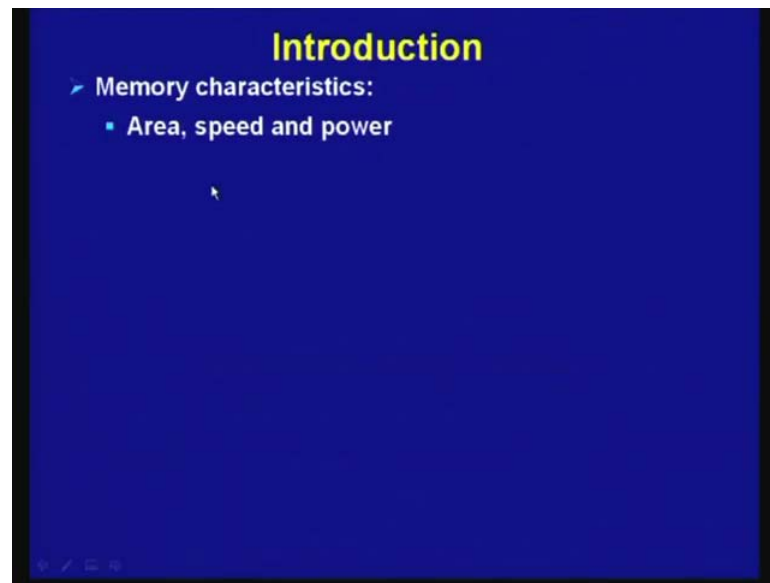
## Introduction

- Why do computer designers need to know about memory technology?
  - Processor performance is usually limited by memory bandwidth
  - As IC densities increase, lots of memory will fit on processor chip
    - Tailor on-chip memory to specific needs
      - Instruction cache
      - Data cache
      - Write buffer

Ajit Pal IIT Kharagpur

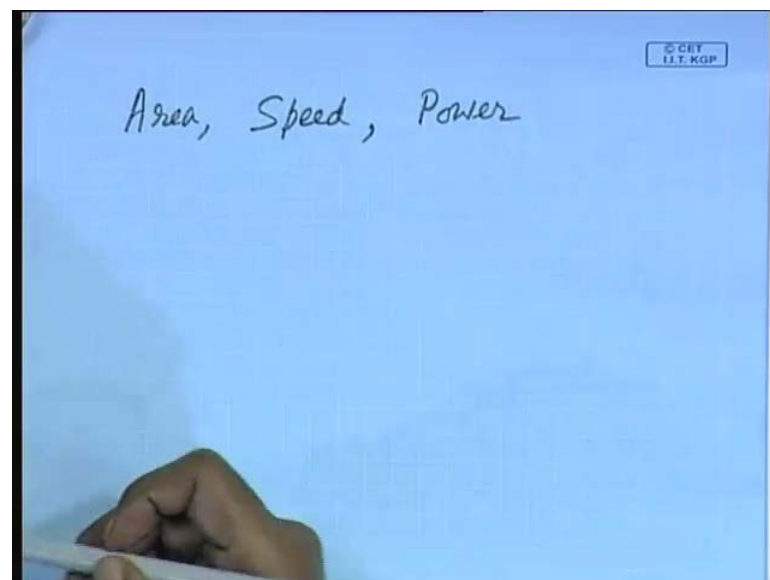
And question naturally arises, why do computer designers need to know about memory technology? Why does this question arise? The reason for that is, processor performance is usually limited, by the memory band width. We have seen that as IC densities are increasing, lots of memory will fit into the processor chip, so there will be many memory devices should be present in the processor itself - 1 chip. And also, they will be some memory devices will be off chip. Particularly, you can you can tailor the on chip memory to specific needs and you have seen the use of instruction cache, which can be on chip. Data cache, which can be on chip and another type of buffer, known as write buffer, that is also used to improve the performance and which is also on chip.

(Refer Slide Time: 02:55)



Now, let me very briefly give you the characteristics of memory. How do you really define or characterize memory devices? There are 3 important parameters. As you fabricate the memory devices, using semi conductor technology, number 1 is the area.

(Refer Slide Time: 03:17)



How much chip area it occupies? Then speed. How fast it is? Which is usually specified in terms of access time and the power, power dissipation is also very important in the present day context.

Because of several reasons: number 1 is you know, in imbibed applications, the devices are battery operated. So life of the, I mean usefulness of a particular embed system, is depend on how long the battery will survive. On the other hand, in case of your desktop and also in work stations and server, power deception is also important, because we have know that the performance, the reliability of the device is dependent on the power dissipation. So, these are the 3 important characteristics, which are important to in the context of memory devices.

(Refer Slide Time: 06:13)

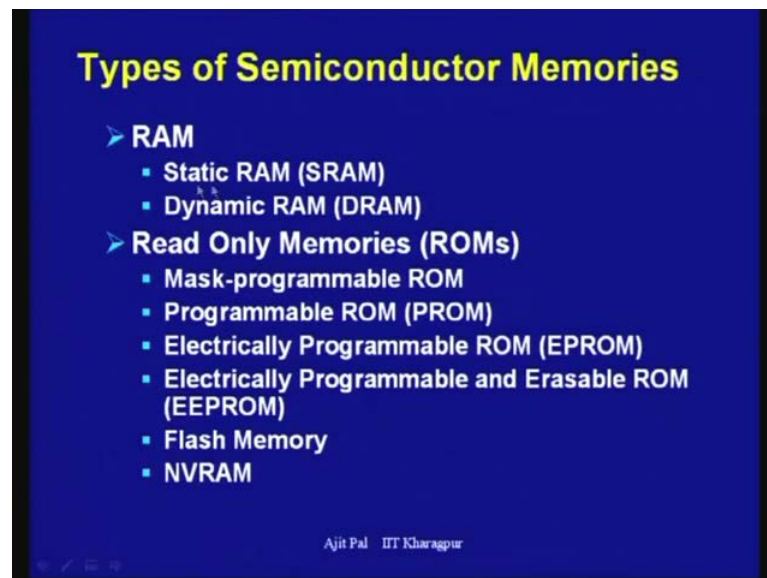


However, there are different ways by which we classify memory devices, particularly semi conductor memory devices. Number 1 of them is basic operation modes, the way they operate and in this context the memory devices can be divided into 2 categories: read-write or read only. In the read-write category, you have got SRAM- static RAM, dynamic RAM, EPROM, EEPROM, flash memory and so on. On the other hand, in the read only category, you have got mask-programmable ROM.

And there is another way you can categorize, that is your data storage mode, in which the data is stored. In this context, the basic category is volatile and non-volatile. Volatile means as long as power is there, information is available; as the power is removed information is lost. So, some memories like static RAM and dynamic RAM, which are volatile in nature. Similarly, you have got non-volatile type. Again there are 2 basic categories under this, under this non-volatile type. Read-write, where which you can

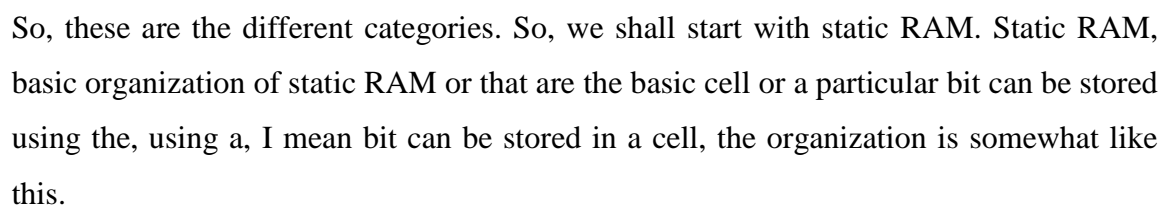
perform read as well as some write flash memory, EPROM, EEPROM. I shall discuss about them in more details. Then read only category, mask programmable ROM. Then another way in which you can categorize is by access mode. Access mode is, there are 2 basic ways. 1 is known as random. Random means irrespective of the location, the access time is fixed. Access time is not, does not change irrespective of the location of memory, within that chip. So, random access, that is why we called it random access and ROM, static RAM, dynamic RAM, EPROM, EEPROM, flash memory, they all these belong to random access memories. On the other hand, there are some non random access memories like serial access memory and content addressable access memory- CAM. So, they belong to nonrandom memory category, because the access time is dependent on the location of memory, with the, within the chip. So with this back ground, let us see what are the types of semiconductor memories we use in our present day computer systems.

(Refer Slide Time: 07:05)



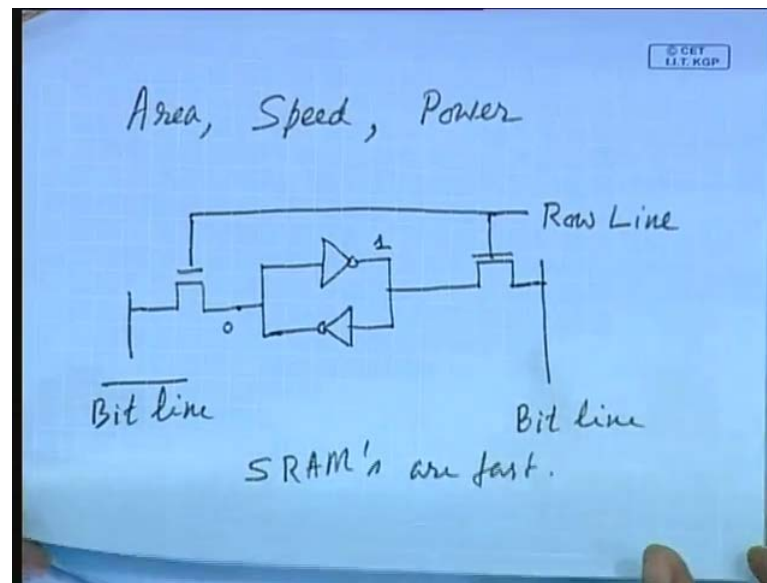
Basically as I said, you have 2 categories; random access memory and read only memories. And random access memory, there are 2 types. Static RAM, SRAM or dynamic RAM – DRAM. On the other hand, in the read only memory category, we have different names like mass-programmable ROM, programmable ROM. Mask programmable ROM is called ROM in short and programmable ROM is called PROM in short. And electrically programmable ROM, in which you can do the programming and re programming is allowed in both. In programmable ROM, it is programming can be only once, but electrically programming can be done more than once.

(Refer Slide Time: 08:37)



So, these are the different categories. So, we shall start with static RAM. Static RAM, basic organization of static RAM or that are the basic cell or a particular bit can be stored using the, using a, I mean bit can be stored in a cell, the organization is somewhat like this.

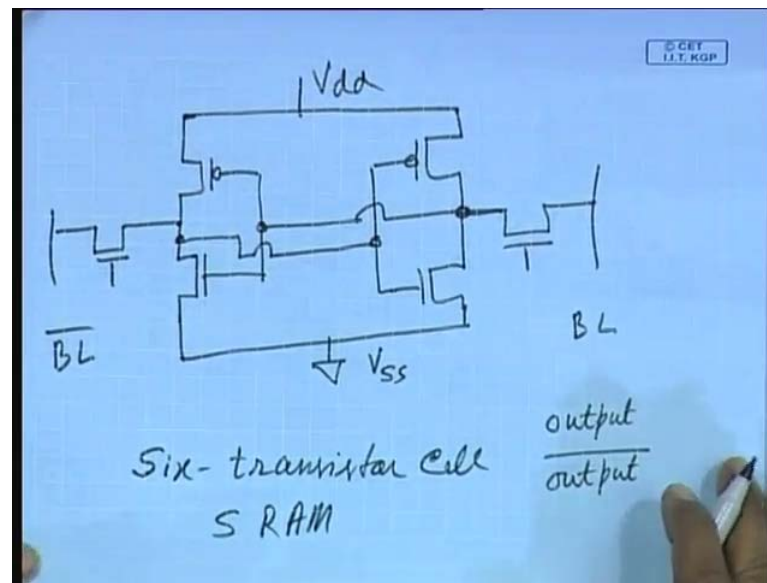
(Refer Slide Time: 08:50)



The way it is realized is, 2 invertors connected, are connected back to back, then you have got 2 transistors. And when this is connected to what is known as row line and these are connected to what is known as bit line and this is bit line bar okay. So, this is bit line and bit line bar. Now, in this case you can see, there are 2 invertors connected back to back.

This invertors can be realized by using CMOS technology, as usual. And suppose this whenever say, I mean this is 0, this will make it 1 and this 1 will drive this particular invertor and it will make it 0. So, there will be a kind of re generative action, that means as a particular point becomes 0, the output of invertor become 1, that 1 drives the other invertor, which forces the other point to become 0. And that is how very quickly, the device will switch from 1 state to another state. And that is the reason why the SRAM s are, this type of memory is known as static RAM, are fast. Now you can realize, I said you can relies the static Ram by using CMOS technology. In that case, you will use the invertors, you know, by using 2 transistors that means 2. This is 1invertor. This is a inverter and another invertors is this one.

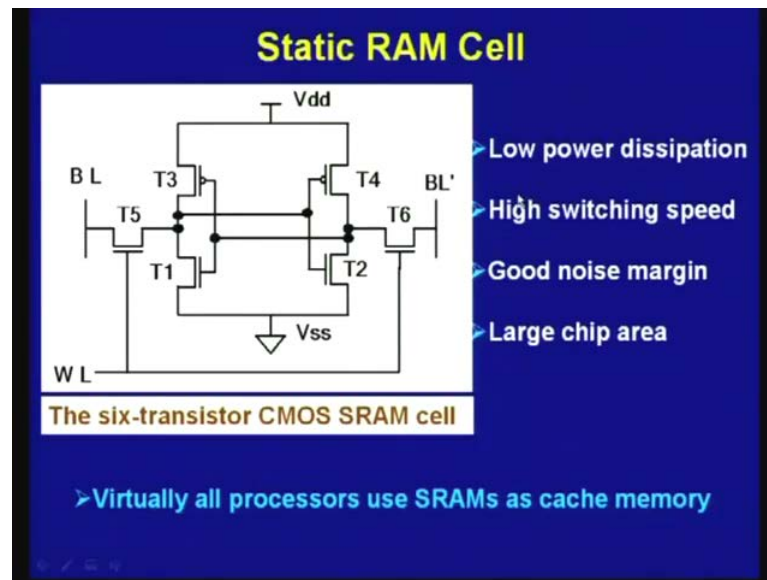
(Refer Slide Time: 12:45)



And then, this is connected to ground or  $V_{ss}$ , the substrate. So, these 2 invertors are connected back to back. This is connected to  $V_{dd}$  and we will use 2 additional transistors to connect, to take the output, which are known as bit line and bit line bar. And this is, this is input, so this has to be connected to this output and this input is to be connected to this output. So, you see 2 invertors connected back to back and 2 transistors are used to take the outputs from this point, so you get both the output and output bar. That means the normal output and it is complement, so both are available for particular cell. And this is the typical 6 transistor cell of static RAM. So this particular Static RAM cell, provides you low power dissipation.



(Refer Slide Time: 12:50)



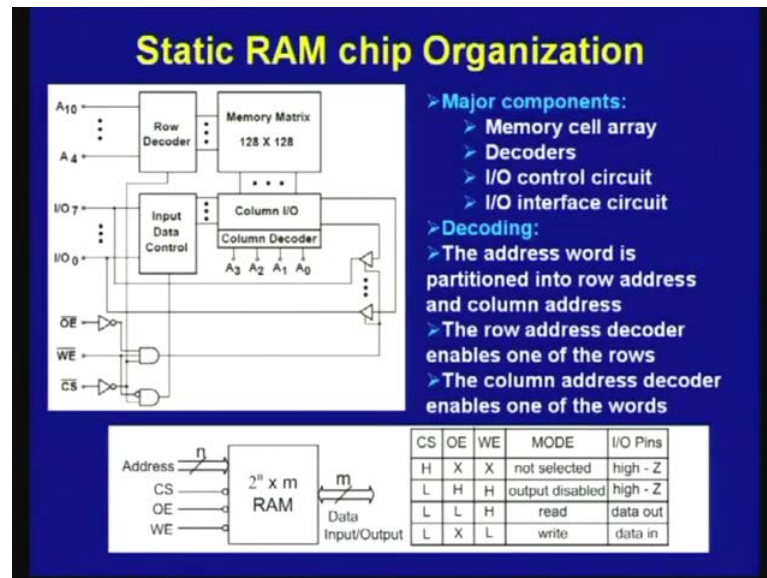
Why low power dissipation? As you can see, since this is a CMOS inverter, when this transistor is on, this transistor is off. Here also, when this transistor is on, this transistor is off. So there is no static current flow. So, as a consequence the power dissipation is quite small, in this in this types of circuits. However, you have to take the outputs by using 2 additional transistors. So, you require 6 transistors and I have already told, it provides you high switching speed. The high switching speed comes from the regenerative action that takes place, which I have already explain. It also provides good noise margin, the reason for that is the outputs switches from rail to rail, that means we did it to 0. I mean 0 you get 0 volt, very close to 0 and the one you get very close to  $V_{dd}$ .

So rail to rail switching you get and that gives you good noise margin. And also, it gives you large chip area, because you know this is a standard CMOS technology can be used and you give the chip area that can be obtain is quite high, I mean large chip in area can be realized. So, that means you can put lot of a cells on a single chip, and now a day's, virtually all processor use SRAM s as cache memory.

I have already mention about the, we have discuss about the cache memory, without mentioning the technology that is being used in analyzing cache memory. And the cache memory is, cache memory uses this SRAM circuits for the realization with the chip. Now, one severally mediation of static RAM is the number of transistors that is required, to realize a cell. So, for storing a single bit, you require 6 transistors. And obviously

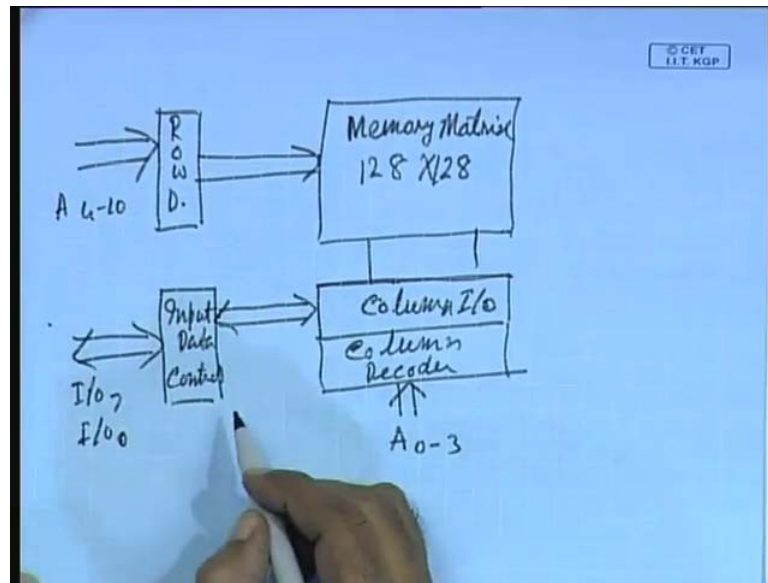
whenever you are thinking in terms of 2MB, 2 megabyte or 6 megabyte or now a days you know computer system have got several gigabyte of main memory. In such a case, the total number of transistors will be very large and we have to reduce the, I mean we have to find out some alternative way by which the packaging, I mean the number of canisters required is more and this is how you can realize a RAM chip.

(Refer Slide Time: 15:32)



You can see, you have got a memory matrix 128 by 128. And this is the row decoder, this is the column decoder. What is being done, the memory is organized in terms of in 2 dimensional matrix. So you have got, say 128 by 128 memory cells are present here.

(Refer Slide Time: 18:04)



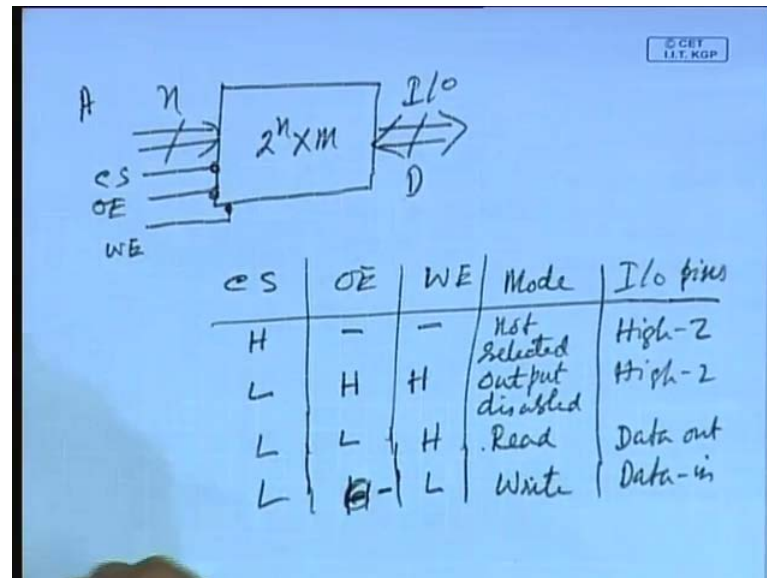
So, this is your memory matrix. And then, here you apply some of the address lines, say, some of the address lines, this is known as row decoder and row decoder is output applied. And also you have got column decoder. Here is your column decoder and to which you apply some other address lines, that means the lower address line you have apply here, say A<sub>0</sub> to A<sub>3</sub> and here you apply A<sub>4</sub> to A<sub>10</sub>, that means and this is the column I/O. This column I/O, here you have got sense amplifier and other things which communicates, which receives information, from the, which interacts with the this memory matrix.

And then you have got the input data control, input data control here you have the outputs are taken here; so here you will get say I/O<sub>7</sub> seven to I/O<sub>0</sub>. This means, you are storing 8 bit data in this, so 8 bit data comes out through these lines. And of course, you will require some additional inputs like as it is shown in this diagram, output enable, write enable, chip selects. These are use for controlling the devices. That means and so the major components that is present here is memory cell array, this is the memory cell array. Then you require 2 decoders; column decoder and row decoder, I/O control circuit. This is the I/O control circuit column I/O control circuit and the I/O interface circuits.

So, this is column I/O interface circuit is connected here, so through the same lines you get the input as well as you get the output. So, these lines are by directional. On the other

hand, the address lines are unidirectional. And you have got several control signals, output enable means whenever this output enable is, output enable is lower, this is low active when it is 0, then only output is available otherwise it is tri-stated.

(Refer Slide time: 21:36)



So, this chip can be represented in this way; here you have got  $n$  at these lines. In addition to, you will be having chip select, you will be having output enable and also write enable. These are all low active and here you will be having  $2^n$  into  $M$ , assuming that they are  $M$  lines available at the output. So, this is the basic organization, so this are the address lines, this are the data lines and here you can see that the control within this chip is performed with the value of this signal  $CS$  output enable and write enable. And when this chip select is high, as I said, it is low active represented by the symbol; when it is high irrespective of these lines, you will get the output will not be selected so you will be, it will not be selected. And I/O pins, these are essentially the I/O, pins data lines will be in the high impedance state, so high end.

Now, only when the chip select is low, this the low active and then, you can have other variations like output level is high, right level is high. In this case, output is disabled. Again in this case, it will be high  $Z$  and when the output level is low and this is high, then you will get the output. So, this is essentially read mode operation, that means the output, at the output you will get the data that is being stored, so this is data out. And similarly, whenever it is low and this is low, this is high. Output enable irrespective of this, that

means this can be independent of this write in enable. Then it is a right mode and you will get data in. So, this is how the chip is controlled and the organization is shown. So, the address is word is partitioned into row address and column address. Row address decoder enables one of the rows and the column address decoder enables one of the words and that is how the communication takes place, with the outside world. So this is the basic static ram chip organization.

(Refer Slide Time: 21:58)

**Static RAM Read/Write Operations**

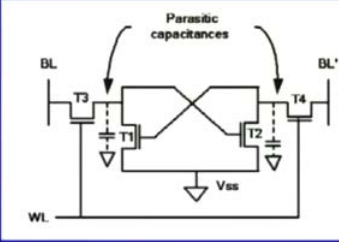
- **Write:**
  1. Drive bit lines ( $\text{bit} = 1, \overline{\text{bit}} = 0$ )
  2. Select row
- **Read:**
  1. Precharge bit and  $\overline{\text{bit}}$  to Vdd
  2. Select row
  3. Cell pulls one line low
  4. Sense amp on column detects difference between bit and  $\overline{\text{bit}}$

Ajit Pal IIT Kharagpur

And write operations is done by driving the bit lines 1 and 0 and selecting the row and that will perform the writing. On the other hand, read will involve pre charging the I/O line, the bit lines, bit and bit bar to Vdd and both the lines precharged to Vdd. So it is best on precharge logic. And then you have to select the row, by providing the row address and the cell pulls 1 line low, depending on data being stored there, depending on what is stored, it will pull down one of the low. That will be sensed by the sense amplifier on column and sense amplifier is on and column detects the difference between bit and bit bar. So there is kind of differential amplifier, which access sense amplifier and then that difference is reflected at the output and you get a output. So, this is a nut cell, how the static RAM read write operation takes place. And now let us focus on dynamic Ram.

## Dynamic RAM Cells

### ➤ Four-transistor dynamic RAM cell



- In the quest for smaller chip area, four-transistor DRAM cells emerged
- All RAMs are volatile
- Moreover, dynamic RAMs gradually lose information even when power is not turned OFF
- Gate capacitances store information
- To retain information the cells must be refreshed periodically
- There is marginal area advantage over six-transistor static RAM

As I mention, the number of transistor in static RAM is quite large – 6. So, if you want to reduce the number of transistors, then we have to go for some other techniques and that has lead to dynamic RAM cells. So it started with and 4 transistor dynamic RAM cell, as we shall see, this dynamic RAM technology has involved over the years, so starting with 4 transistor cell to single transistor cell. For example, in this particular diagram, you have got 4 transistors.

So, essentially those pull up transistors, that means pull of transistors are not present here, but you have got 2 transistors T1 and T2, which are acting as the basic cell, storing information and 2 additional transistor, those past transistor T3 and T4, for getting the output in bit line and bit line bar. So, in the quest for smaller chip area, 4 transistor DRAM cells emerged and this is also volatile in nature. But where does the information is stored? Information is not stored in the flip flop. This is not acting as a flip flop, information is stored in the gate capacitances. Gate capacitances; for example, although this capacitance is shown here, this is essentially the capacitances of this transistor.

So gate capacitances of these transistors, these are parasitic capacitances, you are not connecting any additional capacitances and this capacitance capacitor is also corresponding to this particular gate. So, in this gate capacitances, you are storing information in the form of charge. So storing charge and obviously one of the capacitors will be discharged, if this transistor is on, if T2 is on, this capacitors will be discharge. And if this transistor is off, this transistor will be charged. So that charge and discharge signal, I mean which one will be charge that will come from this bit lines. That means, if



it is 1 this 1 will come here, that will make this transistor 1, so which will pull down this capacitor, this charge to 0.

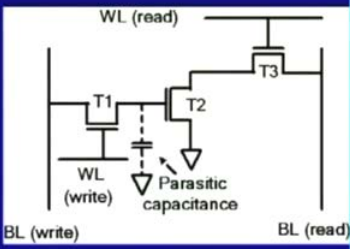
So, information that is being stored in a capacitor, will be present that how long that will be available. If you turn it on, information will be lost, that will make it that makes it a volatile. What say suppose load some information, one is stored in this particular capacitor some charge, one corresponds to storage of some charge, 0 corresponds to no charge. So that is being stored, question is how long it will information will return. It has been found that, the I mean any capacitor has a property known as, it will it will leak the charge will be leaked away. So, here also the information that is being stored in this capacitors, that will be lost because of the leakage of the charge. And that is the reason why another technique is used, that is known as refreshing.

To retain information, these cells must be refreshed periodically. That means, the refresh controller will read the value and as the voltage goes down because of discharge, it will again discharge the capacitor, where one is stored. So whenever you realized in this form there is marginal area advantage over 6 transistor RAM. So you do not have to not get much benefit because you are saving 2 transistors, but that the capacitors would be quite relatively large, so that information is stored. In other words, these transistors will be of bigger dimensions. So, you do not have much advantage and that has led to another type another type of cell, known as 3 transistor cell.

(Refer Slide Time: 27:26)

### Dynamic RAM Cells

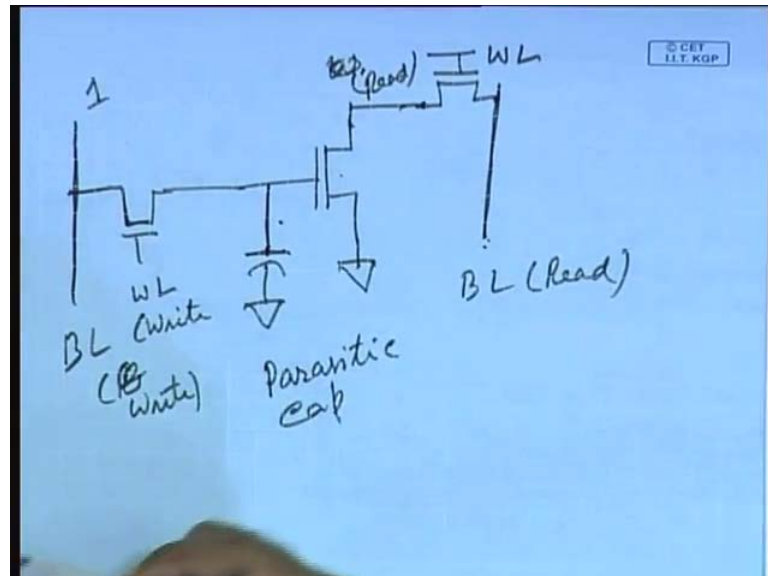
➤ **Three-transistor dynamic RAM cell**



- A 3-T DRAM cell uses only one transistor as storage element
- The gate capacitor of T2 is used to store information
- Two additional transistors are used for read and write access
- It is faster than the 4-T cell
- Reading the 3-T cell is nondestructive
- No additional capacitor is required for storage purpose
- The fabrication process is compatible with that of CMOS

So in this 3 transistor cell, here you are not using 4 transistor, but you have a single transistor and the gate capacitance of this transistor is used to store information.

(Refer Slide Time: 29:57)



And then you are taking the output, with the help of again 2 transistors. This is your, this is your, this is connected to that word line and this is connected to bit line and this is also connected to bit line. So, this is bit line read and this is bit line write. So in this case, the operation is different. Here you are applying write line write signal and for read you are using, this is for write, this is for, this is for write and this is for read.

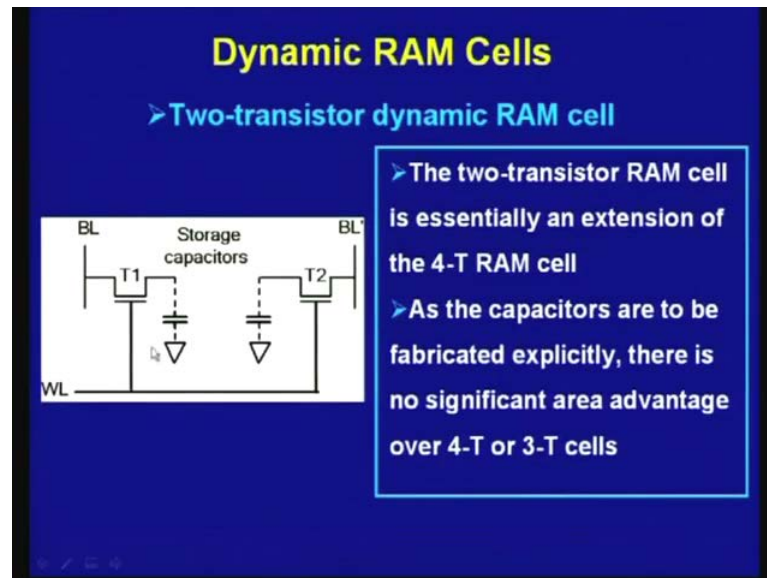
How it is happening? Say, you can see the bit line is connected here. If it is 1 and as you make this transistor on, this capacitor will be charge will charge through this transistor. This is how you can write information in this gate capacitance of this transistor. So this is the paratactic capacitance. And here the charge is stored and one is store here. Now, for reading purpose, you make this transistor is piece transistor on. That means, whenever you are reading this transistor is on, if you one store here, this transistor will be on and so this will be 0. You will get a I mean one store here, then you will get if you perform this read operation, this will be connected here.

If one is store, you will get 0 here, compliment of heat you are getting here. And if it 0 is stored, this transistor is of and so you will get 1 here. So you can see, you get the compliment of that output, corresponding to what is being store here. And in this case, the gate capacity is use to store information, 2 additional transistors are use for read and



write as I have told. And it has been found that it is faster than 4 T cell and reading the 3 T cell is nondestructive. Here you know, since it is isolated by this get, the operation is not destructive. That means, reading can be done with the help of this 2 transistors and also the fabrication process is compatible with CMOS. So, this is another technology, that was use for realizing dynamic RAM cells.

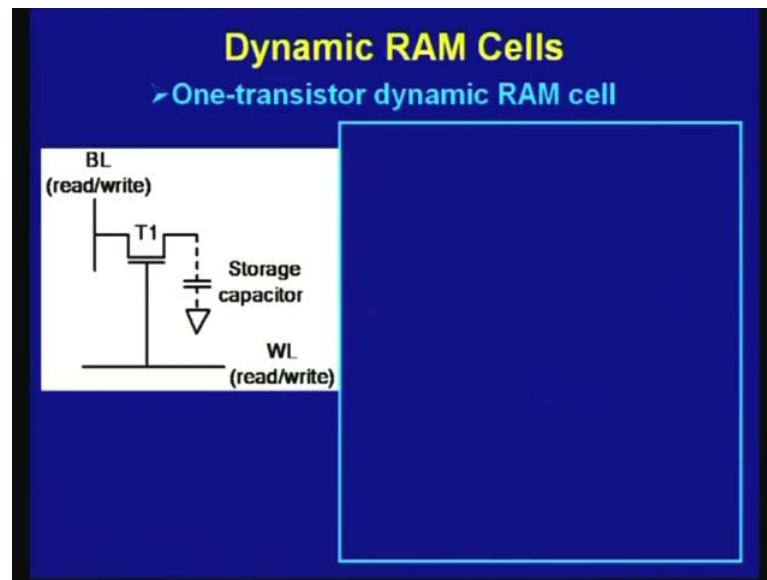
(Refer Slide Time: 30:44)



Now, let us come to a significantly different realization, where you are using 2 transistor dynamic RAM cell. In 2 transistor dynamic RAM cell, you can see there is no a get capacity present here. So you have to explicitly fabricate these capacitors. So, the 2 transistors RAM cell is essential and extension of 4 T RAM cell, with the exception that you are not using the great capacitance to store information. Because you are using some to explicit the 2 capacitors are explicitly fabricated, which will store the information and you get both bit and bit line bar information, depending on what values are store.

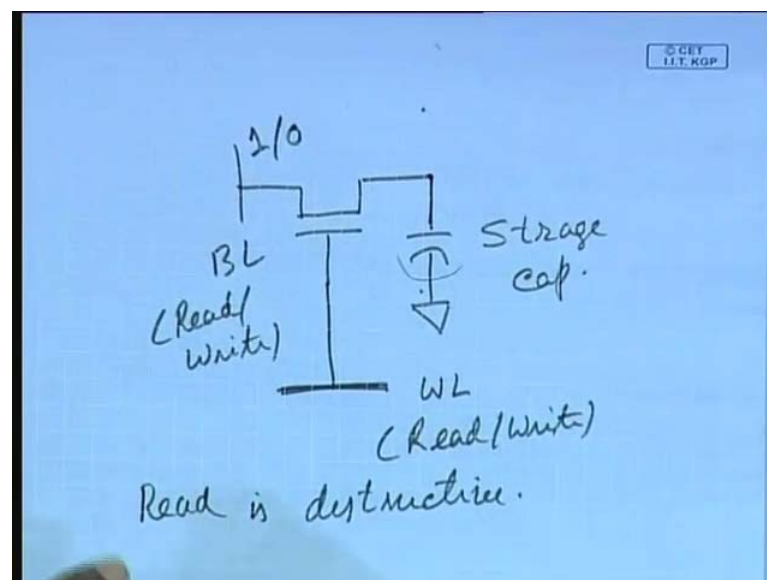
So, if this is 1, then this will be charge and if this is, this bit line is 1, bit line bar will be 0 and that will be this charge capacitor. So both 0 and 1 become complimentary value will be the stored and as you select this 2 transistors, by activating this watt line, then you will get the bit line and bit line bar at the output. So, this is the 2 transistor dynamic RAM cell. However, this was also not very popular, this is not very popular in the present context.

(Refer Slide Time: 32:11)



We have gone for what is known as 1 transistor dynamic RAM cell. So, here you can see you have got only 1 transistor.

(Refer Slide Time: 34:42)



So, only 1 transistor is here and that capacitor which is explicitly fabricated is present here. This is the storage capacitor and you have got 2 or say 2 lines. This is your write line, for read and write. Word line, for reforming read and write and this is the bit line. This is also used for read and write. So, this is word line, this is the bit line. How it works? So, suppose the bit line is 1, you have to store 1 and then whenever you activate

this word line, this transistor is on and it will charge the capacitor. On the other hand, whenever it is 0, charging the capacitor means it is 1.

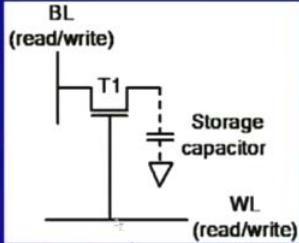
Now, if it is 0, then of course, this capacitors will discharge, through this transistor and there will be no charge, as you write it. So, writing is I mean writing is done in this way, by charging this capacitors or discharging it by turning the, this I mean word line select word lines transistor on. However, let us see how do you perform read. So, whenever you are performing read, then again you are activating this transistors. Now the charge store available here, will be available on the bit line. So it has been this is found down that the this is no longer isolated by a gate.

That means, this particular capacitor is no longer isolated by gate. It is directly connected to the bit line. That means, all the charge that is being stored here, gets transferred to the bit line. Of course, they will be sense amplifier this difference in the charge on bit and word line bar, will help you to get the output, but unfortunately this capacitor will get discharge. So in other words, read is destructive; read operation is destructive.

(Refer Slide Time: 34:50)

### Dynamic RAM Cells

#### ➤ One-transistor dynamic RAM cell



- A significant improvement in the DRAM evolution was to switch to 1-T DRAM cell
- One additional capacitor is explicitly fabricated for storage purpose
- To store '1', it is charged to  $V_{dd}-V_t$  and to store '0' it is discharged to 0V
- Read operation is destructive
- Sense amplifier is needed for reading
- Read operation is followed by restoration operation
- Small chip area (256Mbit on a chip)
- Process technology not compatible with CMOS

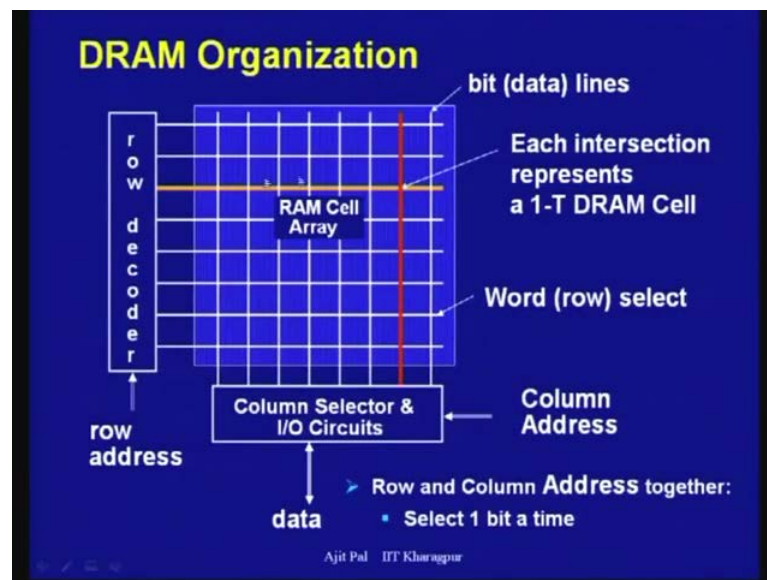
➤ Virtually all desktops or servers used DRAMs for main memory

So, this is the significant improvement in the DRAM evolution, by switching to this 1 transistor dynamic RAM cell. So, only 1 additional capacitor is explicitly fabricated for storage purpose, in addition to single transistor. To store one, we charge to  $V_{dd}$  minus  $V_t$ . So, you are not charging to  $v_t$  because, since it charge to this transistor, as you know there will be voltage drop of  $V_t$ . So, you will get  $V_{dd}$  minus  $V_t$  here and that will it will

be charge to either  $V_{dd}$  minus  $V_t$  or it will discharge to 0 volt. And as I explain, read operations is destructive and you require sense amplifier for the purpose of reading, because this signal has to go to sense amplifier, to amplifier the signal. And so read operation is followed by restoration operation. That means, if there was 1 was store there, you have to write back 1.

You have to restore the information and that is why the dynamic RAM s are slower. And of course, the advantage here is, you require very small chip area. You can go off to say to 256 mega bit on a chip or more. Now a days, you can have large number of transistors on a single chip. And another unfortunate part is that, the process technology is not compactable with CMOS process technique technology. So, up to you know 3 transistor, we have seen the dynamic most technology a circuits RAM s and transistor and Dynamic RAM was compactable with the CMOS technology. But 2 transistor 1 transistor cells are not comparable with CMOS technology. So, you have to go for an incompatible technology to fabricate the chip. However, because of large packaging density, this is very popular and widely used nowadays. And these particular type of dynamic RAM s are virtually used in all desktop, servers as main memory.

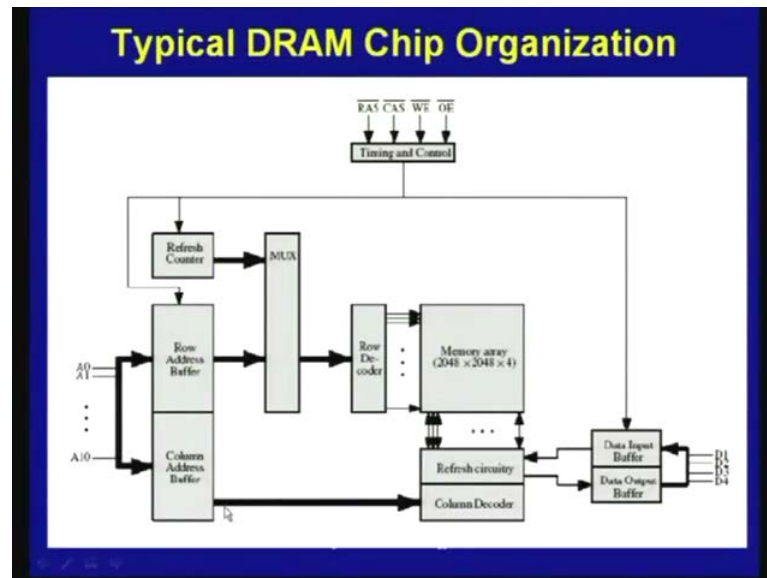
(Refer Slide Time: 37:12)



And this is the typical organization, here also you have got row decoder, column selector and I/O circuit and at the intersection you have got 1 cell, 1 transistor cell and this are the words lines and this are the column lines. And here you get data, so row and column and

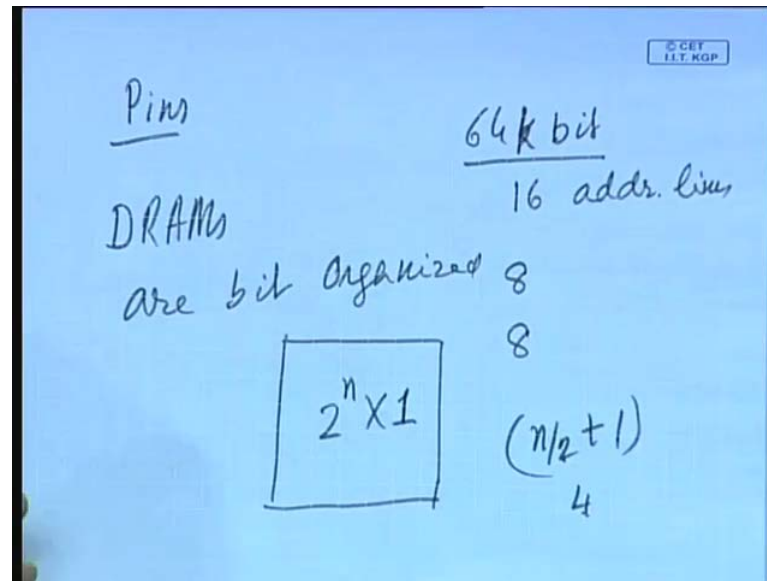
this together are applied and which actually select a particular bit and you get a single bit at the output.

(Refer Slide Time: 37:44)



And you can see, this is the typical DRAM organization of a commercial chip. So, here you are using several address lines,  $A_0$  to  $A_{10}$ . Now, another important change, that is the address is given, same as this lines are used for applying to row address and column address. Earlier we have seen, that row address and column address here you can see the row address line and columns address lines are applied. I mean, which are lower order bit go here and higher order bit goes here. But in dynamic RAM, that is not done. The reason for that is, that you want to reduce the numbers of pins.

(Refer Slide Time: 38:36)



The total number of pins increases, as the size of the memory increases. So, suppose you have got 64 kilo byte or 64 K bit assume it is bit or bytes 64 kilo bit. So, you will require 16 at this lines. So, instead of applying the all this lines together, we use 8 line, then another 8 eight line. So, this is for 64 kilo bit, but whenever you have got much larger capacities, says 256 kilobits or 1 gigabit, then the numbers of lines will be more. So, you are saving 8 eight pins by applying the address, I mean row address and column address separately. So, that is what is being shown here.

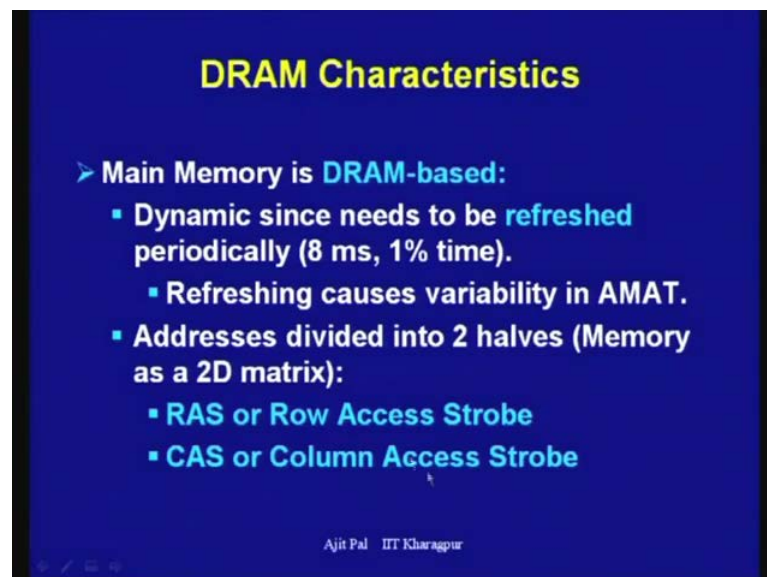
For example, in this case total number of address lines is 20, but you are using row address and column address and here are using 2 separates buffers, 1 for storing the row, another for storing the column. So, row address and column address are coming from the same lines. And as a consequence, you will require 2 signals. Row address and row address, I mean select and column address select. So row address select and column address select, this 2 inputs are there whenever you are applying row address, then you have to active row address select.

And whenever you are applying column address, then you have to activate the column address select. So, in addition, to write enable to and output able 2 additional controls are necessary to control the chip. So and you have got a multiplexer, which is needed for refreshing the memory. That means, the refreshing take place row by row and that is the reason why you are using a refresh counter.

So refresh counter output is multiplex with the row address, either the row address is applied or when you are refreshing then the refresh counter output is applied to the row decoder and which selects one of the rows in the memory and that entire rows get refreshed. Not one at a time, but entire rows get selected. So, this is the refresh circulatory and column decoder output, column decoder column address is applying here to the column decoder and you get the 2 lines data input buffer, data output buffer. And through this, you get the go get the output data lines.

So, in this particular case, it is shown that there are 4 data lines. But in the nowadays, always the DRAMs are bit organized. What do you really mean by that? That means, you will be it is organized in the form of  $2^n$  into 1. So, number of data lines is 1 and again  $n$  is divided by  $n$  by 2.  $n$  by 2 that is row address and column address, so you are total number of pins required is  $n$  by 2 plus 1. I mean for address and data of course, for control and other purposes you will require 4 more, so total numbers of lines will require  $n$  by 2 plus 1 plus 4. So, the number of pins that is required is significant reduce by using this bit organized memory and also by using row address and column address.

(Refer Slide Time: 42:19)



**DRAM Characteristics**

- Main Memory is **DRAM-based**:
  - Dynamic since needs to be **refreshed** periodically (8 ms, 1% time).
    - Refreshing causes variability in AMAT.
  - Addresses divided into 2 halves (Memory as a 2D matrix):
    - **RAS or Row Access Strobe**
    - **CAS or Column Access Strobe**

Ajit Pal IIT Kharagpur

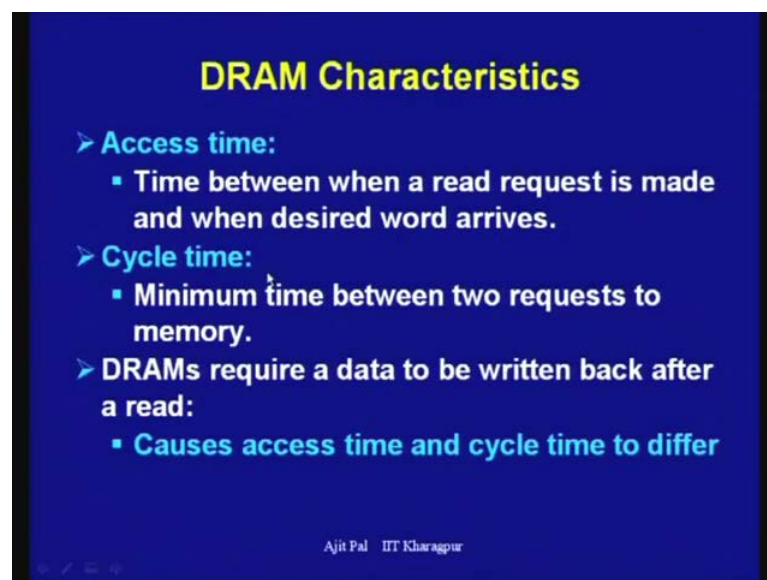
As I mention, main memories always dynamic RAM based, dynamic RAM it is called dynamics, since needs to be refresh periodically. So you may be asking, why the name dynamic. So to make the difference with static, since it has to be refreshed at the inter



periodically 8 millisecond or and that is why the name dynamic RAM. And this refreshing causes variability in average memory access time, so we have to seen in case of static RAM, we are there is no need for the refreshing. So, here there is the additional time that is required for refreshing.

And now that will contribute to the average memory existence because sometimes will wastes for the purpose of refreshing. And as I have already mention, the address is divided into 2 half's - row access strobe or A S and column access strobe; that is column I mean C A S. This 2 signals are require for C A S, row address row access strobe, column access strobe those signal are to be provided for giving the providing the addressing I mean in  $n$  by  $2$   $n$  by  $2$ .

(Refer Slide Time: 43:45)



**DRAM Characteristics**

- **Access time:**
  - Time between when a read request is made and when desired word arrives.
- **Cycle time:**
  - Minimum time between two requests to memory.
- **DRAMs require a data to be written back after a read:**
  - Causes access time and cycle time to differ

Ajit Pal IIT Kharagpur

Now the access time and cycle time is also different, in case of dynamic RAM. So, we know that access time is the time between when re request is made and when the desired word arrives. On the other hand, cycle time is a minimum time between 2 requests in a memory. So, dynamic RAM s require a data written back up after read, we have seen you have to do the refreshing. That is why the access time and cycle time is different in context of dynamic RAM, which is not true in the context of static RAM.

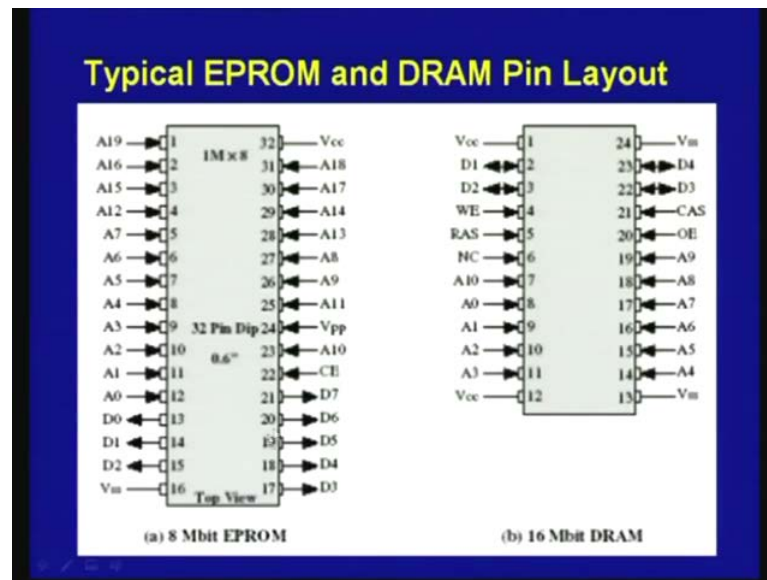


(Refer Slide Time: 44:22)



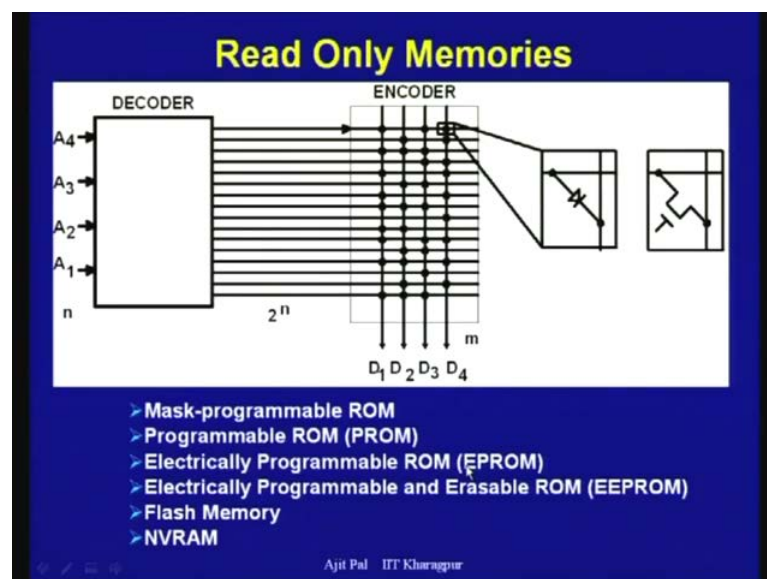
So, this is the brief history of dynamic RAM. It was patented back in 1968 by Dennard. But it was not commercialized, significantly cheaper than SRAM. 1 transistor and 1 capacitor versus 6 transistors and 1 bit is represented by high or low as we already seen. And one very important aspect is static RAM is I mean it is significant slower than static RAM. So, SRAM is used for on chip memory like caches and scratchpads. And dynamic RAM is always is off chip. So, dynamic RAM, it is off chip because of 2 reasons; number 1 is the technology is different. Processor is fabricated by using CMOS technology. Dynamic RAM is not compatible with that .So it has to be off chip, separate chip.

(Refer Slide Time: 45:15)



So, this is the pin layout of this one, so just to compare here you can see 8 megabyte EPROM. So, with the help of 32 pin chip, dual in line pin package, you can get 8 megabit EPROM. On the other hand, you can see only by using 24 pin chip, you can have 16 megabit dynamic RAM. Here as I mention, you have got I mean multiplex address, that A0-A9. Those 2 address lines are applied in multiplex manner selected by RAS and CAS. And of course, there are 4 data lines present here. It is not bit organized, but now a days, since the capacity has increased, it is bit made bit organized ok.

(Refer Slide Time: 46:12)



Now, let us focus to read only memory. So far, we have discussed about the RAM technology. Now let us consider that the read only memory. Now may be asking why do we require read only memory?

You see whenever you turn the computer on, then there will be nothing in the RAM. So how this CPU will work? CPU is a damped device. It has to get instructions from somewhere. So, in other words, from where it will come, it has to be stored in a memory which is, which is a non volatile and that is what is stored in ROM. Just like in you know when a child is born, the child must be born with some built in intelligence and with the help of which the child starts taking, I mean communicating with the environment, mother, relatives and parents and starts trying making the difference. So, it is born some born with intelligence, then subsequently it accrues intelligence from teacher, father and environment nature. So, you require read only memory in your system.

For example, in your computer system, the basic input output system is stored in the ROM. And this is the basic organization of read only memory. As you can see, despite its general name it has got 2 distinct components. 1 is known as decoder, as you require in any memory system and decoder is there in addition to the decoder you require encoder. So, you will require decoder and encoder and you can see the decoder depending on the number of lines, if there are  $n$  address lines, it will generate  $2^n$  lines to the output of the decoder. And at any particular instant, one of them will be active, as you know that is the basic function of a decoder. So, it will select one of the lines and that will go to the encoder.

Now, in the encoder as you can see, the you can have dots corresponding to different lines. Now you can store 0 and 1 in this at the crossing of the lines. In the encoder, you can see this in some places there are dots in some crossings there is no dots, indicating that there is some device; it may be diode or a transistor and sometimes micro wire that is being storage used at each of these junctions. And depending on what is present at this junction, the different types of ROMs are realized. In case of mask programmable ROM, why it is called mask programmable ROM?

You know, in the factory, a mask is created and that decides at what junction there will be duo diode transistors and what junction they will be no duo diode transistors. So, depending on what you have to store for example, here 1, 0, 1, 1 that is being stored and

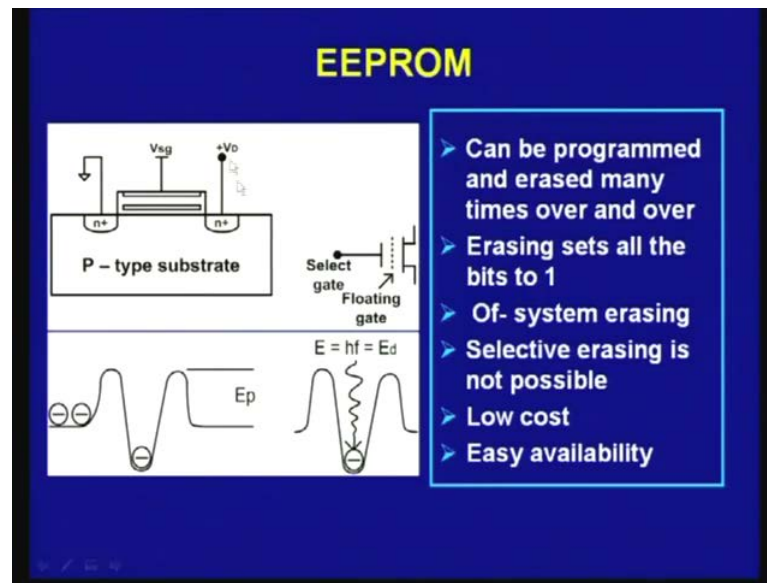
on the second line you have store 0,1,0,1. So, that means presence of a device makes it I mean allow you to store 1; why because whenever the decoder is activated, this line is 1, then that 1 can gets transmitted to the to the column line, through this through this link may be diode or it transistor or it can microware.

So, in the factory this are fabricated that is why it is called mask programmable ROM or simply ROM. Now, in case of programmable ROM, you know the mask programmable ROM is fabricated in the factory. So, this can be done only when it is produced in mass scale, that means you require million search millions of such devices only then you can order a ROM which can be fabricated in a factory. So, that is used only when you require a particular device, particular type of encoding in large numbers. On the other hand, this programmable ROM is user programmable.

User programmable means the user can do the programming, that programming can be done only once. How? Suppose you have got a micro ware connected here. So that micro wire can be burnt by programming. So, with the help of a PROM programmer, you can burn the connection between the column row line and column line and so, it can really be constricted, that is why you can do it only once, that is why in programmable ROM the writing can be done only once. However, reading can be done many times, that is why it is a programmable read only memory. Now the third type, electrically programmable ROM, in this case the programming can be done by electrical means.

So by electrical means the electrical means, the programming can be done, in such a case the device that is being use is little different. The most transistor is used and in the most transistor you know that in the most transistor there is a floating get, I will show in the next slide that will be used. For example, this is the technology that is being used, in case of electrically erasable and programmable ROM.

(Refer Slide Time: 52:06)



So, in case of sorry, this is in case of EPROM it is electrically programmable, but erasing is done by exposing 2 ultra violet light. So this is not really EPROM, but it is EEROM. one you will be there. So, can be program erase actually there are 2 techniques. In the first case, in case of EPROM, what is done; the programming is done by applying high voltage to the select gate and there is some electrons get accumulated in the floating gate And that floating gate and the high voltage that actually overcomes the barrier and electrons can be low put in this, put in this particular position. And that remains in those, I mean near the gate.

And then, for the purpose of, there is optical window which you have most of seen. The electron can be taken out and it will come out of the because of the energy that is being obtain from the from that from the ultra violet lines sources and electrons comes out. So, again it becomes 0, that means whenever you do the programming, you can say 0 is total. And whenever you do erasing, all this else becomes 1.

On the other hand, in case of electrically programmable, will be another variety electrically programmable and erasable ROM, you can do the programming and erasing electrically. How it is done? You can, this is what is written here, can be programmed and erase many times over and over. Erasing sets all the bits to 0 and so you require off system erasing, selective erasing is not possible and this is low cost and available that means erasing is done all the time in case of EPROM.

(Refer Slide Time: 54:29)

### Flash Memory

- A form of EEPROM:
  - However, allows a block to be erased or written in a single operation (in a flash).
- Floating Gate Avalanche-injection Metal Oxide Semiconductor (FAMOS)
- Electrons are trapped in a floating gate.
- Writing a byte requires creating a new block:
  - Old block is copied along with the byte to be written.

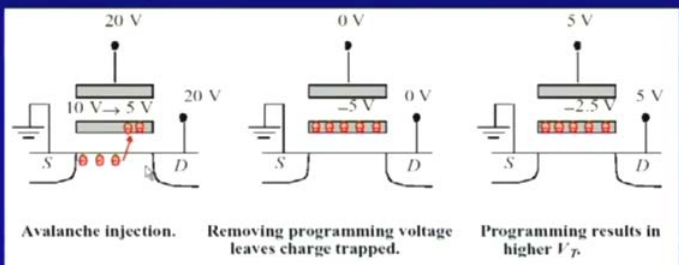
Ajit Pal IIT Kharagpur

But in case of electrical erasable and programmable ROM, which is the version of electric and programmable ROM, you can do the writing and reading selectively. So, this allows a block to be raised or written in a single operation in a flash. That means, instead of a single word, in case of electrically erasable and programmable ROM, writing is done word by, but in case it is done block by block. That is the different between EROM and flash memory. So, floating gate avalanche-injection metal oxide semiconductor is used for the implementation of a cell and electrons are trapped in a floating gate as I have already mention and this is shown here.

(Refer Slide Time: 55:10)

### FLASH Memory

- Performance:
  - Reads at speed of DRAM (~ns)
  - Writes like DISK (~ms). Write is a complex operation.

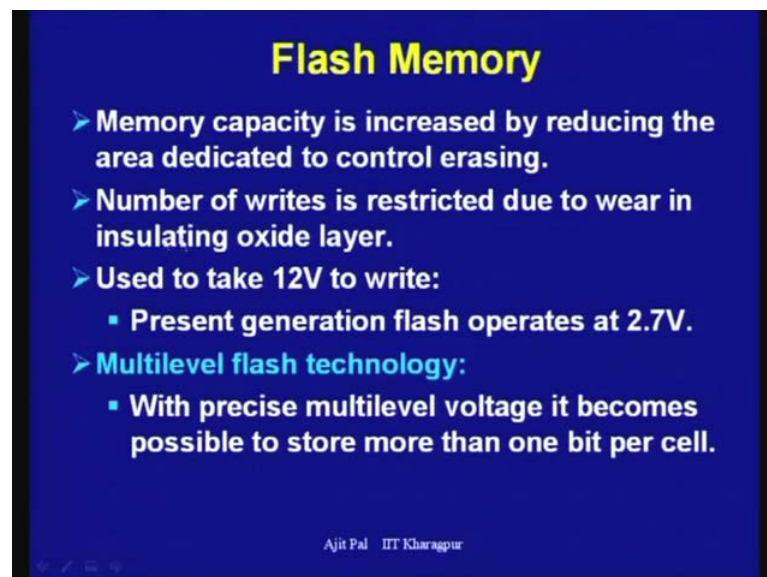


Ajit Pal IIT Kharagpur

So those electrons get trapped, you know trapped in this floating gate, as you can see here. Now by applying high voltage, you are trapping electrons are getting trapped here and that is being stored here. And then, if you remove that high voltage, the threshold voltage of device increases. So, for those transistors, I mean where those electron has been trapped, their threshold is high. For other transistors, where the electrons have not been stored, the threshold voltage will be low. So, that is we will that will be use for the purpose of reading and writing. And writing of bio request creating in new block, old block is copied along with the bit. And so this you can do block by block into pass memory.

So the reading can be done at the speed of dynamic RAM. So, at the speed at the speed of nanosecond do the reading. However, writing is quite complicated because you have to apply high voltage and that high voltage has to be applied for certain duration, so that electrons are get trapped and that takes the order of mille second. So writing is a complex operation even in flash memory. And memory capacity is increased by reducing the area dedicated to control erasing because instead of controlling byte by byte, you are controlling block by block.

(Refer Slide Time: 55:00)

A blue rectangular slide with a black border. The title 'Flash Memory' is at the top in yellow. Below it are five bullet points in white and blue. The last one is 'Multilevel flash technology:' in blue, followed by a sub-bullet in white. At the bottom, the name 'Ajit Pal IIT Kharagpur' is written in small white text.

### Flash Memory

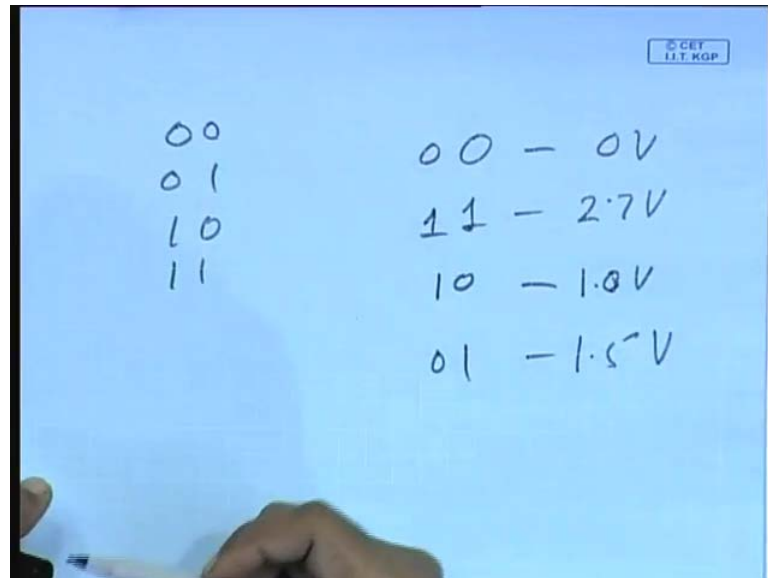
- Memory capacity is increased by reducing the area dedicated to control erasing.
- Number of writes is restricted due to wear in insulating oxide layer.
- Used to take 12V to write:
  - Present generation flash operates at 2.7V.
- Multilevel flash technology:
  - With precise multilevel voltage it becomes possible to store more than one bit per cell.

Ajit Pal IIT Kharagpur

So, number of writes is restricted due to wear in insulating where oxide layer, use to take to 12 fold to write. So present generation operates at 2.7 volts, that means when you are

in using the read only mode voltage that you require is 2.7 volt, but for the purpose of writing you require 12 volt. And now a days, you can go for multi level flash technology.

(Refer Slide Time: 57:27)



What did you really mean by multiple flex technology? So, this is the very interesting, we know that for 0 it can be 0 volt and for 1, it is let us assume 2.7 volt. Now what can be done, you can if you are allow to store an intermediate voltage, may be 1.3 volt, then what you can do, that intermediate voltage can be used for the purpose of storingsay 0 0, 0 1, 1 0, 1 1, 2. I mean 2 bits can be simultaneous store by storing by multi level voltages in those capacitors.

And different levels voltages and instead of storing only 1 bit, it will possible to store more than 1 bit, that means this may correspond to, this may correspond to 0 0. This may correspond to say 1 1. This may correspond to 1 0 and some order like that. That means intermediate voltages, you can store say 1 it can be 1.0 or it can be 1.5 that may correspond to 0 1. I mean just I am writing arbitrary, but this is the basic idea of multi level a flash memory technology. So this is possible, but not yet commercially realized.



(refer Slide Time: 58:38)

**Flash vs. EEPROM**

- EEPROM can write to one location or byte at a time:
  - Flash writes multi-KB blocks.
  - As a result flash memory is faster.
  - Also flash can be written in-system in contrast to EEPROM.
  - The control circuitry required for erasing is much less leading to higher capacity.

Ajit Pal IIT Kharagpur

So, here is a comparison between flash memory and EPROM. I have already explain that flash memory is faster because you are writings in terms of blocks. And flash can be written in system, contrast to EPROM. In EPROM cannot be written in the system, but flash memory can be written in n systems. So, controls architecture require for raising is much less that into high capacity of flash memory, as I have already told. Anyway so with this we have come to the end of our discussion, on different types of technology that is used in your realizing memory devices.

And in my next lecture, I shall discuss about how you can reduce the this memory access time essentially that will help you to reduce the miss penalty. So, just like improving the cache memory performance, we will like to improve the main memory performance. So what are the techniques that can use for improving the main memory performance that I shall discuss in my next lecture.

Thank you.