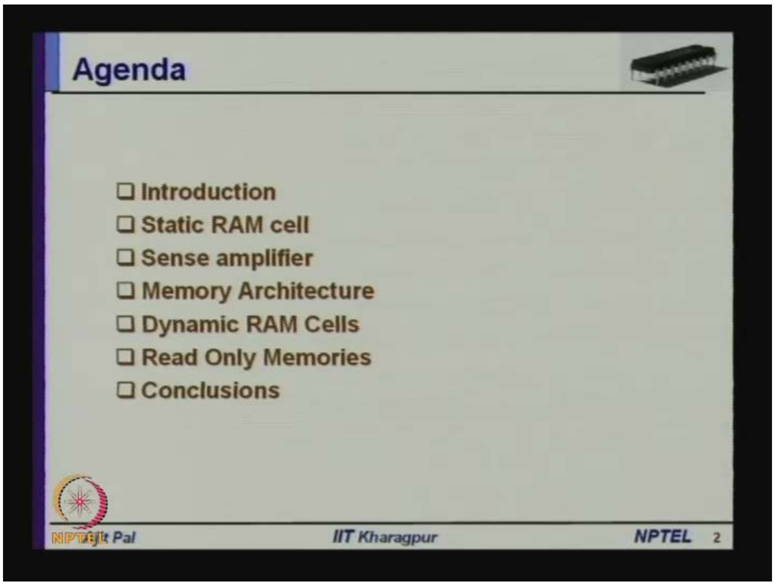
Low Power VLSI Circuits and Systems
Prof. Ajit Pal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture No. # 16
MOS Memories

Hello and welcome to today’s lecture on MOS memories. Since, the dawn of electronic era, memory devices have been found to be an integral part of electronic circuits and systems. You cannot think of any system without some memory, particularly the semiconductor memory has some kind of enabling can be considered as enabling technology of the computer technology; that means the computer technology has evolved based on the evolution of memory devices. So, without discussing about memory devices, our discussion on MOS circuits is not complete. So, today I shall try to give an overview of MOS memories, obviously it is a vast subject, you can get a complete book on MOS memories, but in a single class, a single lecture I shall try to give an overview of MOS memory devices.

(Refer Slide Time: 01:32)



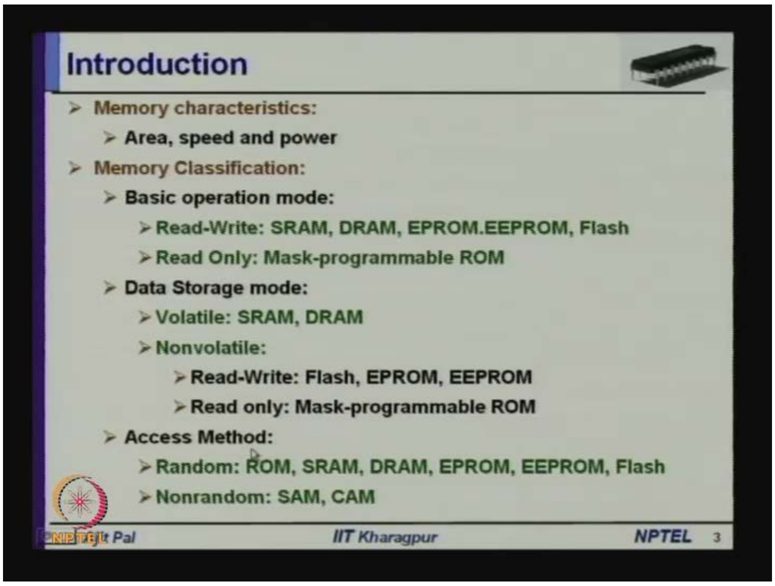
Agenda

- □ Introduction
- □ Static RAM cell
- □ Sense amplifier
- □ Memory Architecture
- □ Dynamic RAM Cells
- □ Read Only Memories
- □ Conclusions

NPTEL t Pal IIT Kharagpur NPTEL 2

And here is the outline of my lecture; introduction, it will be followed by static RAM cell, how a RAM cell can be realized by using MOS transistors. Then use of sense amplifier, then memory architecture, MOS memory architecture. Then we shall focus on dynamic RAM cells, read only memories, and I shall conclude my lecture with some concluding remarks.

(Refer Slide Time: 02:06)



Introduction

- Memory characteristics:
  - Area, speed and power
- Memory Classification:
  - Basic operation mode:
    - Read-Write: SRAM, DRAM, EPROM.EEPROM, Flash
    - Read Only: Mask-programmable ROM
  - Data Storage mode:
    - Volatile: SRAM, DRAM
    - Nonvolatile:
      - Read-Write: Flash, EPROM, EEPROM
      - Read only: Mask-programmable ROM
  - Access Method:
    - Random: ROM, SRAM, DRAM, EPROM, EEPROM, Flash
    - Nonrandom: SAM, CAM

t Pal IIT Kharagpur NPTEL 3

How do you characterize a memory. It has been found that there are three important parameter, which can be used to characterize a memory devices; area, speed and power. Area is important from the viewpoint of its physical implementation by VLS I technology, smaller the area per bit, you can accommodate more and more devices or more or more and more memory on a single chip. Obviously, if you can accommodate large number of devices on a single chip, cost per bit reduces. So, area plays a very important role and with the sinking size of MOS memory devices, MOS devices, the number of memory devices that you can put on a single chip is increasing.

For example, now a days you can have 256 megabit chip, I mean that is dynamic RAM chip. Then speed, is another important parameter is speed, speed of operation plays a very important role, the performance of electronic circuits is increasing. The computers and other systems are operating at higher and higher speed, so unless the memory devices are available which can operate, which can communicate at that speed, you cannot really build a system using memory and those processors. So, speed of operation which is usually specified in terms of access time, so this access time has reduced from micro second to nano second, and less than nano second in the present day.

So, speed of operation is another important parameter, which specifies the performance of a memory device. Then finally, the power is becoming increasingly important, because MOS memories are used in many battery operated portable systems, like cell phones, p d s, MP 3 player, camera, and other things. Obviously, memory constitutes a significant portion of that, for example if you look at any present day processors, you will find that more than 50 percent of the real state is occupied by memory, and obviously power dissipation of memory will play an important role, that is the reason why power dissipation of the memory is important, and you have to develop memory devices that will consume lower power. Now, how do you classify memory devices. The classification can be done in various ways, particularly in three distinct ways.

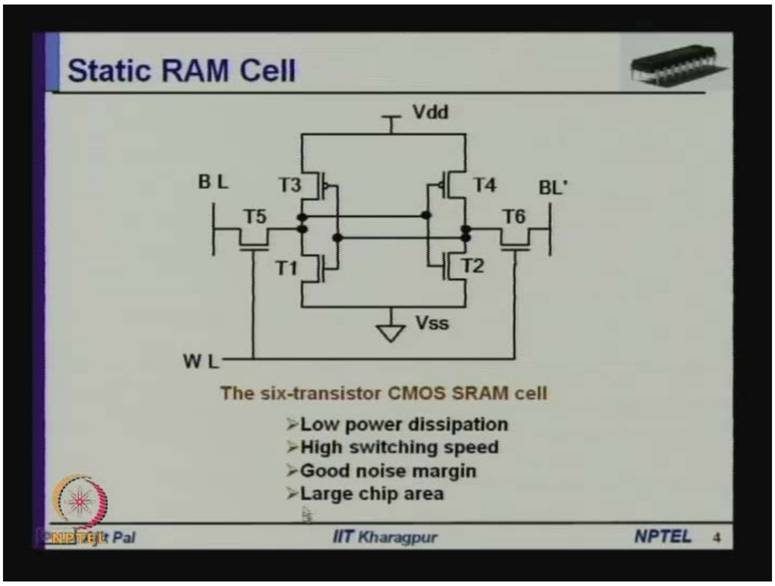
The first one is based on operation mode, operation mode; mode means, whether you can perform only read or you can perform read and write. So, based on that you can do classification there are memory devices, which are read only, that means in the normal mode of operation, you can only read from them. On the other hand, there are memory devices which are read write, that means you can perform read operation, as well as you can write perform write operation in normal mode. So, based on this you can classify in two categories read-write and read only, under the read-write category you have got SRAM, static RAM, DRAM dynamic RAM, EPROM erasable programmable ROM, and then electrically erasable programmable ROM EEP ROM, and flash memory, these are all read write memory devices. On the other hand, mask programmable ROM or there is another one which is which is not written here; fusible link ROM, these are read only, that means they are programmed once, either at the time of manufacturing or user can do it using a special device known as prom programmer, but that can be done only once.

So, those are read only, and of course, the other types for example, EPROM, EROM they can be pre-programmed, you can perform read and write, but later on we shall discuss how exactly writing can be done on them. There is another way you can classify, is based on data storage mode, data storage mode means based on that, that means how it is stored and how long it remains there, in this way you can categorize into two types; volatile and non-volatile, volatile memory devices will actually store information, as long as power is there, as soon as power is turned off information is lost; that means volatile memory devices can retain power as can retain information, can store information as long as power is there.

On the other hand, the non-volatile category which are read-write type and the read only type can return power, even when power is shut down, power is turned off and we can see SRAM static RAM and dynamic RAM belong to the category of volatile memory. Memory and non-volatile types are read-write, in the read-write category you have got flash memory, EPROM, EPROM and also mask programmable ROM. So, these are non volatile memory devices; that means if the power is turned off information will not be lost. So, in a computer system or any other system, we will find there is need for both volatile and non-volatile memory, both are required in a system, certain information are kept permanently and certain information are stored as they are used, during normal mode of operation, so both are required in a system.

Finally, final categorization is based on access method. There are two basic ways by which you can categorize; one is random there are some memories which are random access, random access means you can read from any part of the memory. So, you may have for example, 16 kilobyte, you can read any byte and the access time is same, so these are random access. On the other hand you have got nonrandom access, for example serial access memory, or content addressable memory, SAM stands for serial access memory and CAM stands for content addressable memory. In these cases, the access is not random, for example whenever you read data from floppy or tape, floppy disk or tape; it is inherently serial in nature. So, there you have to read in a sequential manner, that is why they are called non-random. In case of content addressable memory, actually the internal memory is searched and wherever there is a match, between the content, between the key and the element is being stored that data is read; so, in such cases these are nonrandom in nature. So, these are this is how you can classify the memory devices.

(Refer Slide Time: 10:31)



Static RAM Cell

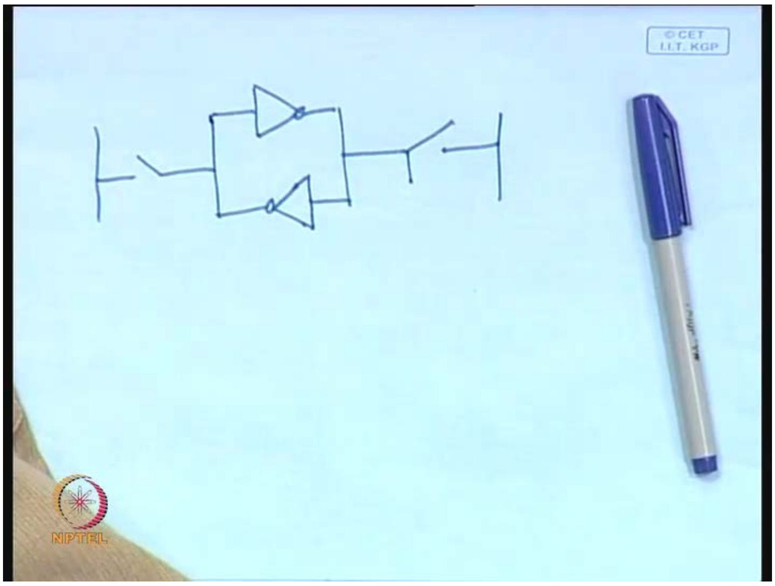
The six-transistor CMOS SRAM cell

- Low power dissipation
- High switching speed
- Good noise margin
- Large chip area

t Pal IIT Kharagpur NPTEL 4

Now, let us focus on static RAM cell, which is one of the most important components of any computer and other electronic systems. A static RAM cell can be realized by using 6 transistors as it is shown here.

(Refer Slide Time: 10:53)



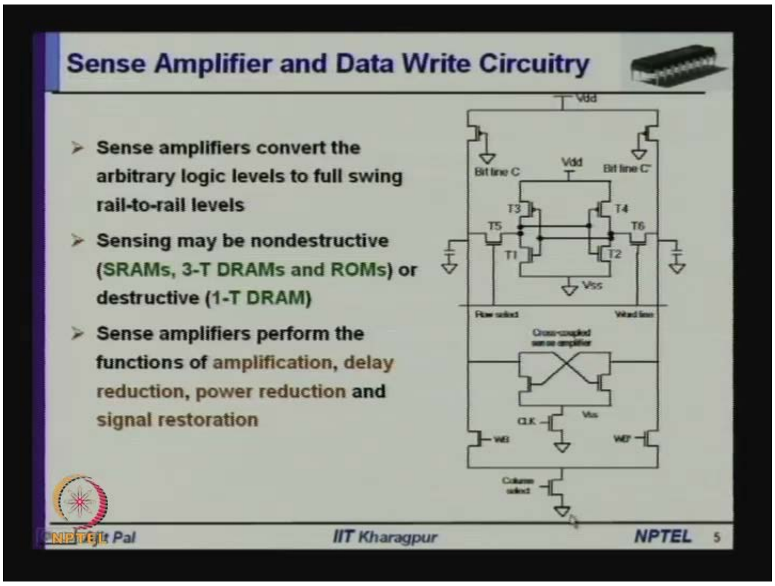
Essentially, it is realized in the form of two inverters, two inverters connected back to back and there are two switches, you can say there are two switches, of course these are all realized by using electronic circuit by MOS transistors. So, you have got two inverters connected back to back, and two switches through which you can perform read and write operations. So, this is the basic building block generic circuit and this is how it has been realized. So, you can see this is an inverter which is realized by using T 1 and T

3, and another inverter is realized by using T 2 and T 4, and the output of the this inverter is connected to the input of this inverter, and output of this inverter is connected to the input of the other inverter. So, they are connected back to back, and these two transistors T 5 and T 6 they are used for performing read and write operation. You see both the gate of both the transistors T 5 and T 6 is connected to a line known as ward line, later on we shall see, the access is performed in terms of ward and bit line. So, ward line or it is coming from row line also you can say, wherever this is one, this particular line is selected these two transistors are on.

Then, you can access data from the information which is being stored. So, in normal mode as long as power is there, theses these inverters connected back to back will be in one of the two states; that means output of this inverter realized by using T 2 and T 4 will be one, and on the other hand the output of the other inverter realized using T 1 and T 3 will be 0, or vice versa. Now, that you can read with the help of T 5 and T 6 and information will go to the bit line, and from the bit line it can be transferred to the outside world, through some interface circuit, that I shall discuss later, and as you can see, because of this configuration this 6-transistor CMOS static RAM cell, has got the following features.

Low power dissipation; low power dissipation is arising, because as you know CMOS inverters are inherently low power, because there is no static power dissipation and as a consequence, when these inverters are realized by using CMOS, they consume lesser power, and also high switching speed is achieved by using this type of configuration. And as you know there is good noise margin of CMOS inverters, as a result static CMOS circuit, I mean CMOS RAM s static RAM realized by using CMOS inverters, gives you good noise margin. However, because of the use of large number of transistors, 6 transistors per cell; that means to store one bit of information, you will require 6 transistors, and obviously it will occupy large chip area. So, the large chip area is occupied by this by its cell, and as a consequence the number of memory cells or number of bits that you can accommodate on a chip cannot be many. So, static RAM cells static RAM has lower smaller chip, I mean large chip area, so number of devices, number of bits that can be accommodated is small. So, this is the basic structure of a static RAM cell.

(Refer Slide Time: 14:58)



Sense Amplifier and Data Write Circuitry

- Sense amplifiers convert the arbitrary logic levels to full swing rail-to-rail levels
- Sensing may be nondestructive (SRAMs, 3-T DRAMs and ROMs) or destructive (1-T DRAM)
- Sense amplifiers perform the functions of amplification, delay reduction, power reduction and signal restoration

t Pal IIT Kharagpur NPTEL 5

Now let us have a look at how the reading and writing operation take place, actually in addition to a cell, I mean a cell in which the information is written bit a particular bit, you will require additional circuit, additional circuit particularly the sense amplifier, and read write circuitry is required, of course these are common to a large part of the circuits; that means this is using the same sense amplifier you can read from a large number of memory bits, so that means this part is common and also this part is common; that means apart from these 6 transistors, other transistors are common to other parts of the circuit.

Now, how the reading and writing can take place, to let us explain how the writing occurs, writing can take place by enabling these two lines right bit W B and W B bit here, and proper value is applied to these lines, it come from outside, so this will be applied here to the bit lines, and whenever let us assume we want to write 1 in this particular bit. So, whenever you want to write 1 in this particular memory bit, what will happen as you can see, 1 means this is your bit lines C bar and bit line, this is bit line C. So, 1 will be available here, so V d d will be available here and 0 will be available here. Now, whenever you want to perform read operations this ward line will be selected, this ward line will be selected, and these 1 and 0 will be applied to these bit lines; that means voltage close to V d d will be applied to this line, and close voltage close to 0 will be applied to this bit line.

As a result this row select line or ward line is now selected, so T 5 and T 6 are on; that means whenever you are, when this line is 1, what will happen, this will make this transistor T 2 on T 2, this will make T 2 on, because this will make T 2 on, because this is high, so this will make T 2 on, and this in turn will make, that means the output will be 0 and this will make T 3 on. So, as a consequence this will be 1 will be written here, T 3 on means, this will be 1 and this is 0. So, this is how writing will take place, on the other hand when this is 0, you know this 0 will turn this transistor off, so this will you know is supporting, I mean the one is supporting the other. So, you are applying both the signals and very quickly, because of regenerative feedback very quickly the change occurs and you get 1 and 0 at the output of this inverter, and the output of this inverter.

Then this signals ward line signal is withdrawn, right signals are withdrawn and information will be will remain in this cell as long as V d d is maintained. Now, how the read operation is performed, read operation is little complicated, but I can explain that in simple terms, what is done, read operation is preceded by a kind of pre-charging, pre-charging is done with the help of these transistors, this transistor at the at the top, so these 2 transistors are turned on, and these bit lines are charged to more or less half the supply voltage; that means V d d by 2, so they are charged to V d d by 2, pre-charge to V d d by 2. So, these transistors these capacitors, which are essentially relatively large will charge to V d d by 2.

After that, for reading operation performing read operations, this row select the ward line is activated, and when the ward line is activated what will happen, depending on what is told here in this particular cell, if this is one, this one, one will be connected to this line, that means charge will transfer from this point to this bit line, and if this is zero charge will transfer from the bit line to this I mean to this to this transistor T 2, that means there will be some kind of imbalance, here the voltage will rise, and here the voltage will decrease.

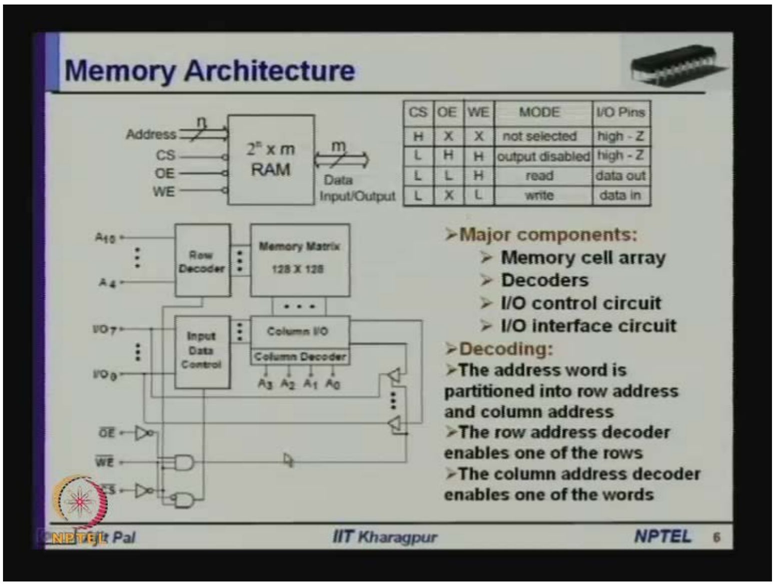
So, that is what will happen, if one was stored in this particular capacitor, and that will actually influence, this is acting as a kind of differential amplifier. So, this cross copper sense amplifier acts as the differential amplifier, and when the clock is applied what will happen, this will sense the difference and accordingly it will do several functions like amplification, delay reduction, normally the change at these bit lines is very slow, because of large capacitances, because this is connected to a large number of devices.

And this amplification is done with the help of this cross coupled sense amplifier, small change of the order of mill volt, can be sensed by the sense amplifier, and this will do the amplification not only it will do the amplification, it will do delay reductions very quickly change over will take place.

And also power reduction occurs to for the changeover, as you know if it is done quickly the power dissipation in CMOS circuit is dependent on how quickly the output changes, as you know if the output changes slowly, then there is a short-circuit power dissipation, but if the change occurs quickly then the short-circuit power dissipation is reduced, and as a consequence this because of delay reduction, there will be reduction in power dissipation, and also it will restore the signal level to rail-to-rail voltage; that means the voltage level that you will get at these two points will be raised to V d d and ground. That means normally, because of charge sharing you know what is happening here, charge sharing pre-charge followed by charge sharing, and charge sharing is leading to increase a voltage at one one and one bit line and the complementary bit line it is going in the other direction, and that is being sensed by the sense amplifier, it amplifies it, restores the voltage, and it does it quickly and now you get rail-to-rail voltage one at the bit line and zero at the bit line bar.

On the other hand if you write zero in that if the bit is reading, I mean bit is opposite. Then of course on the bit lines values will be opposite. Then the output is transferred after sensing this output will be transferred to outside world, through some sensing circuit. Now, in this particular case in case of SRAM, this sensing is nondestructive; that means the information that is being stored here is not lost, because it is a flip-flop kind of thing and may be the output will be connected, this output of this will be connected to bit line, this output will be connected to the bit line, and current will be supplied by either by the p MOS transistors and cant will be slinked by the n MOS pull-down transistors, but information will not be lost, as a consequence it is nondestructive. And as we shall see not only SRAM there are other types of dynamic RAMS, where also this sensing is nondestructive. However, later on we shall see, it is destructive in one transistor dynamic RAM which we shall discuss little later. So, I shall not come to this topic once again, that is why I am covering this. So, this part is different for different types of memory devices, this is for static RAM and other types of RAMS, this part will be different, but the sense amplifier and other parts will be same.

(Refer Slide Time: 23:53)



Memory Architecture

CS	OE	WE	MODE	I/O Pins
H	X	X	not selected	high - Z
L	H	H	output disabled	high - Z
L	L	H	read	data out
L	X	L	write	data in

- ➢Major components:
  - ➢ Memory cell array
  - ➢ Decoders
  - ➢ I/O control circuit
  - ➢ I/O interface circuit
- ➢Decoding:
- ➢The address word is partitioned into row address and column address
- ➢The row address decoder enables one of the rows
- ➢The column address decoder enables one of the words

[illegible]t Pal IIT Kharagpur NPTEL 6

So with this discussion, now we can move to more, I mean as we are following bottom of approach. Now, let us consider chip level, it is not a single bit. Whenever it is packaged in a chip, you have got large number of bits, so memories can be organized in the form of bits, bit organize say 256 bit, or it can be organized in terms of wards or bytes. Say, 4 k into 8 bit, so in this case we shall call it byte organized. So, irrespective the irrespective of whether it is bit organized or byte organized; the reading, writing, operations, these are all same. Except as you know, you will see that, this reading operations whenever it is byte organized or ward organized, you will see you will perform reading simultaneously of one ward.

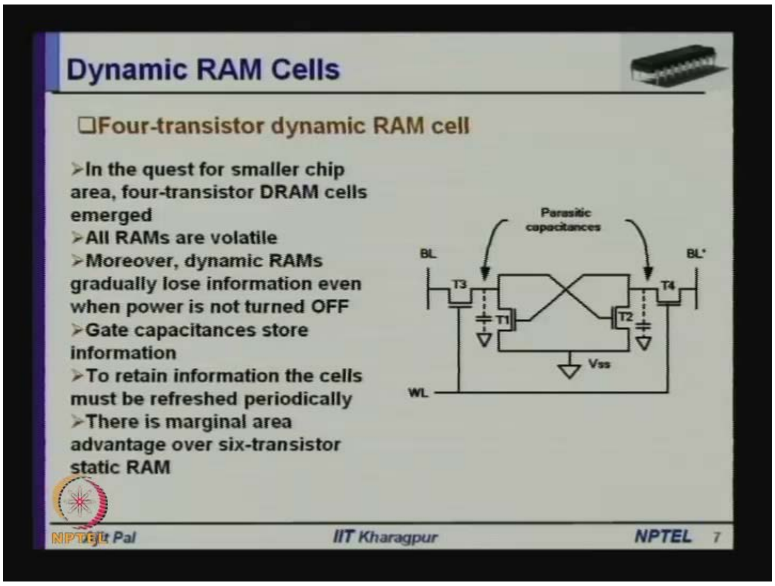
Anyway, here is an example, so here a chip has been organized in the form of 2 to the power n into m; that means m bit per ward and you have got 2 to the power 9 wards. and how do you select a particular ward, ward is selected with the help of address lines, and address is applied to the chip, this is a memory chip, address is applied to it, and along with some chip select signals. So, depending on whether you are performing read or write operation, for example, this is write enable; that means for demand whenever you are performing write, then you will be zero, and chip select has to be applied when a particular chip is being used, and as you can see, you can read all the m bits simultaneously. And depending on this control signals, how various operations are performed is shown here, when chip select is high irrespective of output enable and write enable signal, you will see the output is in high impedance state, high z or high impedance state, and whenever chips select is selected and if both the lines are high, then of course output is disabled, again it is in high output high impedance state.

On the other hand, when output is enabled and chip select is also enabled both are low, then you can perform read operation and data will be read through these lines, through the internal circuit, data will be available on these output lines. Similarly, when write enable is enabled, irrespective of whether output enable is present or not, the data that you applied to it will be written into the memory cells, and internally this is the organization of the memory. So, you can see the address lines are divided is partitioned; first of all you can see these are the measure components, memory cell array which I have already discussed, memory matrix, then you will require decoders, and you can see here there are two decoders; called row decoder and column decoder. The reason for having two decoders is to reduce the complexity of the decoders, instead of applying all the inputs to a single decoder, then the decoder complexity will be very high, but by using two dimensional selection of memory bit or memory ward, that complexity of the decoders is reduced.

So they will require lesser chip area, whenever it is organized in this manner, that is the reason why in all cases memories organized in two dimensional form, two dimensional memory matrix, and you can see lower order at bits are applied to the column decoder. On the other hand, higher order bits are applied to the row decoder. So, in this case you have got 128 by 128 array, and of course memory is little more eleven bits; so 2 kilobyte, so 2 kilobyte will require and organized in the form of 8 bits that is the reason why it is organized in the form of 128 by 120 bit matrix. However, there are 8 bits per ward, so 8 values will be available simultaneously. So, you can see as you apply row address, row decoder is going to these lines, this row select line.

So, as you know the function of a decoder is to select only one of the 2 to the power n lines. Similarly, column lines will be selecting these lines bit lines, bit lines of a particular ward, it may be 8 bit or 1 bit, depending on whether it is bit organized or byte organized. Then depending on whether you are performing read or write operation, the data will be written here. If you are performing read operation, then the data will be from the memory matrix, will go to the sense amplifier, then from the sense amplifier it will go to these interface circuit, and it will be available to the output line. On the other hand if it is a read operation data will go, it will go through these lines. This input data control data will be written into these cells, the way I have explained. So, this is how you can see you have got control circuitry, input data control then column I O control circuit. I am not going to the details of this circuits, but you can understand from this basic block diagram, how exactly they will operate. So, this is the overall structure of a memory chip. So, I have shown a RAM chip it is not different in case of ROM chip, except there will be no reading facility, conventional reading facility in a ROM chip, so having discussed a static RAM chip, a cell, then the sense amplifier, and writing mechanism, then a detail of a chip.

(Refer Slide Time: 30:15)



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

NPTEL ...t Pal IIT Kharagpur NPTEL 7

Let us now focus on dynamic RAM; historically the dynamic RAM was proposed to reduce the number of transistors per cell. We have seen that in static RAM the number of transistors is 6, and obviously it occupies large area, how the no of transistors can be reduced. First, what was done the p MOS transistors was removed, and the structure was somewhat like this. Now, you have got a four transistor cell, so what is done those p MOS transistors are no longer present here, so this collector is collector of this transistor is connected to the base of this transistor.

So, here these are cross coupled and these capacitors are not really external capacitors, these are all parasitic capacitors of the transistors. For example, this is the gate capacitance of T 1, this one and this is the gate capacitance of this one and some parasitic capacitances. So, there is no explicit capacitors which are present in dynamic RAM cells.

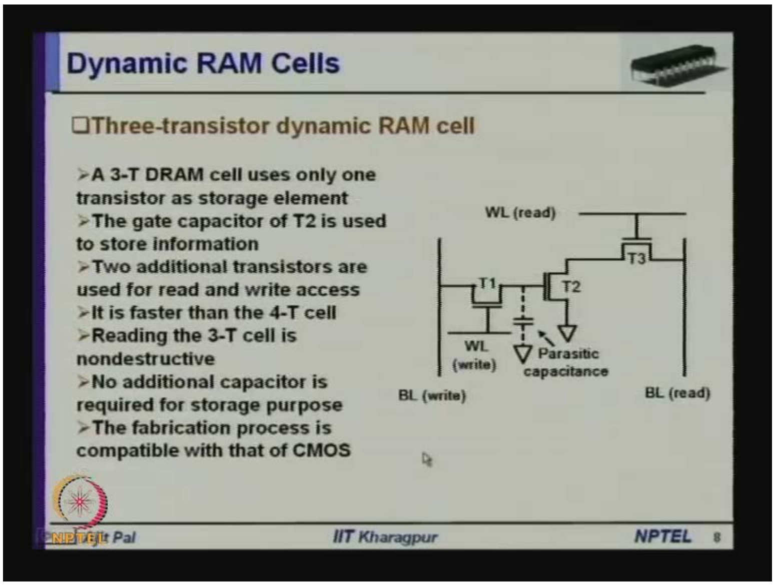
Now, as I mentioned here, in the quest for smaller chip area four-transistor dynamic RAM cells emerged.

Now as I mentioned all RAMS are volatile, including dynamic RAMS, static RAM as well as dynamic RAM; that means the power is withdrawn information will be lost, but here there is one additional problem, additional problem is in case of dynamic RAMS gradual, the information is lost even when power is not turned off; that means power is present here, but because of the leakage problem these capacitors will leak, any capacitor leaks, that means information's is to what is stored they are in the form of charge will be gradually discharged.

So, the problem this problem can be overcome by periodic refreshing, what do you really mean by periodic refreshing. By periodic refreshing means it will read the information, will sense the information. Then again write it, as you write it second time again the voltage builds up on the capacitors. So, that means refreshing, involves, reading and then writing back into the same memory cells. So, as you rewrite again the voltage levels are boosted. So, this kind of periodic refreshing is necessary to retain information in cells.

And as you can see here, in case of four-transistor dynamic RAM cells, although there is some reduction in the area, but unfortunately the area reduction is marginal, so there is marginal area advantage over six-transistor static RAM cells, and not only because of this, whenever you are going for dynamic RAM, you have to keep provision for periodic refreshing, so you require additional hardware whenever it is a dynamic RAM, and as a consequence by shifting from six-transistor static RAM to four-transistor dynamic RAM, there is no significant reduction in area. So, this was not really very popular, but because of historical reason I have discussed it.

(Refer Slide Time: 34:00)



Dynamic RAM Cells

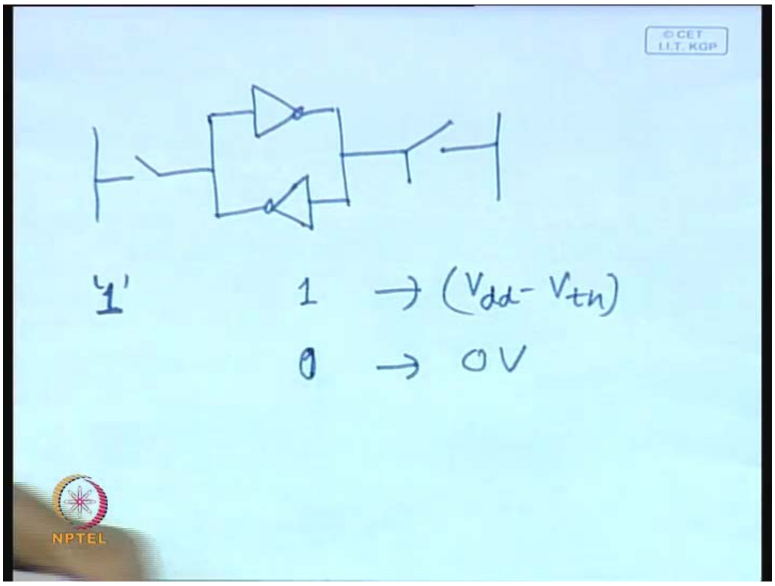
❑Three-transistor dynamic RAM cell

- 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

t Pal IIT Kharagpur NPTEL 8

Now, let us consider the second step, and second step was three-transistor dynamic RAM cell. So, three-transistor dynamic RAM cell were proposed, where one it is instead of having 2 bit, 2 arms, two transistors only, one transistor is used as you can see T 2 for storing information. So, the gate capacitance of T 2 is used to store information, again it is a parasitic capacitance, you are not adding any external capacitor there, and this is being used to store information. However, you will require two additional transistors T 1 and T 3 for the purpose of writing and reading; that means whenever you have to write this ward line is selected, it is activated, this transistors are turned on, and depending on the value on the bit line this capacitor will be charged, that means if one is present here this will be charged too V d d minus V t, as you know there will be difference of V t as it is charging through a n MOS transistor.

(Refer Slide Time: 35:11)



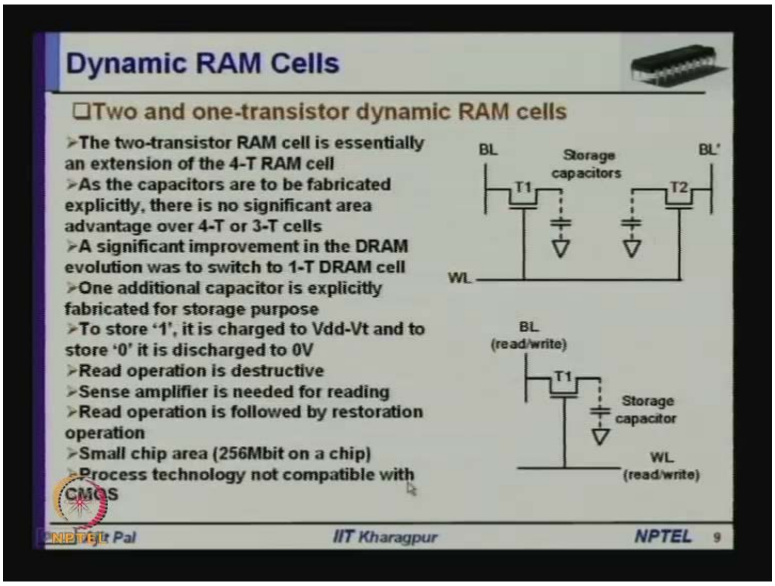
That means whenever it is charged to one, I mean whenever you are writing one it will be this capacitor will charge to V d d minus V t n, where V t n is the threshold voltage of the n MOS transistor. On the other hand, whenever you are trying to write zero, then it will be zero volt. This capacitor this capacitor will discharge through T 1, because bit line will be now held to zero, whenever you are writing, so this capacitor will discharge through T 1. So, this is how writing operation takes place through this transistor T 1, by selecting it with the help of this ward line W L.

Now, how do you perform read operation, to perform read operation this transistor is turned off, however the T 3 transistor is turned on, for the purpose of reading. So far the purpose of reading this transistor is turned on, and as you read it, as you turn it on. The state of this transistor is decided by the charge on this capacitor; that means if one is stored in this capacitor, then this transistor will turn on, and as it is turned on this will be discharged, that means whenever you are charging, when it is one then zero will be written here, because this transistor is on and it will be discharged.

So, the bit line here you will read zero whenever one is written here. So, it has to be inverted to get the actual data. On the other hand if it is zero, then this transistor is off and this bit line will retain the value as I mentioned earlier, the bit line is initially pre-charged, so this remains in pre-charge value, and because this T 2 is now off, and it does not discharge through T 3 and T 2. This is how reading operation take place in three-transistor, three-transistor CMOS circuits. Now, this three-transistor CMOS circuits I mean this dynamic RAM technology is quite popular, because of two reasons, no additional capacitor is required for storage purpose, and the fabrication process is compatible with that of CMOS.

Moreover, reading from a three-transistor cell is nondestructive as I have already told, and it is faster than four-transistor cell, because of these reasons it is faster than four transistor cell, reading the three-transistor cell is nondestructive, no additional capacitor is required for storage purpose, the parasitic capacitor, the gate capacitance is of T 2 is serving the purpose of storage device, and fabrication process is compatible with that of CMOS technology. That means you can fabricate, suppose you are building an embedded system where you require some memory, so this dynamic RAM cell can be fabricated off on chip. So, this type of dynamic RAMS can be fabricated on chip, and in this case the reading is nondestructive. So, this is very advantageous, and that is why this is also used in many situations.

(Refer Slide Time: 38:53)



Dynamic RAM Cells

❑Two and one-transistor dynamic RAM cells

- ➢The two-transistor RAM cell is essentially an extension of the 4-T RAM cell
- ➢As the capacitors are to be fabricated explicitly, there is no significant area advantage over 4-T or 3-T cells
- ➢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 Vdd-Vt 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

NPTEL ...t Pal IIT Kharagpur NPTEL 9

However, in quest for further reduction in the chip area, two and one-transistor dynamic RAM cells were proposed. So, two transistor dynamic RAM cell is essentially an extension of four-transistor RAM cell, where four-transistor RAM cell we have already seen, so those cross coupling is removed, so whenever you remove the cross coupling, those capacitors, those transistors are also removed, you have got only two transistors for reading and writing. Then question arises, where from you get these storage capacitors, and fortunately this case the information is not stored in parasitic capacitors. You have to fabricate explicitly these capacitors, how can you fabricate, as you already know you can fabricate a MOS capacitor, by using with the help of this substrate, you can fabricate silicon dioxide and top of that, you can deposit fall silicon that will form a MOS capacitor.

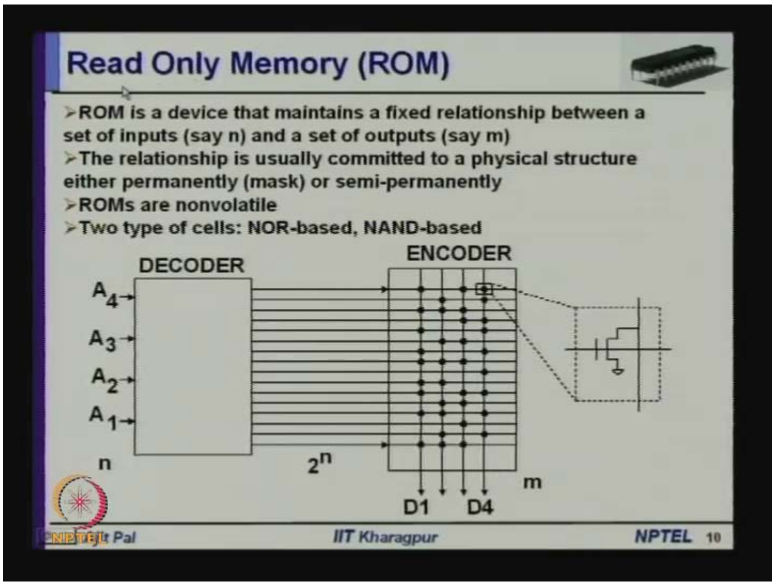
So, just like I mean it is essentially the gate region portion, so you can fabricate explicitly MOS capacitors, and where you have to store information, so this capacitors are essentially fabricated. So, these capacitors are fabricated explicitly there is no significant area advantage over 4 T and 3 T cells. So, these two transistors dynamic RAMS were not very attractive, because it did not give any area of advantage, that is a reason why this two-transistor dynamic RAM cells were not very powerful, and particularly very soon the 1-transistor dynamic RAM emerged. As you can see here, you have got only one transistor T 1, and of course you will require a storage capacitor that has to be explicitly fabricated, and this is the read-write line, and this is the read-write line, this is the bit line this is the ward line.

So, you can see obviously this will require very small chip area, and it has been found that it occupies rapidly one-fourth of the chip area of that of static RAM, 6-transistor static RAM cell, so it is very attractive. Now, to store one it is charged to V d d by V t, so it is not different, and to store 0 it is discharged to 0 volt. So, writing operation is not very different from 3-transistor dynamic RAM cell. However, reading is quite different, so whenever you do the reading you can see this charge will be transferred directly to the bit line, so this transistor is on.

So, this charge will be transferred to this, and you know kind of charge sharing occurs, and this capacitor is not isolated from the bit line. So what will happen, this is a kind of destructive operation that means information that is being stored will be destroyed after you perform a read operation. Then what you have to do, after you have performed a read operation, you have to again write back the data, same data. Since, read operation destructive, you have to perform, you have to restore the read that data by, I mean read operation is followed by restoration operation to restore the data, and also you will require sense amplifier for reading as I have already explained.

These are all required; I mean sense amplifier is required in all cases. And as I mentioned, the chip area is very small in this particular case, it will occupy very small chip area, because only one transistor and a MOS capacitor is required per cell. Unfortunately, processed technology is not compatible with CMOS; that means you cannot fabricate single transistor dynamic RAM on chip, with the CMOS technology with the help of the conventional CMOS technology. That means the static RAM chips has to be separate or off chip, you have to use separate chip for 1-transistor dynamic RAM cells, however the capacity can be very large. Nowadays, you can get a dynamic RAM chip occupying 256 bit or more on a single chip, and that has to be interfaced to the processor, through some bus. So, this cannot be used for on chip applications, but for off chip this can be used, so this is the dynamic RAM cell.

(Refer Slide Time: 44:11)



Read Only Memory (ROM)

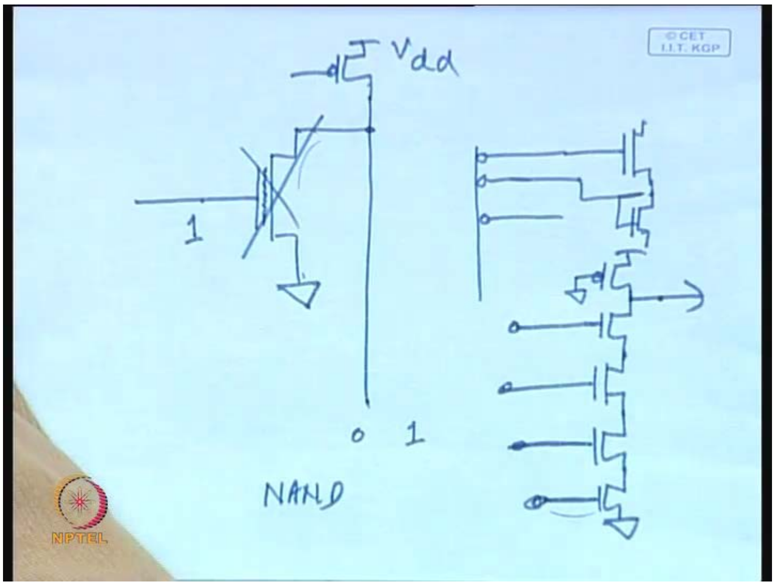
- ROM is a device that maintains a fixed relationship between a set of inputs (say n) and a set of outputs (say m)
- The relationship is usually committed to a physical structure either permanently (mask) or semi-permanently
- ROMs are nonvolatile
- Two type of cells: NOR-based, NAND-based

Ajit Pal IIT Kharagpur NPTEL 10

Now, let us switch to another type of memory that is read only memory. So, read only memory is a device that maintains a fix relationship between a set of inputs and a set of outputs. So, despite its grandiose name, it is functionally very simple. If you look at the function of a ROM read only memory, it has essentially got two components; a decoder and an encoder as it is shown in this diagram. So, you have got a decoder, function of a decoder is known, whenever you apply some input, there are n inputs one of the 2 to the power 2 n inputs will be active, active means one remaining bits will be zero, so that is a function of a decoder, so one of the 2 to the power n lines will be 1.

Now, in the encoding part, you can write some desired information, so this relationship is usually committed to a physical structure, either permanently or semi-permanently; that means you can store information in these encoder part that can be permanent, it will be done only once at the time of fabrication, that is known as mask programmable ROM, or there can be some fuses, those fuses can be blown only once. And so during the remaining lifetime of the device, it will be used only for reading. And another very important property is ROMs are nonvolatile; that means if you withdraw the power the information is lost, I mean not lost its information will be retained by this encoder, in this encoder part. So, here is an example of how the information is stored here. Information is stored in special type of MOS transistor.

(Refer Slide Time: 46:20)

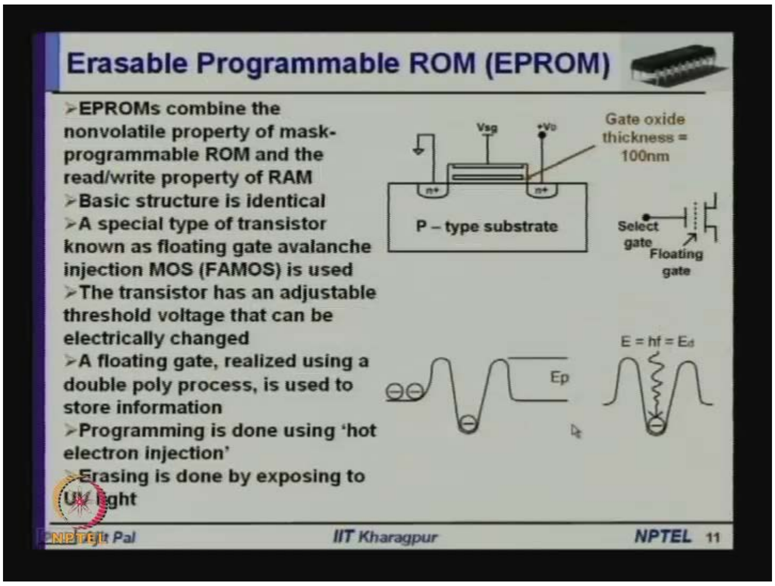


So you have got a MOS transistor, where information is stored. So, this side is ground, this is coming from the decoder, and this is this is connected to the encoder line. Now, here what you can have, a not type of configuration this is connected to V d d, so what will happen in this case, if this is one then this transistor will turn on, this is not required, so this will be grounded and I mean this line will be grounded and you will get zero here.

On, the other hand this transistors may not be there, and when if one is there, obviously this will not be changed so you will get one; that means a MOS transistor is fabricated, wherever the input will affect the output, and wherever it is not necessary, I mean one is to be stored, this MOS transistor is not fabricated. So, you can see here at this junction there is no MOS transistor, so if you if this line is one you will get one on d 2. On the other hand that means if the input is 0 0 0 0, then this top line will be activated and you will get zero here, on D 1 you will get one here on D 2, you will get zero on D 3 and zero on D 4. Similarly, if you apply 0 0 0 1, then the second line will be activated and then you will get 1 0 1 0 1 0 at the output.

So, it is acting as a kind of combinational circuit and at the time of fabrication at the junctions, I mean between encoder and decoder part at these cross points, you can have a transistor or you may not have a transistor, depending on what you are writing and this is permanent. So, this is the typical you know mask programmable ROM, so at the time of fabrication you will fabricate a MOS transistor or not fabricate. And this is actually known as nor-based encoder you can have a NAND-based encoder the other way, in that case you know decoders are will generate zeros, and in case of NAND-based, you know the transistors will be in series, this is how let me draw several. Let’s assume there are four lines, so these are zero active, and here you have got a pMOS transistor, and output is taken from here in a NAND-based cell; that means from the decoder the four lines will be zero active, and you will have four n MOS transistors and a pMOS transistors, which will be permanently on and output is taken from here. So, this is the typical NAND-based ROM cell, where you are fabricating, the encoder part is fabricated in a form of NAND, rather than in the form of NOR as it shown here.

(Refer Slide Time: 49:59)



Erasable Programmable ROM (EPROM)

- EPROMs combine the nonvolatile property of mask-programmable ROM and the read/write property of RAM
- Basic structure is identical
- A special type of transistor known as floating gate avalanche injection MOS (FAMOS) is used
- The transistor has an adjustable threshold voltage that can be electrically changed
- A floating gate, realized using a double poly process, is used to store information
- Programming is done using ‘hot electron injection’
- Erasing is done by exposing to UV light

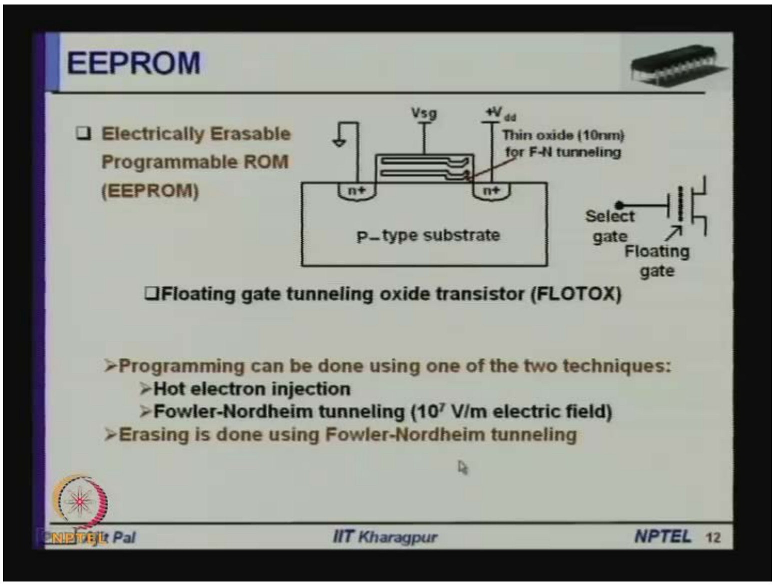
...t Pal IIT Kharagpur NPTEL 11

How do you realize erasable programmable ROM, erasable programmable ROM can be fabricated with the help of a, what is known as FAMOS, or floating gate avalanche injection MOS. So, here a at MOS transistor where you have got two gates; one is select gate, another is floating gate. So, you can see here the floating gate is between the substrate and the control gate, how it works. So, EPROMs combine the nonvolatile property of mask programmable ROM and the read-write property of RAM. So, you can perform writing also in this type of cells, so basic structure is identical to that of the ROM I have shown, special type of transistor known as FAMOS is used in place of those transistors.

The transistor has an adjustable threshold voltage that can be electrically charged; that means whenever you have to, I mean this floating gate what can be done by using a technique known as electron injection process, if you apply high voltage to the gate and also between this source and drain. The electrons will be attracted, hot electrons are generated and that is being captured in that floating gate, as it is shown here. So, floating gate in the electrons will go there and it will be trapped.

So, electrons will be trapped in these regions and as a consequence what will happen, after the voltage are withdrawn, those electrons will remained trapped for years, maybe for ten years, and then if you apply this standard gate voltage, this transistor will not turn on, because the threshold voltage is increased, whenever electrons are trapped in these floating gate. So, essentially what you are doing, you are fabricating transistors of two types; one is of low threshold voltage, another is of high threshold voltage. High threshold voltage is above the V d d. So, this programming is done using hot electron injection as I have already explained, but erasing is done by exposing to ultra-violet light. So, whenever there is a optical window in the erasable programmable ROMs, through which you can expose it t o ultra-violet light. So, these photons will transfer energy to those electrons, and they will come out of these, you know those that floating gate and this is how erasing is done. And erasing take place for all the transistors that is present in a ROM.

(Refer Slide Time: 52:43)



EEPROM

- □ Electrically Erasable Programmable ROM (EEPROM)

Vsg
$+V_{dd}$
Thin oxide (10nm) for F-N tunneling
n+
n+
P–type substrate
Select gate
Floating gate

□ Floating gate tunneling oxide transistor (FLOTOX)

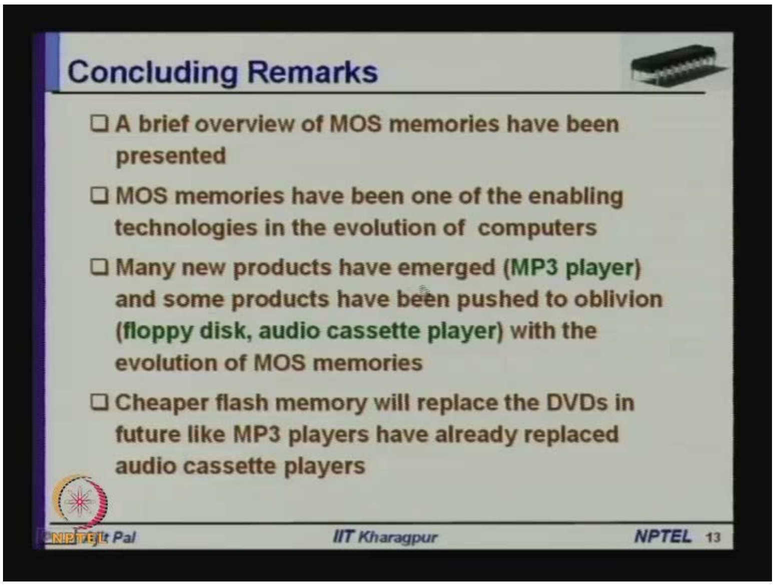
- ➢ Programming can be done using one of the two techniques:
  - ➢ Hot electron injection
  - ➢ Fowler-Nordheim tunneling (10^7 V/m electric field)
- ➢ Erasing is done using Fowler-Nordheim tunneling

NPTEL t Pal IIT Kharagpur NPTEL 12

Finally, EEPROM Electrically Erasable PROM, there the reading and writing operation is done, by using a new technique known as Fowler-Nordheim tunneling. So, what is done here in this case you can see, the gate oxide thickness is very small; 10 nanometer. In the previous case, the gate oxide thickness was of the order of 100 nanometer, but in this case, it is of the order of small portion, you can see it is of the order of 10 nanometer. So, in this particular case, whenever a high voltage is applied on the floating gate, then electrons will tunnel through this region to the floating gate.

And this is how the tunneling is done, and writing and reading can be done by grounding, either the gate voltage or the this source voltage; that means whenever gate is applied high voltage electrons will tunnel to the floating gate, and if the source is grounded the electrons will tunnel back to the source. So, this is how tunneling take place in one direction or the other. And this can be used to implement electrically erasable programmable ROM, and both the programming can be done, either by hot electron injection or by Fowler-Nordheim tunneling, but erasing is always done by using Fowler-Nordheim tunneling. So, with this let me conclude this lecture with some concluding remarks. So, I have given an over a brief overview of MOS memories, I have presented a brief overview. Now, MOS memories as I told are acting as kind of enabling technology of the present day computers. So, the present day computers are evolving with the evolution of MOS transistors.

(Refer Slide Time: 54:40)



Concluding Remarks

- ❑ A brief overview of MOS memories have been presented
- ❑ MOS memories have been one of the enabling technologies in the evolution of computers
- ❑ Many new products have emerged (MP3 player) and some products have been pushed to oblivion (floppy disk, audio cassette player) with the evolution of MOS memories
- ❑ Cheaper flash memory will replace the DVDs in future like MP3 players have already replaced audio cassette players

NPTEL t Pal IIT Kharagpur NPTEL 13

Then what is happening, because of that many new products have emerged, because of this evolution. For example MP 3 player and some products have been post to oblivion; like floppy disk, audio cassette player, those are becoming absolute. Those are no longer present here, all are using MP 3 player and this is primarily, because of this technology, this EPROM or you know this is based on that flash memory is based on this technology. And as I mentioned cheap flash memory, has replaced the audio cassette player and floppy disks, and cheap cheaper flash memory will replace D V Ds in near future, like M P 3 players have already replaced the audio cassette player. So, this is not the question of it but it is the question of fame. So, time will come, when again the D V D players will be will again go to oblivion, and you will be using those, you know those like MP 3 player you will be having, those flash memory based DVD players. So, those are coming in the future. So, with this happy note for future, let me end my lecture and in the next lecture I shall discuss about sequential circuits. Thank you