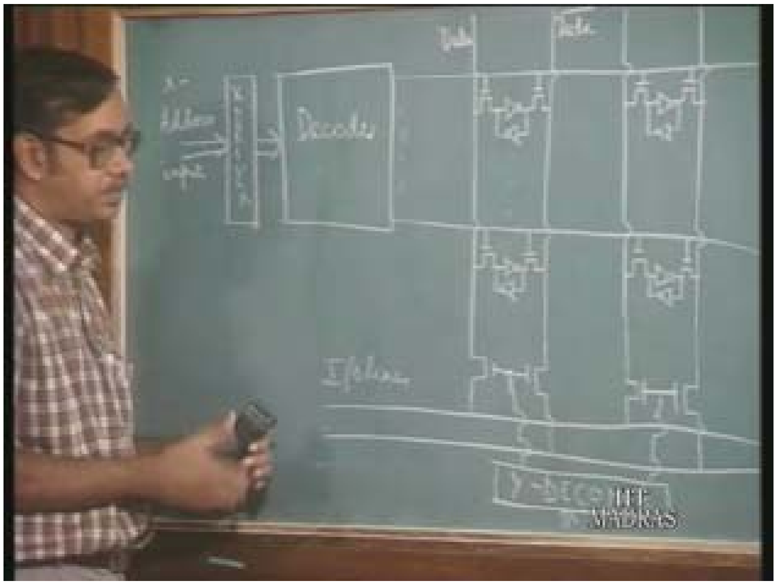
Digital Integrated Circuits
Dr. Amitava Dasgupta
Department of Electrical Engineering
Indian Institute of Technology, Madras
Lecture No # 33

CMOS SRAM

We shall continue our discussion on memories. Last class we introduced the CMOS memory cell so continuing on that. This CMOS memory would consist of the address inputs here the x decoder, you have the driver, decoder. These are the x address inputs and from the output of the decoder you have a number of lines. If you have say 8 address inputs here, you will have 2 to the power 8 lines. This gives raise to 2 to the power 8 rows. I am just drawing the two rows here and then you have the different columns. This is one particular column where you have the memory cell. The memory cell basically in a CMOS case consist of two pass transistors like this and two CMOS inverters connected back to back. You have similar cells along this and this is one column. I am just showing two rows, similarly you have other columns. You have again just exactly similar cells and here also.

Now what is done is when depending on the x address inputs, one of the output lines of the decoder goes high and which selects the pass transistors corresponding to that particular rows. All the cells in that particular row are selected. It selects one cell in each of the columns and that particular cell is connected to the data lines.

(Refer Slide Time: 05:37)



You have two lines actually one is called the data and the other is the data bar line that is because of this particular configuration here, you have an inverter if one line is high, the other line will be low.

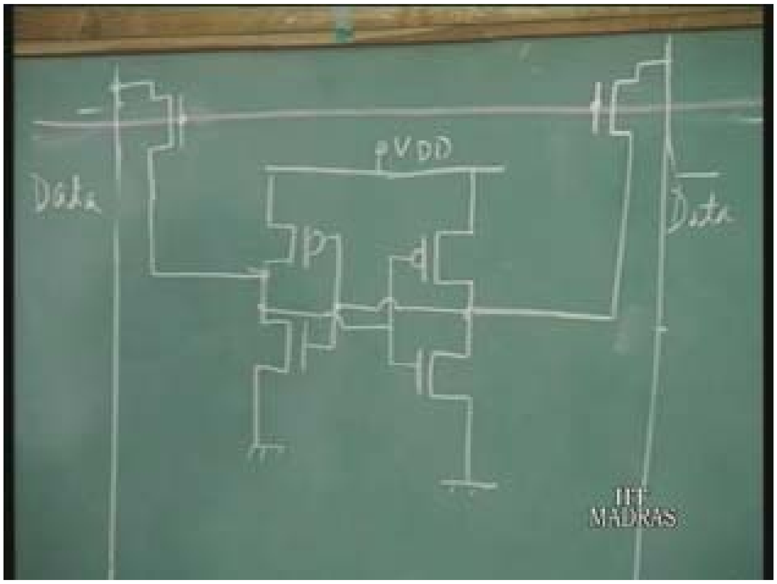
Now the next task is to select one of the columns that is done again here using pass transistors. You have the I/O lines input output lines, these pass transistors are connected to these I/O lines and this input to this pass transistors comes from the y decoder. The y decoder which receives the y address as inputs and the outputs pass on to this gate of these pass transistors. Again depending on combination at the input here, only one output of the y decoder is going to be high, it basically selects one particular column. Only the output of one particular column is connected to the I/O lines. Depending on the input address, one row will be selected which means that one particular cell in each column is going to be selected. The information available on the data lines corresponds to that particular cell and with the help of the y decoder again we are selecting one particular column. The output of that particular column is available on the I/O lines. Depending on the combination of the x and y inputs one particular cell is selected, that is how it works.

Now you can imagine that if we have a large size memory say for example if you have say 256 k memory, bit memory that means there will be 18 address lines so 9 x address lines and 9 y address lines and there will be 2 to the power 9 rows and 2 to the power 9 columns, a 512 rows and 512 columns. These word lines and bit lines are going to be extremely long and they have a large capacitances and resistances and there is going to be a lot of delay in charging and discharging these lines. Let us take a particular example, I will just draw the particular cell in the magnified form. The cell looks like this. You have the inverters here, the CMOS inverters this is one inverter and this is another inverter here and what is done is in the cell output of one inverter is connected to the input of the other. This is the cell, this is $V_{DD.}$, this is ground. You have the word lines running like this. You have the pass transistors which are activated by the word lines and which connect the output of this particular cell to the data line. This is data line say and this is the data bar line.

Now you have the word lines running all through the memory array. Now this word line is connected to the gate of this pass transistors and we know that in a MOS technology, the gates are usually made up of poly silicon, these word lines are usually run on poly silicon. In fact I would better draw it in this form to make it clearer, this is the poly silicon word line say I am drawing it in red to distinguish and this is the data line. The transistor is formed at this input, this is the transistor and the gate is connected here.

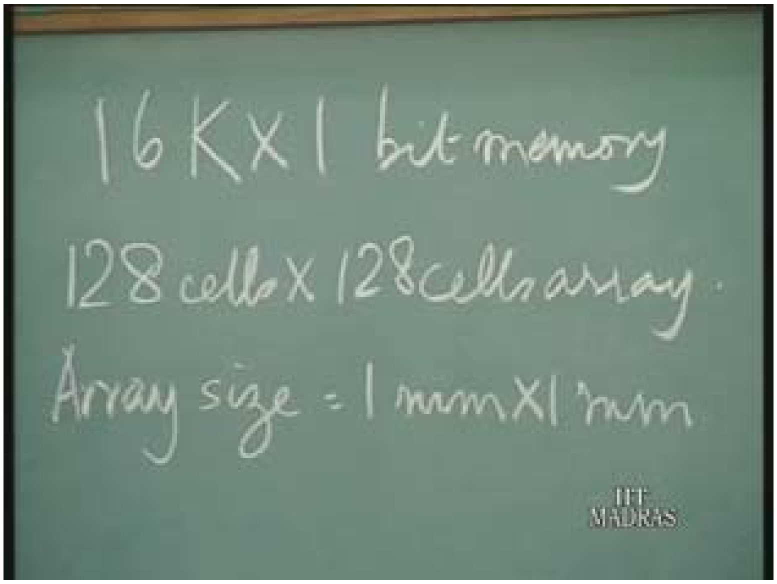
This is the transistor. Basically what it does is when this word line goes high, this data line is connected to the output of this cell. This word line runs all through the cell, it's a very long word line and large part of the delay is due to the charging and discharging of this data line. When the output of the decoder changes then this voltage on the data line should change and if the word line is of large capacitance, the capacitance has to be charged.

(Refer Slide Time: 09:51)



Let us take a small example to find out the delays, how one may calculate. Let us assume take a small problem, small memory say 16 k into 1 bit memory and array is 128 cells into 128 cells array, array size is 1 mm into 1 mm. Poly silicon line which is the word line is 1 micron wide so that is the 1 micron line. Select transistors one micron that is the size of the select transistors, channel length of one micron and width of one micron. Poly silicon sheet resistance is equal to 10 ohms per square, capacitance to substrate for poly line equal to 0.2 femto farad per micron squared and gate capacitance 4 femto farad per micron square. Given these details, can we estimate the delay required to charge the word lines. That's a small problem let us see how we are going to tackle it.

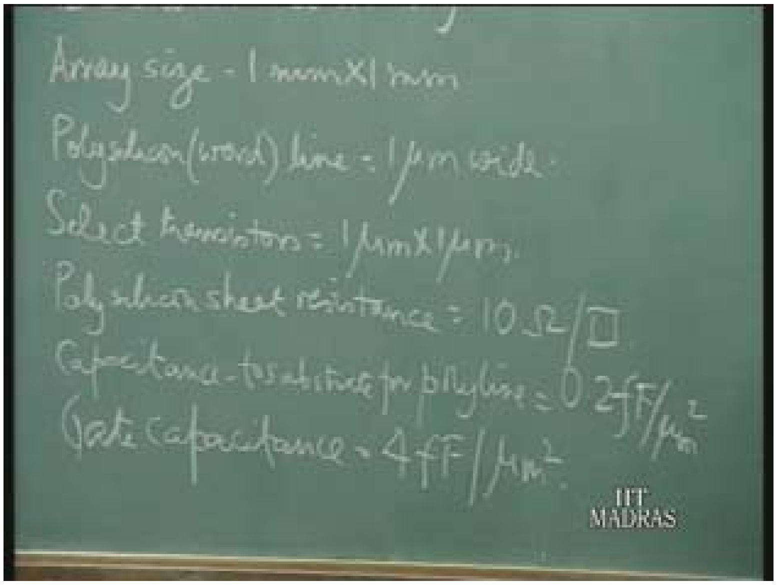
(Refer Slide Time: 11:04)


16 K x 1 bit memory

128 cells x 128 cells array.

Array size = 1 mm x 1 mm

(Refer Slide Time: 12:44)


Array size = 1 mm x 1 mm

Polysilicon (word) line = 1 μm wide.

Select transistors = 1 μm x 1 μm.

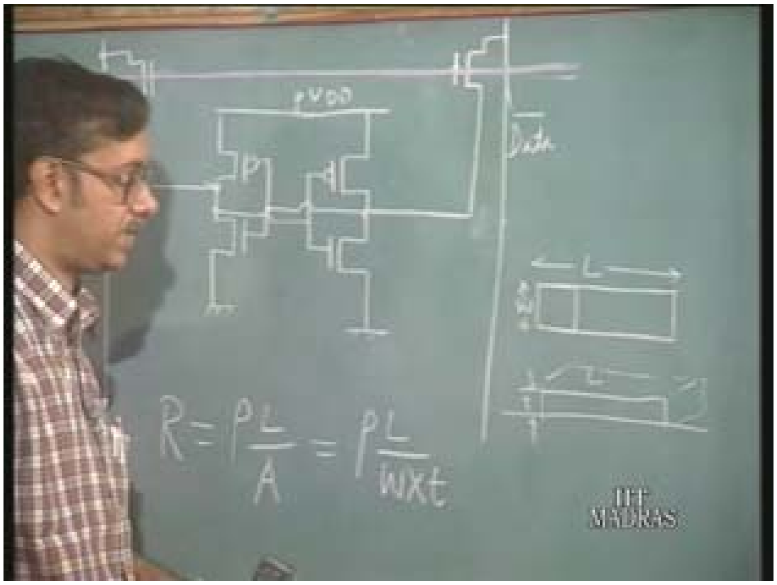
Polysilicon sheet resistance = 10 $\Omega/\square$

Capacitance-to-substrate for polyline = 0.2 $fF/\mu m^2$

Gate capacitance = 4 $fF/\mu m^2$.

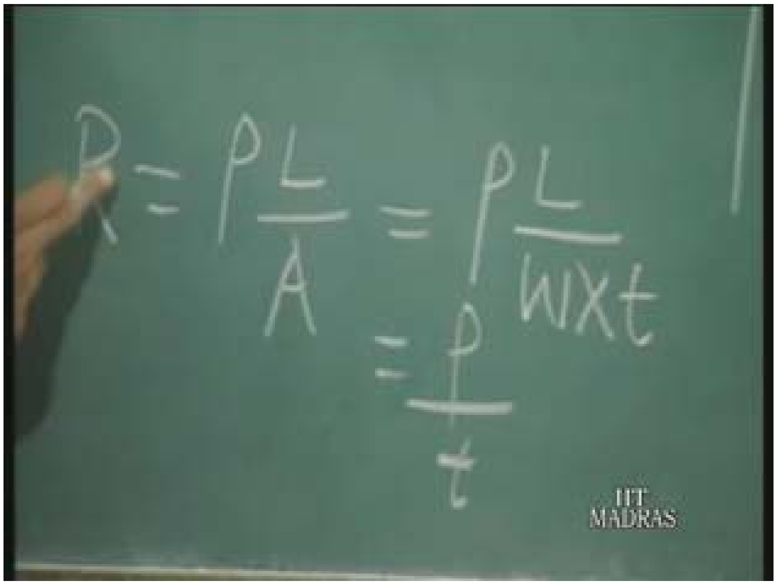
First this is a poly silicon line, you know the word line. The details about this is given, poly silicon sheet resistance is equal to 10 ohms per square. For those of you, who are not acquainted with this term of sheet resistance, see a resistance of any line is given by resistivity, into the length divided by area. Now if you have a line, if you look at the top view of the line say this is the top view, this is the length of the line and if this line has width of w and if this is made of a sheet of material of thickness say t, if the material thickness is t. Suppose if you take the cross sectional view this is L, this is the thickness t and this may be the width w. The resistance can be given as rho L divided by w into t.

(Refer Slide Time: 14:29)



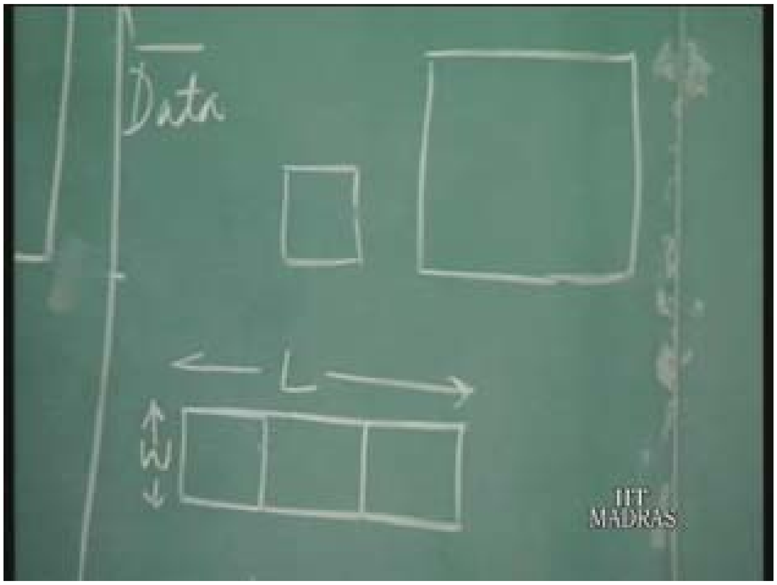
Now this is the length, this is the width of the line and this is the thickness. Now if we assume a square here, one square of this material then the length is equal to the width. If you take a one square of this material and length is equal to width, this becomes equal to rho by t that is the resistivity by thickness. If the thickness of the material is constant then we can define a resistance for a square of the material. Important point to note here is that this resistance of one square of the material is independent of the size of the square. If you take the same thickness of the same material and you have a small square or you have a large square, the resistance from one end of this line to the other is the same irrespective of if you have a big square or a small square. It does not depend on the size of the square.

(Refer Slide Time: 14:47)



In this case the total resistance is now going to be given by how many squares you have. If you know the sheet resistance which is nothing but the resistance of one square of that material, the total resistance is going to be number of squares. Since here you have 3 squares, the resistance of this particular line is going to be thrice the sheet resistance. That is how we determine the resistance of a line.

(Refer Slide Time: 15:41)

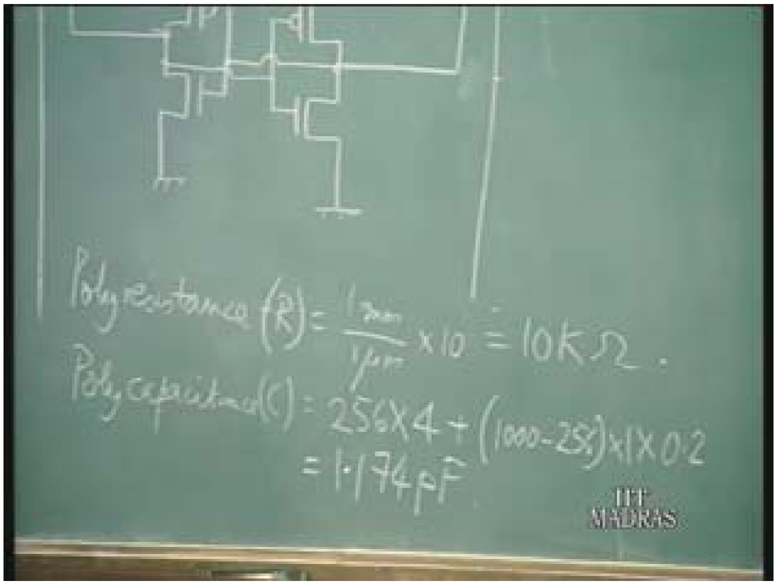


Now here it is given that the poly word line is one micron wide and the array size is one mm into one mm. It just runs for a length of one millimeter and it’s one micron wide. The poly line number of squares is one millimeter divided by one micron which is one thousand. Poly resistance R say is equal to one millimeter by one micron that many squares you have. The length is one millimeter, width is one micron, a thousand squares into 10 ohms. The total resistance of the poly line is 10 kilo ohms.

Now what about the capacitance? The total poly line is again one millimeter long but it runs partly if you look at this figure, this is the poly line. It is going to run partly on the field oxide which is the thicker oxide. It has a smaller capacitance and that is given by the capacitance to substrate of the poly line. The capacitance to substrate for the poly line is 0.2 femto farad per micron square. That is the capacitance for this part of the poly line whereas when it is the part of the MOSFET, you have the gate capacitance which is higher because the oxide thickness is less. Partly this poly line runs over thinner oxides which forms the gate and partly over thicker oxide for the remainder of the line.

If you have a 128 cell into 128 cell array, there will be 256 such transistors. There are 256 such transistors because each cell has got two select transistors. Each transistor is one micron by one micron. The total capacitance is going to be poly capacitance say c is partly due to the gate capacitance. The gate area is the 256 transistors. So 256 into one micron into one micron that is 256 micron square into, the gate capacitance is 4 femto farad per centimeter square. So 256 into 4, this is the total capacitance due to gate plus… for 256 micron it is running over gate oxide. For the reminder of one millimeter it is going to run on field oxide plus one millimeter that is 1000 micron minus 256 micron that is the field oxide into one micron is the width. This is the area into point 2 femto farad per micron square. This is the total capacitance. The total capacitance works out to 1.174 Pico farad. This is the resistance and this is the capacitance.

(Refer Slide Time: 19:39)



Poly resistance (R) = $\frac{1mm}{1\mu m}$ x10 = 10KΩ.

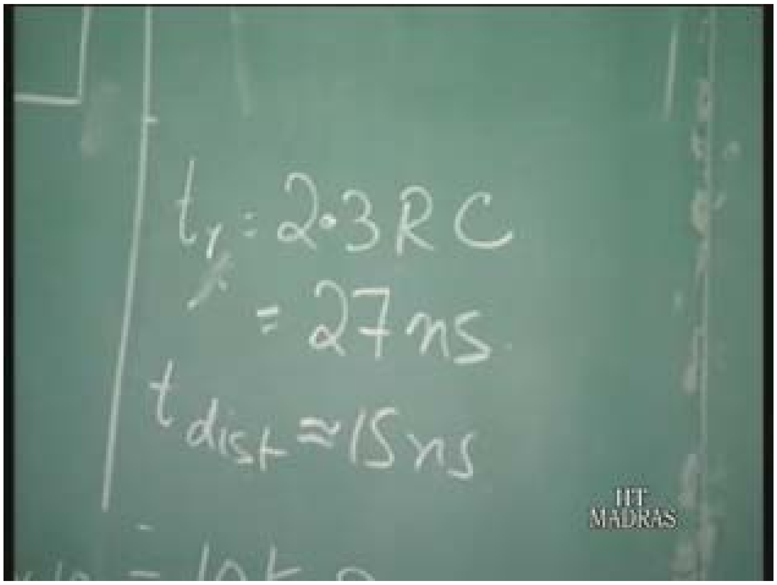
Poly capacitance (C) = 256X4 + (1000-256)x1X0.2

= 1.174pF

The total estimated raise time if you consider just RC type of network that is the time required to raise up to 90% of the final value is given by 2.3 RC. If you take this R and C, this works out to be 27 nanoseconds. Of course this is rather a very large over estimate of the actual delay because this considers R and C separately whereas this resistance and capacitance is distributed throughout the line. If these are used as lump parameters here but in actual practice they are distributed. If we use a distributed parameters, the actual delay will be around 15 nanoseconds. If we use distributed parameters, this is much less than this 27 nanoseconds.

This $t_{r.}$ for lump parameters, $t_{dist.}$ will be around 15 nanoseconds. This is the actual value 15 nanoseconds. This gives a rough estimate of the delay by assuming lumped value of resistance and capacitance. That is one resistance and one capacitance but since in actual practice, the resistance and capacitance are distributed, one has to do an analysis using distributed resistance and capacitance where if one does such an analysis, the delay would be approximately 15 nanoseconds. This is the order of delay of this word line. This is an appreciable part of the delay of a memory cell, the time required to access a particular memory cell. Obviously we have already studied BiCMOS drivers and we know the BiCMOS is capable of driving large capacitances instead of CMOS drivers, we shall be using Bi CMOS drivers. We shall take it up when we look at BiCMOS memories in the next class may be.

(Refer Slide Time: 21:18)



$$t_r = 2.3RC$$
$$= 27ns$$
$$t_{dist} \approx 15ns$$

Now this is one aspect of the delay that is charging or discharging the word line. We require circuits which are capable of charging or discharging large capacitive loads to drive this word lines. Now once we select the particular cell say suppose this particular cell is selected then when we are reading the information of the cell then the information stored on this cell should affect the data and the data bar lines. So that we get the information regarding the cell on the data and data bar lines and we are able to read the information of the cell. This means that this cell must be able to charge or discharge this data and data bar lines.

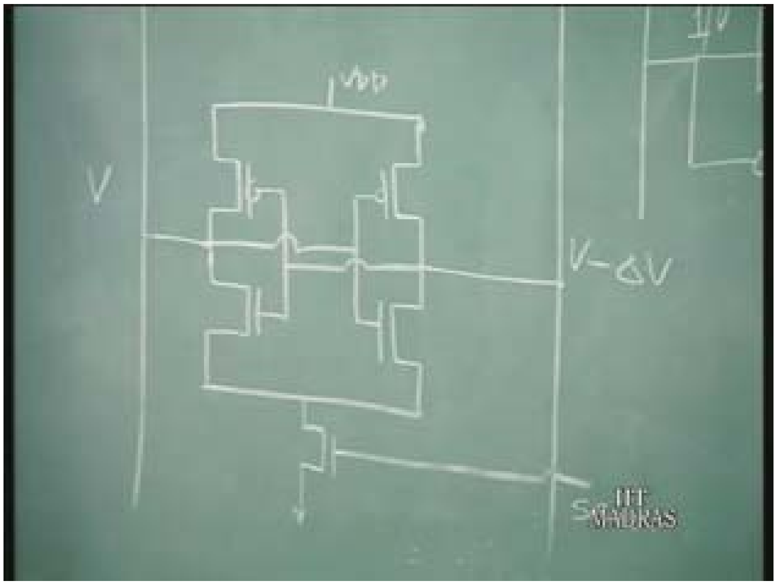
Now these data and data bar lines also represent large capacitive loads because they are also running all along the memory array and these lines will be as long as the word lines. These data and data bar lines must also be charged, it can also result in a large delay. Now the problem here is that these data and data bar lines must be charged or discharged with the help of these transistors in the memory cell. Now since we know that say for example if you have a 256 k memory, we have 256 k such cells and this memory array takes up the major chunk of the area of the total memory and these transistors in the memory cell must be a small size transistors because if we are going to make them big transistors then the area of the cell becomes very large. This transistors necessarily must be small transistors because they take up a major, I mean if you just increase one transistor to twice the size then the whole memory size is going to go up by factor of 2.

One must make these transistors small. At the same time if you have small transistors it is difficult to drive the data lines using such small transistors that is the problem. How do we solve this problem? The way to solve it is to use what are called sense amplifiers. The sense amplifiers will be used actually to charge or discharge the data lines. They are going to sense what is stored in the memory and based on that they are going to make one of the data lines go high and the other data line go low.

Basically the charging and the discharging of these large capacitive loads that is the data lines are done by the sense amplifiers and since we are going to have only one sense amplifier for example for one column, it doesn't matter we can make them quite large transistors. We can use quite large transistors because the number of sense amplifiers is going to be quite small. For example if you have again 256 kilo byte memory which means that 2 to the power 18 cells, 25600 cells, we have 2 to the power 9 columns. If we put one sense amplifier per column, we just have 512 sense amplifiers. 512 compared to 25600 is quite a small number. We can afford to make this sense amplifiers quite large and that the data lines can be charged and discharged at a faster rate.

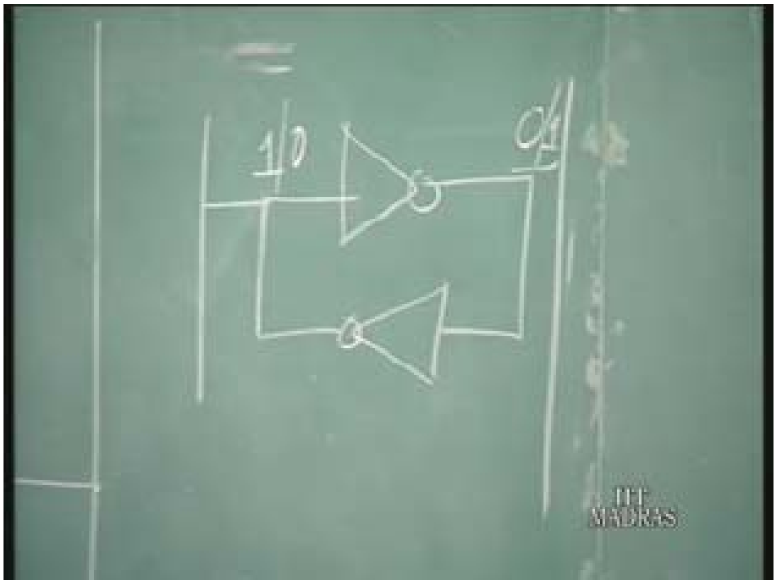
What is done is now instead of depending on the transistors in the cell to charge or discharge the data lines, we will use sense amplifiers to do so and these can be made quite large. What we have is we have here sense amplifier or each column we have a sense amplifier connected to the data lines.

(Refer Slide Time: 30:17)



Now what is this sense amplifier? Let us see. The sense amplifier is also nothing but two inverters connected back to back. Just draw it in a magnified form. Suppose this is the data line on which the cells are here, the sense amplifier consist of again 2 CMOS inverters. This is one inverter, this is another inverter and the output of one inverter goes to the input of the other as usual and these are connected to the data lines, this goes to V.DD. and the source goes to an input here which is the sense input and this goes to ground. This is the sense amplifier. Now this is nothing but two inverters back to back. If you have such a structure basically two inverters connected back to back like this, this is one inverter and this is another inverter and this is connected to two lines. You can have two possible combinations that is either this is 1 and this is 0. If this goes to 1 this will go to 0 and this 0 1 or you can have the other combination this is 0 and this is 1. There are two possible combinations which are stable.

(Refer Slide Time: 29:12)



Now if you have such a combination it will always tend to go to one of these two possible combinations and there is a large positive feedback involved. If it is maintained in the unstable condition, it will always tend to go to one of the two states which is favoured. Now what happens is if you consider this, if we maintain these two the data and data bar lines at the same voltage and then if by mechanism or by some reasons, one of those voltages are changed by a small amount.

Suppose this is v and this is v plus delta v, this voltage is higher compared to this and if we then turn on this sense amplifier since this voltage is higher, this will always tend to go to the higher value that is $V_{.DD.}$ and this one will go to ground. This one will go to logic 1 and this one will go to 0. On the other hand if this is v and this is v minus delta v that is this as the lower voltage here, what will happen is this side will tend to go to 0 whereas on the other side which is higher voltage will tend to go to 1 because of the large feedback, positive feedback here in this circuit any discrepancy here between the two voltage is going to be magnified or amplified and one line will go to 0 and the other line will go to 1 and this configuration will go to one of its two possible steady states.

This sense amplifier is included in this cell here. What we must do is we must initially have equal voltages on these lines. What we do is we do a pre-charge, both these lines the data and data bar lines are maintained in the pre-charged condition. What we have is here a pre-charge circuit which is nothing but two transistors here. These are two PMOS transistors and here you have a voltage $v_{.p.}$ which is the pre charge voltage and this is the pre charge line. Again a two phase clock is used. When the pre charge input here goes low, these two PMOS transistors they turn on and this pre charge voltage here which is available here is now available in the data and data bar lines. The data and data bar lines are both pre charged to this initial value $v_{.p.}$, pre charge voltage. (Refer Slide Time: 32:00)

When this pre charge input is low then both the data and data bar lines have the same values. Now when we select a particular cell the word line goes high, particular cell is selected. When a particular cell is selected what happens is as we said it can be in one of the two possible combinations. That is either if we look at the cell which is nothing but again two inverters connected back to back. If this is a cell here, there can be only two possible combinations. In one combination, if this is logic say 0 and this is 1 (Refer Slide Time: 33:52) then it means that this transistor is on, NMOS transistor is on and this is off and here if this is one, it means CMOS is on, this is off. This is one particular combination. If this is the particular combination what is going to happen is this line because the NMOS transistor is on tends to get discharged through the on transistor.

This voltage here on this line is going to tend to get reduced whereas this line, this PMOS is on so there is going to be a charging current and this voltage, this data line here tends to get charged up. If this is the condition of the cell then what is going to happen is one line is going to be pulled down and the other line is going to be pulled up. There is going to be a small difference in the voltages between the data and the data bar lines. Suppose the data and data bar line is charged to $v_{.p.}$ then one line will go above $v_{.p.}$, this will go below $v_{.p.}$ this side and this one will go above $v_{.p.}$. If the state of the cell was different then it would be in the different direction. Depending on the state of the cell, a voltage difference is created.

Now then what happens is the sense amplifier is turned on. Once the small difference in voltage see because the cell uses small transistors, it will take a long time for any large appreciable voltage difference between the lines to develop. When a small difference in voltage is developed say 0.1 or the sense amplifier is turned on. If this is the sense amplifier, when this transistor is turned on, this sense amplifier is activated. If this is off, the sense amplifier is off. When this sense amplifier is activated what happens is this is again going to go to one of the two states. If there is a small difference, initial difference say this is $v_{.p.}$ plus delta v and if this is $v_{.p.}$ minus delta v (Refer Slide Time: 35:21). This voltage being larger than this voltage, $v_{.p.}$ plus delta v being larger than $v_{.p.}$ minus delta v here. There will be a tendency for this voltage to go up to $V_{.DD.}$ that is the high voltage and this voltage to go down to ground.

Depending on the initial state of the data and data bar lines seen by the sense amplifier, one line is going to go high and the other line is going to go low. Depending on the information stored in the cell, one line is going to go high and one line is going to go low. That is how the sense amplifier is used to speed up the operation of the charging and since this is a big transistor, they can charge or discharge the data lines faster compared to the small transistors. Basically again, just to go through the operation of this memory, what is done when you are reading? Depending on the x address and the y address, a particular row is selected and a particular column is selected. When a particular row is selected, the cell is connected to the data lines and depending on the configuration or the status of the individual transistors in this cell, the data and data bar line voltages are going to change. The data and data bar voltages are kept at a pre-charged value, they are equated initially.

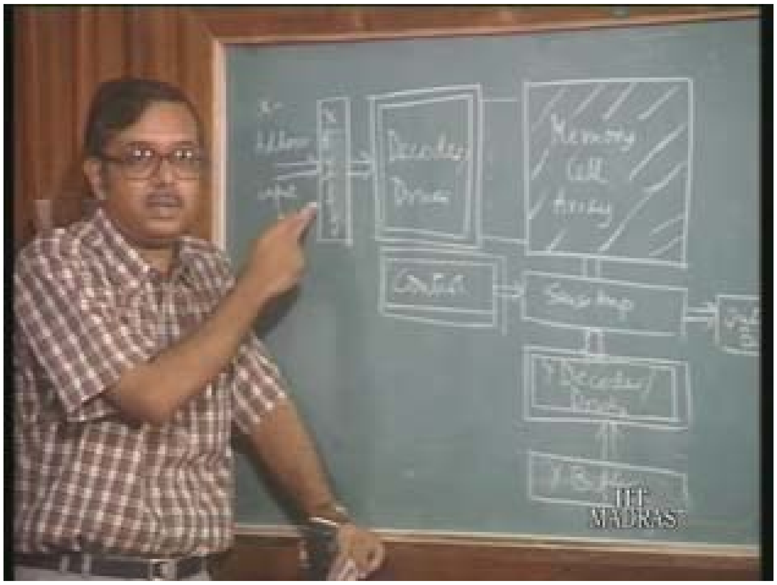
One line is going to go up by small amount and the other line will go down by a small amount then once a small difference is created, a sense amplifier is turned on. This is important that the sense amplifier is not turned on right away because what happens is if the sense amplifier is kept on initially then what happens is due to small noise voltages the sense amplifier may tend to go in one of the two directions, making the data line go to one of the two high or low directions. Once the small difference in voltages created due to the cell voltage, the condition of the transistors in the cell then the sense amplifier is turned on which makes one of the data lines go high and the other one go low, this happens in all the columns because one cell is selected in all the columns.

Now due to this configuration here, due to the presence of this pass transistors and the y decoder selects only one of the columns. One of the column is selected which is connected to the input output lines and the information on that particular column only is passed on to the I/O lines. For writing mechanism what is done is this I/O lines can be used to write information into the cells by forcing voltages on these data lines. If one line is made to go high and the other line is made to go low then you can basically write these into a particular cell. This is how a CMOS memory cell is going to work.

We have already seen that the BiCMOS has certain advantages over CMOS and this is especially in the areas of driving large capacitive loads and here we have seen as an example that the word line has large capacitance and for example a BiCMOS can be used to drive this large word lines. In fact the BiCMOS can be used in a number of areas in this memory, the static RAM cell which have been discussed. In fact BiCMOS memories have become quite popular and they have much improved performance when compared to the CMOS memory.

Now we shall take up the case of BiCMOS static RAM and see the particular areas where you introduce Bipolar transistors to speed up the operation of the particular static RAM. When you have BiCMOS technology available you would introduce bi polar transistors into all areas of the static RAM. You will have to introduce this bi polar transistors judiciously in certain areas. Now if you just look at the block diagram of the static RAM, you have the x address inputs, the driver decoder then you have the memory cell array and then you have the sense amplifiers and then you have the y decoder. These are the y inputs here, in fact in BiCMOS usually the interface circuits are the bi polar circuits. You have buffers here to convert the voltage levels for example if you are using ECL interface, from ECL to CMOS. You have a y buffer and you will also have x buffer here, I will put x decoder driver, y decoder driver and you have x buffer. This is the block diagram of the memory array.

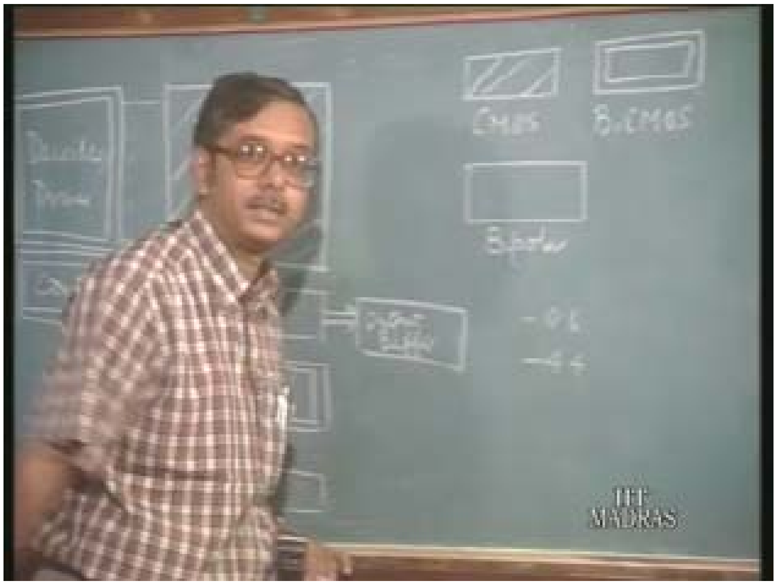
(Refer Slide Time: 48:24)



Now let us see which areas we would use and also you have some control circuitry here which actually is required to choose whether you want to do a read operation or a write operation and then you have the outputs coming, so you will have output buffers. For this which technology would be used in which regions? For example if you take the memory cell array which constitutes the major portion of the device of the cell, in terms of the area, this has to be in a technology which consume less area. Also since the maximum number of transistors are in this memory cell array, the total power dissipation of the cell is going to depend on the power dissipation of the memory cell array. This cell has to be made using such a technology which has low power dissipation and takes up less area. This memory cell array is basically done using CMOS.

The memory cell array will remain CMOS in BiCMOS memory. This is CMOS. The memory cell array even in a BiCMOS is going to be a CMOS. We don’t introduce bi polar transistors in the memory cell array because the bi polar takes up a lot of area and also the power dissipation is going to go up. What we are designing is for a large memory, this would still be CMOS. We know that in ECL for example, if this were made of ECL then the power dissipation would be quite lot and we cannot afford this in a large memory. Now what about this decoder driver? These decoder drivers would be using BiCMOS because these have to drive very large capacitances and we have already seen that this BiCMOS is capable of large driving large capacitance. We make this y decoder driver also using BiCMOS.

(Refer Slide Time: 49:18)



This is the symbol for BiCMOS that is you use basically use BiCMOS gates to drive these large capacitances. The decoder driver will be driving the large word line capacitances and also the y decoder driver has to drive large capacitances. These are basically the BiCMOS. The sense amplification in this case the sense amplifier usually in a BiCMOS memory is done using bi polar transistors. Why? Because bi polar transistors we have already seen has a very large trans conductance much larger compared to the CMOS counter parts and that is used for sensing in this memory.

Now a small voltage can be sensed much more easily using bi polar transistors. A small difference f of less than 100 milli volts can be sensed by a bi polar transistors which will result in a large change in current, which is not true in the case of MOS transistors because you require a much larger difference in the gate voltage to be really sensed. This is because of the larger trans conductance in bi polar transistor. Bi polar transistors are generally used in this sense amplifier. This is completely bi polar, just keep this bi polar. This blocks like this, this is a bi polar (Refer Slide Time: 47:32). Now the control again has a combination of bi polar and CMOS, this in the bi CMOS whereas this output buffers in general because these are supposed to be high speed circuits.

Many of these bi CMOS memories use ECL logic levels, the input and output levels are ECL. You require buffers to convert these levels into the standard CMOS levels because you see the memory cell array is actually using CMOS. What is done is it has to be converted to CMOS levels. In fact here you have negative voltages, ECL level is negative voltage if you are using an ECL level and this has to be converted and this has to work in a CMOS environment here inside. What is done is you have negative voltages but it is still in a CMOS type of levels that is suppose if you want 5 volt memory, this is converted into say in a BiCMOS you know that the output is going to go to V_{DD} minus V_{BE} to V_{BE} which is the lower level.

If zero is going to be V_{DD}, you can represent zero voltage as V_{DD} and you have minus V_{SS}, in fact the voltage levels internal to the circuit can be minus 0.6 to minus say 4.4. This minus 0.6 is considered logic high and minus 4.4 is considered logic low. Basically the CMOS is going to operate in a same way, it is just a level shifting down to a negative voltages. If you have 5 volt, you have a 5 volt full swing for this operation of the memory cell array. Only the voltage levels would be negative instead of positive. Zero volt would be high and minus 5 volt would be low. What we shall do in the next class is we shall study in detail about BiCMOS static RAM and see how a static RAM is realized in the BiCMOS environment.