

Microelectronics: Devices to Circuits
Professor, Sudeb Dasgupta
Department of Electronics and Communication Engineering
Indian Institute of Technology Roorkee
Lecture 60
Memory Design

Hello everybody and welcome to the NPTEL online certification course on Microelectronics Devices to Circuits. This is the last model of this whole course, by this we will be finishing this module or this lecture series on microelectronics. The module name is Memory Design and therefore we will be looking into the last section of the digital integrated circuits and one of the applications is usage of Digital Integrated Circuits from Memory.

So let me show you what is the outline of the course, the outline of the course is that we will be first introducing to you the memory itself and then looking into memory classifications, a basic building block and its architecture and then we will look at the memory core. Core means basically whenever you actually look into a memory typically we have a core here.

This core is the area where we are storing information and you will have peripheral here, you will have peripheral here which will be responsible for extracting or writing information within the core itself.

(Refer Slide Time: 1:31)

The slide is titled "Outline" and contains a list of topics with checkmarks. To the right of the list is a diagram of a memory core. The diagram shows a central box labeled "Core" with a vertical line on the left and a vertical line on the right, both labeled "P". A horizontal line crosses the core, with "P" on the left and "W/R" on the right. A red arrow points to the top of the core with the word "storing". A red arrow points to the bottom of the core with the word "manipulating".

- Introduction-Memory ✓
- Memory Classification ✓
- Memory Architectures and Building Blocks ✓
- The Memory Core }
 - a) Read-only Memories ✓
 - b) Nonvolatile Read-Write Memories ✓
 - c) Read-Write Memories (RAM) ✓
- Recapitulation }

IT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 2

So the core will be actually storing the data and the peripherals will be responsible for manipulating the data which means that should be able to read or write the data. So primarily read and write will be responsible for the peripherals, so we are looking into that. Then we

will look into ROM read-only memory and nonvolatile read-write memories and then read-write memory as well and then recapitulate the whole thing and that is what the whole structure is all about.

(Refer Slide Time: 1:57)

The slide is titled "Introduction: Memory" in blue text. To the right of the title, "Image Processing" is written in red cursive. Below the title, there are three bullet points, each with red underlines and a red circle around the phrase "two major concerns". The first bullet point says: "More than half of the transistors in today's high performance microprocessors are devoted to cache memories, and this ratio is expected to further increase." The second bullet point says: "Memory cells are combined into large arrays, which minimizes the overhead caused by peripheral circuitry and increases the storage density." The third bullet point says: "Reliability and power dissipation are two major concerns of the semiconductor memory designer." At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL ONLINE CERTIFICATION COURSE, and the number 3 in the bottom right corner.

So let me introduce memory here. So more than half of today's transistors when we use in a microprocessor the memory part of it are devoted to cache memories and this is the ratio expected to increase which means that the cache memory is going to increase. The reason is that the computation is also increasing and you need to store large amount of data. I will give you an example.

Whenever you are discussing for example let us say and image processing algorithm, then you are actually taking an image and then breaking it down into pixels and those pixels are getting processed at every stage. So if your image is basically high-density image, the number of pixels will be very large and therefore you require a memory to do that and therefore it is very-very important that the performance of the memory is kept at the optimal level.

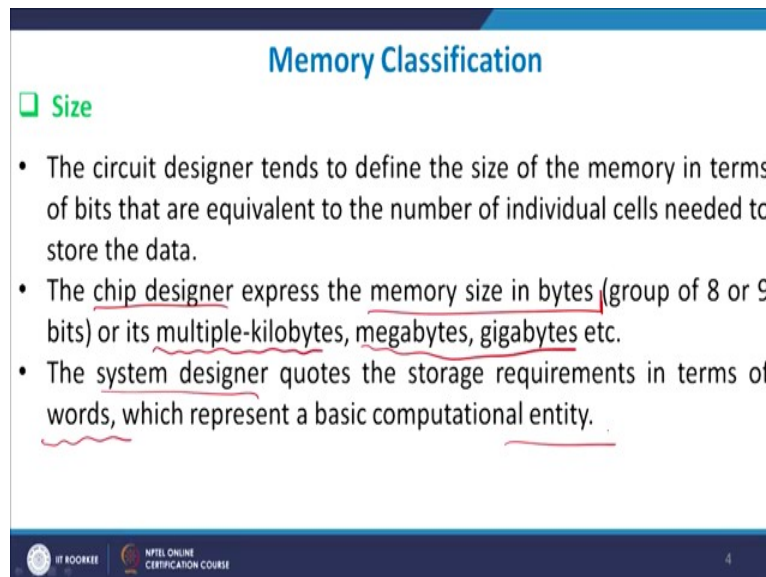
What we try to do is that memory cells are combined into large arrays, so it is an arrayed architecture and why is it so? So that it minimizes the overhead cost by peripheral circuit. So which means that if I have one core and one peripheral then it becomes very non-robust architecture because if there are N numbers of memory core storing N number of bits you require N number of peripheral circuitry.

And that makes life very difficult therefore what people have done is that your core memory and the peripherals are therefore multiplexed across the whole memory. And this increases

the storage density as I discussed with you. 2 important problems which are there in a semiconductor design is primarily, first of all is the reliability and the second is the power dissipation.

And these are the 2 major concerns of memory designer, reliability when I say, I mean to say are we able to reliably retrieve the value of the data from the memory court to the external port or you are able to reliably write down the value of the external data onto the memory, What is your excess time? What is your time taken to retrieve the data? So on and so forth. Let me come to the memory classification where we define that one core of memory will store maybe one bit of data, so if you have a 1 memory core we will see that later on we will store one bit of data.

(Refer Slide Time: 4:22)



Memory Classification

□ Size

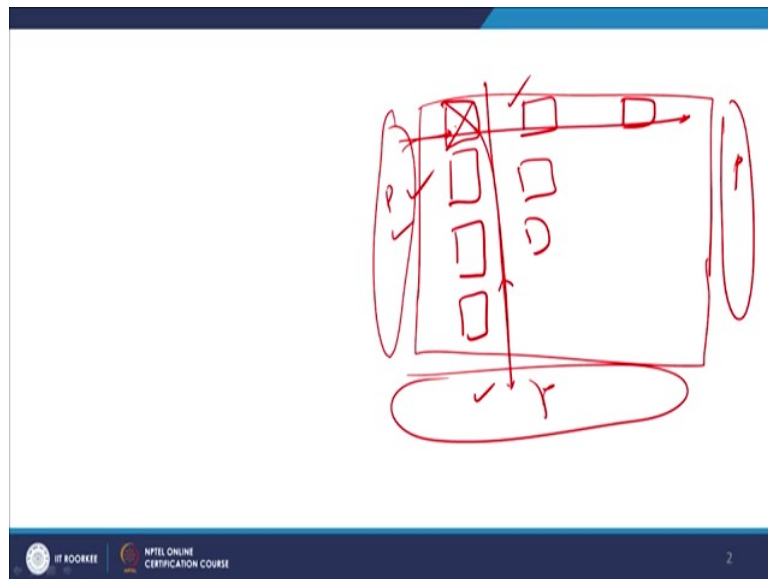
- The circuit designer tends to define the size of the memory in terms of bits that are equivalent to the number of individual cells needed to store the data.
- The chip designer express the memory size in bytes (group of 8 or 9 bits) or its multiple-kilobytes, megabytes, gigabytes etc.
- The system designer quotes the storage requirements in terms of words, which represent a basic computational entity.

IT 4008EE NPTEL ONLINE CERTIFICATION COURSE

Typically chip designers express memory size in bites so group of 8 we represent we can also represent in terms of multiple kilobytes, megabytes and gigabytes right the cache memory. Where a system engineer will, system designer will be storing in terms of words, how many words you are storing which is a basic competition entity for the memory in a sense. So these are the few areas in which we but for our purposes as an electronic engineer or designer we generally referred to as bits, so I will have 64 kb memory can have up to 56 memory kb, I can have 512 kb be memory and so on and so forth.

And therefore the array of the architecture which you require should be quite large in dimensions. Typically if you look at the array, it is typically square array, so if you want to generate a memory core.

(Refer Slide Time: 5:11)



The memory core maybe something like this big and this should be as square as possible. The aspect ratio should be as close to 1 as possible and the memory elements are all arranged in this fashion that is the general trend which you follow And this is the fashion in which you and then you will have peripherals here and you will have peripherals here and you will have peripherals here.

Then you will be recalling the data and you will be writing the data, you will be recalling this data and this data. So by this peripheral you will be switching on this say cell and by this you will be switching on this. So by using these 2 lines I can select any particular cell and generate right a data or read a data.

(Refer Slide Time: 5:52)

Timing Parameters Cont...


- The time require to retrieve from the memory is called Read-Access Time, which is the delay between the read request and the moment the data is available at the output.
- The time elapsed between a write request and the final writing of the input data into the memory is called Write-Access Time.

The diagram shows three horizontal lines representing signals: READ, WRITE, and DATA. The READ signal has two pulses, each labeled 'Read access'. The WRITE signal has one pulse labeled 'Write access'. The DATA signal shows a shaded region labeled 'Data valid' and a point labeled 'Data written'. A 'Read cycle' is indicated by a double-headed arrow above the first read access. A 'Write cycle' is indicated by a double-headed arrow above the write access. Red arrows and lines connect the read and write access points to the data valid period and the data written point.

Source: J. M. Rabaey, A. Chandrakasan and B. Nikolic, "Digital Integrated Circuit," PHI Learning Pvt. Ltd., 2011.

□ **Functions** Cont...

- Based on memory functionality it is classified as Read-Only memory (ROM) and Read-Write Memory (RWM).
- RWM uses active circuitry to store the data, so it belongs to the class of volatile memory, in which data is lost when the supply voltage is turned off.
- ROM belongs to the category of non-volatile memories. Disconnection of the supply voltage does not result in the loss of the stored data.
- The EPROM and E²PROM provides the facilities of both read-write functionality but comes under the category of non-volatile memories.

IT KOOBEE  NPTL ONLINE CERTIFICATION COURSE 6

Now typically the time required, so there are certain timing parameters. The time required to retrieve from a memory cell is defined as read access time. What is read access time? Already data is there in the memory core you send a read request and there is a certain time interval between the read request and the data coming to the output that difference in time is referred to as read access time or we say RAT.

So if you see, these are the read signals which you see. So the read signal goes high means you have sent a logic or you have sent to the signal to read the data from the memory core and the data has been valid only at this particular point. So the data has actually appeared in the output. So in time domain this is basically my read access time. Now the time elapsed similarly between right request and the final writing of the input data into the memory is defined as write-access time.

And therefore you see when I give write access this is my write going high and the data is written at this particular point till this much point then we define that to be as the write excess time. So this is the right cycle and this is the read cycle, so the idea is to make the read and write cycle closer to each other which means that the frequency of operation will increase and you should be able to do that.

But the problem is you cannot reduce the read and write cycle beyond particular limit and the reason being they will be always interconnect delays and so on and so forth which restricts the high frequency operation of the memory. So basic timing we have understood, we have also understood the basic concept of the size of a memory. Let me excellent you based on memory functionality what are the issues available to us.

So our read-only memory and we have read-write memory. Read-write memory uses active circuitry to store data. So it belongs to the class of volatile memory. Volatile memory is what? When you switch off the supply voltage the data is lost and we refer to that as a read-write memory as such. Whereas ROM belongs to the category of nonvolatile memory which means that the disconnection of the supply voltage does not result in the loss of the store data.

So for example whenever you switch on a computer or any of your devices for example even your mobile phone, tablets and laptops the operating system, the OS is actually loaded on the ROM. Because you do not want the OS to be installed every time you switch on the system, it should be already pre-installed in ROM. So even if you switch off the power, when the power is gone it is stored there and when you switch on the power the OS gets activated from ROM and you are able to do it.

There are 2 types of other one and this is EPROM electrically programmable ROM and electrically erasable PROM programmable read-only memory and these are actually examples of nonvolatile memory and these are the few examples which are there for the memory.

(Refer Slide Time: 8:51)

Access pattern Cont...

- Most of the memories are random-access memories, in which memory location can be read and written in a random order.
- Some memories are restricts the order of access, which results in either fast access time, smaller area or a memory with a special functionality. Examples of such memory are FIFO, LIFO etc.

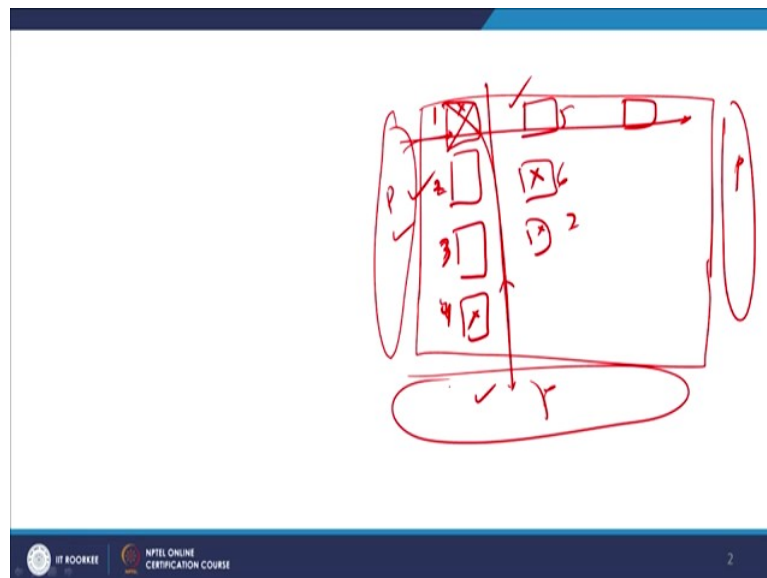
Read-Write Memory		Non-Volatile Read-Write Memory	Read-Only Memory
Random Access	Non-Random Access	EPROM E ² PROM FLASH	Mask-Programmed Programmable (PROM)
SRAM DRAM	FIFO ✓ LIFO ✓ Shift Register ✓ CAM ✓		

Source: J. M. Rabaey, A. Chandrakasan and B. Nikolic, "Digital Integrated Circuit," PHI Learning Pvt. Ltd., 2011.

7

Now how do you access the memory and that the most important part as far as timing and frequencies are concerned. Most of the memories which you will study is the random access memory means which is basically that I can read or write any cell in a random order.

(Refer Slide Time: 9:14)



It is not that for example the diagram which are have drawn suppose this is cell number 1, 2, 3, 4, 5, 6, 7 it is not that I will only be able to read 1 then 2 then 3 then 4, no. it is all random, so I can even read 6 first and then I come to 1 and then I come to 2 and then I come to number 4 and so on and so forth. So this is the random access operation of the memory.

(Refer Slide Time: 9:30)

Access pattern Cont...

- Most of the memories are random-access memories, in which memory location can be read and written in a random order.
- Some memories are restricts the order of access, which results in either fast access time, smaller area or a memory with a special functionality. Examples of such memory are FIFO, LIFO etc.

Read-Write Memory		Non-Volatile Read-Write Memory	Read-Only Memory
Random Access	Non-Random Access	EPROM EEPROM FLASH	Mask-Programmed Programmable (PROM)
SRAM DRAM	FIFO ✓ LIFO ✓ Shift Register ✓ CAM ✓		

Source: J. M. Rabaey, A. Chandrakasan and B. Nikolic, "Digital Integrated Circuit," PHI Learning Pvt. Ltd., 2011.

7

However there are certain memories which actually restricts the order of access but then the gain which you get out of it is, you get a faster access time and smaller area. For example FIFO and LIFO are some of the examples. So FIFO is first in first out and LIFO is last in first out. So if you look at read/write memory which is basically your volatile memory.

We have 2 we have got SRAM static random access memory and DRAM which is dynamic random access memory. So this is a random access and in the nonrandom access we have FIFO, LIFO, shift register and CAM. So content addressable memory, now you see for example shift register. You can have 4 type of shift register. You can have parallel in, parallel out, PIPO, you can first in first out, you can have first in last out, so on and so forth.

So there are 4 types of shift register which are all giving you a nonrandom access to you whereas in nonvolatile memory which is there, we will have EPROM, flash and EEPROM then read-only memory will have ROM and PROM. This is PROM which is programmable read-only memory and these are all read-only memory are obviously as I discussed with you non-volatile memory.

(Refer Slide Time: 10:59)

Memory Architecture and Building Blocks

- To implement an N-word memory where each word is M bits wide, then the most intuitive approach is to stack the subsequent memory words in a linear fashion.

Source: J. M. Rabaey, A. Chandrakasan and B. Nikolic, "Digital Integrated Circuit," PHI Learning Pvt. Ltd., 2011.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE

Let see what are the basic memory architecture because people have been doing it for quite a long time. And what people have done is, that I can just show you here is that you store information in terms of bits and these bits are stored in this word. So there are storage, so you see this mart here. Your brown colored mart is basically a cell which stores one bit of information.

And similarly for 8 bits of information we define this to be as one word here. And similarly you will have large number of words between this point to this point. So there are m bits, let us suppose there are m bits and there are n number of words. So there are n number of words with each word having m-bit and you can have therefore M into N is the total number of bits which you can store in this case.

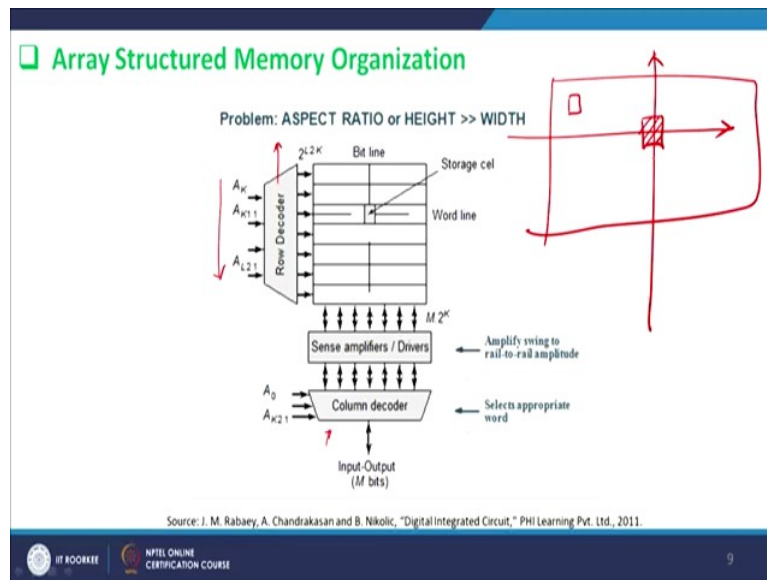
So what people do is that, they put a decoder here in the input side. What does decoder do is, that it basically tries, so if there are say 8 number of select lines S_0 to S_7 Say there are S_0 to S_7 select lines you are choosing which word you want to read or write then there are 8 lines which is available to me then I require a 3 is to 8 decoder. So there will be 3 address lines which you will be inserting.

So you see K is basically your $\log N$ to the base 2. So K is equals to $\log N$ to the base 2 Where N is the number of word which you see. So if the number of word is say 8, so $\log 8$ you will get, so K will be equals to $\log 8$ by $\log 2$, so this you can break down in 2 to the

power 3. So this will be $3 \log_2$ by \log_2 and therefore \log_2 gets cancelled out, K equals to 3. So I get K equals to 3 means you require 3 words here.

3, A_0 , A_1 and A_2 as the decoder input bits, so you require this into consideration.

(Refer Slide Time: 13:06)

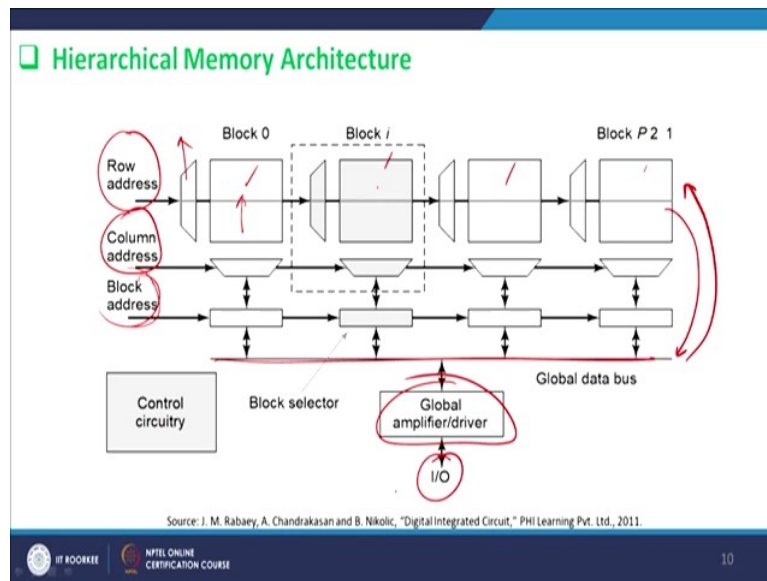


Similarly, typically you can have Row decoder as I have discussed with you with values of A here and you can also have column decoders which is shown here and this column decoders will be attached with sense amplifiers, you do not have to worry what is this but column decoders and Row decoders allow you to choose a particular bit within the array structure of the memory code.

So you have an array structure of memory code. So what I am trying to tell you is that I have any array structure of memory port and let us suppose I want to choose this then I will switch on this column, this row and this column. When I choose using the address decoder I am able to choose this one now I can do reading and writing. So decoders helps you to choose a particular cell and then I can read and write over the particular cells.

Cell amplifiers are drivers which are primarily responsible for making the circuitry fast because you want to read a cell, so I will require that the voltage should come to the output side at the very fast manner and that is required. It also allows you to give you a rail to rail swing. So V_{DD} to 0 swing will be available through this sales amplifier design. This is the hierarchical memory architecture, if we go step-by-step, hierarchy.

(Refer Slide Time: 14:18)



Then we will see that we have Row addresses, we will have column addresses and therefore this row will have Row decoders here and therefore they will be Block 0, Block 1, Block 2, Block 3 so on and so forth and these column addresses will be responsible for taking the column of the block, particular block I will insert the block address here and these will be responsible for talking with these blocks for extracting it.

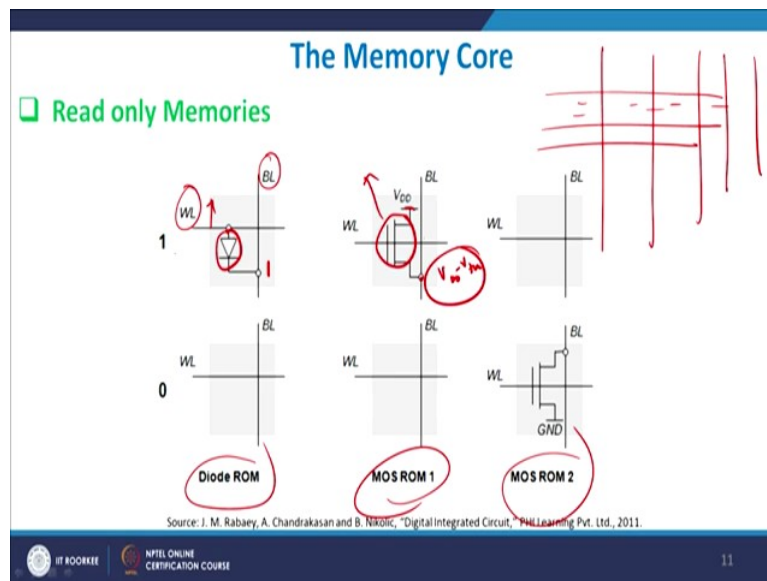
Then the global bus will be responsible for extracting the data or writing the data onto the system which will be driven by this global driver or amplifier and of course there will be control circuitry, there will be a clock which will be driving it and so on and so forth and it will be finally governed by the IO. So if you look very carefully to the whole hierarchical architecture, it is governed by multiple IOs here.

Then you will have amplifier here and then you will have a global data bus over which the data is being fed from external world into the memory or from the memory into the external world and then you have got Row addresses, column addresses and block addresses to ascertain from where the data is to be written or extracted. Say you want to go for the seventh block Ward number 3 bit number 7, let us suppose.

Then these 3 information will be fed by a block, column and row addresses and then you can fetch the data from the memory architecture.

Let us look at the memory code this is only the read-only memory which is basically a nonvolatile memory which I have discussed with you.

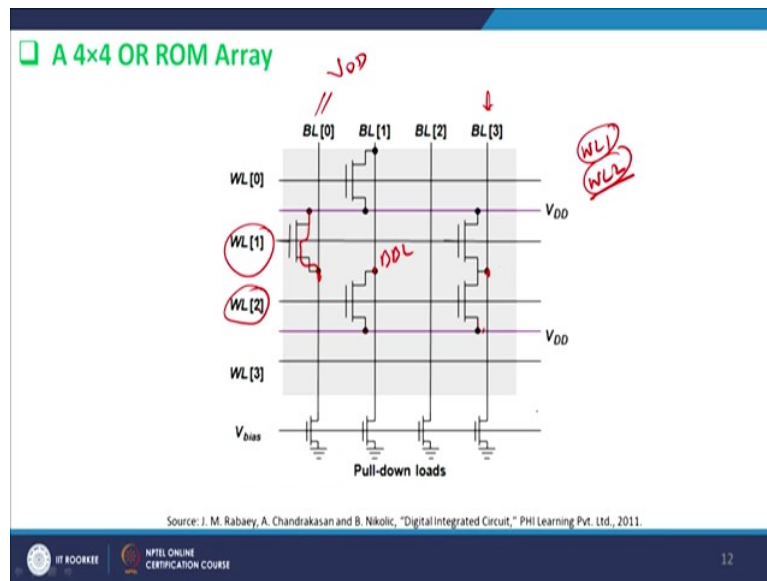
(Refer Slide Time: 15:46)



And therefore for example if the word line is high these diode switches on and one appears on the bit line a very straightforward and simple. Similarly when the word line height this MOSFET switches on, this V_{DD} appears here at a particular point of course there will be one threshold voltage drop of the transistor but that is immaterial at the stage. Similarly, so this is a diode based ROM.

So I have a bit line here, I have a word line here. So word line runs like this for selecting a particular row and the bit lines are responsible for selecting the particular column. Similarly you will have MOS ROM 1, so this is the MOS ROM 1 and this is again the MOS ROM 2 where you have ground and bit line and so on and so forth you can have various architectures with the memory core.

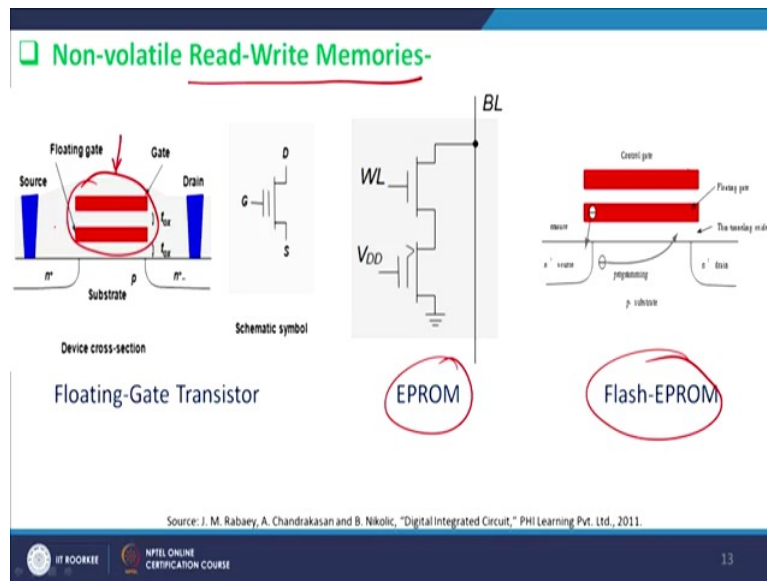
(Refer Slide Time: 16:31)



Let us look at the 4 by 4 ROM array. In this ROM array it is all MOS based ROM array and therefore let us suppose your WL[1] is high then what it does is, this V_{DD} through these appears on bit line, so your bit line equals to V_{DD} . Similarly if your WL[2] is high then I ensure that V_{DD} is written onto BL[1], similarly if WL[2] is high then I will ensure that this V_{DD2} is written on this bit line

Similarly if WL [1] is also high and WL [2] is also high, 2 word lines really they become high, if either of these 2 lines are high then I will get a high-end BL [3] and there what I do therefore is, I can therefore extract the information at particular cross-section for the output world and that is what we have been doing in this case.

(Refer Slide Time: 17:23)



Let us look at the nonvolatile read/write memories. Nonvolatile is switch off your memory will be lost. You'll be losing its data across it. One of the example is the floating gate and then you have an EPROM and you do have a Flash-EPROM. I will request you to study all these if you are interested in this area study, good book on memory as far as these read/write memories are concerned.

We do not have time to go into details of each memory architecture to give you an idea but these are actually your nonvolatile memory so if you switch it off even, so for example this floating gate. If you look at the floating it whenever you switch it off the data is actually stored in this floating gate as a capacitor, so ideally it stores data for infinitely long duration of time and that is the example of a nonvolatile read/write memory.

(Refer Slide Time: 18:18)

Read Write Memories (RAM)

- Storage in RAM memories is based on either positive feedback or capacitive charge.

a) Static RAM

Source: J. M. Rabaey, A. Chandrakasan and B. Nikolic, "Digital Integrated Circuit," PHI Learning Pvt. Ltd., 2011.

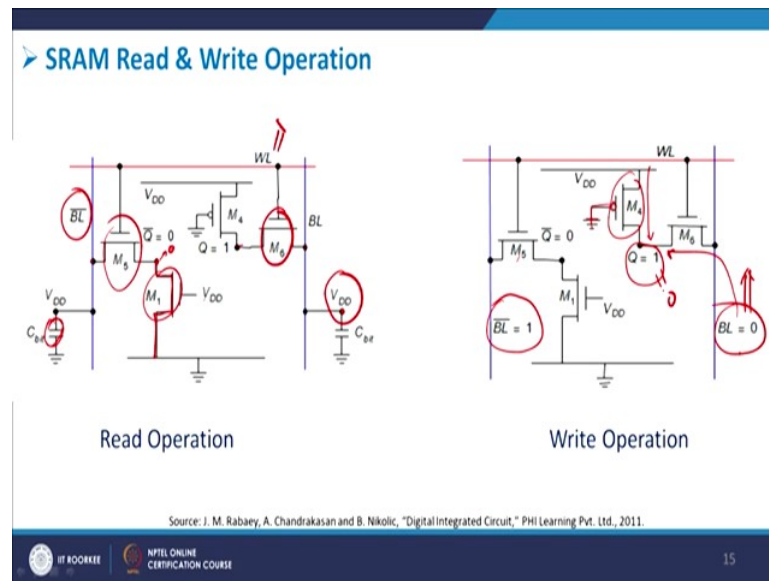
14

Let us look at the static RAM and one of the first example of a static RAM is basically your 6T cell or 6 transistor cell. So if you look very carefully this M_5 and M_6 are referred to as access transistors and this is basically cross coupled inverter so if you just do one small job here it looks something like this, it looks something like this. So this is your word line and this is your bit line and bit line bar. So this is your bit line bar and this is your bit line.

If I store 1 here automatically 0 will be stored here. 0 will feed into this inverter make 1 here. So this behaves as a latch which means that I am showing one bit of data here in this latch right? Now if I want to write or read a data I just have to switch on my WL. When I switch on WL equals to 1 then M_5 and M_6 switches on and then this bit line and bit line bar or bit line and bit line bar gets access to the value of Q and Q bar here.

So if it is 1 here, your hundred percent sure 0 will be here. The 0 will appear here and 1 will appear here. So if you want to read a cell, you have written a cell, suppose Q equals to 1 and Q bar equals to 0. Simply make the word line high, once you make the word line high the access transistor switches on because it is NMOS transistor. As it switches on the Voltage goes onto bit and bit bar line and you are able to therefore read the data from that particular cell.

(Refer Slide Time: 19:50)

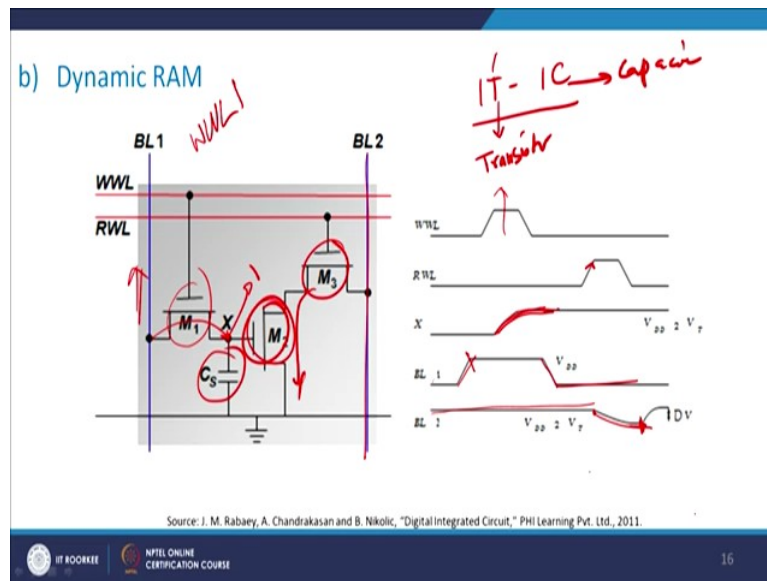


As I have discussed with you the read operation is something like this that you have Q here word line goes high, switches on this M_6 and you are able to transfer this Q onto bit line bar with the one voltage, threshold voltage drop, this appears across this therefore you have a V_{DD} appearing here. Whereas when you have bit line bar getting activated this goes on, M_5 goes on.

Since \bar{Q} equals to 0, so I have pulled on getting activated because V_{DD} was attached to it. This is on and therefore this goes to 0. When this goes to 0, 0 also appears as bit bar line and it is stored in this C bit here, fine. So you have a C bit stored here and that is the read operation. If I come to the right operation what we try to do is at the same exactly the same as the previous case the only thing is that you are already bit line bar here 1 and bit line equals to 0.

You again switch on WL because you want the access transistors to be on and then you allow the 0 to be written through this arm. So you had initially Q 1 0 here, now if this goes to 0, if it goes to 0, this is switched on and this V_{DD} appears here, so you have to make it bit line voltage slightly higher to overwrite those value and therefore this will be converting into a 0. So this is the right operation for SRAM memory.

(Refer Slide Time: 21:01)



Now another example is basically my DRAM which is dynamic RAM and it is consisting of also referred to as 1T-1C design which is one transistor, so this is one transistor and one capacitor and therefore if you see very closely even WWL is high, which you see, this is high. This high means M_1 is switched on, when M_1 switches on The bit line is suppose 1.

The bit line is 1 and bit line to, so I have got bit line 1 is high at this stage and bit line 2 is remaining constant, let us suppose and the RWL which is the read line has gone now high so it has high here, so when WWL is high it switches on M_1 and therefore the value of voltage here starts to rise, why does it start to rise is that, initially if your BL 1 is high and your M_1 is switched on this voltage will be written on this X and CS will gets charged and that the reason you see an exponentially rising function at X And that gives you arising function.

Now let us suppose you want to read a cell, so read this particular cell then you make it read high, once you make it read high then let us suppose the bit line has gone down then you will see that at this particular point bit line 2 which is this one Bit line 2, so when your reading goes high this M_3 switches on you had stored data here one, this will switch on your M_2 and what will happen is, this voltage will fall to 0 and therefore you see that your bit line 2 is going down to 0 And that way you can actually have the operation of dynamic RAM.

So dynamic RAMs are therefore responsible for 1T 1C cell, so this capacitance CS is able to store the data till the next read or write cycle comes into picture So this is what the basic timing diagram of this is.

Let me recapitulate for what we have done for this memory design. You can actually look into there are certain very good books on memory, you can have a look into it. We can discuss it on discussion forum also regarding the books and open sources from where you can get this data. What we have discussed is that, we have discussed the memory time, what is the meaning of various timing definitions, look at the access pattern and the application.

We also looked into the fact that we have looked into the basic read-write memory, volatile memory and ROM which is basically a nonvolatile memory. Finally we went for RAM and we looked into 6T cell, 6 transistor SRAM cell and we looked also into 1T 1C DRAM cell which is basically a volatile memory and example of that and how to read and write an information from a cell

So these basic concepts I have tried to make clear in this particular module, I hope you have enjoyed this module and we look forward to meeting you during the NPTEL course structure when this is released. Thank you very much.