Organization
The basic element of a semiconductor memory is the memory cell. Although a variety of electronic technologies are used, all semiconductor memory cells share certain properties:

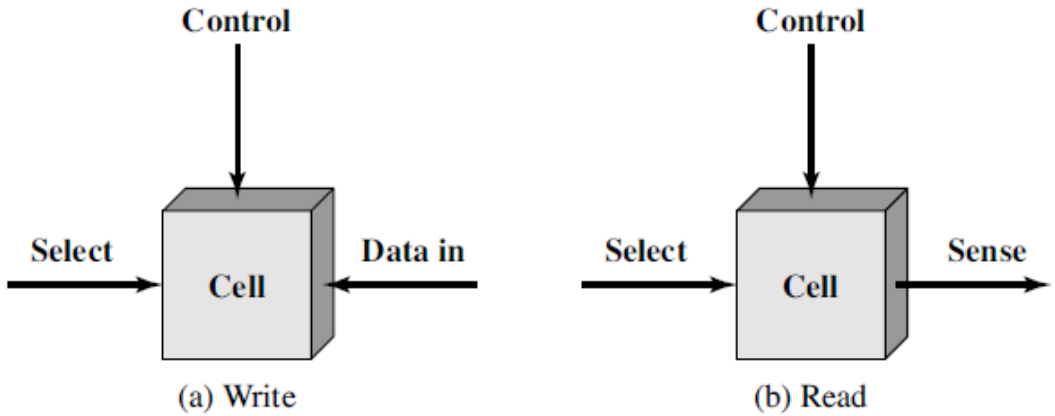
• They exhibit two stable (or semistable) states, which can be used to represent binary 1 and 0.
• They are capable of being written into (at least once), to set the state.
• They are capable of being read to sense the state.

Most commonly, the cell has three functional terminals capable of carrying an electrical signal. The select terminal, as the name suggests, selects a memory cell for a read or write operation.

The control terminal indicates read or write. For writing, the other terminal provides an electrical signal that sets the state of the cell to 1 or 0.

For reading, that terminal is used for output of the cell's state. The details of the internal organization, functioning, and timing of the memory cell depend on the specific integrated circuit technology used and are beyond the scope of this book, except for a brief summary.

For our purposes, we will take it as given that individual cells can be selected for reading and writing operations.




One distinguishing characteristic of RAM is that it is possible both to read data from the memory and to write new data into the memory easily and rapidly. Both the reading and writing are accomplished through the use of electrical signals.
The other distinguishing characteristic of RAM is that it is volatile.

A RAM must be provided with a constant power supply. If the power is interrupted, then the data are lost. Thus, RAM can be used only as temporary storage. The two traditional forms of RAM used in computers are DRAM and SRAM.

# DRAM AND SRAM

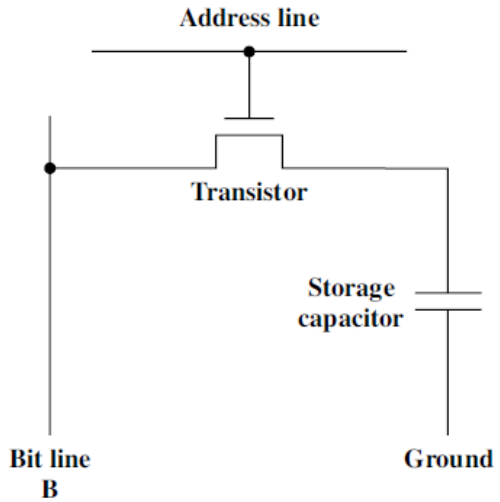
DYNAMIC RAM RAM technology is divided into two technologies: dynamic and static. A dynamic RAM (DRAM) is made with cells that store data as charge on capacitors. The presence or absence of charge in a capacitor is interpreted as a binary 1 or 0. Because capacitors have a natural tendency to discharge, dynamic RAMs require periodic charge refreshing to maintain data storage. The term dynamic refers to this tendency of the stored charge to leak away, even with power continuously applied.

Figure shows a typical DRAM structure for an individual cell that stores 1 bit. The address line is activated when the bit value from this cell is to be read or written. The transistor acts as a switch that is closed (allowing current to flow) if a voltage is applied to the address line and open (no current flows) if no voltage is present on the address line.

For the write operation, a voltage signal is applied to the bit line; a high voltage represents 1, and a low voltage represents 0. A signal is then applied to the address line, allowing a charge to be transferred to the capacitor.

For the read operation, when the address line is selected, the transistor turns on and the charge stored on the capacitor is fed out onto a bit line and to a sense amplifier. The sense amplifier compares the capacitor voltage to a reference value and determines if the cell contains a logic 1 or a logic 0. The readout from the cell discharges the capacitor, which must be restored to complete the operation.

Although the DRAM cell is used to store a single bit (0 or 1), it is essentially an analog device.The capacitor can store any charge value within a range; a threshold value determines whether the charge is interpreted as 1 or 0.

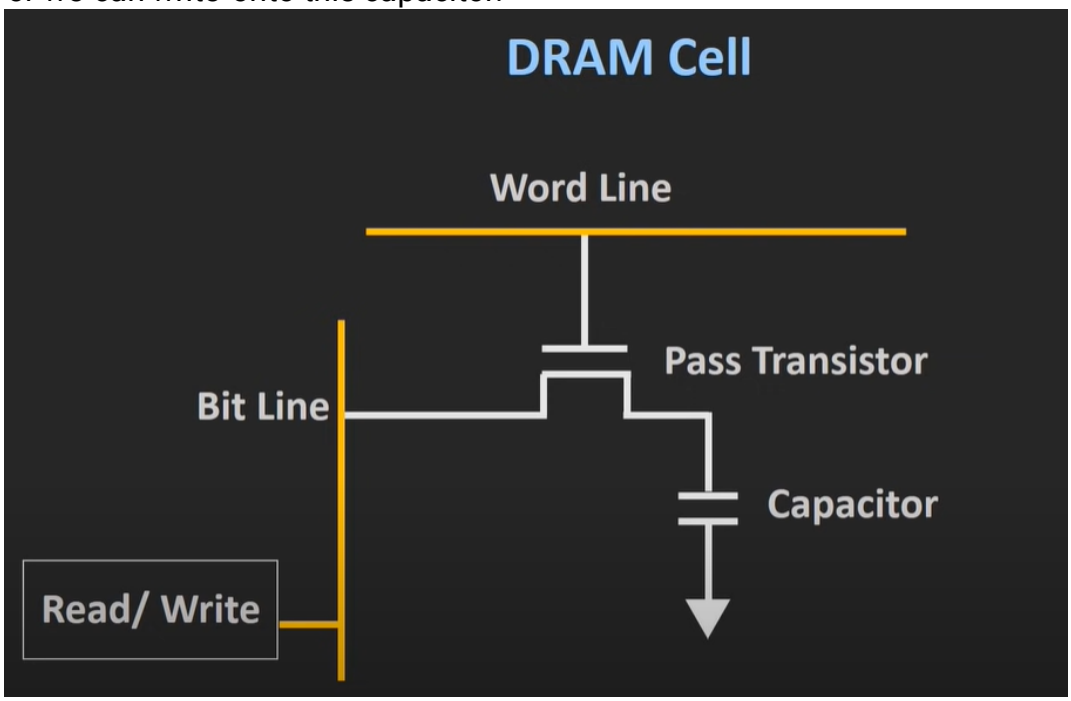



(a) Dynamic RAM (DRAM) cell

**WORKING OF DRAM:**

So, in case of the DRAM cell, the memory bits are stored in the form of charge cross this capacitor. So, by charging and discharging the capacitor, we can know that whether the bit that is stored inside this capacitor is logic 1 or logic 0.

So, now in case of this DRAM cell, we can access this capacitor by using this pass transistor. So, when this pass transistor is turned ON, then we can read the capacitor data or we can write onto this capacitor.

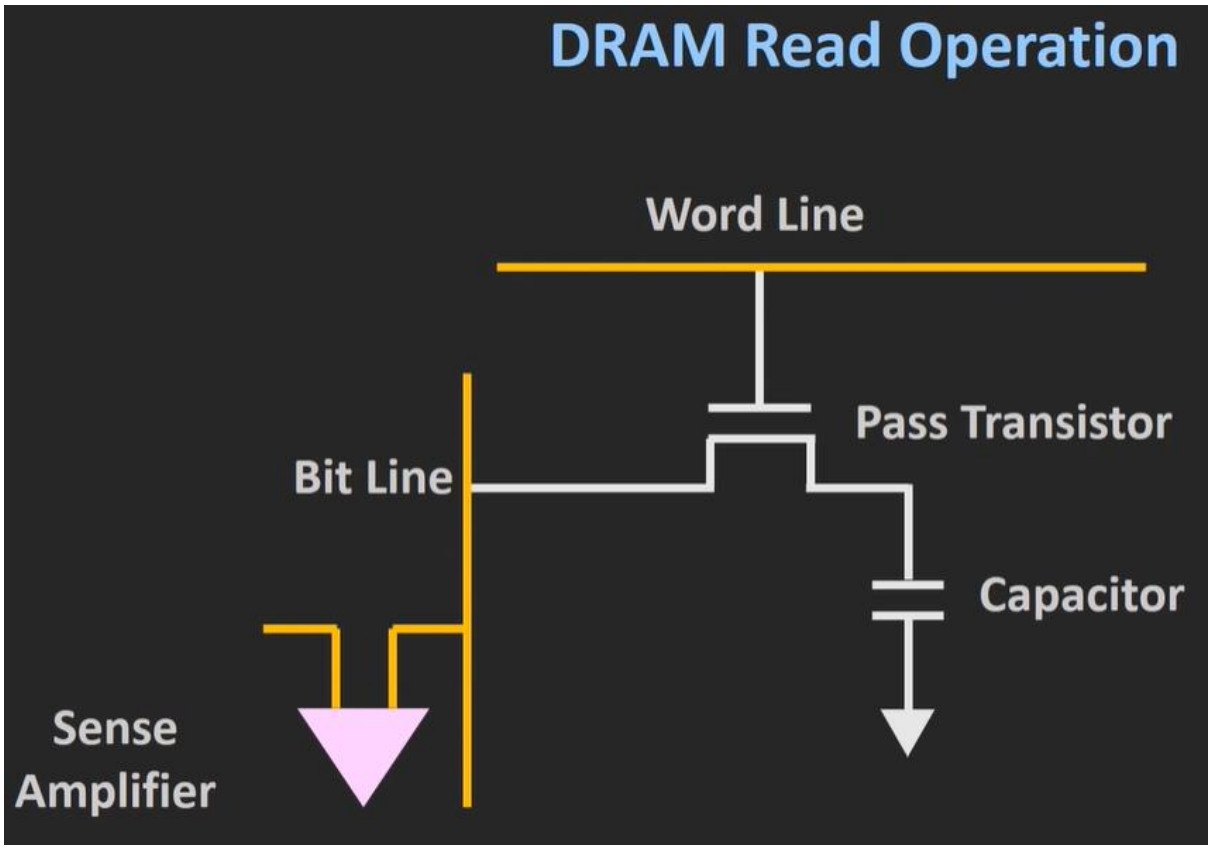



And when this pass transistor is OFF, then the charge across the capacitor should remain as it is.

So, in the ideal case, this capacitor should not lose its charge. But in the actual case, if you see, there will be some leakage current and because of that, the capacitor will lose its charge gradually.

And that is the reason, this dynamic cell requires the periodic refresh cycles. And that is the reason behind it, why this memory is known as the Dynamic RAM.

So, now as we know about the internal structure of this dynamic RAM, let us see, how the read and write operations are being performed on this dynamic RAM.




So, like I said to **read the data** of this DRAM, first of all, we need to turn on this pass transistor. And that can be done by applying the voltage to this word/address line.

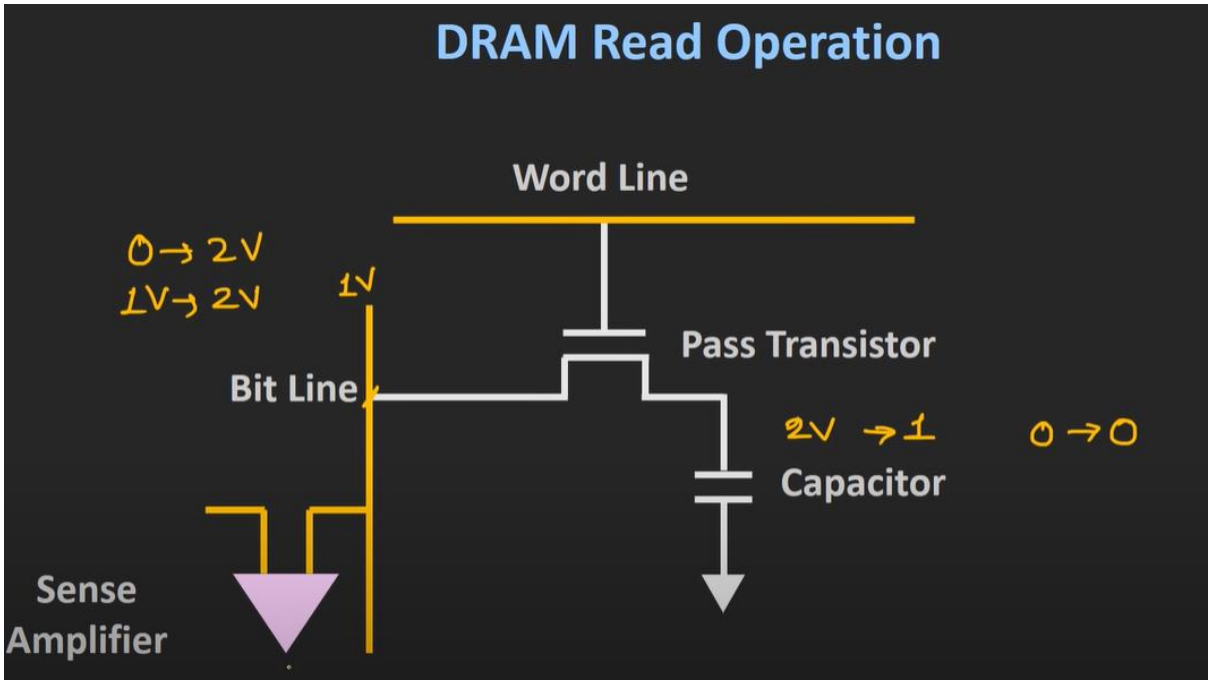
So, once the voltage is being applied to this word line, then the charge across this capacitor will be available at this bit line.

So, just by using the sense amplifier, we can read the voltage that is available at this bit line.

So, now if you observe this **read cycle**, in this read cycle once this pass transistor is ON, then the data or the charge across this capacitor will be available at the bit line.

So, eventually, this capacitor will lose its charge during the read operation. So, such kind of read operations is known as the destructive read operation. And to avoid that we need to perform the refresh operation after every read cycle so that the capacitor can get its original charge. So, now to increase the read operation speed this bit line is charged with the finite voltage.

Now, let us understand, why this precharge is very important. So, let's say, initially, the voltage across the capacitor is 2V. And the 2V represents the logic 1. And the 0V represents the logic 0.



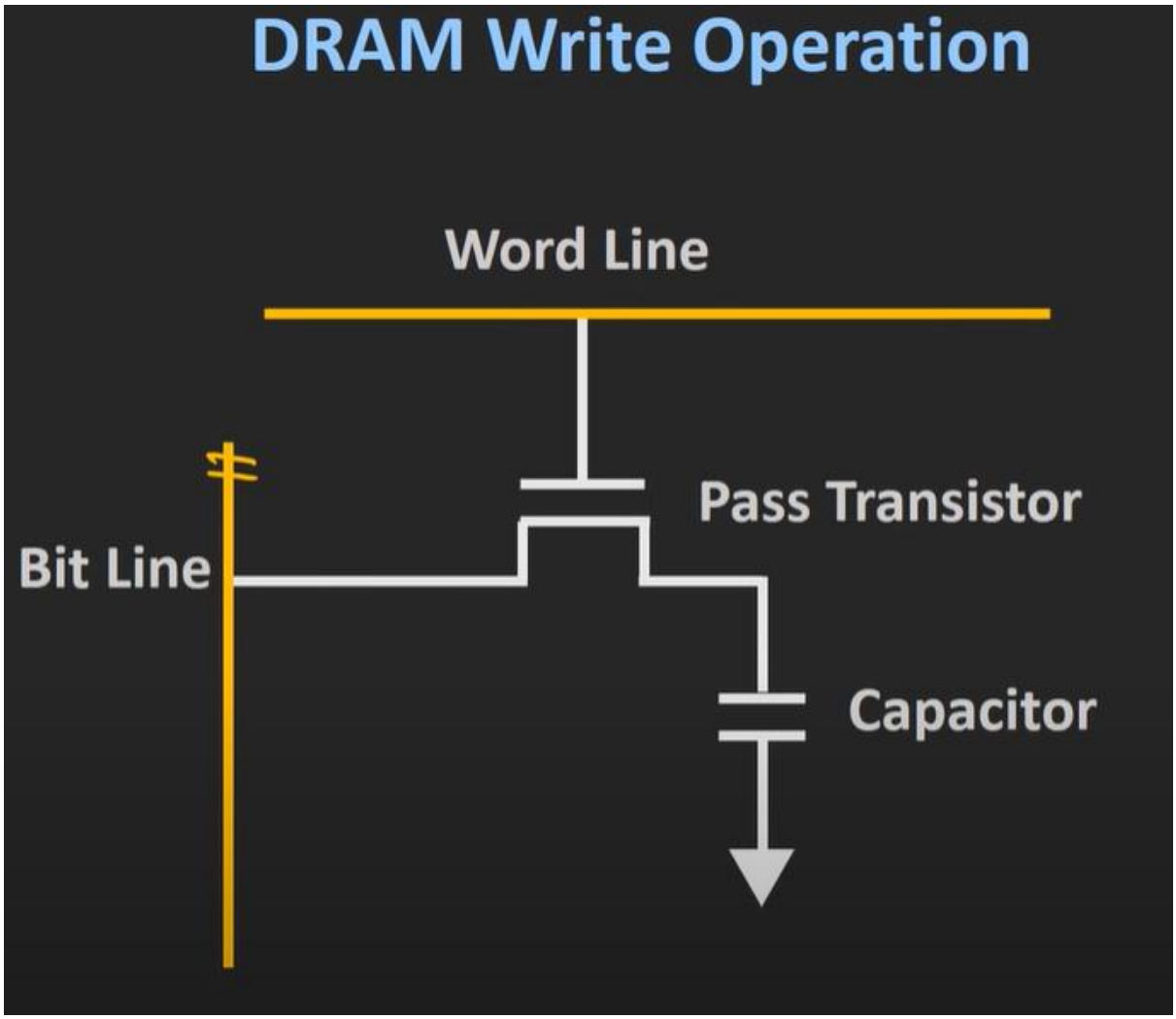

So, once the read operation starts, this pass transistor will be turned ON and the voltage across this capacitor gradually will be available at the bit line.

So, the bit line will start charging from the 0V to the 2V And once it will reach from 0 to 2V, then this sense amplifier will get to know that the voltage across the capacitor is 2V or logic 1.

But instead of that suppose if we have precharged this bit line to the 1V, then the time that is required to reach from 1V to the 2V will be almost half.

So, in this way, by pre-charging this bit line by some finite value, we can reduce the read time or we can increase the read speed.

So, in this way, by using this method, we can reduce the read time of this dynamic RAM. So, similarly now let's see how the write operation is being performed.




So, during the **write operations**, all the bit lines are being precharged with some finite value. And for the particular bit line on which we want to write, the bit voltage is applied to that particular bit line.

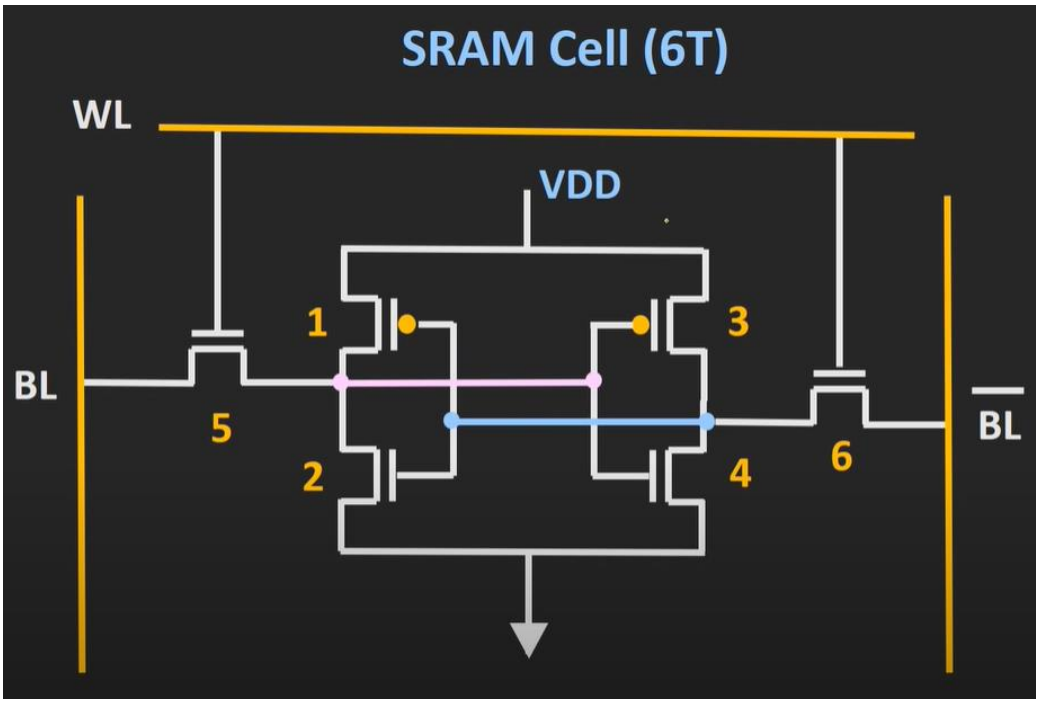
So, let's say, we want to charge this capacitor then the bit voltage will be applied to this particular bit line. And then after this pass transistor is turned ON.

So, whatever voltage that is available at this bit line, will be transferred to this capacitor. And in this way, we can write on this DRAM.

So, this is how the read and write operations are being performed on this DRAM. So, now as this DRAM involves the capacitors, so the reading and writing speed of this DRAM depends upon the charging and discharging time of this capacitors.

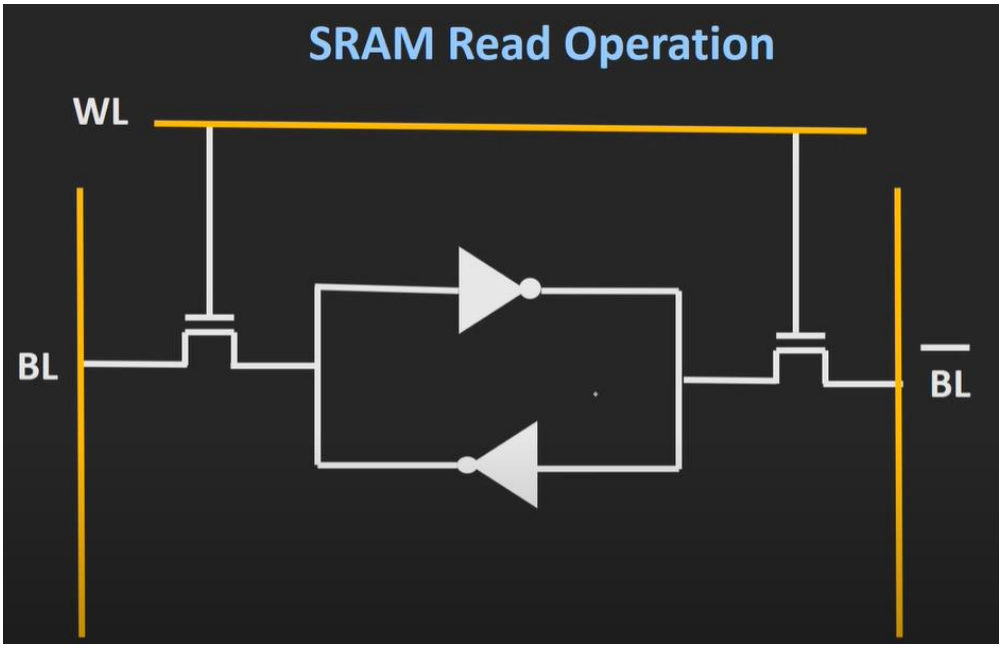
**WORKING OF SRAM**

Similarly, let us see the internal structure of this SRAM and let us see how reading and writing is performed on this SRAM.




So, if you see the internal structure of this SRAM, it consists of 6 transistors. So, out of the 6 transistors, the two transistors are the pass transistors which will give access to the bit lines, while the remaining four transistors are the two cross-coupled inverters. So, here this transistor 1 and 2, is the first CMOS inverter pair and the transistors 3 and 4 are the second CMOS inverter pair.

So, if you see the simplified circuit, then the simplified circuit will look like this.



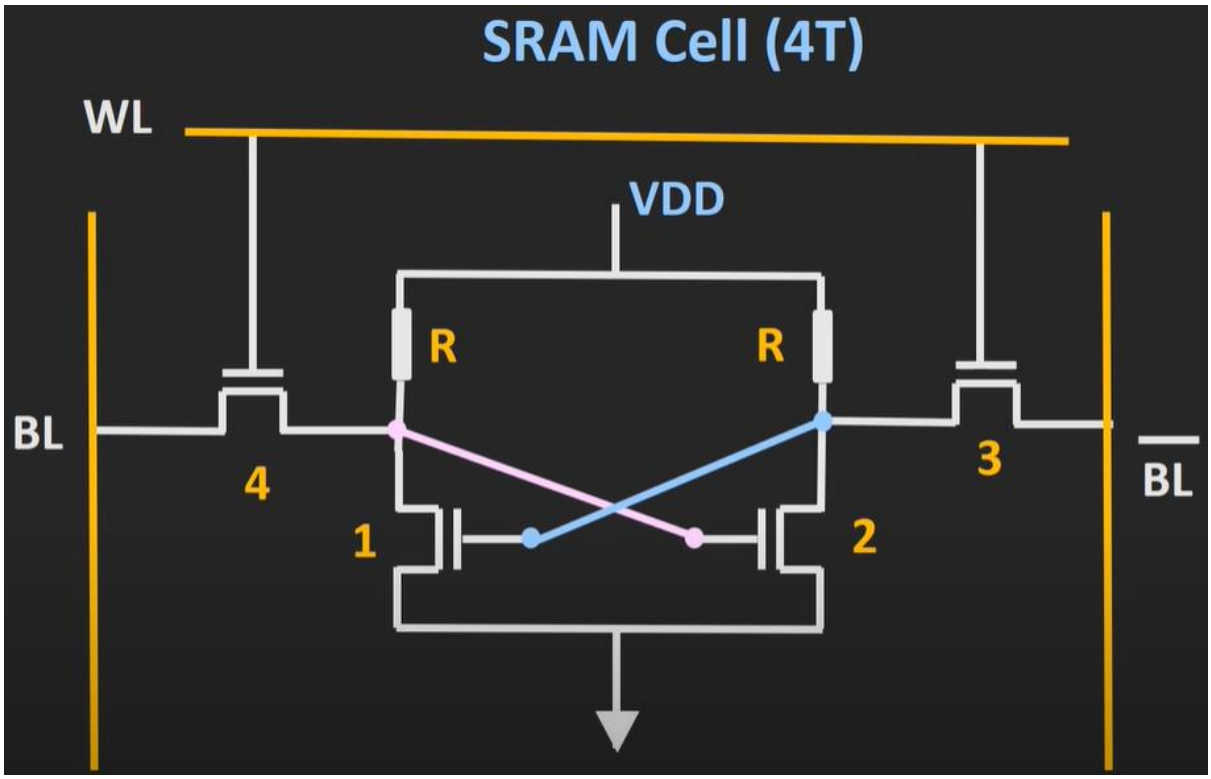
So, in case of this SRAM cell, the memory bit is stored between this two cross-coupled inverters.

So, let us say if we have latched logical 1 then at the output of the first inverter we will have logic 0.

And again at the output of the second inverter, we will have logic 1. So, as far as the power is supplied to this SRAM, the logic 1 will be get circulated between these two inverter pairs.

So, unlike in case of the dynamic RAM cell, we do not require any kind of refresh cycles during this SRAM operation. And that is the reason, this SRAM is known as the static RAM.

So, apart from this 6 transistor design of this SRAM cell, we also have 4 transistor design.




So, apart from this 6-transistor design of this SRAM cell, we also have 4 transistor design. So, in case of this 4 transistor design, the p-MOS are replaced by the high impedance resistors.

So, in this way by using the 4 transistor design, we can reduce the number of bits that are required for the 1 bit of storage. But the disadvantage of this 4 transistor design is that the continuous power will be get dissipated across this resistors.

So, whenever we require the less power consumption, then the 6T design is more preferred over this 4 transistor design.

As in the DRAM, the SRAM address line is used to open or close a switch. The address line controls two transistors (T5 and T6).When a signal is applied to this line, the two transistors are switched on, allowing a read or write operation. For a write operation, the desired bit value is applied to line B, while its complement is applied to line . This forces the four transistors (T1, T2, T3, T4) into the proper state.For a read operation, the bit value is read from line B.

Now, if you see this random-access memory or RAM, it is the essential part of any computing device. So, whether you see the laptop, or mobile or any gaming console you will find this RAM on all these devices.

**Generation of DRAMS**

- ## Asynchronous DRAM

  So, if you see the very old generation of dynamic RAM, they were the asynchronous dynamic RAM.

  So, it means that the RAM is not synchronized with the CPU clock.

  Now, the disadvantage of this type of RAM was that CPU does not know the exact timing at which the data will be available from the RAM on this input-output bus.

- ## Synchronous DRAM (SDRAM)

  This problem has been overcome by the next generation of RAM, which is known as the synchronous DRAM.

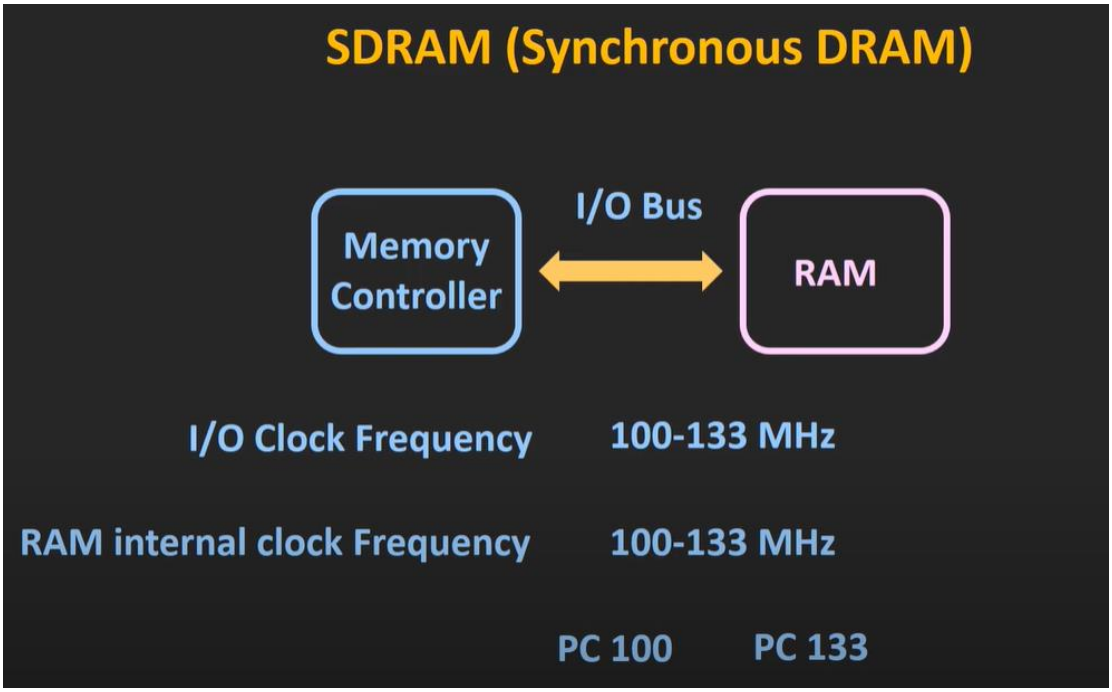
  So, in case of this SDRAM, the RAM is synchronized with the CPU clock.Now, the advantage of this SDRAM is that the CPU or to be precise memory controller exactly knows the timing or the number of cycles after which the data will be available on the bus.

  So, CPU does not need to wait for the memory access. And because of that we can increase the memory read and write speed. So, now we before we understand about this RAM, let's see the different terminologies which are quite often used with this dynamic RAM.

  So, if you see this RAM, we have total two types of different frequencies. The first is the input-output clock frequency and the second is the RAM internal clock frequency.

  So, this input-output clock frequency is the frequency at which the data is being transferred between the RAM and the memory controller.

And the internal clock frequency of the RAM is the frequency which is being used by the RAM for the internal operations.

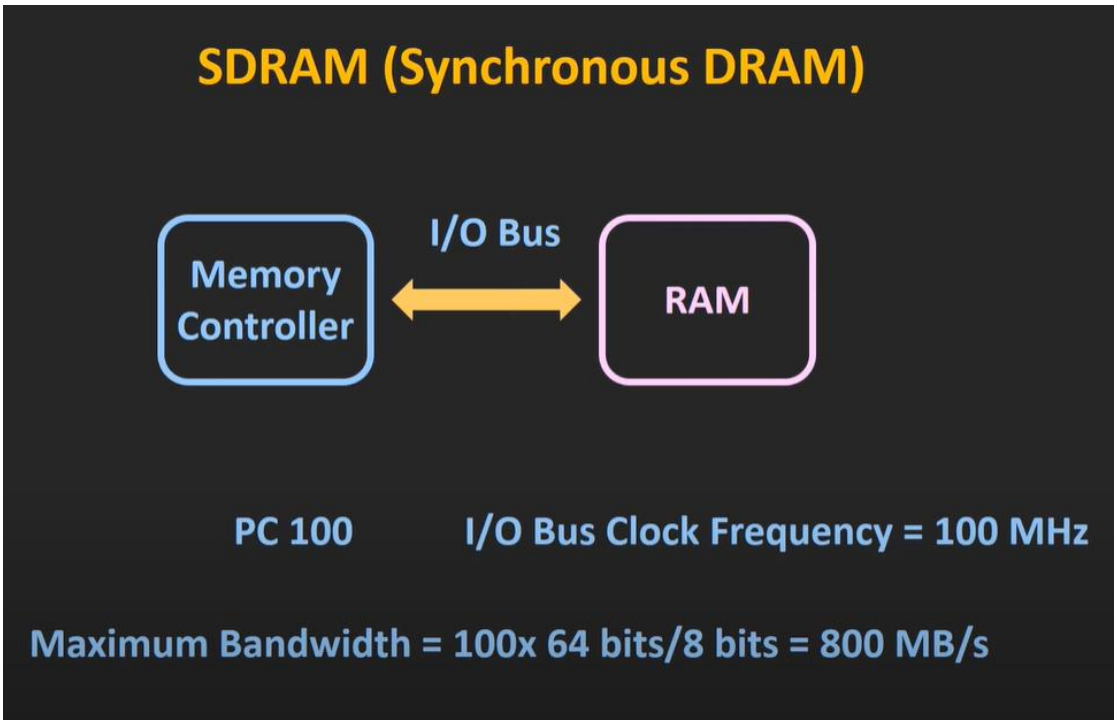



So, now here in case of this synchronous DRAM this input-output clock frequency and the internal clock frequency of the RAM are same.

So, suppose if the internal clock frequency of the RAM is 100 MHz then the input-output clock frequency is also 100 MHz.

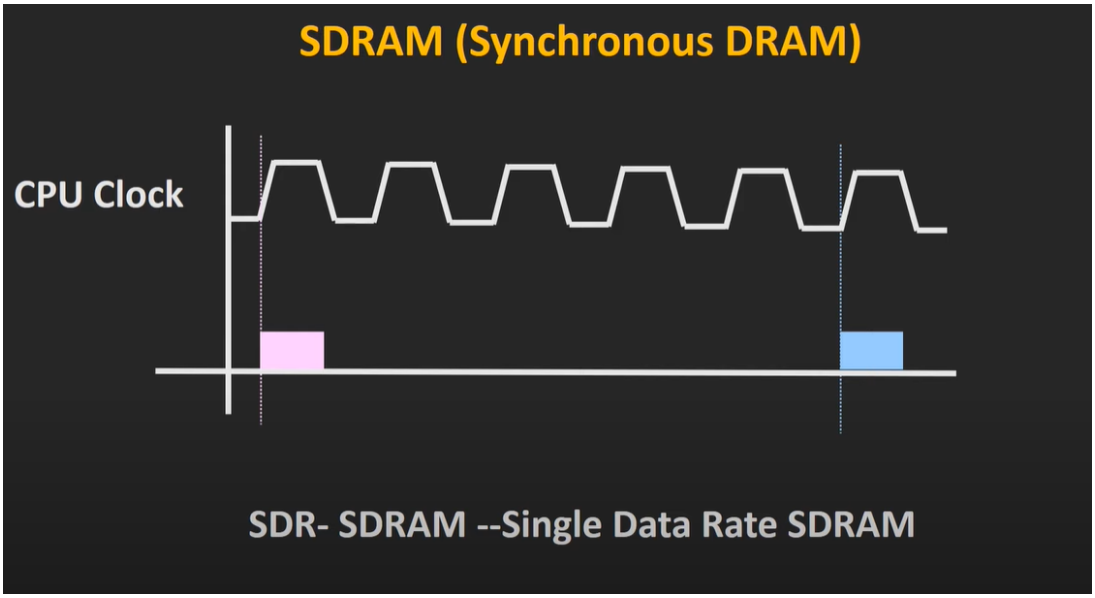
And generally, if this synchronous DRAM the operating frequency is in the range of 100 Mhz to 133 MHz.

So, suppose if you find the PC-100 on the SDRAM module, it means that the input-output bus clock frequency is 100 MHz.




And data that is being transferred between this RAM and the memory controller is at the rate of 100 mega transfer per second.

And if this bus is 64 bit wide, then the data rate in terms of the bits per second will be equal to 100 MHz into 64 bits. And if we convert it into the bytes(per second) then it will be divided by the 8 bits. That is 800 Megabytes per second.




So, now the synchronous DRAM modules are operated at 3.3V. Now, this SDRAM or synchronous DRAM is also known as the single data rate SDRAM.

Because in this RAM, the data is transferred at the every rising edge of the clock cycle.

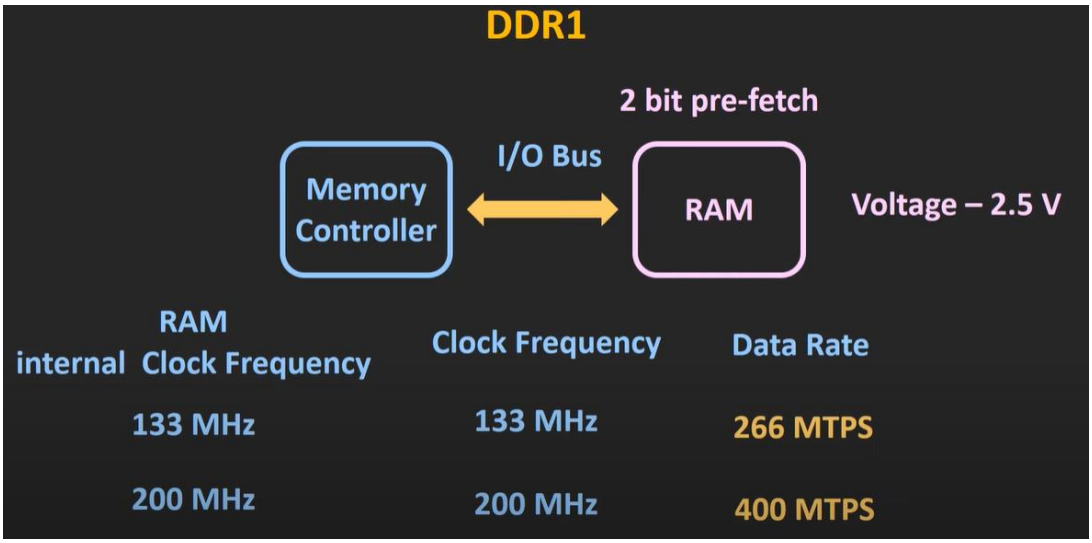
But if you see the next generation of synchronous DRAM they are known as the DDR RAM. Because in case of this DDR RAM, the data is transferred twice during the clock cycle.

That is during the rising edge as well as during the falling edge. So, in this way, the data is being transferred twice during each cycle. And that is why it is known as the double data rate or DDR SDRAM.

So, if you see this DDR RAM, there are different generations in this DDR RAM. Starting from the DDR1 up to the DDR4. And nowadays, the memory that we use inside the desktop, laptop or mobile that is either DDR3 or DDR4 RAM.

So, let's see the different generation of this DDR RAM one by one.

- **DDR1**




So, the first generation of DDR RAM is known as the DDR1 RAM.So, compared to the SDRAM here, the voltage has been reduced from 3.3V to the 2.5V.

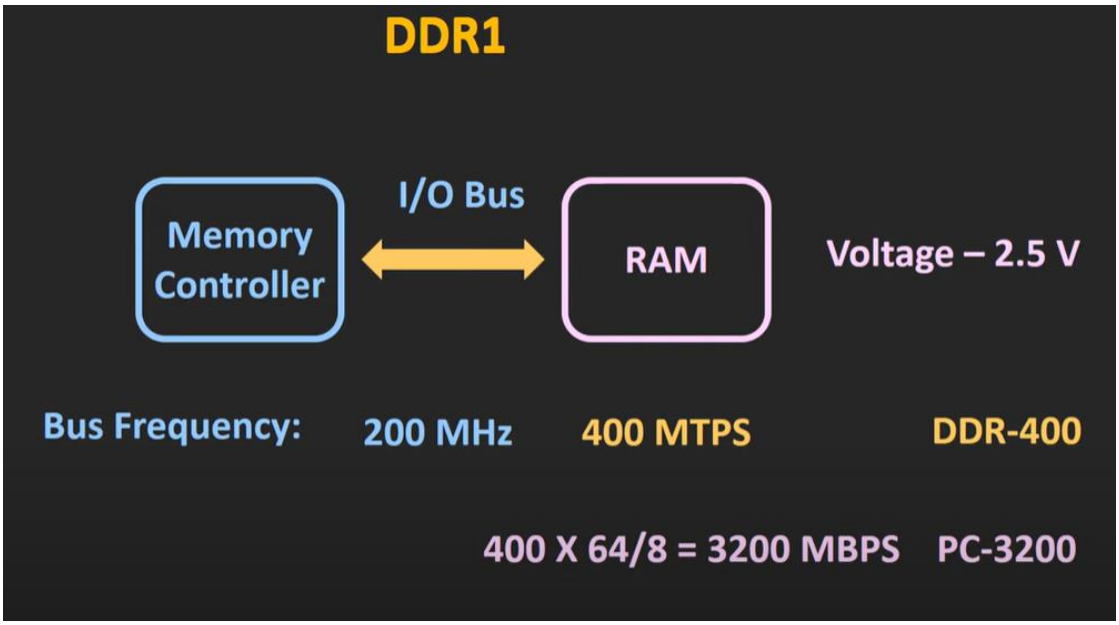
So, like I said earlier in case of DDR RAM, the data is being transferred both during the riing as well as the falling edge.

So, we can say that in the single clock cycle, instead of 1 bit, 2 bits are being pre-fetched. So, that is generally known as the 2 bit pre-fetch. Now, here in DDR1 RAM, the internal clock frequency, as well as the input-output bus clock frequency, are same.

So, generally, this DDR1 RAM is operated in the range of 133 MHz up to the 200 MHz. But if you see the data rate at the input-output bus, it will be double compared to the clock frequency.

As, in case of this DDR RAM, the data is transferred both during rising as well as the falling edge. So, suppose if you are operating this DDR1 RAM at 133 MHz then you will see the datarate as 266 Mega transfer per second.

So, suppose if the bus frequency is 200 MHz then the data transfer rate will be equal to 400 Mega transfer per second. And if the input-output bus is 64 bits wide, then the data rate in terms of the bytes per second will be equal to 3200 Megabytes per second.




Now, these DDR RAMs are generally denoted by the term DDR followed by the transfer rate of this RAM. And if you see the DDR1 module or DDR1 stick, then on that stick you will find that the term which is used as PC-3200.

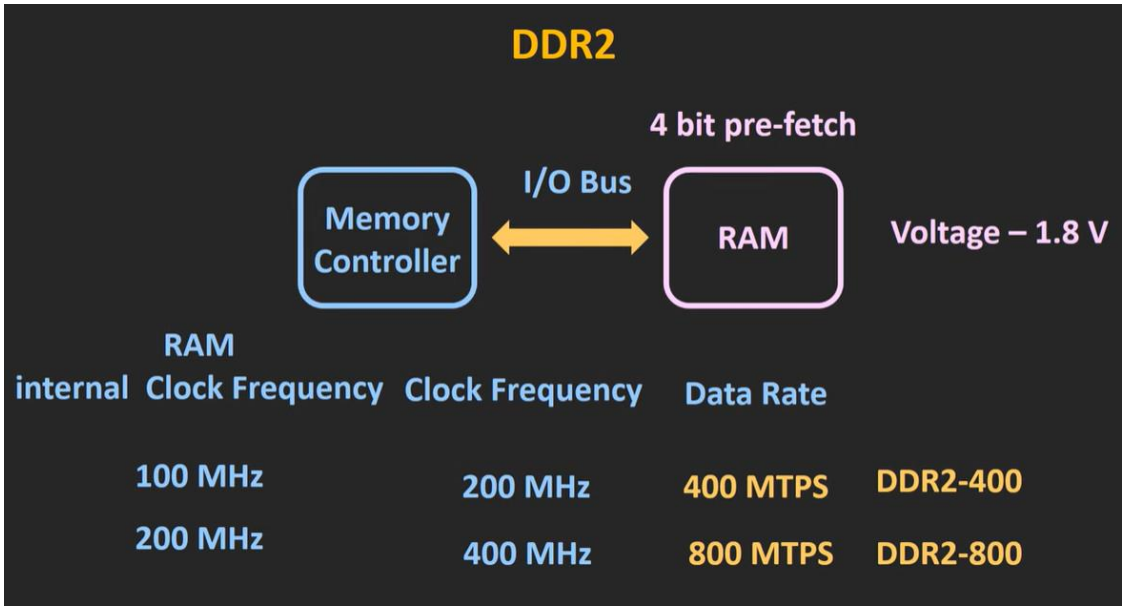
It means that the maximum speed or the maximum bandwidth which can be achieved by this DDR1 RAM is 3200 Megabytes per second. So, after the first generation of DDR1 RAM, the second generation of DDR RAM is DDR2 RAM.

- **DDR2**

So, now in case of this DDR2 RAM, it is operated at 1.8 V instead of 2.5 V. And if you see the internal RAM clock frequency, the internal RAM clock frequency is same as
the previous generation.

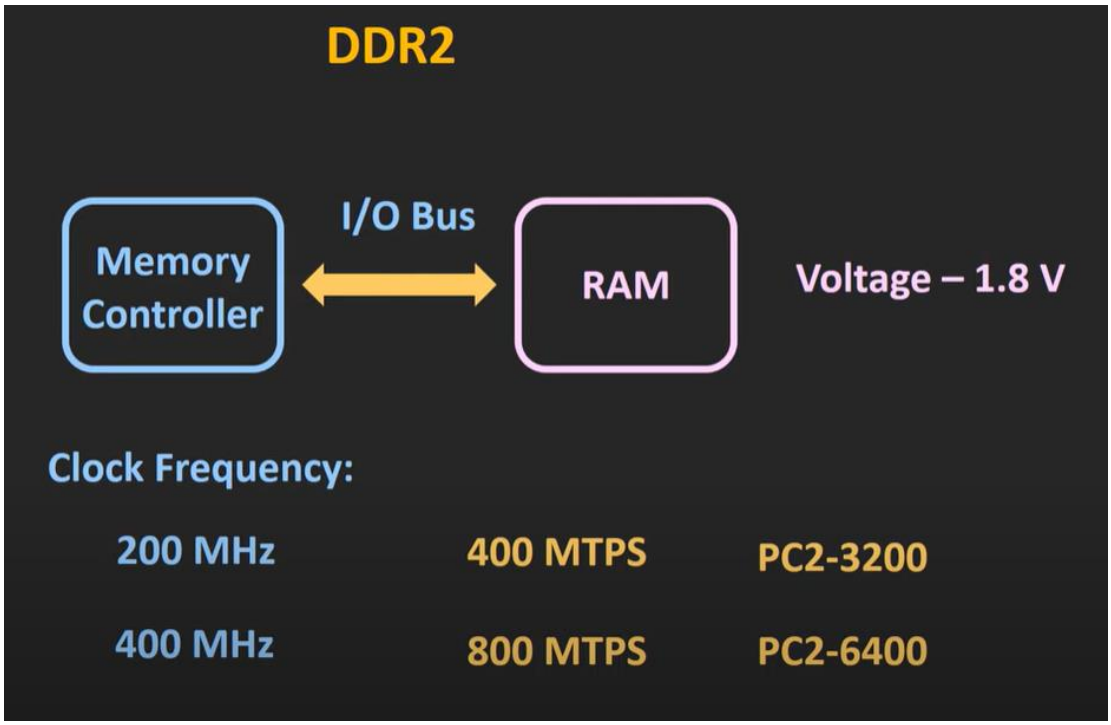
But here the data rate is double compared to the last generation. And that is being achieved by increasing the number of bits that is being pre-fetched during each cycle.

So, in case of this DDR2 RAM instead of 2 bits, here 4 bits are pre-fetched during each cycle. Or in a simple way if I say, in case of this DDR2 RAM, the internal bus width of this RAM has been doubled. So, suppose if the input-output bus is 64 bits wide, then the internal bus width of this RAM will be equal to 128 bits.




So, in this way, in a single cycle, we can handle double amount of data. And to handle the same amount of data, the clock frequency of this input-output bus should be get doubled. So, suppose this DDR2 RAM, is operated at 100 MHz internal clock frequency then the input-output bus should have the clock frequency of 200 MHz.

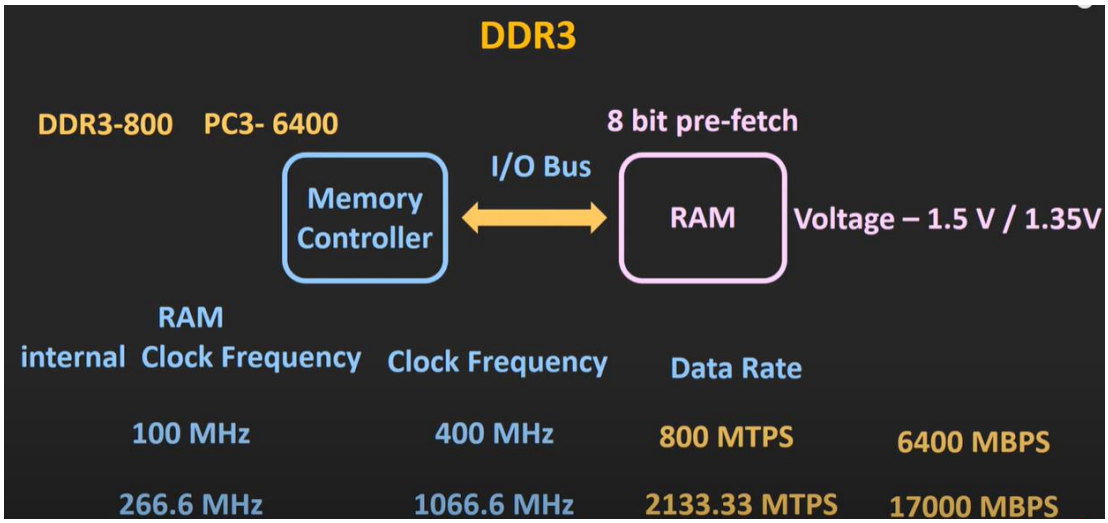
And in case of this DDR RAM, as data is transferred both during rising and falling edge, so the data rate will be doubled compared to the clock frequency, that is 400 mega transfer per second.



And in terms of the DDR terminology, it can be written as DDR followed by the transfer rate. So, now suppose if DDR2 RAM is operated at 400 MHz clock frequency, then the data rate will be equal to 800 mega transfer per second. And in terms of the terminology, it can be written as DDR2-800.

And if you see the DDR2 module, on the module it will be written as PC2-6400. Which means the data rate in terms of the bytes per second.

- **DDR3**




So, after the second generation, the third generation of DDR RAM is the DDR3. So, in case of this DDR3 RAM, the voltage is further reduced from 1.8V to the 1.5V.

Now, if you see the internal clock frequency of the RAM, it is slightly improved compared to the last generation.

But the data rate that you can achieve with the same frequency has been doubled compared to the previous generation.

Because in case of this DDR3 RAM, here the number of bits that is being pre-fetched has been increased from 4 bits to the 8 bits.

Or in a simple way, if I can say, the internal data bus width of RAM has been increased 2 times compared to the last generation.

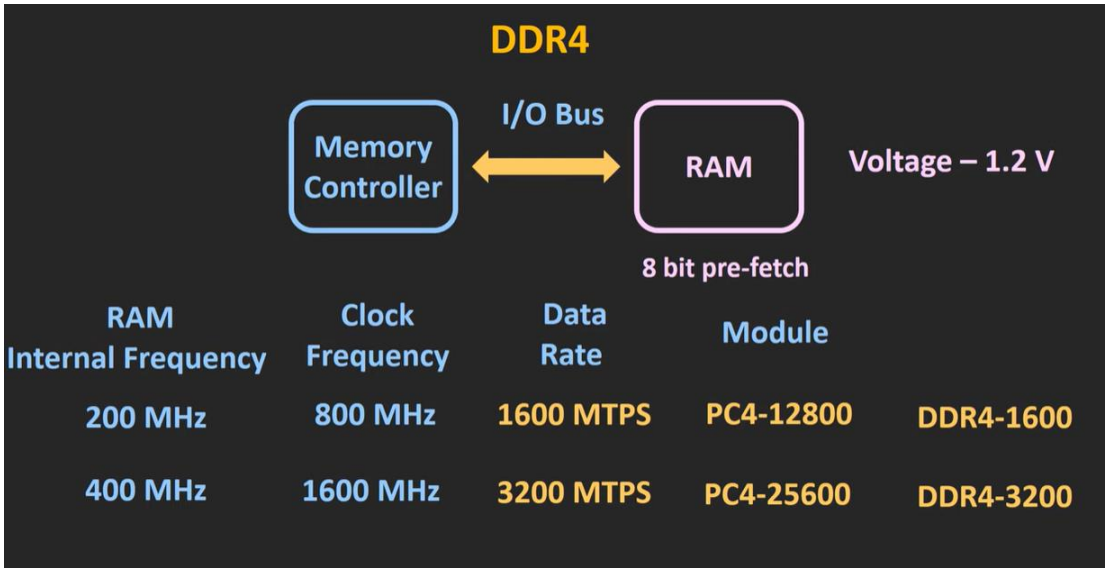
So, in case of this DDR3 RAM, suppose if the internal clock frequency is 100 MHz, then to match the data rate, the input-output bus should be get operated at the 4 times the clock frequency that is 400 MHz.

And the transfer rate that you get will be equal to 800 mega transfer per second. So, now on any DDR3 module, if you find the term DDR3-800 followed by PC3-

6400, it means that this RAM is DDR3 RAM which is operated at 1.5 V and the clock frequency of this RAM is 400 MHz.

And the maximum transfer rate which can be achieved is 800 mega transfer per second And the maximum bandwidth of this RAM is 6400 Megabytes per second.

- **DDR4**




So, after the third generation of DDR3 RAM, the next generation is DDR4 RAM. Now, here in case of this DDR4 RAM, again the operating voltage has been reduced from 1.5 V to the 1.2 V. And here again, in case of this DDR4 RAM, the number of bits that is being pre-fetched is same as the previous generation.

That is 8 bits per cycle. But now in case of this DDR4 RAM, the internal clock frequency of the RAM has been increased.So, if you are operating at 400 MHz then the clock frequency of the input-output bus should be 4 times, that means 1600 MHz.

And the transfer rate will be equal to 3200 Mega transfer per second. So, if you see the module, on the module you will find the term that is PC4 followed by 25600.

That is the speed in terms of Megabytes per second. And in terms of the DDR terminology, you will find it as DDr4 followed by the 3200. So, now so far whatever discussion that we carried out we have assumed that the input-output bus width is 64 bits.

But if we increase this bus width, let's say if we double the bus width then the theoretical data rate that can be achieved will be getting doubled.

So, whenever this RAM is used in such mode, then it is known as the dual channel mode. So, now suppose if you have two options, let's say one 8 GB of DDR4 RAM which is used as single channel mode.

And two 4 GB of DDR4 RAMs which are used as dual channel mode, then the bandwidth which can be achieved with this two 4 GB of DDR4 RAM will be better compared to the single channel 8GB of DDR4 RAM.

So, now so far we have seen the different generations of this dynamic RAMs in terms of their speed and the operating voltages. Now, let's also see the different packages in which these dynamic RAMs are available.

So, very older generation of dynamic RAM was available in the Dual Inline Package. Then after the next generation of RAMs were available in the single In-Line Modules.

So, in case of this Single In-Line module, memory chips are soldered onto the one PCB, and the pins are available on the single side of the PCB.