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

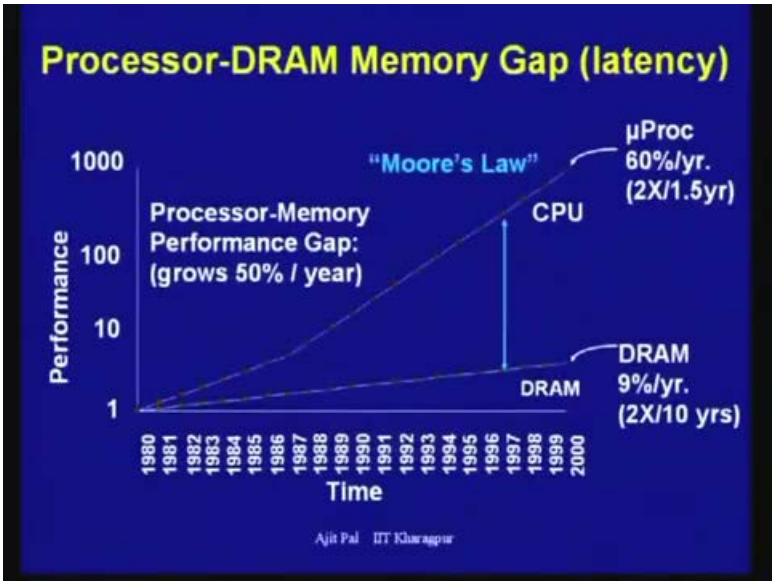
Lecture - 28
Main Memory Optimizations

Hello and welcome to today's lecture on Main Memory Optimizations, in the last lecture we have discussed about the different the organizations of memory, different types of main memory that are used, and how they are organized. Just like we have discussed various gas optimization techniques, today we shall focus on various techniques for optimizing the main memory.

Now, as we know one of the very critical bottom neck in high performance processing, particularly when you are trying to interface high performance processer to the main memory. As we know, dynamic RAM is used as the main memory, and that situation is has not change over the years, although this static RAM is size of static RAM is increasing, cost of static ram is increasing.

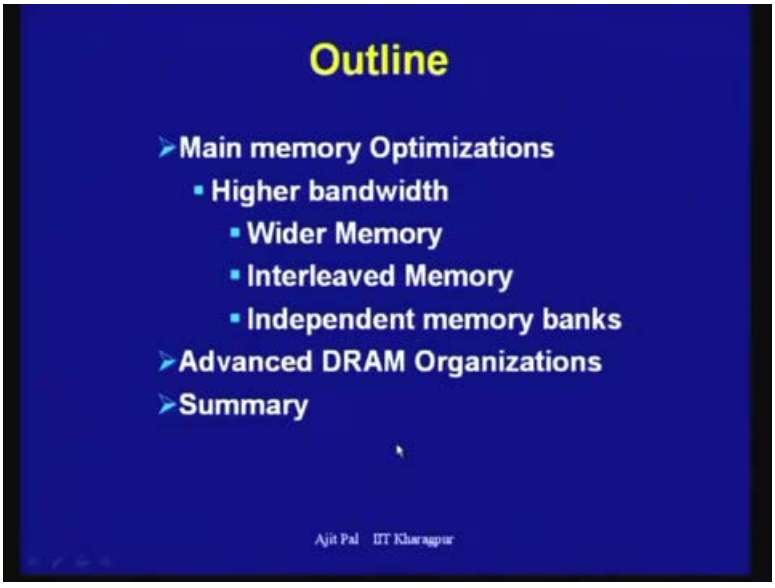
But, as main memory dynamic RAM is the choice of the day, main reason is cost, and second reason is high packaging density that is possible in dynamics RAM you can have a single package with 256 kilo bytes or more on a single chip. So, this has yet to be use of dynamic RAM as the main memory almost in virtually all systems, so whenever we say main memory optimization essentially you shall see, how you can improve the performance of dynamic RAM memory systems.

(Refer Slide Time: 02:40)



This is that performance gap I mentioned, and this gap is increasing over the years because they processers performance is improving at the rate of 50 percent per year, dynamic RAM performance is improving at the rate of 9 percent per year. So, this gap is increasing how to brief gap that is the main question.

(Refer Slide Time: 03:03)

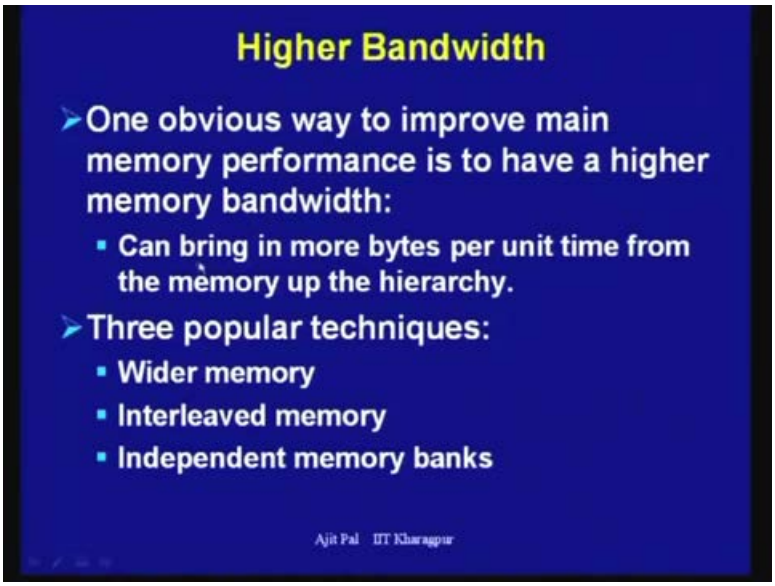


And we shall try to address it by using different techniques, number 1 technique that we used is to increase the bandwidth higher bandwidth. That means, the rate at which data transfer will take place between the processer and the main memory, how we can

increase that rate of data transfer that is known as higher bandwidth. And that can be achieved in 3 different way which we shall discuss, one is the use of wider memory, second is the use of interleaved memory, and third is the use of independent memory banks.

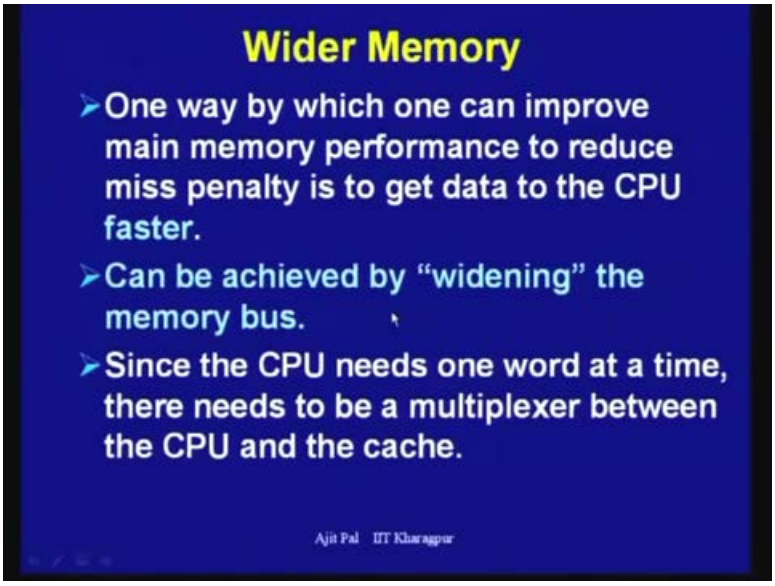
So, this are the three commonly used techniques for increasing the bandwidth of the main memory systems. And after that we shall discussed about advanced dynamic RAM originations the many innovation have taken place to improve the organizations of dynamic RAM with the objective of making it faster. And so that the perform the gap between the processer speed and rate at which the field data can be provided by dynamic RAM is reduced.

(Refer Slide Time: 04:21)



So, one obvious way to improve main memory performance is to have higher memory bandwidth, by that we mean increasing the memory bandwidth that will bring in more bytes per unit time from the memory up to the hierarchy.

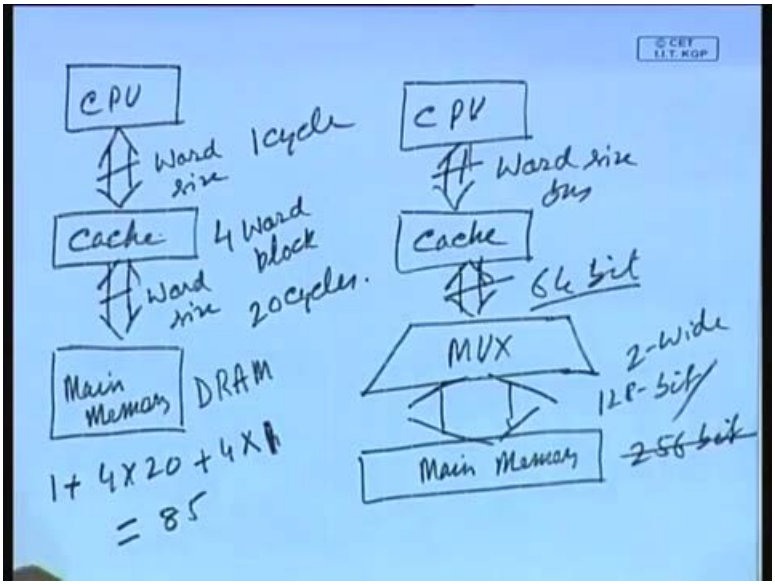
(Refer Slide Time: 04:41)



So, first we shall focused on wider memory wider memory mean, we shall increase the we shall make the memory bus wider, normally you know the width of the bus is same as the size of the words of the processer. So, if it is a 32 bit processer with of the bus is 32 bit if it is 16 bit processer with of the buses 64 bit, now we would like to make a wider bus say may be that will transfer more number of words, and that can be done to improve the performance.

Since the CPU needs one word at a time, they are needs to be a multiplexer between the CPU and the cache. However, as we try to do, so I mean increase the width of the main memory bus, you will see the that you will requires some multiplexer because if the processer in many case we will need in terms of words, it will fetch a word by word because the cash memory, main memory bus is only in terms of words.

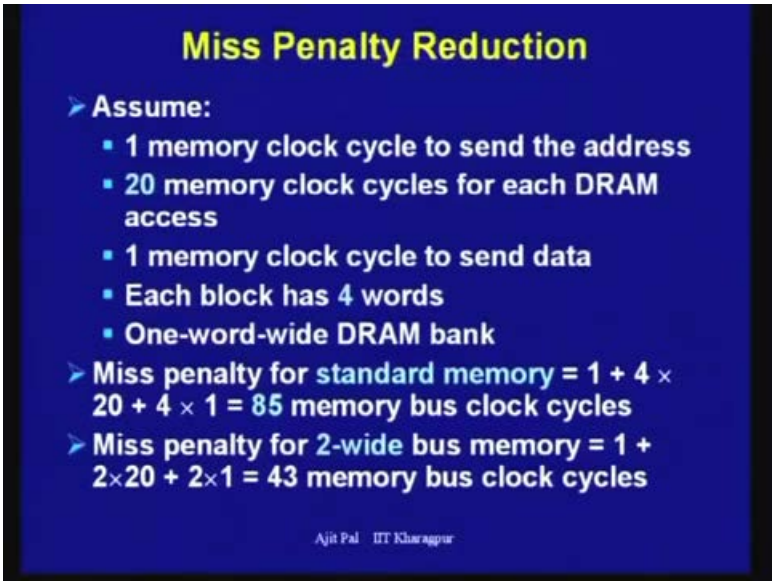
(Refer Slide Time: 06:04)



So, we can explain it this way normal situation we have got your CPU here, there is a bus, and this bus is connecting to the cache memory. And bus with this same as the ward size, and here also you have got another bus that is connecting to the main memory, and having the same width of the bus. So, this is the ward size bus, this is also ward size bus, now what we are trying to do is to have a wider bus between cache and main memory.

So, CPU will be there and having the ward size bus, and here you will have cache memory, then you will require multiplexer between this bus which will transferring terms of wards and the wider bus that will coming from the memory. So, here you will be having a multiplexer, multiplexer can be 2 to 1 or 4 to 1 are depending only width of the main memory. So, main memory is now is having wider bus, and this bus will transfer at a much I mean more number of bits will be transfer. So, if this is a 64 bit, this can be 128 bit or 256 bit depending on the width that there is a choice we can use.

(Refer Slide Time: 08:02)



So, let us consider how the this penalty is reduced because of the use of the wider bus, so for that we assume that one memory clock cycle is needed to send the address. That means, the CPU has to send the address, and that address will be to send at that address associated with other I mean along with other control signal that will require 1 cycle.

Then 20 memory cycles for each dram access; that means, is whenever the data transfer is taking place between the this is the main memory dynamic RAM, after receiving the address it will take 20 cycles not 1 cycle, but 20 cycles. And 1 memory clock cycle to send the data; that means, part 1 memory cycle data can be send from the after reading it from the main memory to the processer, so that is what is being soon.

And we have assume that we are using a block size of 4 words, note really wards; that means, the cache memory is having 4 ward block. That means, you have to transfer even if this it is ward size you have to transfer 4 wards before you can provide control to the CPU, even in the situations. So, in a such a situation whenever you got 1 word wide and DRAM bank that is transferring between the cash memory and main memory. Let us see what is the miss penalty for standard memory first 1 clock cycle that is needed to provided the address.

Then 4 clock cycles in 20 that is required to transfer 4 words of data from the main memory to the cache. So, 4 ward that will be require; that means, 1 plus 4 into 20 for transferring from the main memory to the cache, and after that you know you will

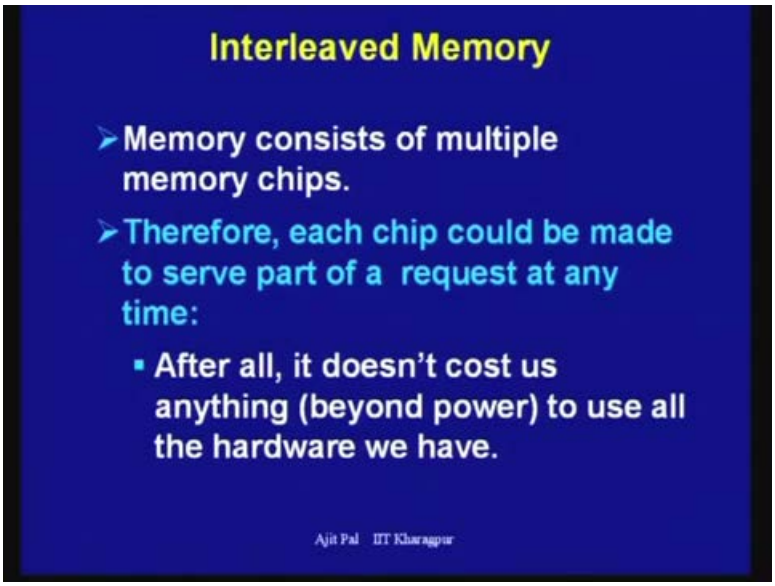
require 4 more cycles to transfer from the 2 to right it into that cache memory. So, this is the total time you will require 85 cycles memory of cycle of 1 plus 4 into 20 plus 4 into 1, so 85 memory is a bus clock cycles and needed in the standard situation.

Now, let us assume that we have made the main memory bus 2 wide, 2 wide means this is say if it 64 bit, this is 120 bit not to 256 this is 2 wide. That means, we can transfer at a time 2 wards simultaneously, and by doing that you can see the how the reduction is miss penalty occurs. So, this case you will require 1 plus 2 into 20 because you are sending 2 wards at a time, so you will require 2 cycles, so 2 into 20 and again 2 cycles upon clock cycle that is for transferring the data to the cache.

So, this is the miss penalty that is required 1 plus 2 into 20 plus 2 into 1 that is 43 memory bus clock cycle. So, by having wider memory we can see it is roughly I mean there is a speed of about I mean close to 50 percent 85 by 43 that is very close to I mean may be 100 percent increasing performance, you can see it transfer rate becomes double almost double. So, this is the miss penalty reduction by using a 2 wide bus, if you make it 4 wide then this will be 1 into 20, and this will be again 1 into 2, so it is still improve made a possible.

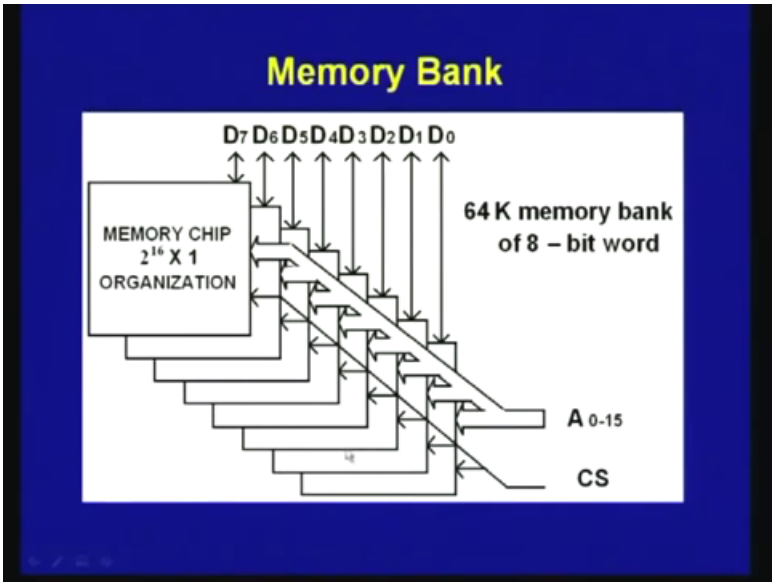
But, you know more wide is the bus they it is very costly to implement that is why the width of the main memory bus is not made very wide may be 2 wider at most 4 wide.

(Refer Slide Time: 12:30)



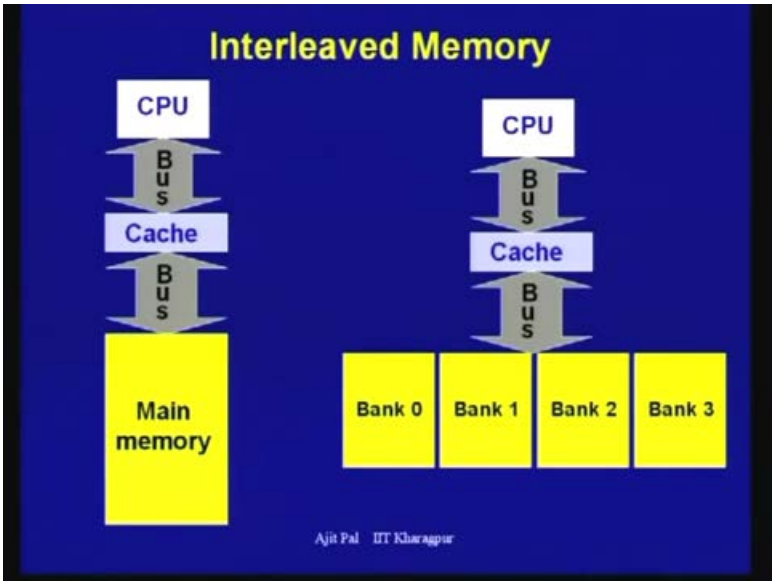
Now, let us consider another situation, where we can use interleaved memory, what do you really mean by interleaved memory. Here as we know main memory consist of multiple memory chips therefore, each chip could made to serve part of request at any time. So, what we are trying to expose wide here, although we are considering a single memory systems, we know that memory system comprises a large number of memory chips. So, each chip may be provide only 1 bit or 8 bit at most, but not more than that, so why not exploit that and; that means, several memory chips are there, and that parallel assume can be exploited.

(Refer Slide Time: 13:21)



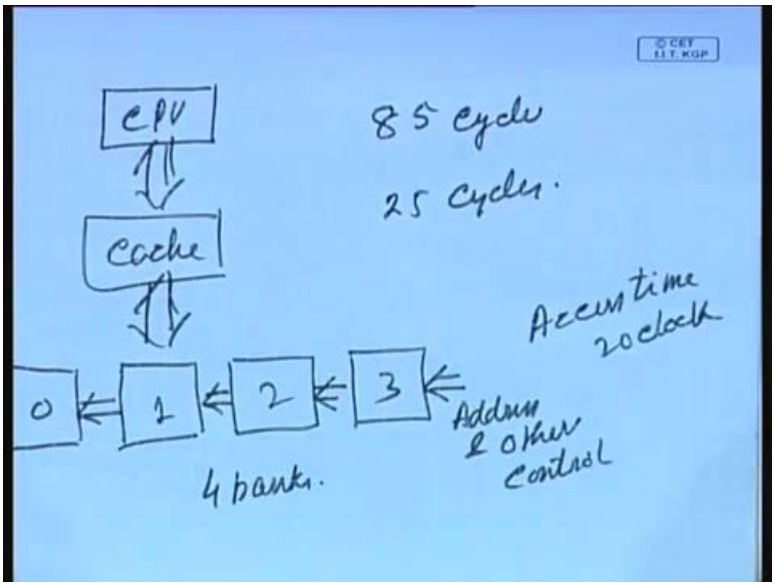
And how we shall see and for example, this is that typical memory bank here it has been assume that your memory chip is organized as 2 to the power 16 into 1. So, this is the how it is organized, so it is the memory chip is bit organized, so what you are doing you are transferring the address simultaneously to all the memory chips, a chips signal to all the memory chips. So; that means, all of them will get ready at the same time to provide the data on the data bus, so this is a 64 K memory bank of 8 bit ward.

(Refer Slide Time: 14:02)



Now, this can be extended this idea can be extended to realize multiple memory bank interleaved main memory system.

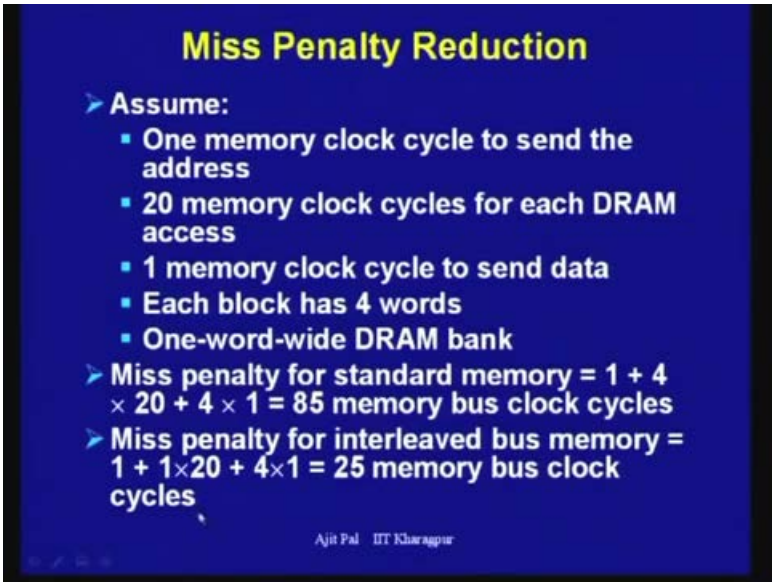
(Refer Slide Time: 14:16)



So, here what has be done the instead of having a single bank we are having the CPU that is interacting with the cache memory. Now, between cache memory and main memory we have got several banks, so this is bank 0, this is bank 1, bank 2, bank 3, so let me assume we have got 4 banks, what is being done here the cache memory will provide will apply the address, and other control signals simultaneously to all of them.

So, since it will be supply simultaneously to all of them, they will require 20 clock cycles to get ready I mean to provide data that is the access time, as you have seen access time is 20 clock cycle. So, after 20 clock cycle all the memory banks will be ready, now after that you can transfer one after the other, you did from memory bank 1, then you read for memory bank 2 then you read for memory bank I mean memory bank 0, then memory bank 1, then memory bank 2 and memory bank 3. So, this one after they are ready together he will transfer one after the other that is what is being done in a interleaved memory.

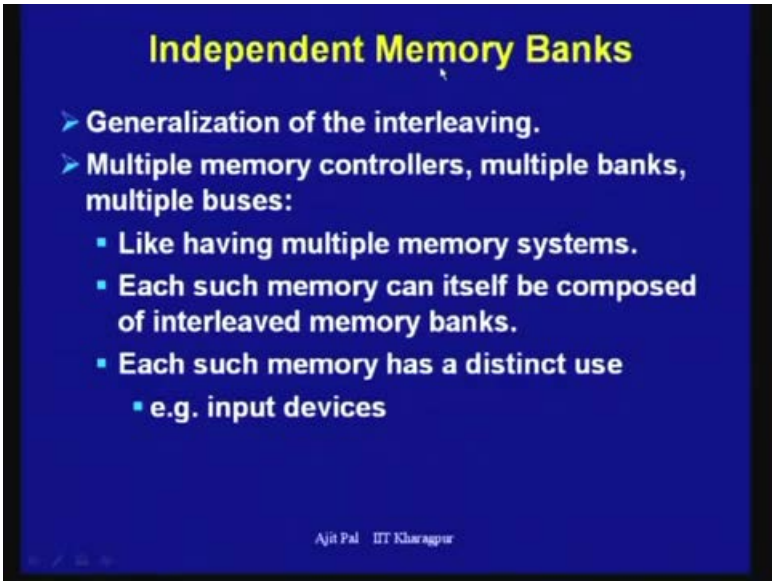
(Refer Slide Time: 15:50)



So, how the miss penalty is reduced, so in this case you can see you require 1 clock cycle to supply the address, to send the address then 20 clock cycle are required for the memory banks to provide data. Then of course, you will read one after the other, so we require 4 clock cycles and total of 25 memory bus clock cycles then needed to transfer the 4 wards to the main memory to the case.

So, you see the reduction in this case is 25 memory bus clock cycles instead of 85, so instead of 85 clock cycles in the normal situations you will require only 25 memory cycles. So, there is a significant improvement in performance you can say speed up is roughly 4 times, so you can use this interleaved memory.

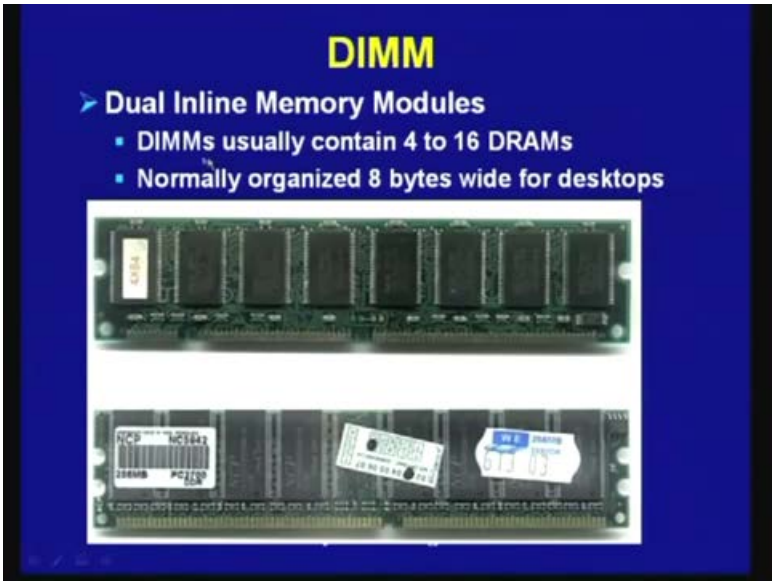
(Refer Slide Time: 16:56)



So, this idea can be extended two have independent memory banks, so this is a generalization of the concept of interleaving. So, multiple memory controllers can be used, multiple banks can be used, multiple buses can be used, so like having multiple memory systems. As if you have ward multiple memory systems, and each of them will work independently and in simultaneously in parallel to provide the respond to the CPU.

And each memory system can itself be composer of interleaved memory rank, so what we are considering here we are considering independent memory banks each of them can be interleaved in side, and each such memory has a distinct use. For example, in case of input out devices, where you will require independent memory banks from where a it will be red simultaneously they will be a responding to the request from the processer, so this is the concept of independent memory banks.

(Refer Slide Time: 18:08)

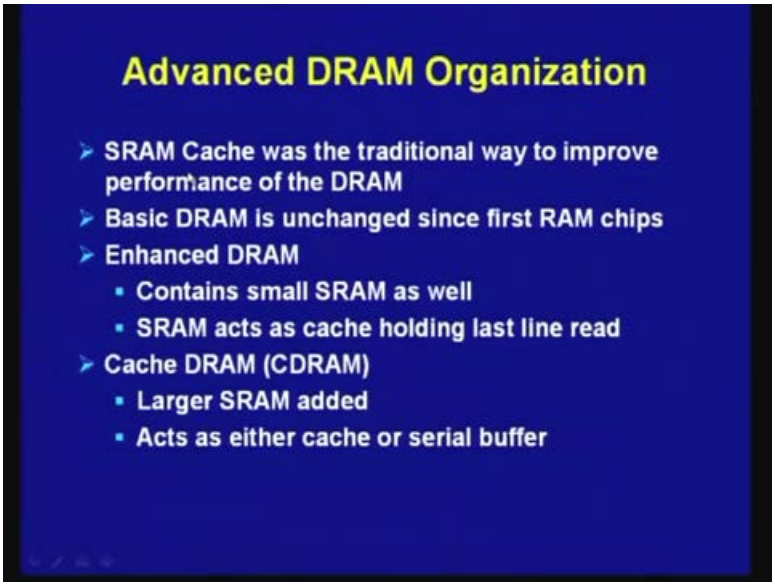


Now, I mean the dynamic RAM's the way they are organized this shown here, there are organize in the normal of dual in line memory modules. So, those both are known as DIMM's forms usually contain 4 to 16 DRAM's, so you can see here you have a got a 4 into 1 DRAM chip that provides you 4 into 64 kilo bit. So, in this way you can have 1, 2, 3, 4, 5, 6, 7, 8, 8 into 4 32 bit byte bus you can have, and this are used for realizing now a days in your desktop, and servers and workstation as the main memories system. So, each RAM will provide 4 bit then you can have 4 to 6 DRAMS on a single printed circuit board providing the, I mean 8 bytes wide pass for the desktop, and then you can have several search DIMM's in your system.

(Refer Slide Time: 19:24)

Now, we have seen how you can enhance the data transfer rate, now we shall consider advance DRAM organizations dynamic RAM has gone through a number of innovations to provide less and less access time or in other words to provide and data active faster rate.

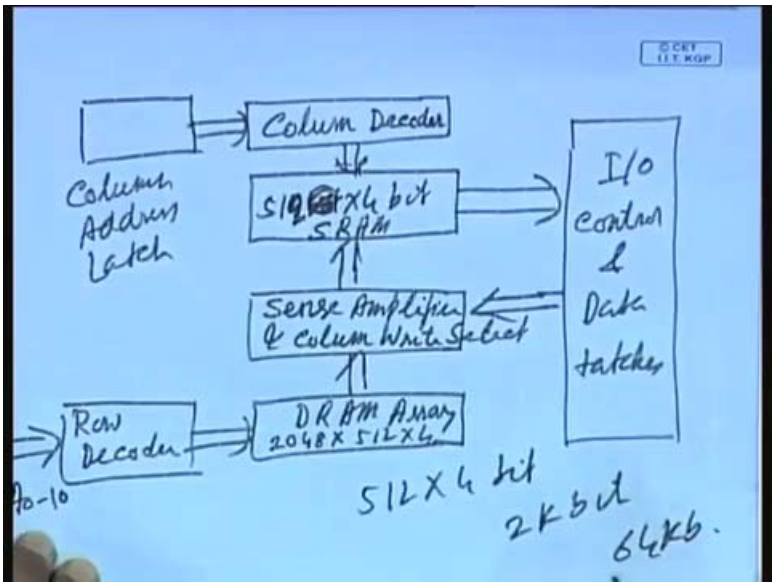
(Refer Slide Time: 20:00)



And we shall consider them one of the other, the first technique that is many use is to have SRAM cache use of a static RAM cache memory built in a spat of the dynamic ram. So, SRAM cache was the traditional way to improve the performance of the

DRAM, so basic DRAM is a dynamic RAM is a unchanged since the first a RAM chip was hence.

(Refer Slide Time: 20:41)



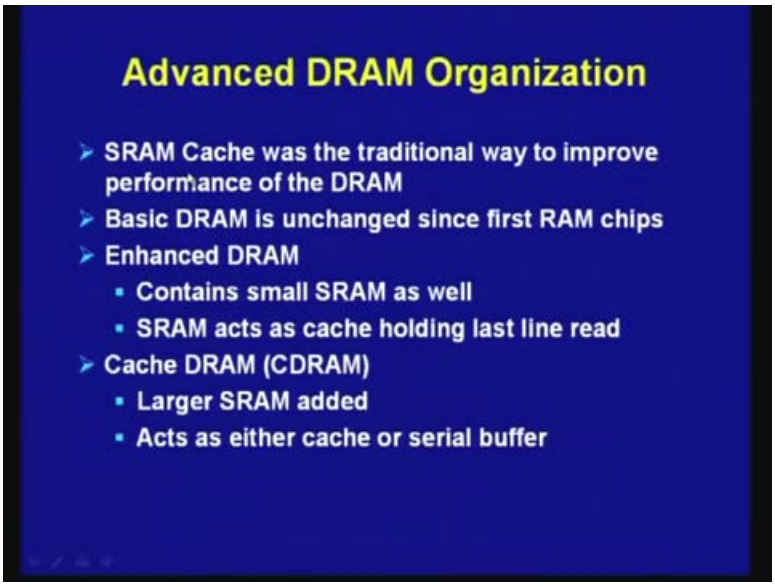
So, how it is being done let me explain by drawing a diagram and how the static RAM is in cooperated as part of the dynamic RAM. So, as you know you have got column decoder I am not drawing the entire DRAM memory chip, but part of it highlighting the incorporation of static RAM this is your column decoder. Then you have got your this is where you put your 514 bit into 4 bit SRAM, this is the static RAM 500 into 4 bit that is actually available on a single row.

And then you have got the other things like the typical function is that is require in static RAM, and dynamic RAM's and sense amplifier then and column right select signal and then here you have got your DRAM array. So, DRAM array is 2048 into 512 into 4, so this will be 512, so this is the how the DRAM is organized in 2 dimensional array from where it goes to decent amplifier and column write select.

And form there a particular single row you can see you have got 2048 rows 2 K rows, and 1 row is transfer and that is available in the form of case in the static RAM. So, this is from here it goes to the external world, so here is your input output control, and data tautology. So, this is for external control and of course, you will require the row decoder here, so decant the address is a 0 2 a 10 level with address come here, similarly here you will have the column address latch.

So, main idea here is what is happening instead of reading it from the dynamics RAM, you are latching one row here. And then by changing the column address you are reading it I mean you are transferring it to the cache memory, so that will make it faster because you know because once you have transferred to the static RAM the access time will be much lower compare to reading it from the dynamic RAM that is a basic idea.

(Refer Slide Time: 24:22)

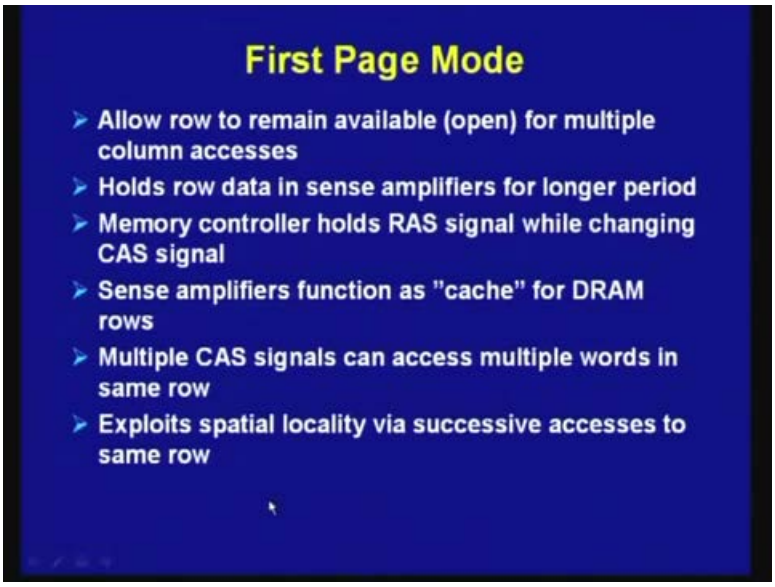


And, so you have got small RAM and SRAM access cache holding the last line, so because of the principle of you know locality that the you will be reading sequentially from sequential address location do which are available in your single row. So, from that single row you will read one of the other, now extension of the idea is to have cache DRAM. So, instead of small DRAM you will be having a large static RAM, say here you have added only 512 into 4 bit. So; that means, 2 kilo bit instead of that you can have 64 kilo bit.

So, larger cache can be used and which you really act you know cache; that means, first you will transfer from the dynamic RAM to the large static RAM. And that large static RAM you will read it one after the other as long as it is available there, so the way the cache memory works you transfer from the main memory to the cache memory, and from cache memory it is transferred to the processer in a same way to work only different here is that from dynamic RAM it is going to the static RAM, and from static RAM it will be read by the external world.

So, this is the cache DRAM concept that is used in many situations, but this enhance RAM this is very popular.

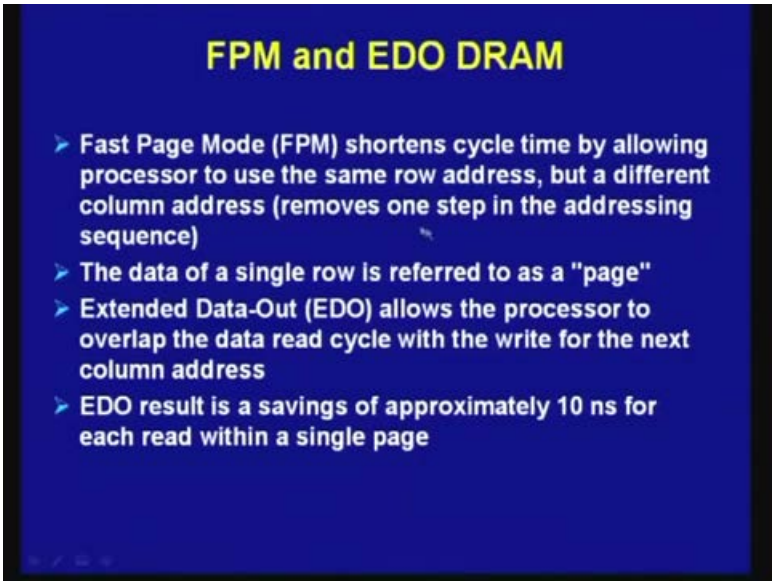
(Refer Slide Time: 26:00)



Now, that an hence DRAM it operates based on the idea of first page mode, so allow row to remain available for multiple column at access as a I mentioned and holds row data in sense amplifiers for longer period. So, sense amplifier in this case is acting as a kind of cache, so we are not using this static RAM, we are not using it, but what you are doing we are transferring one row of data. And this sense amplifier at the output of the sense amplifier one row is available, and which will act as kind of latch.

So, in this case without using static RAM, you are able to read it, but in this case you have to restrict to a reading from the at the output of the sense amplifier, and by change the by holding the row address, select signal while changing the column address selection signal. So, the sense amplifier function is cache for DRAM rows, so multiple column address selection signal can access multiple words in the same row. Again this exploit specials locality via successive access to the same row, so basic idea is same, but in this case you are not using static RAM.

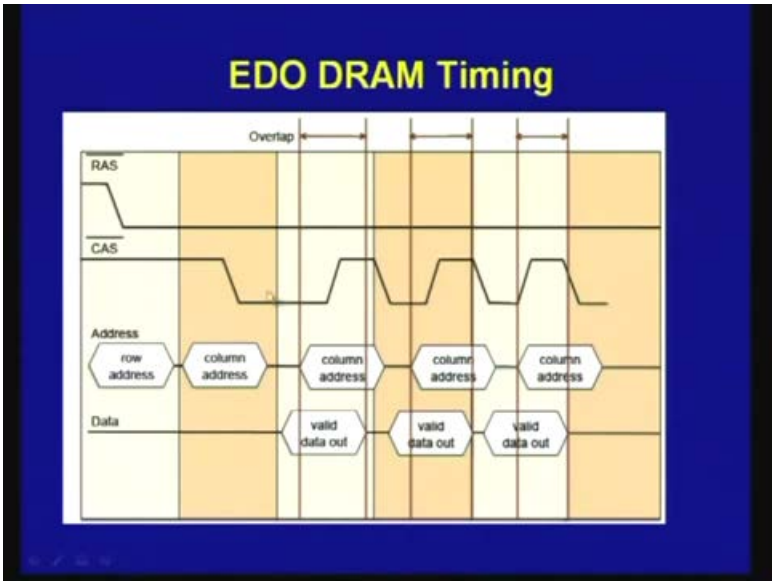
(Refer Slide Time: 27:29)



Then another concept is first page mode shorten cycle time by allowing processer to use the same row address, but a different column address as you have discuss. So, this removes one steps addressing sequence, one step of addressing means you have seen normally avoid this say provided, then column address is provided only at the both address are available, data is added from the dynamic RAM. But, in this case only row address is provided then by changing only the column address, you will be reading one of the other.

So, that is the difference and the data of a single row is refer to as page, and extended data out allows processers to overlap data read cycle with the write for the next column address. So, in this case we are overlapping operations data read with the write for the next column address, so EDO result is a savings of approximately 10 nanoseconds for each read within a single page, single page means in this case you are referring to one row.

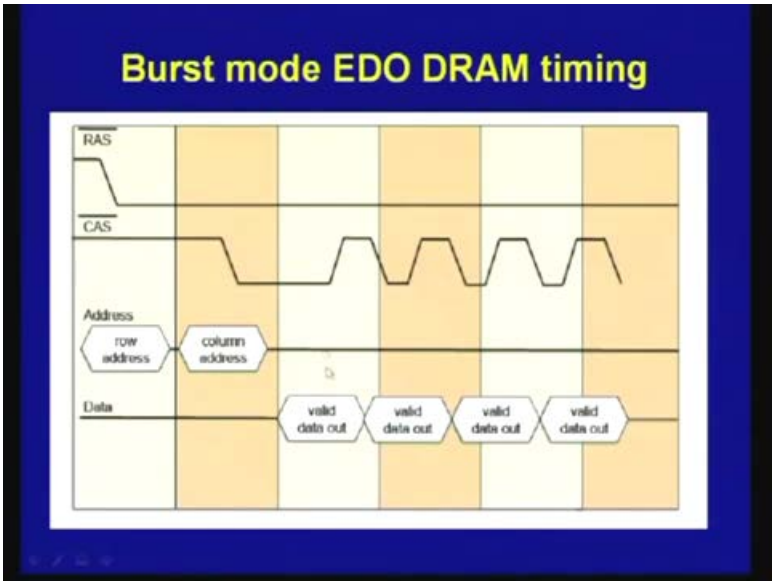
(Refer Slide Time: 28:39)



And if we look at the there is timing diagram this will be clear, as you can see this is the row at the select signal, and that has been I mean kept tables. Now, you are changing the column address selection signal, and you can see this is the row address, and the row address actually it is row address scope not this select. So, row address scope signal will latch that this address in the row address buffer, and the column address stop signal will latch the column address.

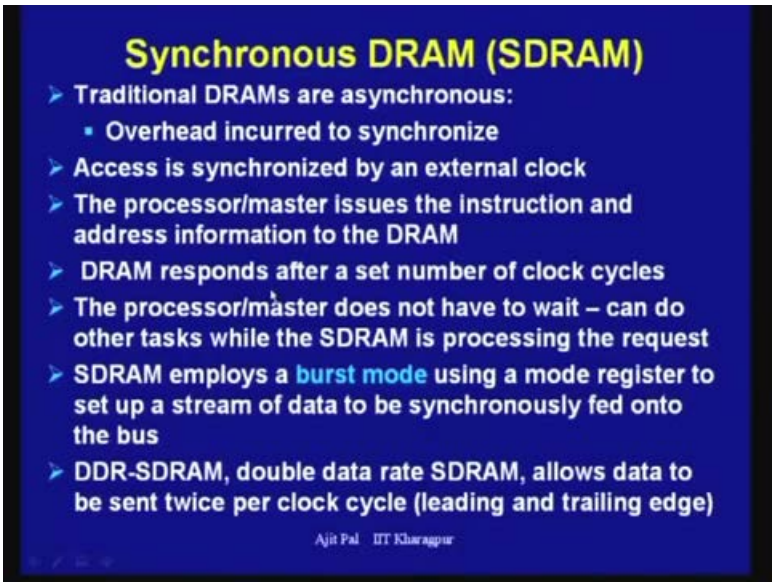
Then the data will be available after sometime, so this data would you are reading this data, and this data would and at the same time you can have another address that will column address that can be a read by using activating another columns address scope signal. So, this data reading and column address generation is over getting overlapped, and which is a essentially read to the data to be read in the next clock cycle. So, you can see this is how by using EDO that extended data would this allows overlapping of data read cycle with write to the next column address. So, write to the next column address and the data out both are taking place simultaneously.

(Refer Slide Time: 30:14)



Now, you can also use bus mode in the EDO RAM timing, so in this case basic idea is you have generated a row address. And that row address strop signal has latch it that remains in the buffered then you will keep on changing the column address one of the other. So, here the column address you have applied and again you will subsequent a column address selection signal generated. And a bus of data is coming out one of the other in each cycle from the same row, and that data is getting transferred, this is the basic idea of bus mode of EDO dynamic RAM timing. So, this will also make dynamic RAM faster.

(Refer Slide Time: 31:05)



Now, let us consider Synchronous Dynamic RAM, SDRAM which is now a day's used in almost all work station, desktops and everywhere. As you know traditional dynamic RAM's are synchronous, what do you really mean by a synchronous, by synchronous we mean what happens in case of whenever we access traditional dynamic RAM, the address will be the address along with different controls signals will be made available the dynamic RAM chip.

That means, row address strophe, column address strophe, and various things and after that the RAM chip within the RAM chip various operation will take place. There are you have seen there are very large capacity that various large capacitances present inside the dynamic RAM chip that those bit lines, and various other I mean row lines. And so those capacitance get charge, then they will be essential amplifier that sense amplifier will sense the data, and that will give you transferred through the input output to the outside world.

Now, this operations will take some time what this CPU does the doing this period, so there is some multiple width cycles, I mean you will require multiple wide cycles before you can did data. For example, 20 clock cycles as I have told is the access time for traditional dynamic RAM; that means, the CPU will wait has to weight 20 clock cycles to read one word of data from the dynamic RAM. So; that means, if you are reading sides successively one after the other you have to spend 20 clock cycles, then another 20 clock cycles and so on.

Because, it is taking place into a synchronous manner, now in synchronous DRAM this is overcome by providing an external clock. So, access is synchronized with the help of this external clock, so the processer or you may call it master issues the instructions, and address information to the dynamic RAM. Then dynamic RAM response after a set number of clock cycles, so here it is synchronized; that means, the processer will generate address, and clock cycles at which is synchronized by clock.

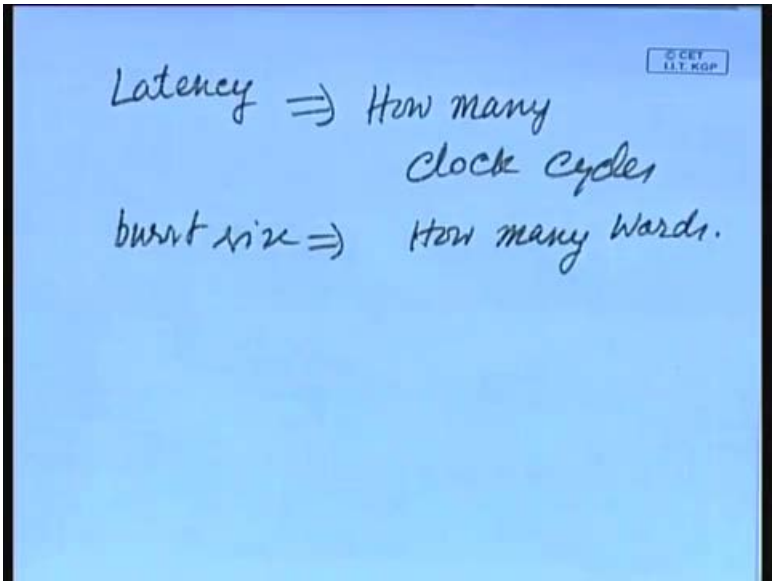
Then there is a kind of synchronized that external clock will do the synchronization, after fix number of clock cycle the RAM will be ready. So, the processer does not in this case what happens, the processor of the master does not have to wait, so can do other task while the SDRAM is processing the request. So, reference between the previous case and

in this case is earlier, the processer was continuously checking the status of the dynamic RAM, whether it is ready or not and weight cycles where been generated.

But, in this case the processer is no longer waiting for the data it is simply after providing the address and control signals, it gets busy in other things. So, it can do it can perform other operation or tasks, in the main time the static this SDRAM will that dynamic RAM will get ready. And SDRAM employs bus mode using a mode register to set up stream of data to be synchronous fed on to the bus, then after that set numbers of times the SDRAM synchronous DRAM will provide data in a bus mode.

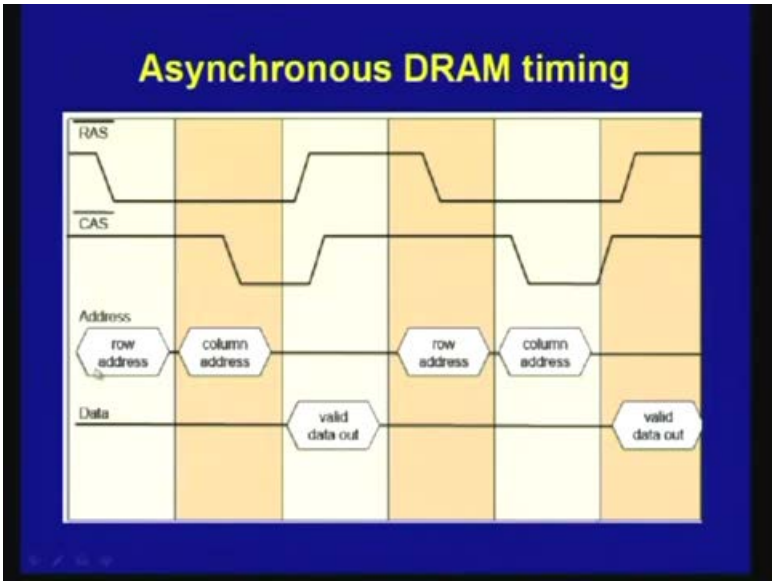
Bus mode means, in each clock cycle one word, one word, one word in this way the data will come, and this is again synchronized by the by the clock. Now, here we are making use of a register known as mode register, what is the role of this mode register, mode register is set up stream of data to be synchronously fed mode register will contain two information, well you can uses can actually set the mode register which will decide what is the latency.

(Refer Slide Time: 35:56)



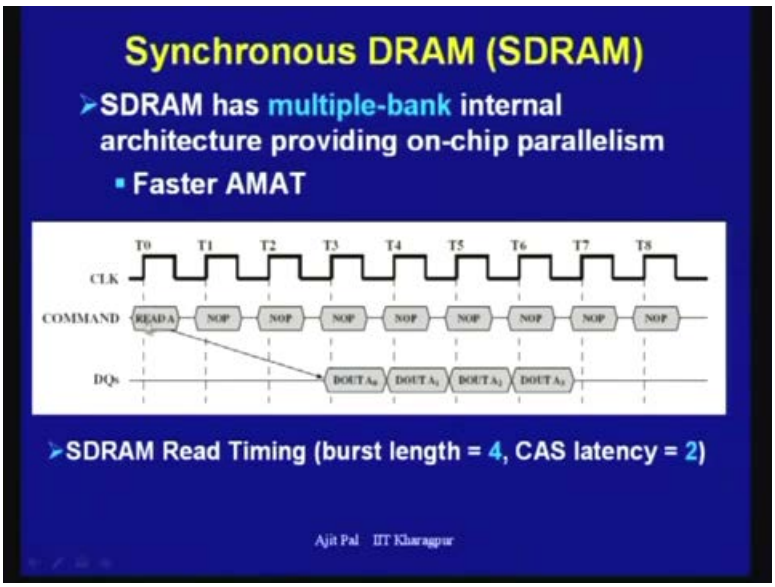
That means, how many clock cycles clock cycles the dynamic RAM will require to get ready to provide data. And another thing is the bus size, how many words to be transferred in a single words, so this two can be said with help of this mode register and that the stream of data can synchronously fed on to the bus.

(Refer Slide Time: 36:43)



As it shown in this diagram, it can see here the processer is providing two information, row address and column address with the help of this signals. And after that this is the asynchronous dram timing which had already explain.

(Refer Slide Time: 37:02)



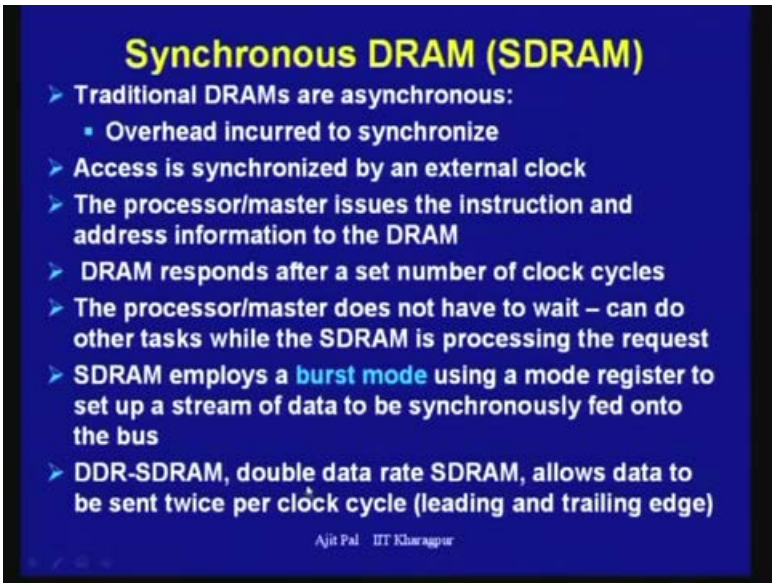
Now, let us consider the synchronous DRAM timing, so the synchronous SDRAM is we will sent this comment read A, this is the address generated. And after that it is weight it is keeping few clock cycles, this is know up; that means, the it is generating 3 clock cycle in this case though let in see is 2. So, it will generate these know half cycle

continue to generate know up after providing this address, and then data will be available after a fix latency.

So, this is the column address latency after the this columns address of signal is obtained, after 2 clock cycles the data will available. So, this is the latency after this latency after 2 clock cycle, the data is now available, and now in this case bus length has been said to 4, so you will get 4 words of data from a 0, a 1, a 2 and a 3; obviously, from conjunctive addresses. So, this is how you will transfer data one after the other. So, instead of this asynchronous the RAM timing I mean way the you are reading 1 word, then another word and for each of them you have to spent 20 cycle.

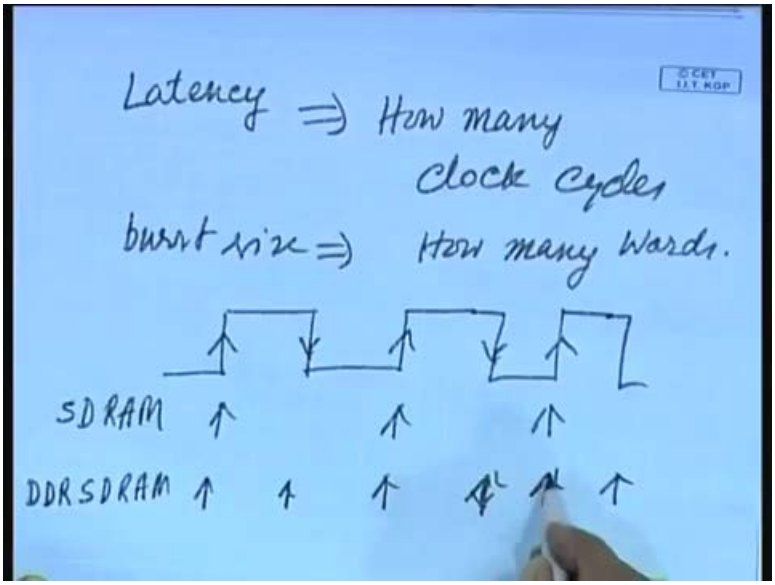
Here you can see initially there is a latency, but latency decided I mean is depended in to from the cache memory access time sorry dynamic RAM memory access time that you are using and accordingly you can modify the mode register. So, the SDRAM has multiple bank internal architecture providing on chip parallelism, so you asking how you are able to provide this one of the other that is possible because of use of multiple banks there is provided inside the chip. So that means, here you have got 4 banks and for each bank you are getting 1 data I mean 1 data in 4 conjugative cycles. So, that is how it has been made faster and; obviously, this will lead to faster average memory access time.

(Refer Slide Time: 39:25)



Now, there is another innovation that was provided which is known as DDR SDRAM DDR stands for Double Data Rate. So, double data rate is the RAM, double data rate synchronous dynamic RAM in this case it allows data to be sent twice per clock cycle.

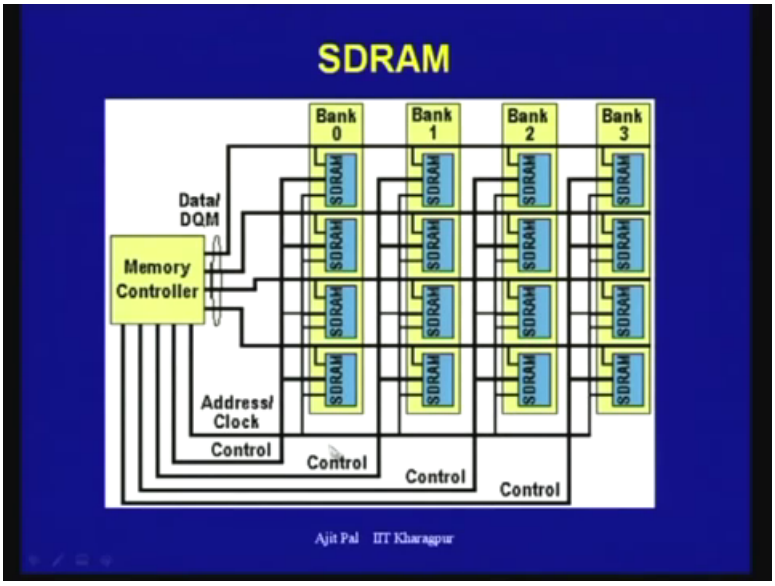
(Refer Slide Time: 39:47)



So, normally you know you have got a clock, usually data is transfer either by using this positive edge or by using this negative edge. One of the edge is use; that means, on this edge on one data is transferred then this edge another data is transfer that is the normal DDR that happens, you can normalized that happen after that initial delay. So, instead of that in case of DDR double data SDRAM this allows data to be sent twice per clock cycle.

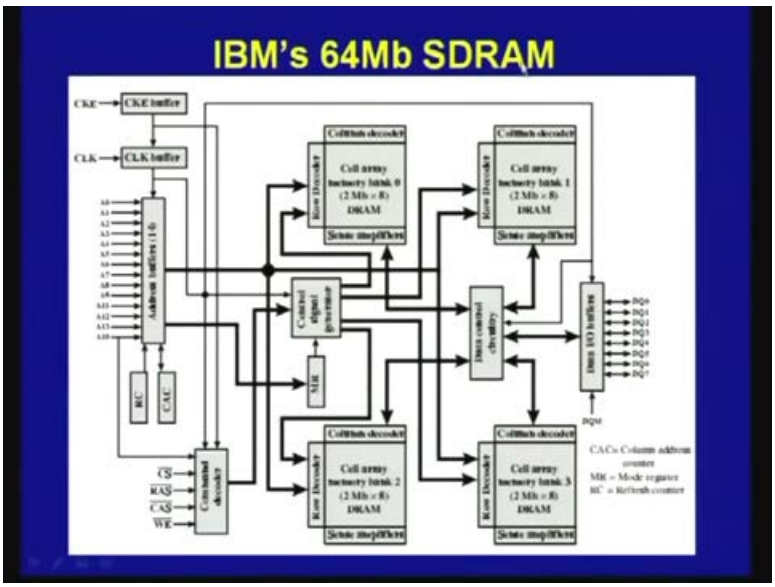
So, you are using leading edge and trailing edge, both the edges are used to transfer one order of data; that means, you are transferring data here this is normal SDRAM, now in your DDR SDRAM will be able to transfer at this edge, as well as at this edge at this edge, as well as at this edge and also in this edge and so on. So; that other means you have data rate is becoming double because you are transferring on both edges, so this is another innovation that was incorporated in a dynamic RAM by data has help to provide data at a faster rate.

(Refer Slide Time: 41:16)



So, this is the typical SDRAM organization I have got a memory controller, and as I mention you have got multiple banks. So, here 4 banks has shown, and 4 banks are getting the I mean different control signal coming out from the controller.

(Refer Slide Time: 41:39)

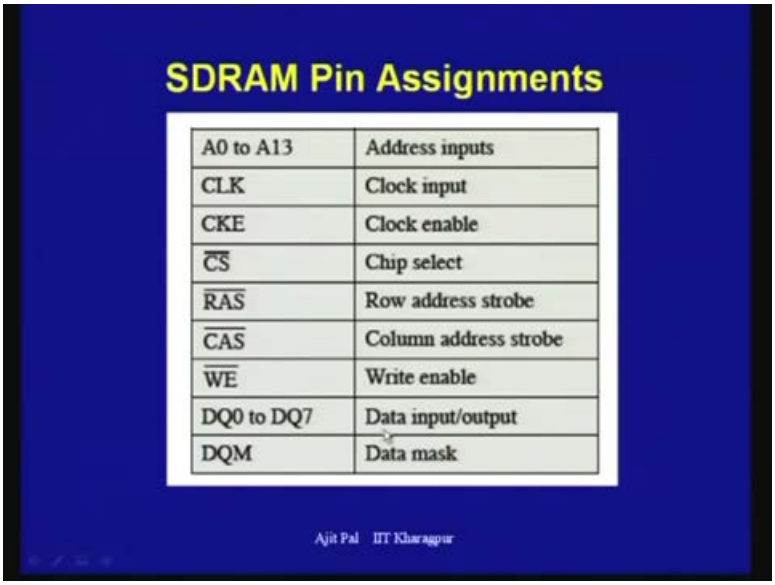


And actually this is very simplified diagram a more realistic diagram or more for a real life process memory that is you developed by IBM, IBM 64 M b s DRAM is shown here. Here, you have got 14 bit address offer then you have got different 4 banks of cell array as you can see this is the cell array 2 M b into 8 here also 2 M b into 8, 2 M b into 8, 2 M

b into 8. So, in this way you have got 2 plus 2 8 M b into 8 that gives you 64 M b mega bit of dynamics RAM stored.

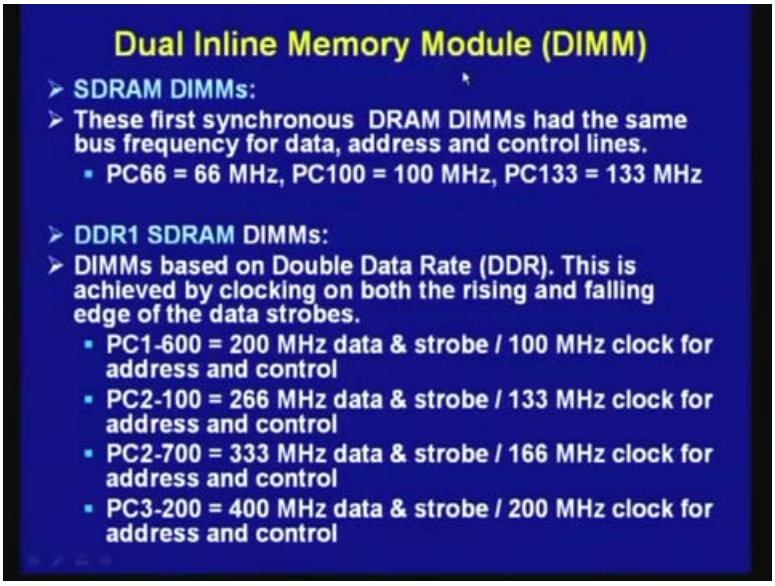
And various controls signals generator, data memory control, circulatory and here is you data higher strophe. So, here you are reading data 8 beta time; that means, here you have got 8 bit bus for external transfer, so using this IBM 64 M b SDRAM you will be able to get 8 bit at a time.

(Refer Slide Time: 42:55)



And you can see various pins assignment of this SDRAM, you have got A 0 to A 13 address inputs then you have got clock input, clock enable signal, chip select signal, read address strobe, column address strobe, write enable and data input output DQ0 to DQ7, DQM data mass. So, these are the various signals which are shown here for external interfacing.

(Refer Slide Time: 43:32)

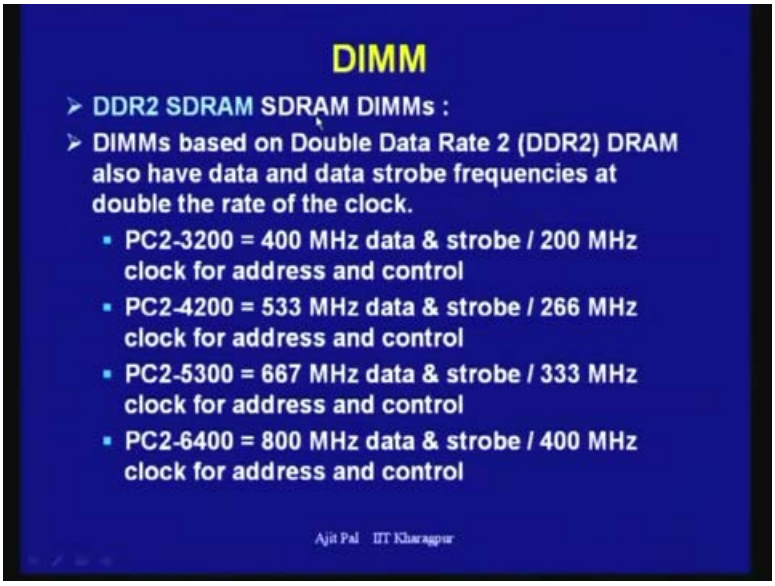


So, this SDRAM is very popular then you can make that dual line memory module can be made by using this chips. So, you can see the first synchronous and DRAM DIMM's has the same bus frequency for data and address and control; that means, here you are using I mean not using the double data rate. So, here you are using single data rate on a single edge, so PC 66 gives you data at the rate of 66 mega hertz, PC 100 at the rate of 100 megahertz.

That means, you will be getting 100 into 8 bit at a time, and if you form that and here PC 133 that will give you 133 megahertz. So, this DIMM,s can be used in parallel to realize the memory system, and here DDR1 is DRAM DIMM's that use as double data rate and this is achieved by clock on both rising and following edge of the data strobes which I have already explain. And again you have got different types, different modulus PC 1600 that gives you at the rate of 200 megahertz data and store.

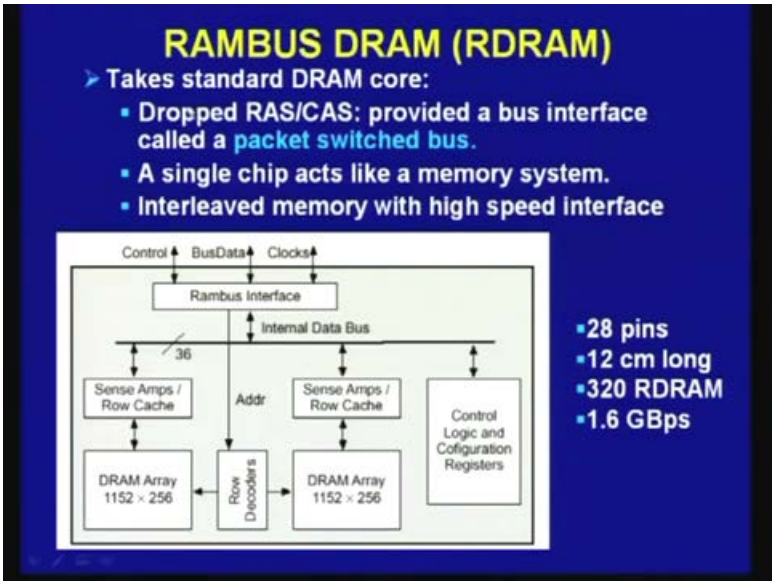
But, 100 megahertz clock and address, so data rate is double rate of clock and address and control. Similarly, PC 200 gives you at the rate 266 megahertz clock data and store, and 133 megahertz clock for address and control, and that PC 2 700 gives you 333 mega hertz data and strobe and 166 megahertz clock and clock for address and control, and PC 3 200's gives you 400 megahertz data and strobe, and 200 mega strobe for address and control.

(Refer Slide Time: 45:31)



And an addition to that you have got DDR2 SDRAM, so you are going from 1 generation into another generation, technology is improving, speed is improving. And as you can see here this being are also double data rate, and it is starting with 400 megahertz that PC 2 series double 400 mega has double data and strobe, but 200 mega clock for address and control. So, in this way close up to PC 2 6400 where you can get at the rate of 800 megahertz data and strobe by and 400 mega clocks for address and control. So, you can see you have got a variety of DIMM's available nowadays commercially available, this are commercial available DIMM's which you can procure and depending on the speed requirement of your computer system, you can use them in your computer.

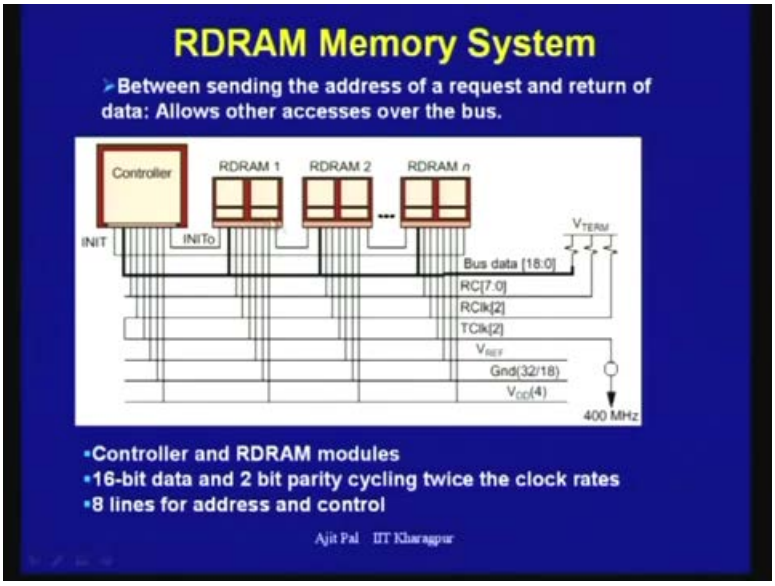
(Refer Slide Time: 46:30)



Another innovation that was in the dynamic RAM it was the huge of RAM bus DRAM, so developed by RAM bus. So, it takes standard DRAM as the core, so inside you have got this standard as the SDRAM as the core, so you have got typical DRAM array, you have got the row decoder, column decoder sense amplifier and various other things that is the required in within a standard dynamic RAM, but you have got a kind of a bus.

So, which is known as packet switched bus, so it provided by bus interface called the packet switch bus, so a single chip acts like a memory system it cannot should be not consider like a memory chip, it is a memory system; that means, lot of control and another functionalizes have been built, built in as a part of the chip. So, the through the bus 28 pin bus it will interact with the processor and it has got 28 pin and various a pinch which are provided shown here.

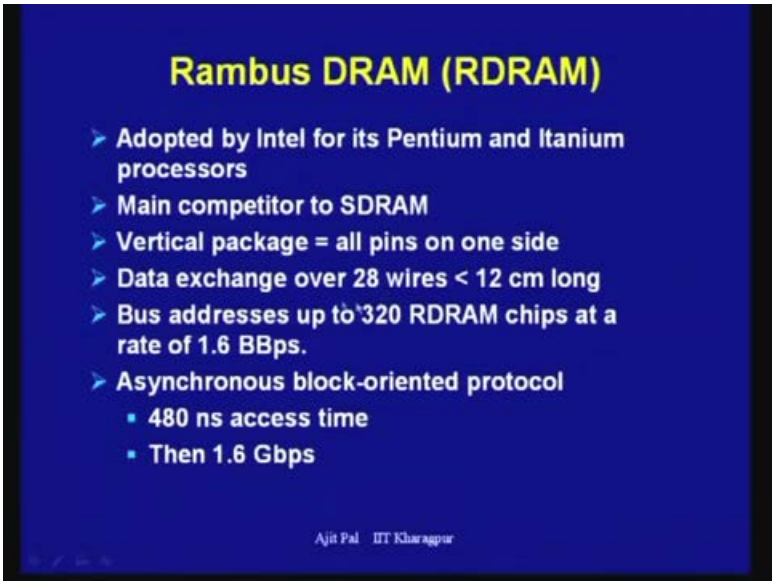
(Refer Slide Time: 47:45)



The data bus is 18 bit that R C 8 bit, R C clock 2 bit, the clock 2 bit and so on, so you can see you can have up to 300 RDRAM's that you can have in a single systems. And you can have a RDRAM number of RDRAM's up to 300 RDRAM's can be you can have in a single system that can be interface your computers systems. So, between sending the address of the request and the written of the data it allow other accesses about the bus.

That means, this is another features that is provided in RDRAM, and also internally does the refreshing. Since you are using dynamics RAM, it will require that the refreshing of the memory that refreshing as you are doing the reading that time writing in many case it gets refresh, while reading also as you read it gets refresh. So, that reading is used for the purpose of refreshing, and controller and this are the various components control and RDRAM module 16 bit data and 2 bit parity cycling twice the clock rates. And 8 lines for address and control, so 8 lines for address on control, so this is the RDRAM memory system that is available.

(Refer Slide Time: 49:25)

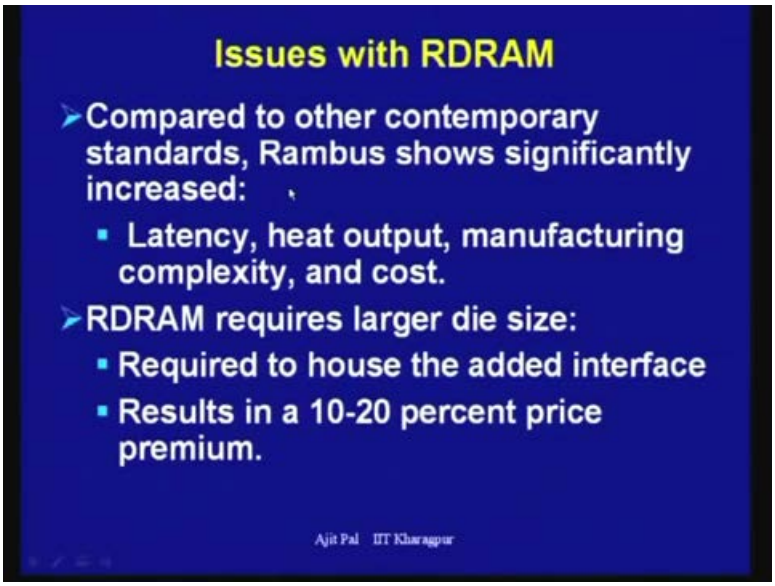


And this was adopted by Intel for Pentium and itanium processers, and is RAM bus DRAM is the main competitor of a SDRAM, and it is available in vertical package. That means, all pins that 28 pins available on one side, data exchange take place over those 28 wires. And that width of that is less than 12 centimeter and but at the bus addresses up to 320 RDRAM chips as I mention wrongly as 300, but it is 320 at the RDRAM's chips at the rate of 1.6 mega bit per second, this will be not B B is G b p s GIGA hertz per second. So, asynchronous block also work in a synchronous block oriented protocol, where the access time is 400 nanosecond, and all then after this access times it gives you at the rate of 1.6 gigabits per second.

(Refer Slide Time: 50:27)

And this is the typical RDRAM with the pins various pins the necessary heat sink integrated heat sink.

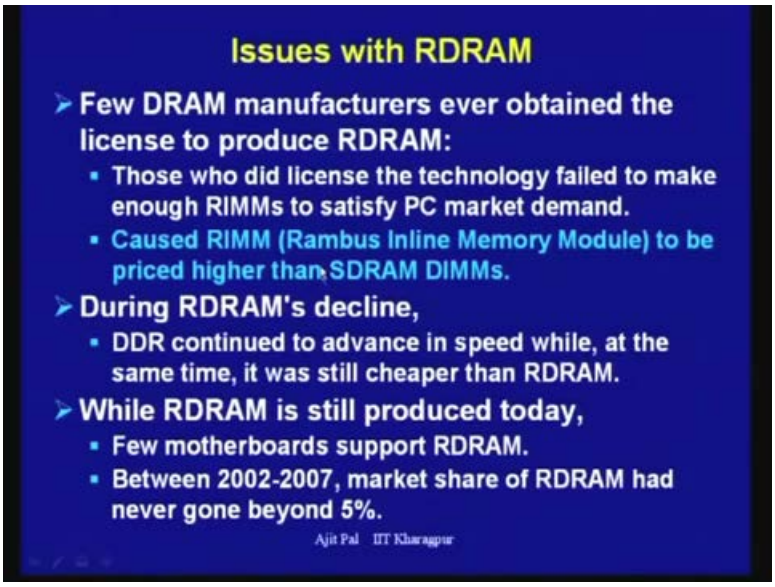
(Refer Slide Time: 50:44)



So, it has been found that RDRAM is I mean compare to other contemporary standards RAM bus shows significantly increased latency, heat output. Because, of a I know that various control circuits built in a part of the RAM bus RAM it generate lots of heats, so there is a long latency, heat output, manufacturing complexity can cost. So, RAM bus norm is much costly than the a SDRAM, and RDRAM requires a longer die size request

to house the added interface results in 10 to 20 percent price premium, so as mentioned is much costly as thin as the SDRAM.

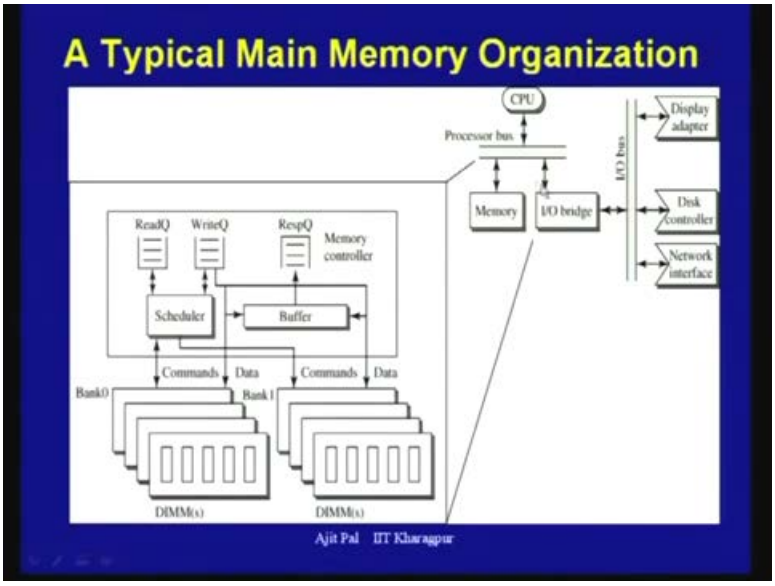
(Refer Slide Time: 51:25)



And this are some other issues divided to RDRAM, few dram manufactures ever obtained the license to produce RDRAM, those who did license the technology failed to make enough RIMM's to satisfy PC market demand. So, this means that it is less popular than SDRAM, and during RDRAM decline DDR continued to advance in speed while the same line, it was still cheaper for RDRAM.

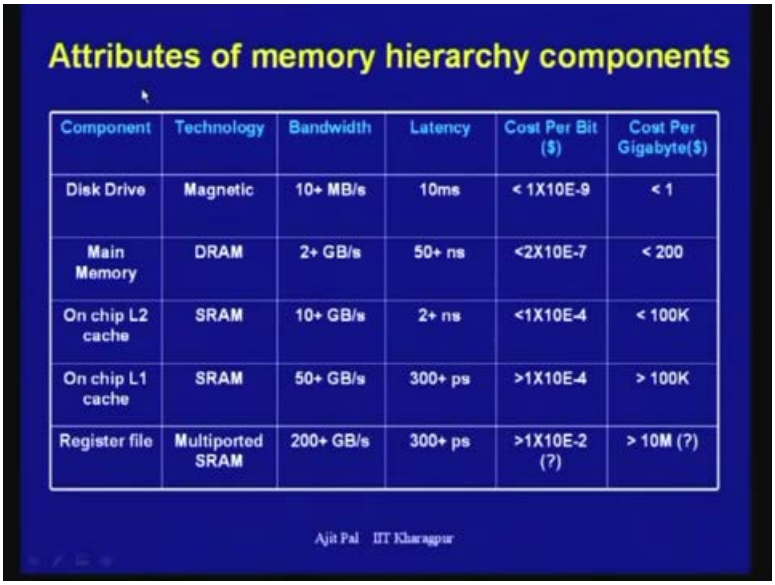
So, while RDRAM is still produce today commercial it is still produce for few motherboards supports RDRAM, and between 2002 to 2007 only 5 percent a market was capture by RDRAM. So, essentially all we are trying to tell is that SDRAM is the more popular than RDRAM, although it is being manufactured.

(Refer Slide Time: 52:32)



So, this is a typical main memory organization here you can see the input output bar space, the processer, this is the processor bus. And this is the memory and input output bridges, and you have got the read Q, write Q, response Q, scheduler, buffer and here are those various banks, bank 0, bank 1, and DIMM's are connected to realize the memory and input output banks.

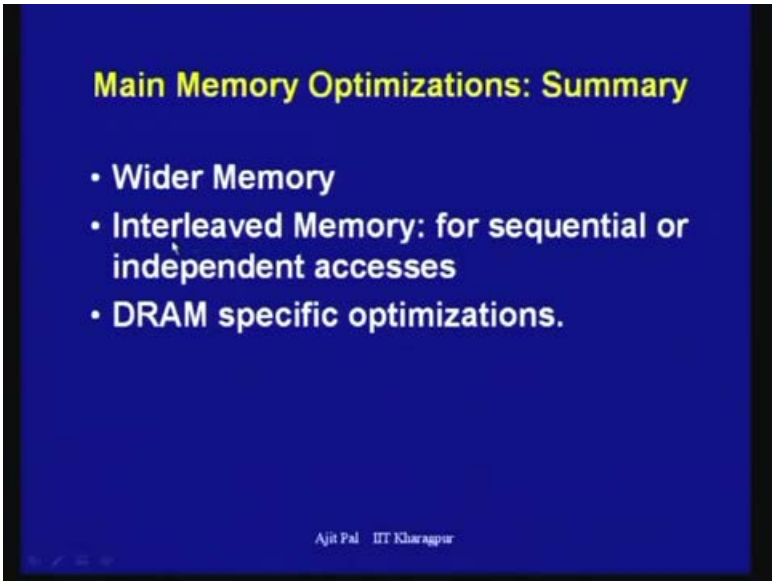
(Refer Slide Time: 53:13)



So, this is the hierarchy that we have already discussed I mean this drive main memory on chip L 2, on chip L 1 register file. So, various technology is that are used magnitude

DRAM, SDRAM, main memory is SDRAM, and bandwidth as you can see 2 gigabits per second 1.6 we have seen now a days it is 2 gigabyte per second. And latency is 15 nanosecond for main memory compare to 2 nanosecond for on chip failed to cache, and cost is significantly smaller then the cache memory. So, later in a next class I shall discuss about the another hierarchy that hierarchy between main memory and the secondary memory, which is known as you know that a virtual memory system. So, in the next lecture we shall discuss about that.

(Refer Slide Time 54:24)



So, to summarized you have discussed about the enhancement of main memory, main memory optimization techniques by using wider memory, interleaved memory and also you have discussed about various dynamic RAM specific optimizations, like the use of SDRAM and the RAM bus RAM. So, with this we have come to the end of today's lecture.

Thank you.