VLSI Technology
Dr. Nandita Dasgupta
Department of Electrical Engineering
Indian Institute of Technology, Madras

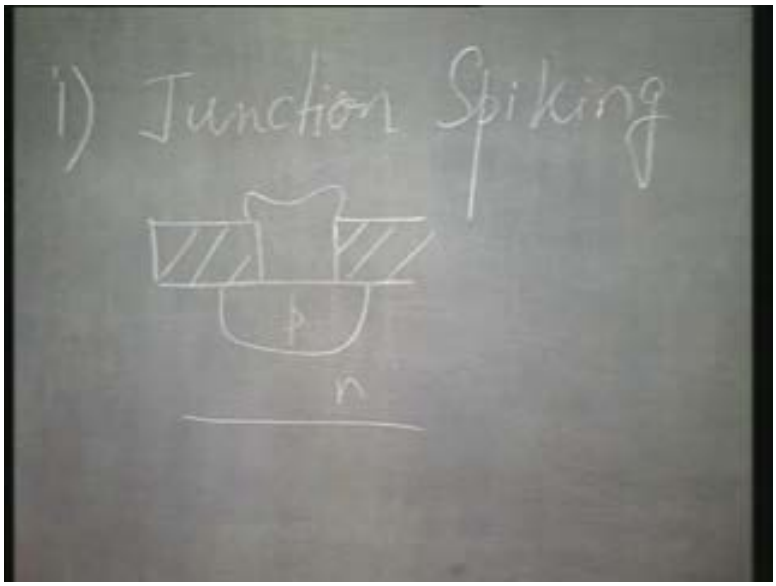
Lecture - 31
Problems in Aluminium Metal Contacts

So, we have been discussing about the metallization in the integrated circuits and you know that by far, aluminium is almost the only metal so far to be used for ohmic contact, no matter whether the region is p-type or n-type and because of this dependence on aluminium, I have already mentioned that it becomes difficult to form an ohmic contact to a moderately doped n-type silicon, right. Aluminium can easily form ohmic contact to p-type region, p region or to heavily doped n plus region. But, it is a bit difficult to form ohmic contact to moderately doped n-type silicon. Now, even though aluminium is almost the only metal to be used in integrated circuit technology that is not to say that it does not have any problems; aluminium does have quite a few severe problems and well, of these major problems, one is called the junction spiking.

(Refer Slide Time: 2:39)

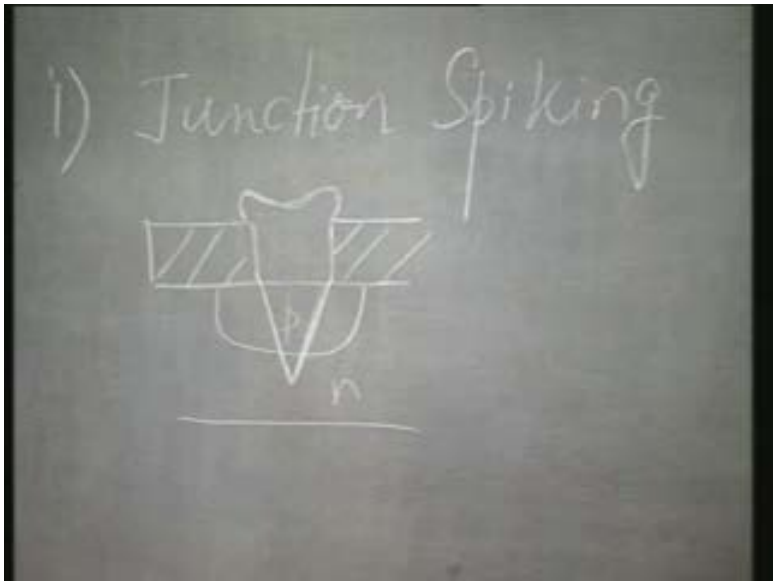
First problem is …. and this problem of junction spiking becomes more and more severe as the device dimensions are reduced that is the contact window becomes smaller and the junctions become shallower. Why so? In order to realize that, you have to first understand why junction spiking is caused? What exactly is junction spiking? I have a junction, a pn junction and I have put metal on top of the p-region, let us say.

(Refer Slide Time: 3:46)



That is to say I have this pn junction. These portions are protected by oxide, so this is my contact window and I want to put metal there on top of this, right. Now, what happens is, if you look at the aluminium silicon phase diagram, then you will notice that at 450 degree centigrade, aluminium dissolves 0.5 weight percent of silicon and 450 degree centigrade is the most common aluminium annealing temperature. That is after you put the aluminium metal, you anneal it at 450 degree centigrade, in order to form a good ohmic contact and at 450 centigrade, this aluminium is going to dissolve silicon. Where does it find the silicon? From this area, right; so, it makes the contact with the substrate, right.

(Refer Slide Time: 5:06)



So, if the contact window is becoming smaller and if the junction is shallow, then there is a possibility that it will dissolve in this fashion and there will be a conducting path going all the way through the junction and forming contact with the underlying substrate. It is dissolving silicon from this region. So, if this junction is shallow, obviously there is more risk of this junction spiking. At the same time, if the contact window is small, then of course, the width of this region is small and therefore, it is going actually vertically down. So, in both cases, as the device dimensions becomes smaller that means it is shrinking in all directions, junctions are becoming shallower as well as the contact windows are becoming smaller and then, this problem of junction spiking becomes a very real problem and all this happens because, aluminium dissolves silicon.

So, what is the remedy? What can we do to circumvent this problem? One possibility is, when you are depositing aluminium, do not deposit pure aluminium; have a little bit of silicon mixed with aluminium, so that the deposited metal already contains some amount of silicon. At 450 degree centigrade, when it starts to dissolve this 0.5 weight percent of silicon, this silicon is supplied from this layer itself; it does not have to take the silicon from the substrate.

(Refer Slide Time: 7:02)

So, one possibility is the metal itself must contain some amount of silicon and usually you have to give sufficient amount of silicon. 1 weight percent of silicon is provided with aluminium, so that it does not need to take any silicon from the substrate. But, there is a small problem. As I told you, aluminium is used to form ohmic contacts not only to p-type silicon, but also to heavily doped n-type silicon, right. Now, imagine the case. I have a heavily doped n plus region, in which I have put this aluminium with silicon to form the ohmic contact and remember I have put a little bit of excess silicon, just to be on the safe side, because I do not want any dissolution from the substrate material.

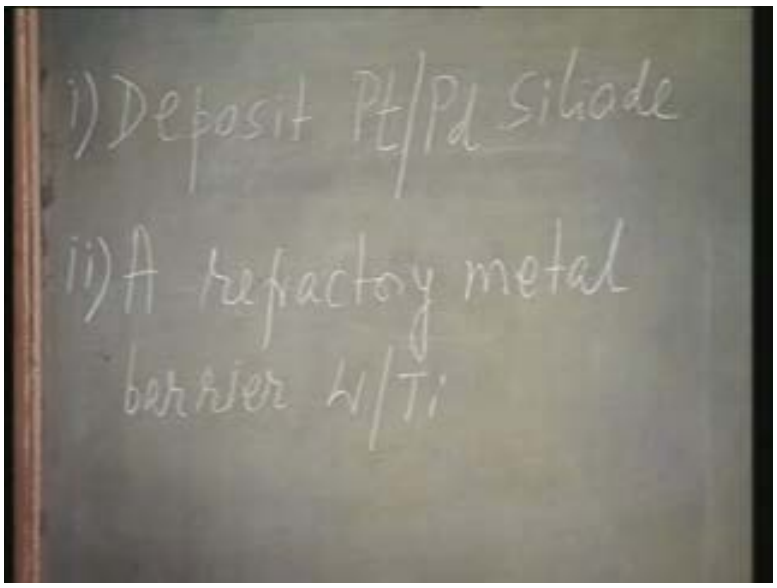
So, what will happen to this excess silicon? It will precipitate and this excess silicon is having aluminium in it. Therefore, it is p-type. So, you may have a non ohmic contact, because this excess p-type silicon may get precipitated. So, instead of having an n-type, contact to an n-type region, you may again end up with having a pn junction. No such problem exists when you are forming contact to p-type silicon. But, if you are forming contact to n plus silicon, then because of this excess silicon getting precipitated, you may again have a pn junction there. It will not be a very good pn junction, but it will anyway result in a non ohmic contact. So, even though it is a common practice to have aluminium

mixed with silicon to be used as the contact metal, it can sometimes lead to problems of having a not proper contact.

A better alternative and therefore obviously a more complicated system will be, instead of using the aluminium directly in contact with silicon, you first use a noble metal silicide, a platinum silicide or palladium silicide; use a noble metal silicide. Noble metal silicides can be formed in various ways. One way will be, you put the noble metal on top of silicon, let it react. This again have the problem that in the shallow junction the noble metal may react with silicon and heat up the entire junction. Therefore, the other better solution is you deposit both platinum and silicon and sinter them, so that it forms platinum silicide there. We will discuss the problems of silicide formation a little later.

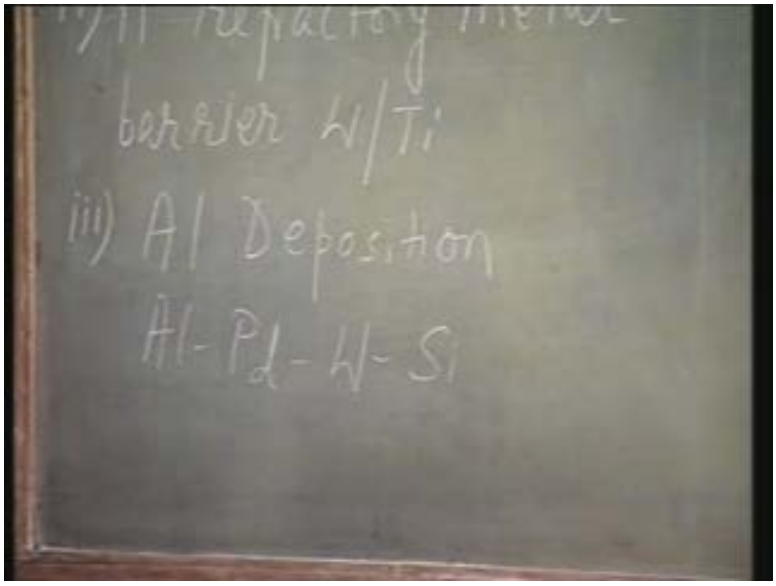
Let me first outline the case.

(Refer Slide Time: 10:37)



So, what we can do is deposit platinum or palladium silicide. Now, even if you put aluminium on top of this platinum or palladium silicide, you know, aluminium may still start eating silicon from the silicide itself. So, usually a refractory metal barrier is placed on top of this, tungsten or titanium and then on top of that, you put aluminium.

(Refer Slide Time: 11:38)



So, now, you are introducing the concept of multilevel contact metallization for more reliable operation. Multilevel - that is it has three levels, as you can see. First, you have put a noble metal silicide and then you have put a refractory metal, so as to prevent the diffusion of aluminium and then on top of that you have put the aluminium. Obviously this is a more complicated process, more expensive process, but it results in better reliability of the contact. So, a very common composition is aluminium-palladium-tungsten silicide; common composition aluminium-palladium-tungsten silicide. So, this is one major problem with aluminium metal contact, the problem of junction spiking.

The other and perhaps the most severe problem is that of electromigration.

(Refer Slide Time: 13:19)

What is electromigration? You see, I have metal lines and electrons are carrying the current, right. They are the current carrier. So, electrons are moving under the electric field. That is you apply a voltage; there is a voltage difference along the length of the metal line, electrons are moving under this applied field. The higher the field, faster is the movement, right; more is the movement. In this process, when the electrons are moving under the influence of this high electric field, they collide with the positive metal ions and transfer momentum to the metal ions.

So, what happens? If sufficient momentum is transferred, then the actual metal ions also start moving towards the direction of the anode, because this is the direction in which electrons are moving. Electrons are moving towards the anode, right; positive electrode. Electrons are moving towards anode and while they are moving, they are transferring some of their momentum to the positively charged metal ions. That is they are carrying the metal ions also along with that. So, there is physical, bodily movement of metal. So, what will happen?

(Refer Slide Time: 15:09)

If I have a metal line running there and if this is somewhere my anode is, then what you may see is a break here, because metals have moved from this point towards the anode and in the anode you may see, towards the anode you may see accumulation, right. So, due to electromigration, you will always find a break in the metal along with accumulation of metal towards the direction of the anode. It is physical transfer of metal from one place to another place under the influence of the electric field. This is electromigration, migration of metal under the electric field and this electromigration is particularly severe in case of aluminium.

Electromigration, the extent of electromigration is quantized by a term called mean time to failure. That is you have formed the device with the metal lines. Under normal operation, how long does it take for the failure to occur? Mean time to failure - average how much time does it take for a failure to occur?

(Refer Slide Time: 17:00)

MTF or mean time to failure and this mtf is found to be proportional to the current density to the power minus 2 or sometimes it is given as minus n, n being close to 2; it is related to this. This part is of course very easy to understand, it is inversely proportional by a power of current density. Obviously, the more the current density, the faster it is going to fail, right. The more current you make the lines carry, the faster it is going to fail. This Q is something analogous to activation energy and in most cases this Q corresponds to grain boundary diffusion. That is the metal line that you are depositing, metal layer that you are depositing, it is not a single crystal metal, right.

It is, most often it is polycrystalline layer that you are depositing; the layer of the metal that you are depositing. Aluminium, it is polycrystalline layer. So, this mtf is found to be closely related to the grain size. The larger the grain size, the more resistant it is to electromigration which signifies that mean time to failure is related to the grain boundary diffusion. If you have larger grains, grain boundary diffusion is less, mean time to failure increases. It is also found that it will depend not merely to the grain size, but the distribution of the grain size; if all the grains are more or less uniform in size, then mean time to failure improves.

Interestingly in this one aspect, the shrinking of device dimensions helps somewhat. As the metal lines become narrower, as you decrease the line width the metal lines become narrower, mean time to failure increase and that is because when the line width is decreased, then aluminium tends to form a single crystal layer. You are depositing it over a very small region; in that small region, the aluminium line is more or less single crystal. So, this is one aspect where diminishing the device dimensions, decreasing the line width, actually helps somewhat since the metal line now is composed of single crystal elements.

A very common way to improve electromigration resistance of aluminium, as I told you that electromigration is particularly severe in aluminium; it has a very small Q, it has a very small Q and that is why mean time to failure is quite low; one way to improve this is to use a little bit of copper along with aluminum. In fact it is quite a common practice nowadays, to instead of using aluminium alone, one uses aluminium copper alloy to improve the electromigration resistance and it is believed that copper actually increases this. Addition of copper actually increases this activation energy, so that the mean time to failure improves and now comes one of the latest developments in the metal connections. See, for a long time people lived with all the disadvantages of aluminium. The particularly severe disadvantage is that of electromigration. But now, there may be an alternative to aluminium metal lines and so, we enter in the copper age in VLSI technology.

I have a reference here of Spectrum issue, January, 1998; you can all take a look at that. I am just going to give the summary of this. Both IBM and Motorola, two of the giants in VLSI technology, fabrication technology, they have independently developed a copper interconnection technology, copper interconnection technology. See, for a long time people knew that copper will have a higher electromigration resistance; at the same time, copper will have lower resistance. See, as the length of the metal line increases, the resistance increases. So, aluminium is a poor choice compared to copper. But for a long time, people could not use copper, because of various reasons. One problem is how do you prevent copper from diffusing in the surrounding region of semiconductor? So, that was one reason why people did not use copper.

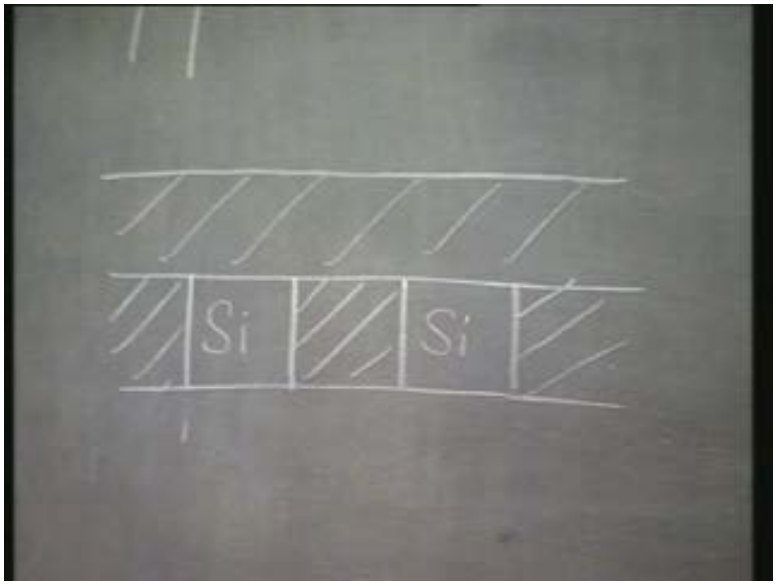
But, in January, 1998 both IBM and Motorola, have announced that their next generation devices will have copper. They are using copper for their submicron CMOS technologies. So, as far as IBM is concerned, their CMOS 7 S series they are going to have all copper interconnects and Motorola is going ahead with the static RAM fabrication using this copper interconnection, which should have been marketed by September. I do not have the latest information whether this has really come into the market, but in January they were projecting that by September it will hit the market. They have already developed the technology in their lab and now they are going ahead for full stream production.

What do they do in the copper interconnect technology?

(Refer Slide Time: 24:34)

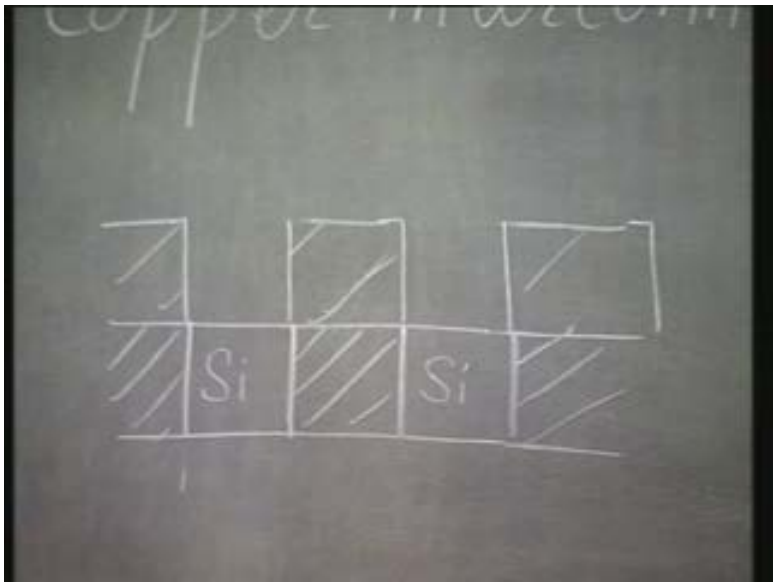
See, first of all you define the region where you want to have the copper connections. That is you have a first level of metal. We are talking about highly sophisticated submicron CMOS technologies, where you have multilevel metallization, right.

(Refer Slide Time: 25:10)



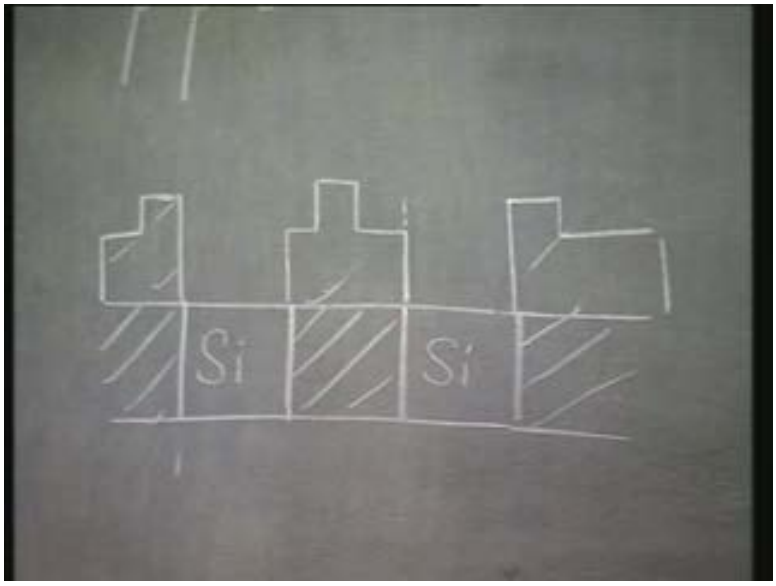
So, first of all you have certain regions in silicon. So, these are my silicon regions or okay, these are my silicon regions; these are isolated by oxide, let us say. First of all you deposit another layer of silicon, another layer of silicon dioxide. Then you do reactive ion etching.

(Refer Slide Time: 25:58)



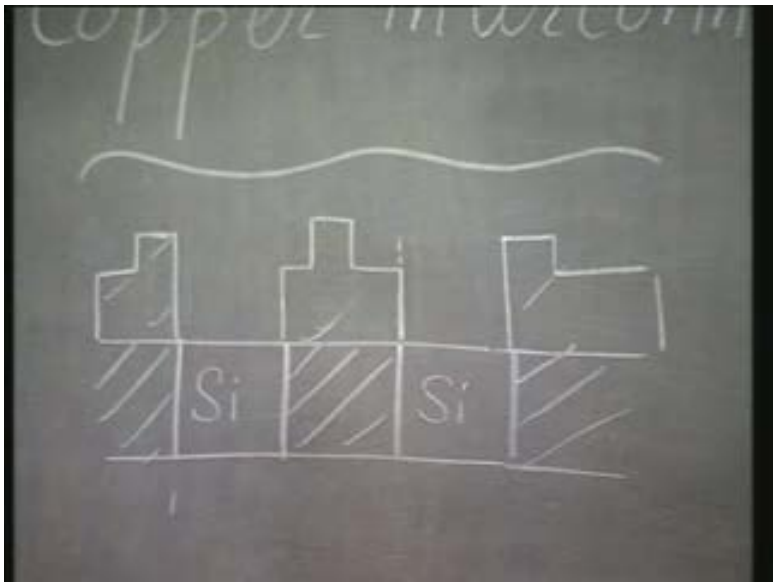
Etch the oxide from certain regions and keep it in the remaining region.

(Refer Slide Time: 26:26)



Then, you do another reactive ion etching in order to realize certain steps in this oxide, so that finally what you have is something like this, somewhat like this.

(Refer Slide Time: 27:05)



In the next step, you electroplate copper.

(Refer Slide Time: 27:26)

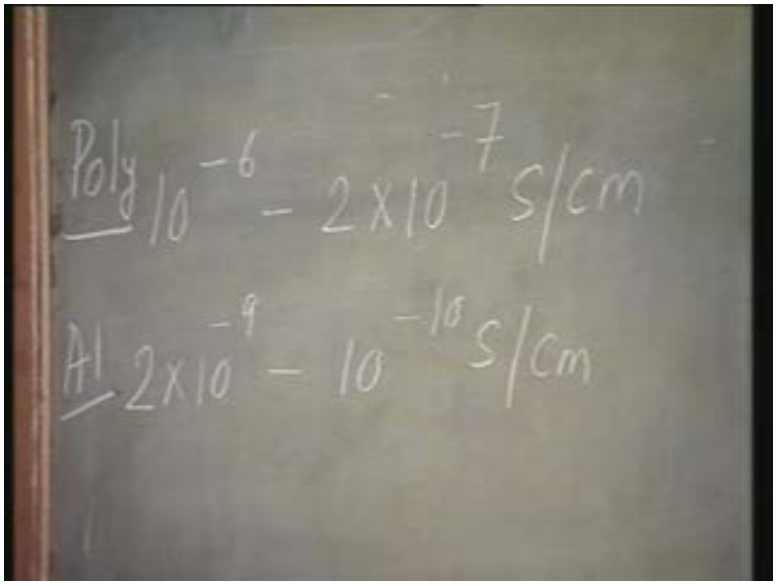
Copper is deposited by electroplating and then, you do a chemomechanical polish, so that copper is retained only in this region; only these regions have copper. So, now you have copper connecting the underlying silicon and from there you can take further connections to other regions. This process incidentally is called the dual damascene process. This has a historical significance. Originally in Damascus, a technology was developed, nothing to do with VLSI; a technology was developed in which they used to cut trenches and fill it with metal.

That was the original damascene technology, nothing to do with VLSI, however. It is very old technology and they have sort of taken that philosophy and used it here; etched trenches in silicon and filled that up with copper by electroplating and then the final step, the most interesting step is that they have chemomechanically polished this metal line, so as to cut back and come to the level, planarity; for planarization and the most significant development in this whole thing is how they have prevented copper from diffusing in the surrounding region, because that was the most difficult problem. Primarily because of that problem, in spite of knowing copper's definite advantages, nobody used copper for so long in VLSI technology. That is how to prevent copper from diffusing into silicon and contaminating all the surrounding area.

Well, Motorola is using titanium nitride to line all these trenches. They are using titanium nitride to line these trenches. IBM is not saying what material they are using; it is a proprietary item. But, they are also using some material to prevent copper from diffusing in the surrounding region and also to improve the adhesion. That is the other very important point. In case of any metallization, the metal must adhere to the substrate. So, Motorola is using titanium nitride in order to achieve both. IBM is also using a proprietary item and by doing this they have reported a two orders of magnitude improvement, as far as electromigration is concerned. So, may be the future of metallization lies with copper. May be the era of aluminium is slowly coming to an end and the era of copper is beginning to start in VLSI technology. So, this is as much as ohmic contact metals are concerned.

If you remember, we said we have to talk about the other class that is the gate metal, which is no longer a metal; we said that we use polysilicon. But again, as the device dimensions are becoming smaller and we are more and more going for higher speed of circuitry, finally it is the RC delays in the interconnections that is going to determine the speed of the circuit. See, the gate and the interconnection we usually use the same material. That is for one level where the devices are being connected with each other. You use the gate and the same metal or same material is used also for interconnecting. Now, it is this RC delay, RC product, which is finally going to determine the speed of the circuit. We will compare the RC delays of polysilicon with aluminium and then we will see. So, the term that we use in order to express this delay is delay per unit length, so many seconds per centimeter.

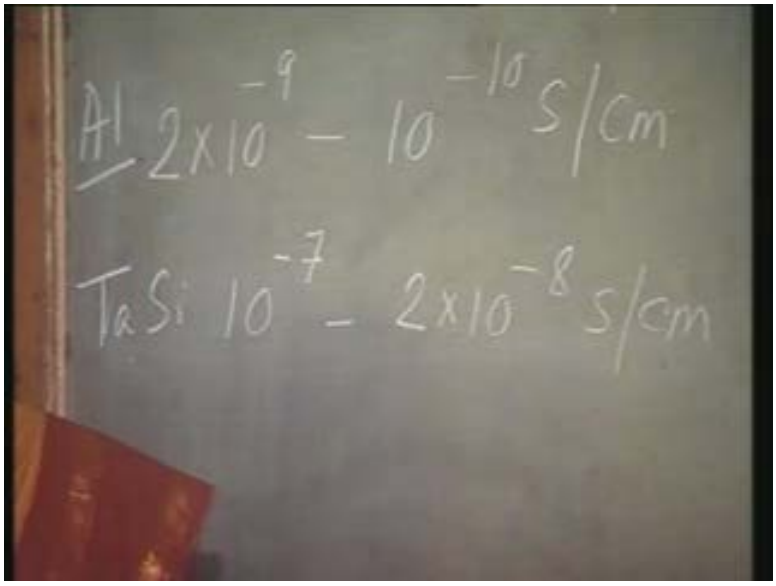
(Refer Slide Time: 32:46)



For polysilicon, this is somewhere in this range - 10 power minus 6 to 2 into 10 power minus 7 seconds per centimeter, when we have a line width of say 0.4 to 5 micron. If we use aluminium, we can go to a much, much smaller delay, 2 into 10 power 9, 10 power minus 9 to 10 power minus 10 seconds per centimeter. So, directly you can see that using aluminium gate would have resulted in faster circuits, but the problem in aluminium is that we cannot use a self-aligned technology, right. If you use aluminium as the gate, then the gate metallization has to be the last step, whereas in self-aligned technology, we first defined the gate and then diffuse or implant source and drain with respect to that. That advantage would be lost. That is why in fact from aluminium, we came to polysilicon gate technology.

But, as the speed of the circuit became an important issue, we find that polysilicon is quite a poor contender as far as RC delay is concerned. That is because very simply, however highly you dope polysilicon, it can never match the conductivity of a metal; as simple as that. So, what is the other choice? The other choice is instead of using polysilicon, if one uses a silicide, we are coming back again to the concept of silicide.

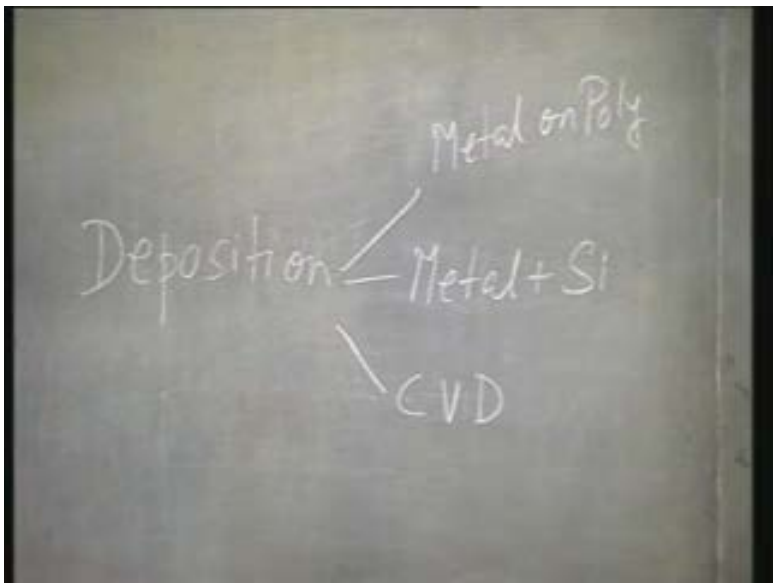
(Refer Slide Time: 35:15)



For example, if one uses tantalum silicide, then one is somewhere in between; almost one order of magnitude improvement compared to polysilicon. So, nowadays, most of the MOS technologies, they retain the advantage of polysilicon, but they do not use simple polysilicon for gate; they use polysilicon with a silicide. So, common choices are tungsten or molybdenum or tantalum or titanium. These are the common silicide choices. Usual practice is you still have polysilicon deposited on the gate oxide and on top of that you have this silicide. This is done, simply because many of these refractory metal silicides they are not very stable in oxidizing ambient. That is if you deposit them straight away on top of the gate oxide without having any intervening polysilicon layer, then these metal silicides will not be very stable.

At the same time, the polysilicon gate technology is well researched. We have a wealth of information about the polysilicon deposition and since, remember gate is the most vital part in the MOSFET technology, you would like to retain all the positive features of the developed technology. That is you know exactly what the polysilicon deposition conditions should be in order to guarantee optimal performance of your device. So, you do not want to disturb that. What one is going to do is, on top of this polysilicon gate we can deposit a silicide. There again you have various choices.

(Refer Slide Time: 37:49)



So, silicide is deposited. One choice is you directly deposit the metal on top of polysilicon. You deposit tungsten or tantalum directly on polysilicon and sinter. You understand the meaning of the word sintering, right. That is you heat it to a high temperature treatment, so that this metal forms a silicide at the region of contact, where the metal and the polysilicon, at the interface it will form a silicide. That is one possibility. Other possibility is you deposit both the metal and silicon either by sputtering or by physical evaporation, vacuum evaporation and let it form the silicide. This is done to be on the safe side. Suppose you put metal directly on poly, sinter it; let it form the silicide.

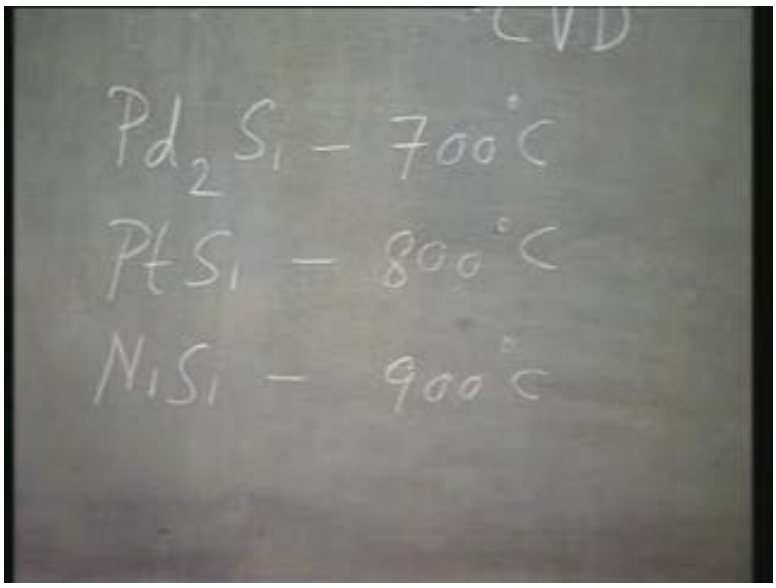
If in the process the entire polysilicon gets consumed, then you again have the problem of having the silicide sitting directly on the gate oxide. In order to avoid that, you are supplying silicon along with the metal, so that it does not eat the entire polysilicon layer. You still have the gate oxide on top of that polysilicon and on top of that the refractory metal silicide. You can have metal directly on poly or metal plus silicon or you can even have a chemical vapour deposition technique, in which you form the metal silicide by chemical reaction, let it deposit on the substrate material. Three most common practices and of these different refractory metal silicide, titanium silicide has the lowest resistivity

and therefore as far as RC product, RC delay is concerned you would prefer to have titanium silicide.

The common practice is, for titanium silicide you let the metal be directly deposited on polysilicon and then you sinter it, because in that case you will have higher grains, larger grains. The only problem with titanium silicide is that the reflectivity is bad. It does not reflect very well, so any subsequent photolithography steps become difficult, because you cannot see it very well. That is the only problem. The other thing to be kept in mind when depositing these silicide is that what is the temperature up to which these metal silicides are stable, because you see that is going to determine the final maximum temperature to which you can subject these devices, once this silicide is formed. That is you want to preserve the advantages of the self-aligned technology.

Why did we use polysilicon? Because, polysilicon could withstand very high temperature; if I am going to use in addition to polysilicon a silicide, then I must make sure that the silicide also can withstand high temperature. Otherwise the advantages of self-aligned technology will be lost. But, most of these metal silicides are quite stable at high temperature, the lowest temperature being for palladium silicide, which is stable up to 700 degree centigrade.

(Refer Slide Time: 42:08)



Platinum silicide is stable up to 800 degree centigrade and nickel silicide is stable up to 900 degree centigrade. If you use tungsten or titanium silicide, they are stable at temperature even above 1000 degree centigrade. So, absolutely no problem as far as the stability is concerned. So, nowadays, most of the MOS tech devices, they are silicided gate technology, silicided gate technology. That is you have the gate oxide, then on top of that you have the polysilicon and then you have a silicide deposited on top using any of these three techniques, depending on which is most suitable and these silicides will have much lower resistivity than polysilicon. It is something in between the polysilicon and the metal.

In fact, there is lot of research going on, because you see ultimately to have a very fast interconnection, you not only have to reduce R, resistance, but you also have to reduce C, R C and what is this capacitance? Essentially this capacitance comes about, because you have two lines, interconnections in between a dielectric. That is the capacitance, right. Most of the cases that is the interconnect capacitance which is finally going to limit the speed, R and C, the product of R and C. Now, for a long time people again were constrained to use only the two available dielectric material that is silicon dioxide or silicon nitride, both of which have a dielectric constant of about 4; 4 to 5. So, a lot of

research is going on to develop alternative dielectric material with lower dielectric constant. You see, if you can lower the dielectric constant, immediately the capacitance is going to be decreased; for the same dimensions, immediately the capacitance is going to be decreased and in the same Spectrum issue, you can see that a new material that is being talked about now is a porous silicon dioxide.

It is silicon dioxide with a high degree of porosity. What does porous silicon dioxide mean? It means silicon dioxide with a lot of pores in it. Now, you see finally, ultimately the dielectric constant of air is 1. So, if you can increase the porosity, the dielectric constant is gradually going to approach that of air and they have succeeded in bringing the dielectric constant down to about 1.8 and because it is, see, a lot of research has been going on; people have developed dielectric material with lower dielectric constant, but the problem is they must be compatible with silicon technologies, right. I mean if I have silicon dioxide or silicon nitride as the dielectric material, what is the great thing about them, that they are compatible with my silicon technology.

So, just developing a new dielectric material is not such a major issue. The ultimate issue is, it has to be compatible with the VLSI technology and in that sense, this porous silicon dioxide is a path breaker, because you are still having silicon dioxide, it is still compatible with VLSI technology; the only difference is by increasing its porosity, you have brought the dielectric, effective dielectric constant down to about 1.8. So, the ultimate fastest interconnections will probably have copper lines with this porous silicon dioxide in between, so as to reduce the RC product. This is the trend which you will probably see in another one or two years.

That brings me to the end of the unit process steps. We have covered all the unit process steps used in VLSI technology in some detail, right. So, next what I propose to do is I will go back now to the two major branches of VLSI technology. That is that of integrated bipolar junction technology and the MOSFET technology. First, we will take the bipolar junction technology, we will look at an integrated circuit bipolar junction transistor, see what are its features and then we will see how the performance of this

transistor can be improved. What process modifications are needed to improve the performance of these bipolar junction transistors and thus we will trace the development of the bipolar junction transistor fabrication process; how from the first old junction isolated npn transistors we have come to today's high speed poly emitter poly base transistors used in emitter coupled logic, how the performance enhancement has been achieved by the advent of process technology.