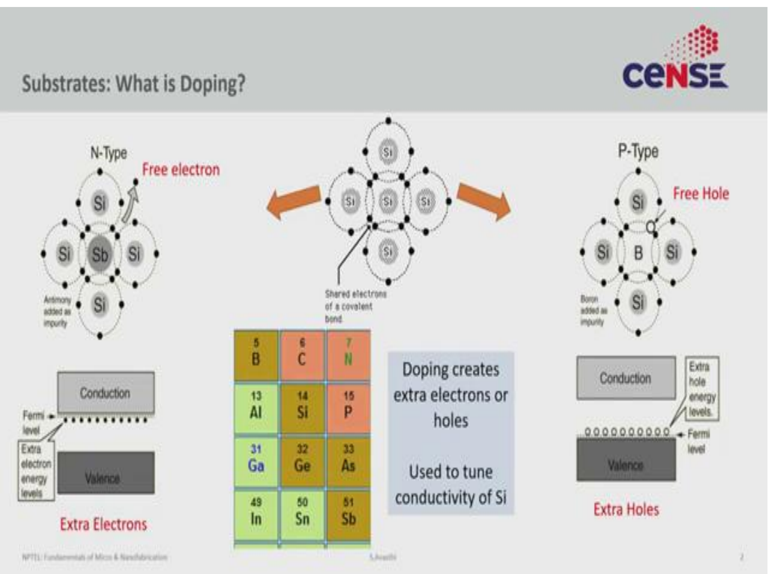
Fundamentals of Micro and Nanofabrication
Prof. Sushobhan Avasthi
Centre for Nano Science and Engineering (CeNSE)
Indian Institute of Science, Bengaluru

Lecture – 08
Diffusion

Welcome back. This is Fundamentals of Micro and Nanofabrication. My name is Sushobhan Avasthi, I am from IISc, Bangalore and today, we are going to talk about the first of the additive steps in microfabrication - diffusion.

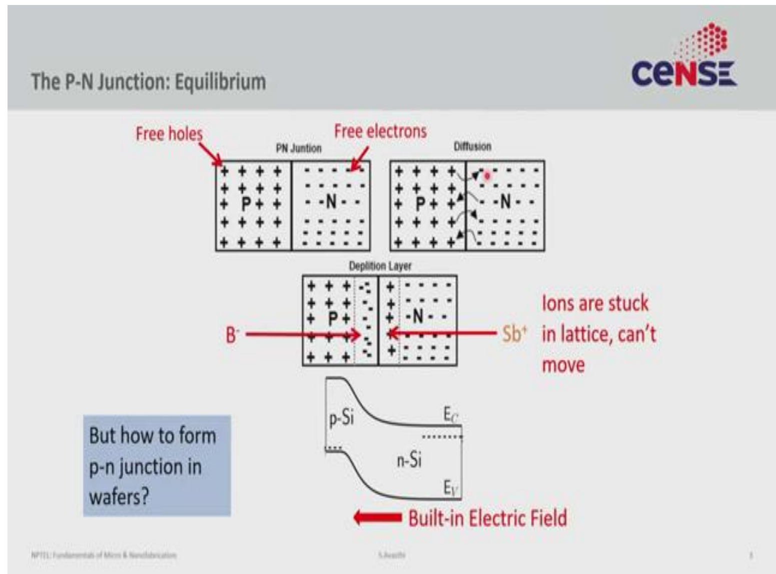
(Refer Slide Time: 00:42)



The fact that you can dope silicon makes it a very good semiconductor. Group 5 element will create N-type doping and a group 3 element will create P-type doping and this allows us to tune the conductivity of silicon. This makes semiconductors so useful.

Now, the question is, how do you create doping inside a material once you have made the crystals? You can always add some dopants at the start of the crystallization, but after the process is done, what can you do to change the doping? One of the ways you can do that is through diffusion.

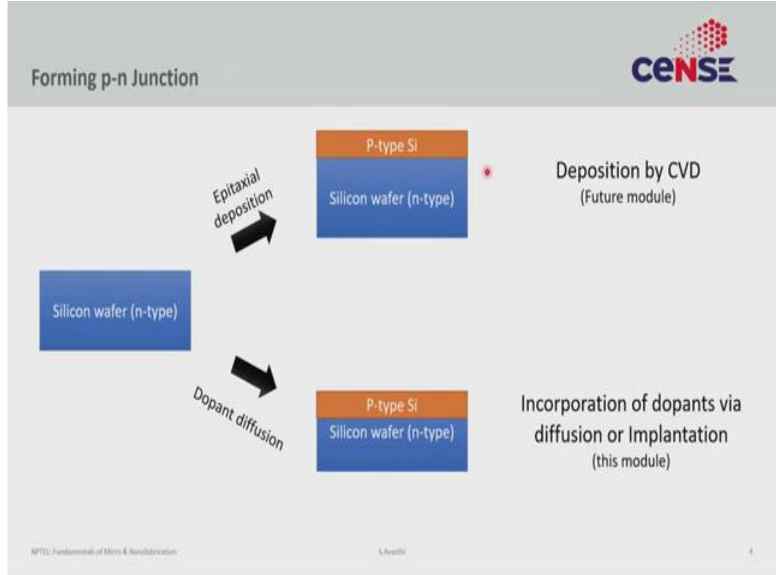
(Refer Slide Time: 01:21)



Diffusion is one of the very important ways in which we make a p-n junction. What is the p-n junction? Join a P-type and N-type material. That creates a gradient for carriers where the holes and electrons go left and right and form a carrier depletion region. This is a p-n junction, one of the fundamental devices that you can actually.

We will not go into the details of what is a p-n junction or why is it useful, just the fact that in order to make a p-n junction, you can use diffusion. Note that the diffusion we were talking about was of carriers after the formation of a p-n junction. In this lecture, we will talk about the diffusion of dopants through the material at a higher temperature.

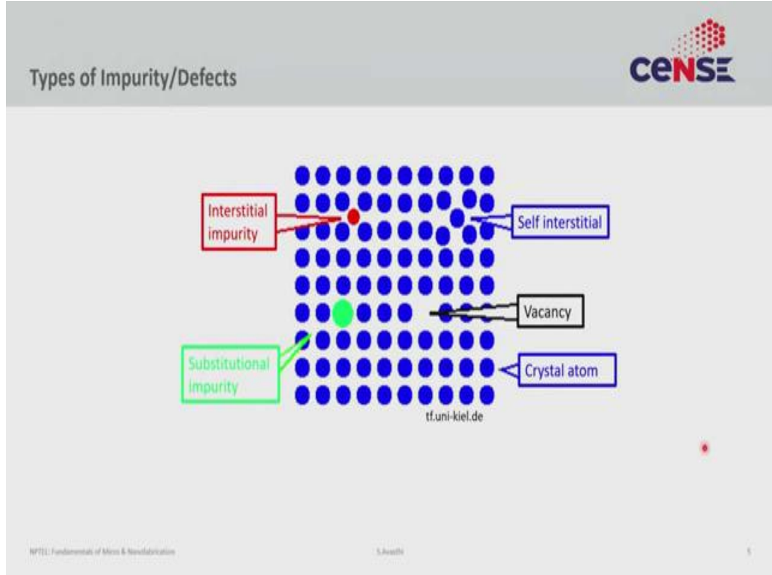
(Refer Slide Time: 02:14)



How do you form a p-n junction? there probably are two ways: deposit a P-type material on top of the N-type material. Take an N-type silicon wafer and deposit to P-type silicon on top. This happens in a chemical vapor deposition process and that would be a future module. Another way of doing this is instead of depositing P-type silicon on top, you can just interpenetrate the P-type dopant, say boron into the N-type silicon. If the concentration of boron is more than the concentration of the N-type dopant, you would form a P-type region on top of the wafer.

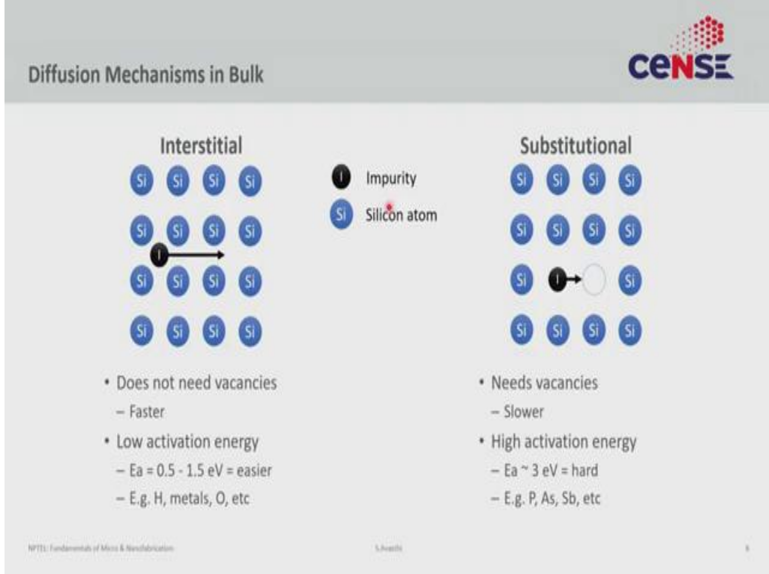
This incorporation of dopants is done by diffusion. It is one of the oldest semiconductor fabrication processes we know. In modern times, we have also started doing implantation which we shall cover later in this module.

(Refer Slide Time: 03:16)



This is a slide you have seen before when we are talking about defects, but basically, any crystalline semiconductor will have a certain amount of vacancies and interstitial at any given temperature just because of entropy. On top of those, you can introduce added impurities in the substitutional or interstitial positions.

(Refer Slide Time: 03:38)



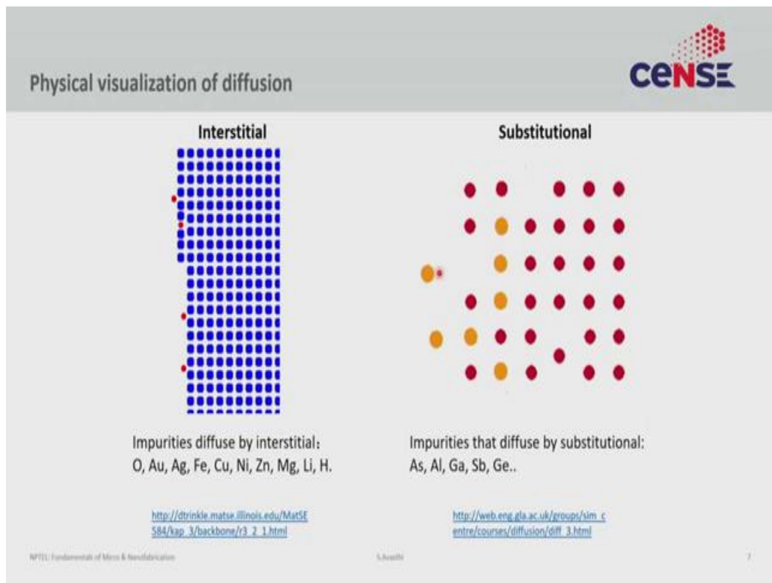
Here is an example of how diffusion can occur. There can be two mechanisms. Take a silicon crystal and introduce an impurity from outside and that impurity can just shift position from one interstitial position to another. This is called interstitial diffusion. It does

not require a silicon-vacancy to move and can move on its own because interstitial positions are present all around in the lattice and hence this process tends to be faster. Thermodynamically, this means that the activation energy (E_a) of this process is low, typically around 0.5 to 1.5 eV. Often this is done by impurities that are small in size. Hydrogen, a lot of metal impurities and oxygen does this in silicon.

Another mechanism is in which the impurities sit at the substitutional site, and the only way it can move if it finds another vacancy close by. Now, because of entropy, at any temperature, there are vacancies floating around in the whole material, but that vacancy must be adjacent to this impurity. So, only when these two are close by can this impurity hop in and that makes this process slower because you are always waiting on the availability of a vacancy to be nearby.

Thermodynamically, that means this process has higher activation energy around ~3 eV. This is mostly done by impurities that have a larger size, so they fit in the silicon lattice position but are unable to move into the interstitial position. Arsenic and antimony are good examples. The primary reason you want to do diffusion is to introduce dopants. So, we focus on materials that create doping - boron, phosphorus, arsenic antimony etcetera.

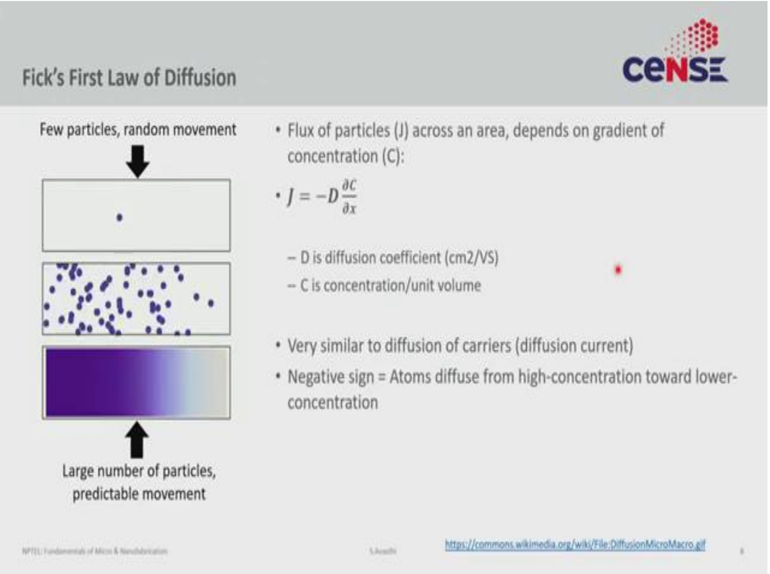
(Refer Slide Time: 05:57)



To help you visualize, here is a slide that shows what interstitial diffusion looks like. The surface is exposed to the red impurity atoms and with time, given enough temperature, these red atoms will sneak in by moving from one interstitial position to another. The other

example is the substitutional diffusion mechanism. On the surface, introduce some orange impurity atom that will move inside the material by diffusing from higher concentration to lower concentration, but through vacancy.

(Refer Slide Time: 06:38)

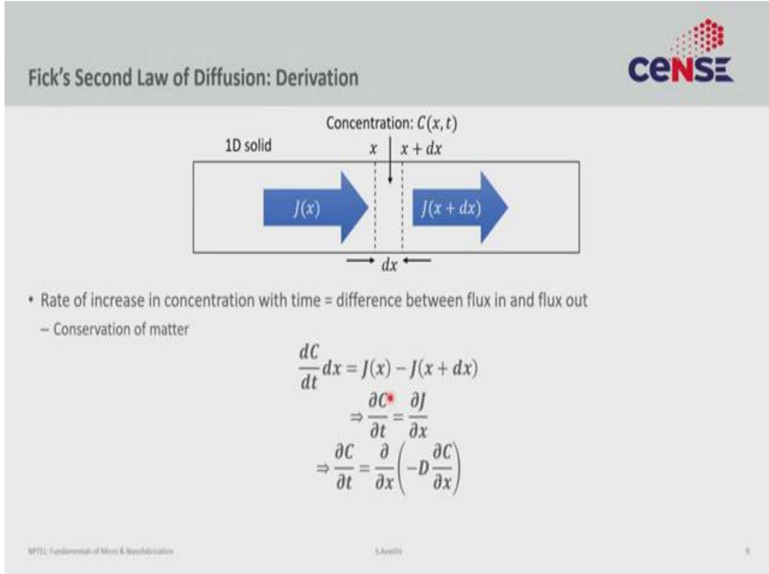


Let's model this diffusion process with Fick's law to get a little more fundamental feeling. Modeling a particle in a box randomly moving around is not very easy. It can be a stochastic and quantum mechanical process and you need statistical mechanics to model this. As you start increasing the number of particles, you get to a point where while the individual particles are doing random walks, the ensemble is doing a certain predictive motion. If the number of particles is so large that it is not possible to talk about individual particles, we only talk about population and in those cases, things start to become much more easy to model. Fick's law is that case and works only when you have a large ensemble of particles when the law of large numbers applies. It does not apply to individual random walks but to the population.

The population would simply move from one direction to the other depending upon a concentration gradient. So, the flux or the number of particles moving $\#/cm^2/s$ depends upon the concentration gradient, from large concentration to smaller concentration. This is how a smell goes from one end of the room to the other end. Small particles created in one corner of the room will go from an area of higher concentration to an area of lower concentration by diffusion.

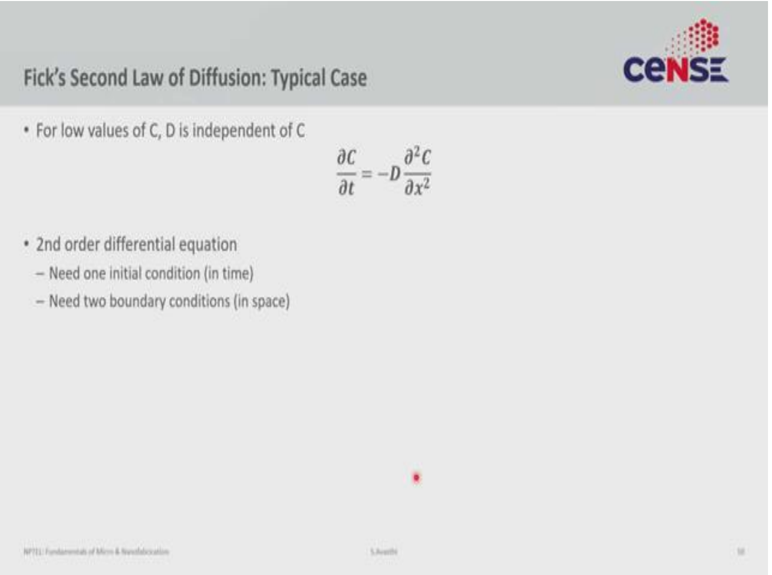
This is very similar to the diffusion of carriers. The only difference the diffusion constant D. The negative sign tells you the direction of the flux; from higher concentration to lower. So, it is negative of the gradient.

(Refer Slide Time: 08:47)



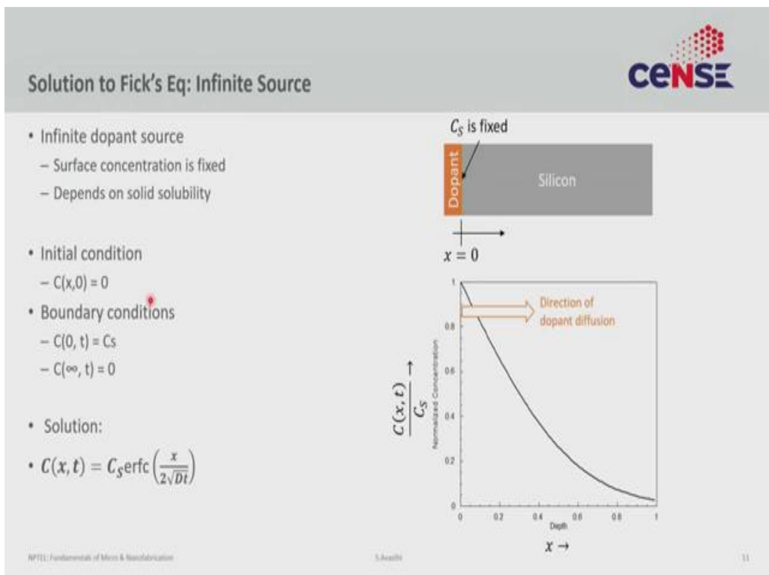
Fick's second law states that matter cannot be created or destroyed. So, the difference between the flux of material coming from the left and the flux of the material leaving must change the concentration in that small box. In 1D, J(x) is the flux coming in from the left, J(x+dx) is the flux leaving and the change of concentration in this small element dx is the difference of these two fluxes. So, the concentration change rate is equal to the double gradient of concentration change with a negative sign and a diffusion constant D.

(Refer Slide Time: 09:44)



If the D is independent of concentration, then you can take it out and this gives you the simplified form of Fick's second law which is typically used in most of the mathematics of fabrication. In order to solve this second-order partial differential equation, you need one initial (time, t) condition and two boundary (space, x) conditions.

(Refer Slide Time: 10:22)

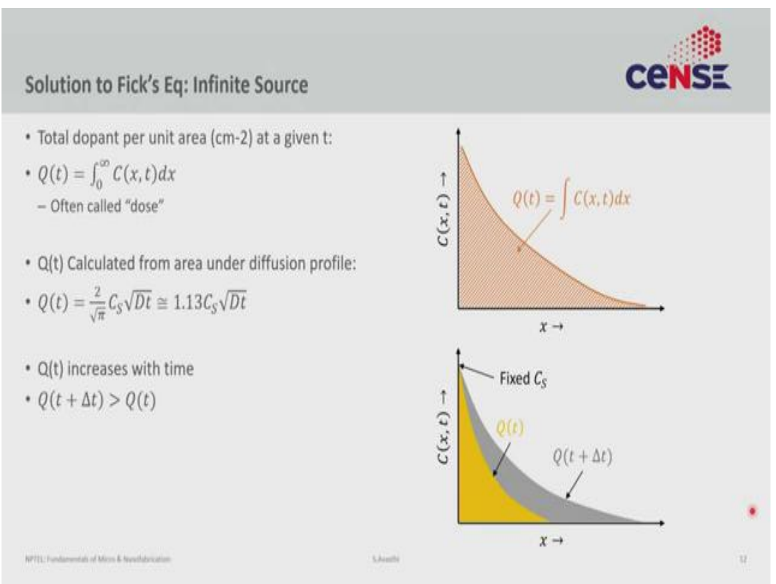


Let us solve it for two generic cases that often occur in dopant diffusion. The first case is the infinite dopant case. Add dopants to one side of the silicon wafer. So, there is a higher concentration of dopants at the interface and no dopant in the wafer. So, the dopants will

diffusion from left to right. Assume that the dopants are in such a plethora that they are nearly infinite. So, the concentration of the dopants will not change much as they diffuse inside the silicon. In this extreme case, the initial condition is very simple. Inside silicon, there was no dopant. So, C(x,0) = 0 everywhere. At the starting point, you have some concentration, but that concentration will not change with time. So, C(0,t) will remain fixed irrespective of time. And very far away, no matter how much you diffuse, you will never have any concentration. So, they will always be gradient and the concentration far away into the bulk of silicon C(∞,t) = 0.

With these conditions, if you solve that partial differential equation, you will get an error function solution. This is a normalized profile as a function of depth. At the surface, where the dopants start diffusing, you have a very high concentration. As you go deeper into the bulk, the concentration falls and asymptotically goes to 0.

(Refer Slide Time: 12:10)

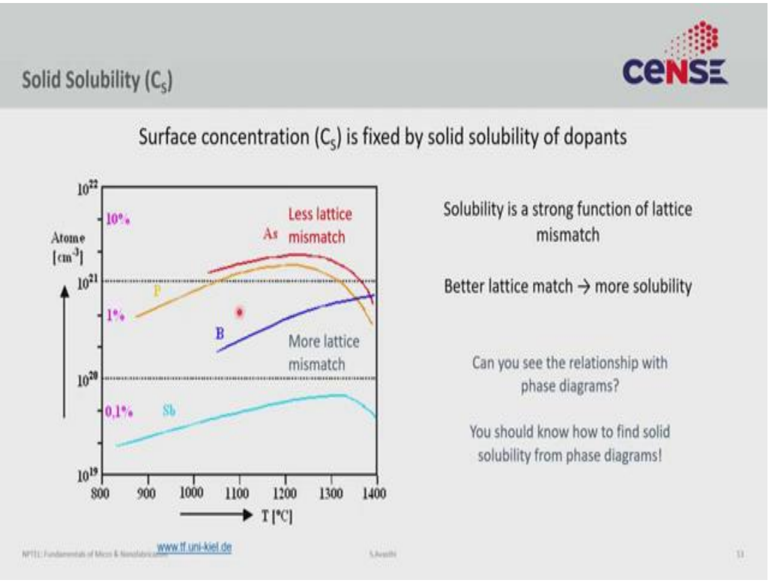


The number of atoms that have diffused in is called the dosage and that is the area under the concentration curve. Integrating it, you can calculate how many carriers have diffused into the material with respect to time. The amount of carriers that have gone inside is proportional to Cs which is the fixed concentration maintained at the surface. If you want more atoms to diffuse in, you should increase the concentration of dopants at the surface. The second thing that you can control is D (diffusion coefficient). If you want more atoms to diffuse in, you must find a dopant that diffuses faster, which has a larger value of D.

Final parameter - time. If you want more atoms to diffuse, make a thicker p-n junction, a thicker doped region then you should just give it more time, but the concentration depends on the square root of both D and time, which is sub-linear. It in order to increase the number of dopants by 10x, you need to increase the time 100x. Sometimes, it becomes practically impossible, which is why diffusion is only useful to create a certain thickness of the doped region. It cannot create arbitrarily thick dopant regions.

The surface concentration is fixed, which is our fundamental assumption that there is an infinite dopant source. The concentration profile maintains its shape but keeps becoming broader and broader. To a good degree, the depth of penetration is $\sqrt{Dt}$ or some function or multiple of $\sqrt{Dt}$.

(Refer Slide Time: 14:31)



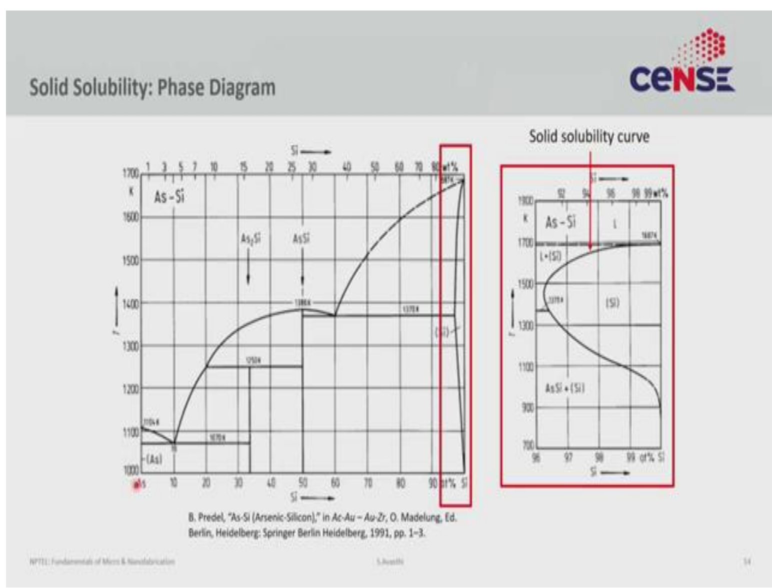
In order to increase the doping, you need to increase the concentration at the surface. What is the maximum concentration that you can maintain at the surface? The maximum concentration you can get is the solid solubility of the dopant in the semiconductor. It does not matter how much material you put on the top, ultimately silicon as a solid can only accommodate a certain amount of impurity inside. The best you can do is solid solubility and it often is a function of temperature. It increases with temperature and at some point, it starts decreasing with temperatures and has a U shaped dependence.

In general, for the various dopants of interest - antimony, boron, phosphorus, and arsenic, the solid solubility limit ranges from around 0.1 % to around 10 %. Arsenic is better

lattice-matched to silicon, and the concentration tends to be higher ~ 10 %. Antimony is not lattice-matched to silicon, so, it does not like getting inside the silicon lattice and the solid solubility tends to be lower.

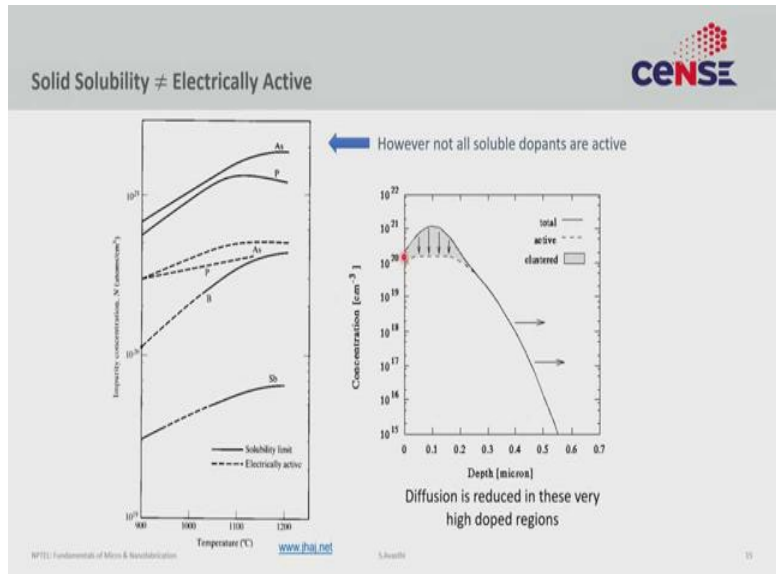
This shape should be familiar to those of you who are familiar with the phase diagrams. Solid solubility is something you can extract from the phase diagram. Phase diagrams per se are not part of this course, but for those of you who did the optional reading the next slide is important.

(Refer Slide Time: 16:04)



In a phase diagram between arsenic and silicon, if you zoom in at the solidus line near the pure silicon part of the phase diagram, it is exactly the solid solubility curve because that is the amount of arsenic that can actually dissolve inside the silicon without phase segregating or a maximum amount of homogeneous dissolution of arsenic inside silicon.

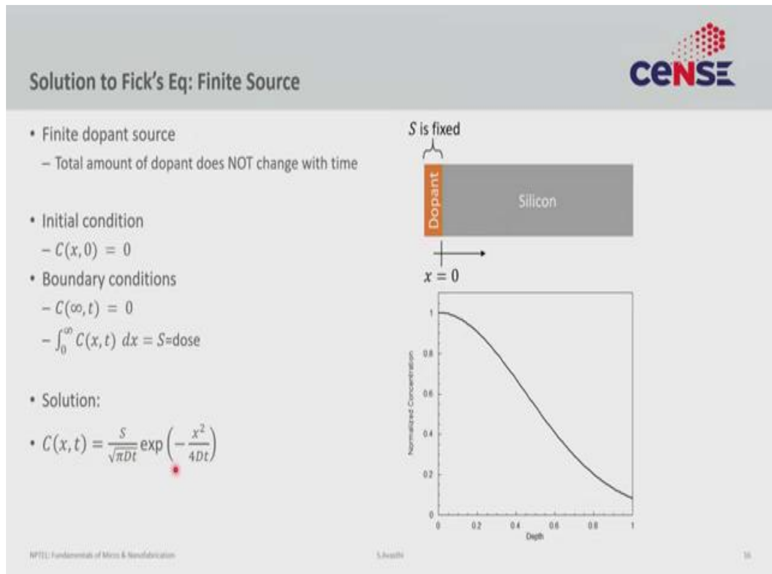
(Refer Slide Time: 16:42)



Solid solubility does not mean electrically active concentration. While doping, the assumption is that dopants would each give an electron if it is an N-type or hole if it is P-type, but that does not always happen. If you look at this slide for example, even though arsenic has around 10^{21}/cm^{-3} or above solubility in silicon, electrically active concentration is an order of magnitude or at least half lower, only $5*10^{20}$/cm^{3}.

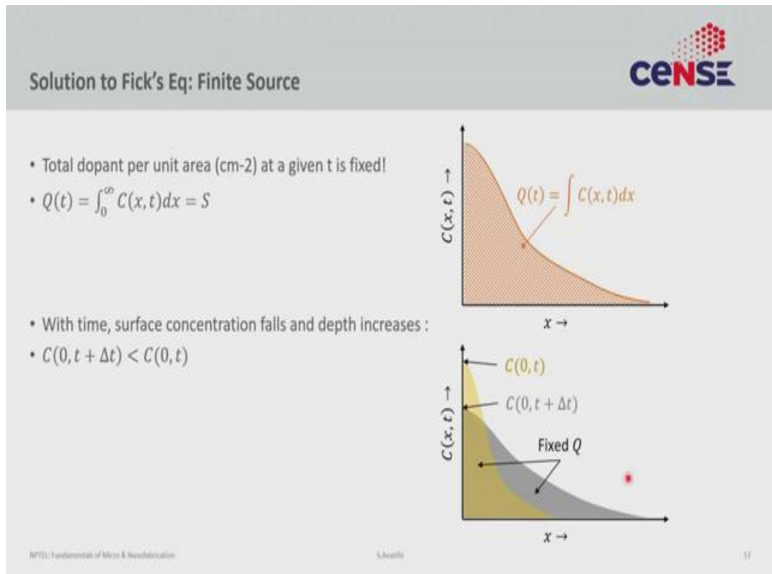
For phosphorous, the difference is even larger but for antimony, the difference is minimal. The effective doping is represented by this dotted line in the slide and is significantly lower in the case of As and P. You have to be a little careful at the very high concentration end as you are limited by solid solubility. And even then, the actual dopant concentration is lower. Some of these excess carriers can also form clusters which are electrically inactive.

(Refer Slide Time: 18:14)



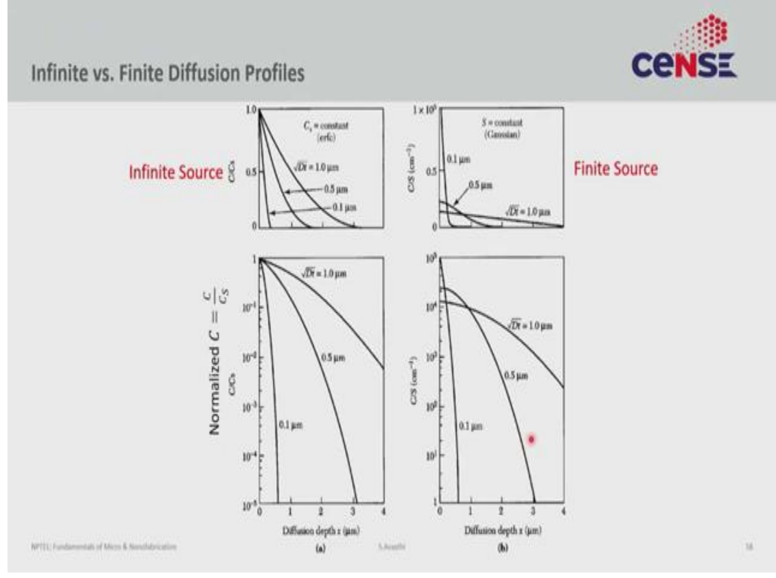
The other interesting case for dopant diffusion, where you do want to solve the Fick's equation (in the 1D system) is the finite dopant source. Assume that you have the dopants at the surface of the silicon that has no impurity, to begin with. However, this time the dopant source is not infinite, but finite. The initial condition remains the same but the boundary condition is slightly different: the amount of dopants available to diffuse in is constant. So, the area under the curve (the dose, S = $\int_0^\infty C(x,t)dx$) must remain constant. With that, the solution to the differential equation is a half Gaussian profile, which is still monotonically decreasing. At the surface, you have a very high concentration limited by the solid solubility or other aspects and as you go deeper into the material the concentration falls and asymptotically goes to 0.

(Refer Slide Time: 19:32)



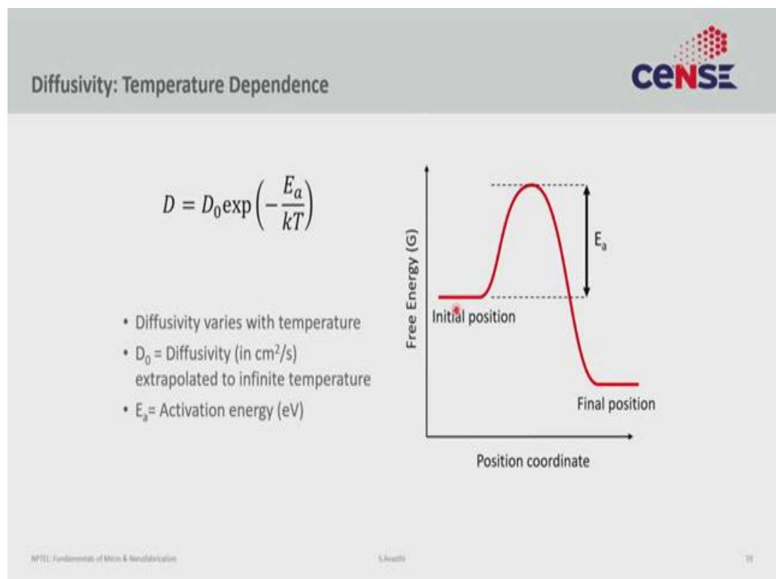
The number of dopants that have gone (dose) unlike the infinite source case is fixed in this case. As you diffuse for longer and longer, the profile remains Gaussian; however, the area under the curve remains conserved. And hence, if the material is diffusing in, the concentration peak must fall. At some point 0, if you start with a very high concentration, it will fall with time and the profile will broaden such that the area under the curve remains constant.

(Refer Slide Time: 20:14)



Here is a comparison of the infinite and finite diffusion profiles. The common theme in both cases is the effective depth of diffusion is some multiples of $\sqrt{(Dt)}$. In order to get twice a thickness of the dopant region, you need to give 4× time or 4× higher diffusion coefficient. In the infinite source, the surface concentration remains fixed while in the finite source, it changes, but the area under the curve remains the same. In one case we have an error function profile and a Gaussian profile in the other.

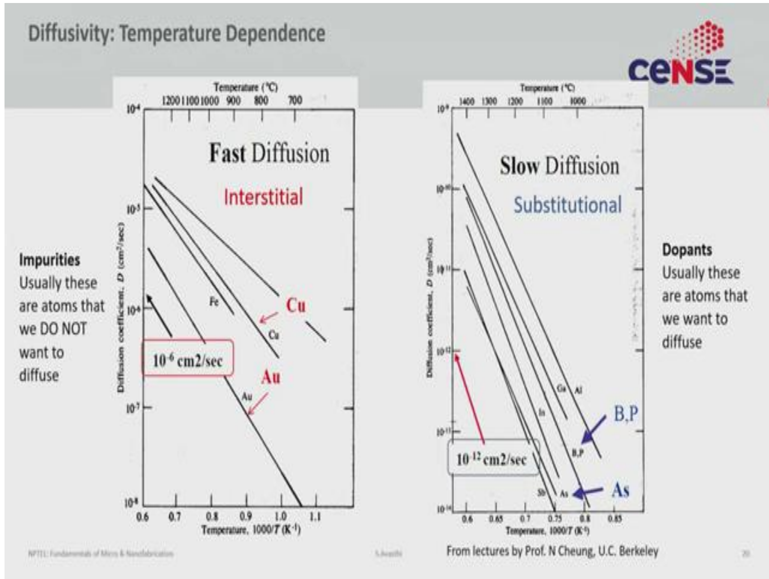
(Refer Slide Time: 20:58)



Let's talk about diffusivity and its dependence on temperature. In general, diffusivity is an exponential function of temperature and that is because, in order for an atom to diffuse, it must go from an initial position to a final position, but in order to make that transition it must overcome an energy hill or a barrier (E_a). Depending upon whether it is substitutional or interstitial, the value of E_a can be different and depending upon E_a being large or small, they will have different temperature dependences.

One way to increase the diffusion coefficient is to increase the temperature because it is a very strong function of temperature. To make the depth twice, you had to increase the time 4×, D increases exponentially with temperature. By a small increase in temperature, D increases by an order of magnitude. So, the dependence of the dopant thickness on temperature is much more strong than on time. The intuition is useful in microfabrication.

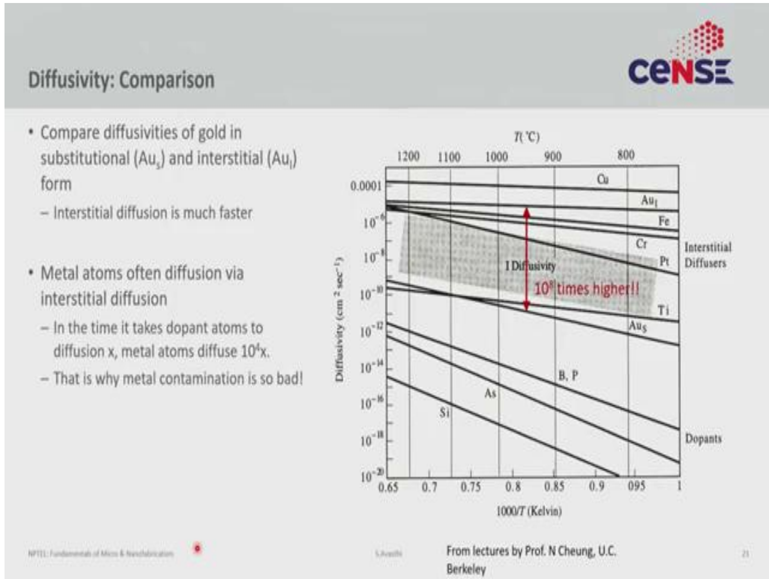
(Refer Slide Time: 02:35)



The exact value of diffusivity depends upon the impurity. Impurities like metals like gold, copper, iron tend to diffuse through the interstitial mechanism and are not dopants. These are also the impurities that we avoid inside a semiconductor, and do not want them to diffuse. However, because they are interstitial their activation energies are lower and the diffusivity coefficients are higher. Typical values are around 10^{-6} /cm^2.

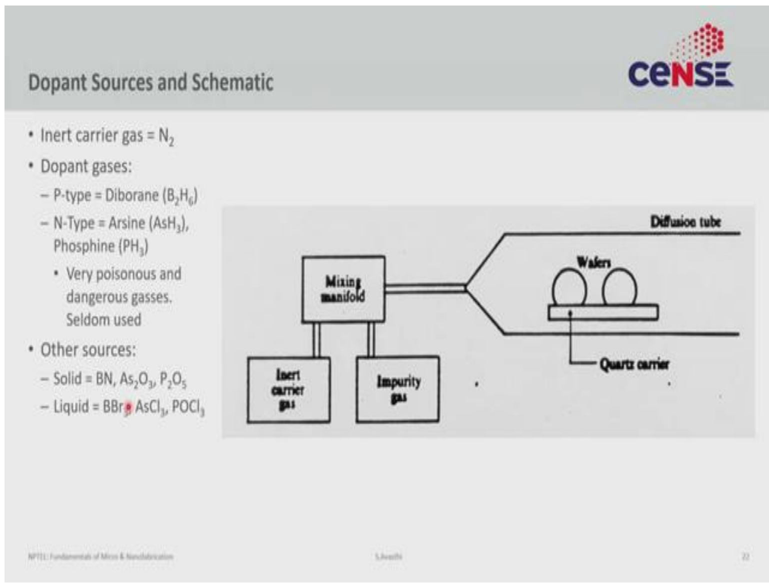
The other type of impurities that we want to diffuse is substitutional (dopants) like boron, arsenic, phosphorus, but because they are substitutional, they have higher activation energies and smaller diffusivities, around 10^{-12} /cm^2. This is 6 orders of magnitude lower. So, the things that we want to diffuse – diffuse slower; things that we do not want to diffuse – diffuse faster and that is a problem. So, during the diffusion process, if you have any metallic impurity coming because the furnace is unclean, or from the metal holder below the wafer, they will go very deep inside the silicon substrate than your dopant atoms. That is why purity so essential. We need to reduce metal contamination as much as possible because any metal contamination will diffuse very fast inside silicon and completely contaminate it and is virtually impossible to remove. So, before any high-temperature step, please be very careful with cleaning.

(Refer Slide Time: 25:12)



Here is a comparison of the interstitial and substitutional diffusivity of gold in silicon. The interstitial diffusivity gold is 10^8 times higher than the substitutional mechanism. Any metal on the surface will diffuse more than the dopants will. So, in the same amount of time that your dopants are going 1 micron, your metal contamination can go hundreds of microns and completely contaminate your wafer.

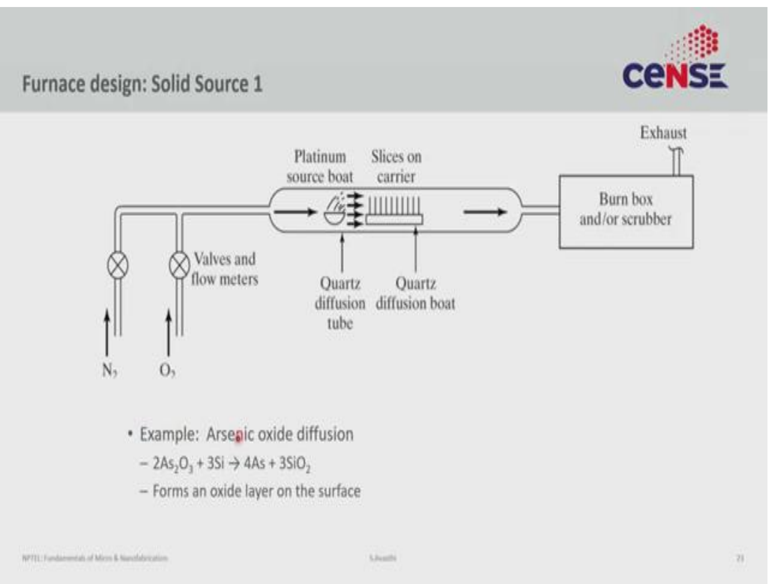
(Refer Slide Time: 26:00)



What are the various sources and equipment that we use for diffusion doping? A diffusion tube is an ultra-clean tube with no impurities inside. After cleaning, you load your wafers

on top of an ultra-clean quartz carrier. Nothing here can be metal. Then you flow some gas that has the dopant. A common dopant gas is a diborane for boron, arsine, phosphine for arsenic and phosphorous. These hydrides tend to be very poisonous and dangerous, so, often not used in practical commercial applications. You can have solid precursor-like boron nitride for boron, arsenic oxide or phosphorus oxide. You can also use liquids like boron halides, arsenic halides or $POCl_3$ that have a higher vapor pressure. Bubble through some inert gas and the liquid forms vapors that are transported to your wafer.

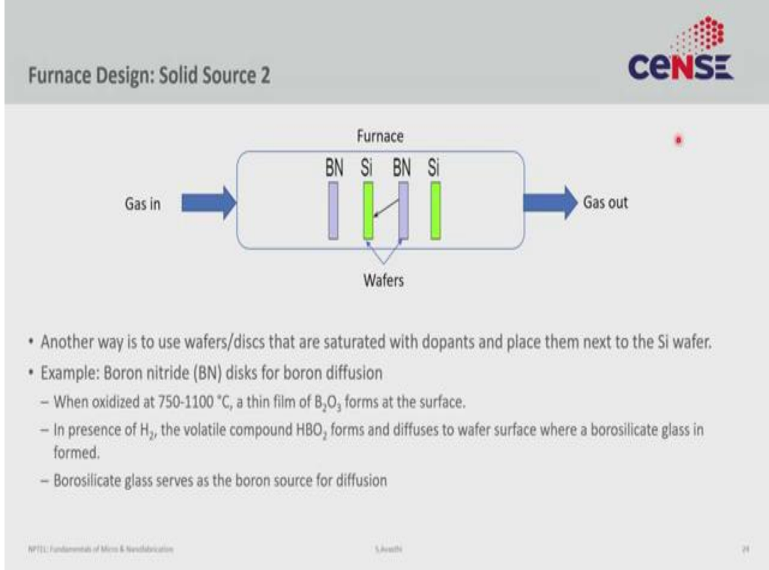
(Refer Slide Time: 27:19)



This is the oldest, probably the first way in which diffusion was ever done. Load silicon wafer, keep a little vial or a boat of the material that you want to dope with say arsenic oxide and then you heat the whole system. This material will go from the boat and get deposited on top of the wafer and then diffuse in the wafer. This is a very old method and does not lead to very uniform diffusion because the nearby wafer sees more dopant than the farther wafers. This is not usually used anymore.

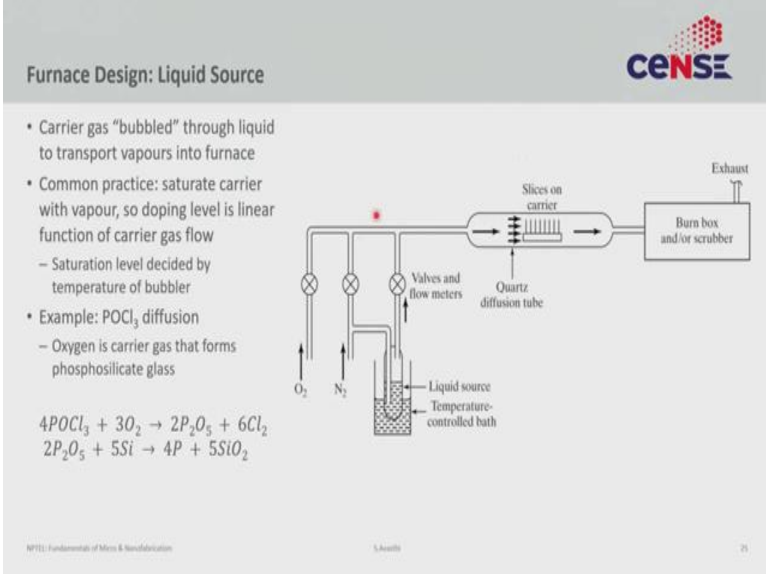
If you are dealing with very toxic things like arsenic, you must have a burn box or a scrubber to remove the arsenic before you let the exhaust out into the open. Any material that is coming out must be treated do you don't contaminate the environment.

(Refer Slide Time: 28:15)



More popular and a more modern method of solid-state diffusion is using discs or wafers. In between two silicon wafers, add a boron nitride wafer and if you heat the boron nitride wafer in oxygen for a certain time, it will form boron oxide that can dope silicon. So, this is some sort of proximity doping that allows much better control and repeatability.

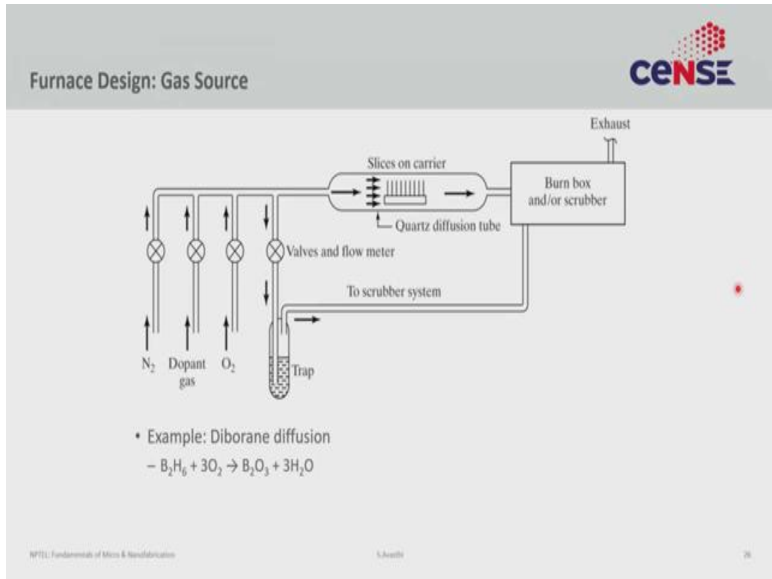
(Refer Slide Time: 28:47)



A lot of liquid sources also very popular for example, silicon solar cells often use liquid dopant. Inert gas is bubbled through a bath of $POCl_3$ and as it goes through the vial, it picks up the $POCl_3$ vapors and gets saturated with those and then those vapors get

deposited inside the chamber where $POCl_3$ reacts with oxygen, forming phosphorus oxide that reacts with silicon to form a phosphosilicate glass. This phosphosilicate glass has the phosphorus that gives the doping. This is very commonly used, but you are forming things like chlorine that are dangerous. In general, you want to have a scrubber and a burn box to capture those before you release the gas into the atmosphere.

(Refer Slide Time: 29:40)



Much cleaner, but a little more hazardous method is to use gas sources, for example, diborane reacts with the oxygen to form boron oxide. Everything here is in the vapor phase. The boron oxide gets deposited on top of the wafer and you get your doping. Once again, you need a burn box or a scrubber to clean all the things that you do not want to dump into the atmosphere.

(Refer Slide Time: 30:05)

Dopant /Type	Misfit Ratio $(\frac{R_{dop}}{R_{Si}})$	Dopant sources	USP
B P-type	0.75	**Solid:** B_2O_3 **Planar disk:** BN **Liquid:** $B(OCH_3)_3$; BBr_3; BCl_3; **Gas:** B_2H_6 **Spin-on:** Carborane derivative	1. Most usable p-type source 2. High rate of diffusion
P N-type	0.93	**Solid:** $NH_4H_2PO_4$; $(NH_4)_2HPO_4$ **Planar disk:** silicon pyrophosphate; LnP_5O_{12} **Liquid:** PBr_3; $POCl_3$ **Gas:** PH_3 **Spin-on:** triphenylphophate	1. Most used n-type source 2. High rate of diffusion
As N-type	~1.0	**Solid:** As_2O_3 **Planar disk:** $AlAsO_4$ **Gas:** AsH_3 **Spin-on:** arsenosiloxanes	1. No stress → no dislocation defects 2. Very low diffusion → shallow junction depth
Sb N-type	1.15	**Solid:** Sb_2O_3; Sb_2O_4 **Liquid:** Sb_3Cl_5	1. Very low diffusion → shallow junction depth 2. Very ideal diffusion characteristics

So, here is an overview of the various types of doping. You can do N-type, P-type using boron and phosphorus; arsenic and antimony for N-type doping in specific cases, but these two tend to be the most common. If you want to understand some of their properties, you have to understand the misfit ratio which is the difference in their size. Boron and phosphorus are smaller than silicon, so, they have a misfit ratio < 1. Arsenic is nearly matched in terms of size with silicon and antimony is larger.

What are the sources for boron? boron oxide, boron nitride, you can have some liquids like halides, some metal-organic liquids, and some gases. There are also some spin-on sources. You can coat these viscous fluids on top of your wafer and then load the wafer. With phosphorus, you can have solid sources like oxide, liquid sources that we already discussed, phosphine and also some spin on derivatives. The USP of boron is it is the most usable of P-type. If you look in this list, boron is the only one that I have listed as a P-type. All other P-type sources such as aluminum or gallium have their own problems.

Because boron is one of the smallest atoms, it is also one of the fastest diffusing one. So, you have to be a little careful if you are trying to make shallow junctions as boron will diffuse very quickly. Phosphorus, the most used N-type dopant is also smaller than silicon and has a higher rate of diffusion. The great advantage of arsenic is that because it has virtually the same size, it does not cause any stress in the film. Remember that substitutional impurity smaller or larger in size can cause compressive or tensile stress.

The fact that arsenic causes neither is often useful in semiconductor fabrication because it does not cause dislocations. However, because it diffuses substitutionally, the diffusion rates are low which can be an advantage in the modern fabrication where shallow junctions are preferred and arsenic gives very shallow junctions. Antimony has even lower diffusion; it gives very shallow junctions and ideal diffusion characteristics. We will talk about some diffusion non-idealities in the next lecture, antimony does not have any of them.

So, with that, we come to the end of this lecture. In the next lecture, we will look at some advanced diffusion processes and some non-idealities.

Thank you.