JUDY HOYT: We're going to begin this lecture on handout number 14. We'll be moving now to chapter 7. This will be our first lecture on chapter 7 on the topic of dopant diffusion and profile measurement. So far, we've discussed a number of major topics, including the fabrication of wafers themselves and cleaning, point defects in silicon. And the last couple of lectures, we've been talking about the details of silicon thermal oxidation, including two-dimensional and stress effects.

During the next few lectures, including this one, we're going to discuss the accurate control and placement of active dopant regions through a process called dopant diffusion. Today I'm going to give an overall introduction to diffusion in silicon from chapter 7. So let's go on to slide number 2. And here I'd like to give an introduction to the basic concepts of why we care about the details of diffusion in silicon.

And what's being shown here is obviously a silicon MOSFET. We see the source, drain, and gate regions. And each one of these regions has a certain resistance associated with it. And that resistance turns out to be dramatically dependent on the placement of the atoms themselves and of the doped regions. And not only that, but the placement of those regions determines many of the so-called short channel characteristics of MOSFETs that we'll talk about.

So as the device shrinks by some scale factor, say k-- we make some dimensions smaller-- the junction depths should also scale by k to maintain the same electric field patterns in the lateral and vertical dimensions. And so that's an important region the reason we need to control the doping profiles. And finally, the doping of other materials, not just the silicon itself, but of the polysilicon gate affects things like gate depletion and limits how well the gate voltage controls the channel potential. So we really need to understand the placement of these atoms.

Let's go on to slide 3. Since we just talked about the idea of the resistance of these different regions, I just want to remind people what the general form for how one calculates the resistance of either a bar or a sheet. We know that the resistance in ohms can be calculated by the product of rho, the resistivity of the material, which has the units of ohm centimeters, times the length of the resistor bar divided by the cross-sectional area through which the current flows. So the resistance in rho L over a.

Shown on the left is a picture of a cube. So let's say we have a cube, and we can calculate its resistance as just being the resistive times the length of the cube divided by this cross-sectional area. This resistivity, where the resistivity of the cube is given by, essentially, the electric field divided by the current density. So now, if we look on the right, and instead of having a cube-- in semiconductors or in silicon, we typically don't have a cube or a chunk of material-- we're usually measuring the resistance of a thin sheet in the near surface region.

And that sheet generally has dimensions of L length and width that are much larger than its thickness. So if we have a shallow junction, as pictured on the right, we can calculate its resistance as follows, again, using the same formula at the top of the page. It's the resistivity in ohm centimeters times the length now divided by the area. But now we're assuming here that we have a square sheet so that the length and the width of the square are equal-- in this case, equal to W.

So in that case, the W's cancel out because we have the length equal to W's. And when we divide by the cross-sectional area through which the current flows, it's equal to W times xj. The W's cancel out. So we end up at the resistance. The sheet resistance is just given by rho over the resistivity over xj. So that's a simple way and a convenient way of calculating resistance of various structures in semiconductor devices.

So let's go on to slide number 4. What we just talked about was for a sheet that was uniformly doped to a given doping concentration. If the doping concentration throughout the sheet is non-uniformly doped in depth, then we can still calculate the sheet resistance, but we need to do an integral. It's equal to-- the sheet resistance is just rho over xj. But at this point, we need to integrate.

So we integrate 1 over the resistivity. So it's 1 over the integral of the carrier concentration N minus the background doping concentration NB-- so that's just the net doping concentration-- times the mobility, which is generally a function of the doping concentration dx. We integrate that over from 0 to xj. You can do that numerically, essentially, for an arbitrary doping profile.

The equation has actually already been integrated numerically for certain special analytic profiles, and we'll talk about those results later. So that's basically how we get from the doping profile to the electrical properties such as the sheet resistance or the resistance of a region. Experimentally, we measure the sheet resistance using a four-point probe setup-- I think we discussed this a couple of lectures ago-- or a Van der Pauw structure, as discussed in your text.

Let's go on to slide number 5. Here, again, I'm picturing that same MOSFET structure with the resistance of the different regions. And in general, as a rule of thumb, of course, we'd like the resistance of the regions that are extrinsic to the device, such as the contact resistance, the source and drain resistance, and the resistance of these extension regions, hopefully should be no more than about 10% of the channel resistance. That is, we'd like to have the intrinsic resistance of the channel dominate the overall resistance of the device because that's what the gate has control over.

So if we apply this 10% criterion, you can write the equation that's shown on slide 5 as follows. That 2 times the contact resistance-- so that's the resistance of the metal contact-- plus the resistance of the source region, as shown here, plus the resistance of the drain, plus 2 times the resistance of the source drain extensions. Remember, these shallow extensions are the shallow xj regions that attach the source and drains, essentially, to the channel region. We'd like the sum of those terms to be less than or equal to 1/10 the channel resistance.

So, in general, as we scale devices, the channel length becomes shorter and the channel resistance goes down. So we similarly need to scale down these extrinsic resistances of the source drain and the extension regions. So to reduce those resistances-- those parasitic regions-- we would like to increase, in general, the junction depth xj. However, there's a problem that if we make the junctions deeper, it will make it easier for voltages at the drain to affect the current flow in the channel because of the way the field patterns are established.

So this two-dimensional spreading of the electric field from the drain can attract carriers from the source, even when the device is supposedly in the off state. So we end up with something called drain-induced barrier lowering if the junctions are too deep. So this results in, for device designers, kind of a fundamental design tradeoff for MOSFETs that is a design trade off between the series resistance versus the DIBL or the ability of the gate to control the current and turn the device on and off.

So if we go on to slide number 6, essentially what we're saying here is that there is a major challenge that we need to keep the junctions shallow so that DIBL and short-channel effects are reduced as we scale. But at the same time, we need to keep the resistance of the source-drain region small so that we can maximize the current drive get the maximum amount of current out of the device. And these are conflicting requirements.

And you can see the effect of these conflicting requirements to a certain extent by examining this chart. This is a chart that I took out of the text. It's a little bit dated in the sense it's from the 1997 Technology Roadmap for Semiconductors. But since it's consistent with what's in your text, we'll go ahead and look at it. If you look at, say, the last three rows in the chart, you'll see the contact region junction depth contact xj from, say, the year 2000 being in the range of 50 to 100 nanometers.

And as we go smaller and smaller channel lengths out further in time, that's scaling down to shallower and shallower junction depths at the channel. So right near the channel or in the source-drain extension regions, you can see that's even much thinner, and that also scales down with time to thicknesses by 2009 on the order of 15 to 30 nanometers-- quite shallow.

And at the same time as we're scaling these junctions to maintain the good electrostatic control, the drain extension concentration-- so that's the doping concentration in the extension regions-- is going up dramatically, say, from 10 to the 19th up to 10 to the 20 or perhaps even higher. And that higher doping requirement is arising from the fact that we're making the sheet shallower.

The junction depth is smaller to compensate for that and have the resistance go down. We really need to up the dopant concentration. And there's a fundamental physical limit on how much dopant we can put in the silicon and how much it will be electrically active. So this is becoming a real problem. We need to find new ways to activate dopants to higher levels if we're going to be able to manage this design tradeoff.

So let's go on to slide 7, again, which is that same picture, to remind you. So, basically, the ITRS requirements in the future really require or dictate that we know the dopant positions in the device with almost atomic-scale accuracy in both two-dimensional and three-dimensional profiles. So being able to scale the device really amounts to, in the front-end processing to a large extent, to being able to control very precisely the shape of the doping profiles where the dopants end up.

And what I'm going to spend some time in the next few slides is giving you examples from the present literature on device scaling, which emphasize or give you some sense of how these doping profiles need to be controlled and what their impact is on device performance. So, again, perhaps you won't understand the detailed device physics, but it's just to give you a flavor for why studying dopant diffusion is such an important topic.

So let's go on to slide number 8 and talk about a topic called the short-channel effect. And this basically takes place when the distance between the source and drain-- that is the channel length L-- becomes comparable to the MOS depletion width in the vertical direction. And then that the source-drain potentials themselves from the source and drain regions end up having a strong effect on the control of the current in the device. So in words, that's a way of expressing the short-channel effect.

And what I'm showing the top equation on slide number 8 is the equation we looked at briefly-- we didn't derive. We just wrote it down in class several lectures ago-- for the threshold voltage. And this is for the threshold voltage in a MOSFET that has a constant doping in the channel-- a very simple profile. It's just constant. And it's a relatively long channel device. So you will not see this short-channel effect. The gate length L is much, much longer than the depletion widths.

And we said we can write down these three terms, roughly, to calculate the threshold voltage. And remember, the threshold voltage is an important parameter as far as the device and circuit designers are concerned. Now, when we get to the short-channel case, which is shown below, that threshold voltage equation has to be modified to a certain extent. And, in fact, the third term, which is represented on the second equation by the bulk charge QB prime over WLC ox. That term ends up being smaller than it would be in the long-channel case.

And this QB prime is smaller, and that ends up affecting-- that third term being smaller affects the Vt. Threshold voltage actually goes down. And this diagram on the lower left is an explanation that people have proposed to explain this. And it's essentially because charge sharing-- that some of the charge under the gate is balanced by the charge from the source-drain regions.

So that effectively gives you an effective channel length. The L prime is actually shorter than the actual gate length. So there's this tendency as we scale devices for the threshold voltage itself to drop or to roll off. And if you want to learn more about this, I've got a reference on the bottom. There are a number of books, but this is one by Taur and Ning where you can look at this so-called Vt roll-off effect.

If we just go on to slide number 9, you can actually see some data on threshold voltage roll-off. The upper plot is for nMOSFETs. And you can see on the vertical axis is the threshold voltage Vt as a function of channel length. And these curves-- one is for in the linear regime with a low drain-to-source bias. The triangle is for the saturation regime with a source-drain bias of 3 volts.

And you can see, indeed, the threshold voltage is dropping or rolling off going towards 0 as we decrease the channel length-- same trend for the pMOSFET. And that's an effect that needs to be controlled as we scale the devices. And the way we partly control that is by controlling the dopant profiles underneath the gate.

So if we go on to slide number 10, it's discussing something called channel doping profile engineering. Essentially, this refers to optimizing the final doping profiles in of the p-type regions in the nMOS that is under the gate or the n-type regions under the gate in the pMOS in the channel region. So this is just a schematic cartoon picture of cross-section of a MOSFET.

And this channel doping profile engineering is referring to these red regions-- both these dark red so-called halo regions that are marked here that go around the perimeter of the source-drain extensions and this lighter red retrograde well regions. These two-dimensional dopant profiles are engineered or designed to minimize these short-channel effects.

So let's go on to page 11 and give you an example on page 11 taking in depth-- going right through the center of the channel-- the profile concentration of the channel of the doping as a function of depth. And there are two different doping profiles that are shown here. One is for a uniform well-- uniformly doped phosphorus. It's reasonably constant doping around the [? 2a17 ?] at the surface.

And the other-- the blue line-- is the so-called super steep retrograde well, where you have certain well doping. And then near the surface, the profile falls off in doping very rapidly. These two profiles, as it turns out, in the channel region give very closely the same threshold voltage for a given device. But you get better leakage current control. So you can scale the device to a smaller L effective.

So, again, this is an example of how you need to control the doping profiles in order to optimize the device design. Let's go on now to slide number 12. And this is sort of an extreme case of scaling, where we're trying to scale the MOSFET gate length down to 25 nanometers dimensions. And in order to do this, if you look back in the literature, there's a paper by Taur, Wann and Frank in 1998-- the so-called the super-halo profile.

And what the super-halo profile is is a fairly sophisticated ion-implanted and then diffused profile, say, for an nMOSFET of boron doping that creates a fairly complicated two-dimensional boron doping profile. And what you're looking at here is the MOSFET. And in the central region here underneath the channel, you see these p-type doping contours. It looks sort of like butterfly shaped. There's one contour here shown that around a doping of 10 to the 19th and another contour shown at about 5 10 the 18th.

And you can see this is designed to put reasonably high boron doping against the source-drain regions to help stand off the field in this from the source-drain regions. And so that's refers to this halo design-- very sophisticated. And there's also the doping of the source and drain regions themselves designed to be quite abrupt in both the vertical and the lateral dimensions.

So what exactly are the effect of these doping profiles on electrical performance? Let's go on to slide number 13, which shows the short-channel threshold voltage roll-off and basically how the Vt varies with channel length. So on the y-axis is the threshold voltage of the device. The x-axis is the channel length in nanometers. And there's a couple of different designs here that are shown.

The dashed line with the open squares refers to the retrograde. So that's not the super halo, but a more classical super steep retrograde profile. You can see it's threshold voltage rolls off quite rapidly as we shrink the channel length below 30 nanometers. So that's not going to work. But the super-halo profile shown by the diamonds and the stars have a much flatter, nearly flat short channel and Vt roll-off characteristics.

And furthermore, the Vt roll-off is not that sensitive to the vertical junction depth, as you can see by comparing the diamonds to the stars. So this lower variation of the Vt with L effective or with channel length allows a larger design window, which we need because there's always going to be some process variations in the channel length across the wafer.

And this enables the technologists to push the channel length down to smaller dimensions. So it's not so much a fundamental improvement in device performance, but it really enables you to manufacture circuits with these shorter channel devices. And again, a lot of it boils down to controlling and detailed doping profiles in the source train and under in the channel regions.

Let's take a look at slide 14, again, is another example of the fact that not only do we care about the doping in the vertical direction, of course, but the dopant profiles for the source-grain dopants themselves in the lateral direction are very important because it's not just the junction depth, but it's the lateral gradient of the source-grain doping. So this is a plot of the threshold voltage versus channel length again. But this time, the different curves refer to different lateral source-drain gradients.

So you've got the top curve refers to a lateral gradient-- so how quickly the arsenic doping profile rolls off at about 2 nanometers per decade. We've got a curve at 4, 8 and 16. And you can see that for lateral gradients larger than about 4 nanometers per decade, the Vt roll-off is just too large. The threshold voltage is approaching 0. You wouldn't be able to make a 25 nanometer MOSFET-- so, again, illustrates the importance of controlling the lateral doping profile and of controlling diffusion processes themselves.

So given that brief introduction to the electrical effects, let me go on now on slide number 15 and talk about dopant diffusion fundamentals. So I've tried to make the point that understanding these profiles in detail is important. And that's what we're going to do in the next several lectures. So what is diffusion? Diffusion is really the redistribution of atoms from regions where they exist in high concentration to regions of low concentration.

Diffusion occurs essentially at all temperatures. But the diffusivity-- or the diffusion coefficient-- has an exponential dependence on temperature. So above a certain temperature is when the diffusion rates really become large. In silicon IC processing, there are two different steps that we refer to in diffusion historically. The first step was so-called predeposition.

And what this refers to is that you had an initial step in which the dopants were introduced into the silicon wafer with a required integrated dose into the substrate. Originally, in the early days, this predeposition step to introduce the dopants was done by diffusion of the dopants into the wafer from a doped glass or by introducing into the silicon by heating in a doped gas ambient.

The pre-dep is rarely done these days in that particular way. In the more modern technology, it's usually done by ion implantation, which is a process that we're going to discuss later and is covered in detail in chapter 8. Let's go on to slide number 16, which very pictorially illustrates the predeposition and the drive-in process.

The second process in creating a region of the wafer with a certain doping is what is typically referred to as the drive-in. This is a subsequent anneal after the pre-dep that then diffuses and redistributes the dopant, giving the required junction depth that you need to get the right resistance and giving you the right profile or surface concentration, hopefully. So, again, schematically we have two processes going on here.

The first one would be the ion implant step or the pre-dep, which would result in the bright orange or the red region. And that has introduced a controlled integrated dose of atoms per square centimeter into the silicon. And then without introducing any additional atoms but keeping the dose constant, we then diffuse in-- or drive in-- that profile. And that would then give us the final junction depth represented by the lighter orange region.

So let's go on now to slide number 17. And here I'm comparing somewhat qualitatively these two different methods of doing predeposition. On the left hand column, we're talking about some of the characteristics of doing ion implantation of the atoms. And on the right hand column we're talking about doing this more old-fashioned solid or gas phase in diffusion in order to do the pre-dep.

Now, so ion implantation-- What are some of its advantages? Which we'll see when we talk about in chapter 8. It's done at room temperature, essentially, so you can mask it with simple materials like photoresist. It gives you, probably most importantly, a very precise control of the number of atoms that are introduced per square centimeter into the substrate. It also gives you very accurate depth control. So those are key advantages.

The problems with it, of course, is as the implant process occurs at high energies, it actually damages the crystal to a certain extent. And we have to heal this by an annealing process. But unfortunately, the damage itself can enhance the diffusion rate. And we're going to spend some time in this course talking about the transient-enhanced diffusion. The dislocations or extended defects associated with the damage can lead to junction leakage, which is not desirable. And you may have some channeling of the implant. We'll talk about how that affects the profiles in chapter 8.

The advantages of the pre-dep by gas phase is there's no damage created by this process. And it can be done in batches. But it has some serious limitations. Usually, you are limited to introducing the dopant at the surface at a high concentration of the solid solubility. It's very hard to achieve low surface concentrations without a long drive-in step. And so it's hard to control the shape of the profile to certain types of shapes. And low-dose predepositions are very difficult, and that's a major limiter. So, as I said, except in very special cases, people typically use ion implantation for the predeposition step.

Let's go on to slide number 18. Again, if we are talking about predeposition, in that case the dopants are typically introduced at their solid solubility limit. So you have some atmosphere of gas, say, that you introduce the wafers at high temperature into this gas atmosphere. And at the surface, you would end up with the solid solubility of that dopant. And just to give you an example here of what some of the solubilities are showing in this plot, solid solubility as a function of temperature.

And say, if you are doping something with boron, for example, and you're heating away for up to 1,000 degrees, the solid solubility is somewhere in the range of 2 to 3 times 10 to the 20. So they are soluble in bulk silicon up to that value. And above that, presumably, they start to precipitate out into another phase. So it gives you an idea of the surface concentration you might get of these dopants if you did a predeposition at that temperature.

Now, if you go on to slide 19, it turns out there's a subtle difference. And a point we want to make in this course is that dopants also have what's called an electrical solubility that is different from the solid solubility that we defined according to precipitation. The electrical solubility refers to the maximum doping concentration in terms of electron density per cubic centimeter that you can achieve with that dopant.

And that generally varies with the temperature, as we see here. This is a plot for the maximum electron concentration doping you can achieve using arsenic in silicon as a function of temperature. And there's a lot of different points on this curve from different measurements in the literature. But they generally follow this roughly this straight line. And so, again, what this is saying is that at 1,000 degrees, if you look at the curve, you can get something like 3 to 4 times 10 to the 20 electrons per cubic centimeter by introducing arsenic into the lattice.

If you introduce more arsenic than that, it may still be below the solid solubility, but you won't get any more electrons. It's not electrically active. It may not precipitate until you get up into the 10 to 21 range. But so there's this intermediate range, say, at 1,000 degrees between 310 to the 20 and 10 to the 21 in which you may not see silicon-arsenic precipitates, but you do not have any higher electron concentration.

That turned out to be a bit of a mystery to people for a number of years. They couldn't see precipitates, but they knew they wouldn't weren't getting the electron concentration above a certain number. And it was subject to a lot of discussions in the literature. In fact, if you go on to slide number 20-- I took this from your text. People eventually came up with a number of models to try to explain how it is you can get more arsenic in the lattice if it does not precipitate. And yet it doesn't contribute electrons to the doping in this range.

And here on the left, I'm imagining arsenic in the lattice-- the pink atom in the silicon lattice-- a relatively lightly doped sample, say, in the 10 to the 20 range or so-- heavily doped, but not too high. And what you see is arsenic generally surrounded by four silicon atoms. And it donates its extra fifth electron that is not conveniently bonded to the conduction band, and you can get a free electron.

Now, on the right side is shown a hypothetical case where you have, say, a lot of arsenic on the lattice. So you might have 8 times 10 to the 20 or 10 to the 21 or something in which people still didn't see a second phase. They didn't see precipitation happening. What you might have, say, four arsenic atoms near each other-- these four arsenic atoms surrounding, say, a vacancy-- a silicon vacancy.

So this AS 4 and V is a complex that people have hypothesized which would enable you to introduce four arsenic atoms in the vicinity of the vacancy. And yet no electrons will be donated-- no free electrons. So these four arsenic atoms would be essentially electrically inactive, and yet essentially on substitutional lattice sites at high arsenic concentration. So this might be one way of accounting for the fact that the electrical solubility is a little lower than the solid solubility.

So, again, we just point this out for a couple of reasons, mainly because it points to the fact that as we increase doping, we don't always get an increase in the electron concentration, and therefore a decrease in resistance. OK, so let's go on to slide number 21. And we're going to consider macroscopic first-- macroscopic models for diffusion. Later on, we'll talk about the more atomistic diffusion mechanisms and effects.

And hopefully, you've read part of chapter 7, and you know that macroscopic dopant diffusion is described by Fick's first law, which describes how the flux or the flow of dopant depends upon the doping gradient. And I'm showing here a schematic sketch of concentration as a function of distance. And one of these curves-- the one with a higher peak concentration-- occurs at some time called t1, an earlier time. And then the second curve that has moved out a further distance corresponds to time t2.

And essentially, what we see is that the flux F-- moving in this case to the right-- is equal to minus a constant called the diffusion constant times the gradient dC by dx. So as you take the slope along that curve, wherever the slope is very steep you get a large flux, and therefore a large movement in the profile. As you get near the top of the profile, the concentration gradient is getting small, and then the flux is a little bit lower. But at the flux at any given point can be given by this equation. When the concentration gradient goes to 0, essentially, the dopant or the atoms are uniformly distributed, say, in the solid, and the flow would stop according to Fick's first law.

So let's go on to slide number 22. Again this F equals minus D the partial C by dx. This is a general sort of flux law which hopefully may be familiar to you. It's similar to Fourier's law or analogous to Fourier's law of heat conduction or to Ohm's law for current flow. In this case, the proportionality constant is called the diffusivity D. It has units of length squared per time or centimeter squared per second.

And as it turns out, we'll see that D is related to the atomic hop rate or the jump frequency over some kind of energy barrier. And this energy barrier is associated with the formation of migration energies of mobile species. D is generally exponentially activated. So it's dependent on temperature in exponential fashion. And in the silicon lattice, by symmetry D is isotropic. So it doesn't depend on which direction you're diffusing.

And of course, the negative sign in this Fick's first law indicates that the flow is down the concentration gradient. So just by drawing yourself a profile, you see that dC/dx is negative. So in order to get flux to the right, you need to have the negative sign

Let's go on to slide number 23. This illustrates a derivation of Fick's second law, which describes how the change in concentration in a small volume element is determined by the change in the fluxes into and out of that volume. So if you take a look at this volume element, which has a certain length to it-- this delta x. And there's a certain concentration change in a certain time period, delta t, in that volume element, which we'll call in delta C.

There's a flux in coming from the left phase, and there's a flux out going out the right phase. And just by a bookkeeping on the upper left, we can write a simple equation that says the change in the concentration delta C in the time period delta t is just equal to the difference in these two fluxes-- the incoming flux minus the outgoing flux-- divided by the distance, delta x.

Mathematically, instead of writing it in terms of these differences, we can write it mathematically as partial derivatives. We can write delta C by delta t as partial C by partial t. So the time rate of change of the concentration is equal to the partial of the flux divided by x. So it's the gradient of the flux. That's what we're saying with Fick's second law. And now for the flux F we can substitute in Fick's first law. It's just F is equal to D partial C by partial x.

So we can substitute that in, and we get this equation in the middle of the slide, which essentially is the differential form of Fick's second law. Now, what we do is we make the assumption at this point that the diffusivity is constant. So it doesn't depend on x. In that case, and only in that case, then we can pull the D out of the derivative. The partial by dx, if we do the chain rule, is 0.

And then we can get the equation at the bottom of the page, which says that the time rate of change of the concentration at any given point is a constant D times the second derivative of the concentration with respect to x. And that is Fick's second law. And again, this particular formulation only applies if the diffusion diffusivity is a constant. It doesn't depend on x.

So let's go on to slide 24 now. And there are a handful of cases, maybe three or so or four, in which it's possible to write down or to derive relatively simple analytic solutions to the diffusion equation. In all the other cases and most of the cases we'll end up using in this course, we'll have to do numerical solutions. And we'll talk next time about how numerical solutions work. But for now, let's look at a couple of special cases where we can solve this equation by hand.

The first case is pictured on slide 24, which is called the steady state. And what that refers to is that, in fact, we have a profile that is not varying in time, in which case we write partial C partial t equals 0. And so then we have a relatively simple equation that D is equal to the second derivative of C with respect to x equals 0. D times that is equal to 0. So we can just simply integrate this equation twice, and we end up with a linear profile over distance.

So if you just take the equation that the concentration C equals a plus bx, and we differentiate that twice, indeed we get 0. So this says that in a steady state solution to the diffusion equation is a linear equation. In fact, when we did the solution of the diffusion of the oxidant through the oxide during thermal oxidation, this is the equation that we actually use. This is implicitly the solution that we were assuming.

And I think you'll recognize-- from the last few lectures, you'll recognize on the lower pictures this steady state solution is either in the left side-- being in the thin oxide regime, we get a straight line that's just a constant number of the diffusion through the oxide. Or on the right hand side, we get, again, a linear profile of the oxygen through the oxide during thermal oxidation. And again, in that case it's a concentration profile that's not changing with time.

So let's go on to slide number 25 and do the first solution of a case that's a little more complicated than that, which is called the limited source case. So we consider that we have the dopant in this region, and it has a fixed dose Q, so a fixed number of atoms per square centimeter. And we're going to introduce it as a delta function at the origin. And then we're going to let it diffuse and diffuse out.

And as it turns out, the C-- if diffusivity is a constant, it diffuses into the shape of a Gaussian profile. And so, basically, the boundary conditions are that we have essentially a delta function at time t equals 0 and that the dose-- the integral of this delta function-- is a constant.

Let's go on to slide number 26. And we find that the solution that satisfies Fick's second law is written down by this equation. And in fact, it's a Gaussian distribution. The concentration is a function of x and time can be given by that constant dose Q divided by 2 times the square root of pi Dt times the exponential of minus x squared over 4 Dt. So that's what's known as a Gaussian profile.

And the important consequence of this are that one, of course, the dose Q remains constant. That means then that the peak concentration-- so the concentration at the origin-- is going to decrease according to the square root of Dt over time. So the peak concentration goes down. And the width of the profile or the diffusion distance from the origin is going to increase according to 2 times the square root of Dt. You see that just by looking at the argument of the exponential there.

So at this distance-- this distance equal to 2 times the square root of Dt-- the doping concentration will fall off by 1 over e. And in fact, what we do is we often call this distance-- we give it a special name. We call the diffusion length L. We typically write it as twice a square root Dt or sometimes just the square root of Dt. It gives us an idea of the width or the broadening of the profile.

So let's go on to slide number 27 which shows pictorially the time evolution of a Gaussian profile. The left hand plot is on linear y-axis and the right hand plot is on a logarithmic scale-- give you a little better view. So, first, let's look at the left. And we there are three curves shown here. The red is for time t equals some time t 0. And the y-axis is plotting the concentration in a normalized fashion.

And the x-axis is in units of diffusion distance. So its units are units of 2 times square root of Dt0. And you can see in going from the red to the blue curve, indeed the concentration has dropped by 1 over the square root of t because, basically, the blue curve is 4 times t0. So it's dropped by a factor of 2. And it's broadened. And same thing by looking at the green curve. You see the same type of phenomenon.

And on the right on the semi-log plot, you just get to see a little bit more detail. You get your eye calibrated for what a Gaussian looks like on a logarithmic scale. So you can see many orders of magnitude down below the peak what the broadening actually looks like. Because in semiconductor processing, linear scales for dopants are not all that useful because, in fact, we often care about how the dopant falls off over many, many orders of magnitude of concentration. So we typically like to use a semi-log type plot.

OK, so we have one solution for one case. Let's go on to slide number 28 and talk about the second case, which is a fixed dose Q, just like we talk about, constant in time. But now we're diffusing near a surface. Before we have the origin, and we assume the silicon was semi-infinite in both directions. Let's say we had this delta function of dose Q as the initial profile. But we're right near the surface of the silicon wafer.

Well, there's a relatively simple trick for solving this. If we can assume that there's no dopant loss through evaporation or segregation at the surface-- that the dopant is contained in the surface-- if there's evaporation, all bets are off, and we have to solve it differently. So we assume that there's no segregation or evaporation. And we also assume that the annealing takes place over a long time so that the initial profile is reasonably can be reasonably approximated by a delta function compared to the final profile.

If those two assumptions are hold, then we can essentially solve it by assuming that we have virtual diffusion-- that we have a symmetric diffusion with an imaginary delta function of equal dose Q on the left-hand side. So we can solve it by using, essentially, the prior solution, pretending that the medium is semi-infinite.

So, in fact, if we go on to slide number 27, that same the graph is shown at the top. Effectively, what this means is that we have a dose of 2Q introduced into a virtual infinite medium by symmetry so that the concentration profile is given, again, just by the same Gaussian we had last time. But instead in that formula wherever we had Q, we replace it with 2Q from the prior solution.

So the surface concentration now goes Q over the square root of pi Dt. But it's very similar to what we had last time. And so this is what the equation looks like for diffusing with a fixed dose into a surface where we have no loss from the surface. Again, it's a Gaussian profile.

So let's go on to slide number 30. And the third case, essentially, that we can solve analytically is called the case of an infinite source. And what this is essentially an infinite source of dopant which is made up of small slices, essentially each diffusing as a Gaussian. So what we look at in this plot looking at the black line-- we have a concentration C being constant everywhere x is less than the origin.

And then there's a step function at the origin. And then it's 0 everywhere greater than the origin. So we have the step function at x equals 0. And that's the initial profile in black. And what we're going to find is that the diffuse profile looks like the red. The step function gets rounded to the left of 0. And some of that dopant then has diffused into the right-hand side at x greater than 0. And it gives us that diffused profile.

So how are we going to solve for this? Well, actually, we do it by using the solution we had obtained previously and essentially by a linear superposition of solutions for each of these thin slices. So we break up this infinite source on the left-hand half plane into a series of very, very small thin slices, each of which has a certain dose. And its dose, by definition, is just the concentration C times delta x-- the width of that little thin slice.

And in fact, after some time t, we know how to write down for that little slice what the profile looks like. In fact, it's a Gaussian. So, now, if each slice then of all these slices can be added up, their Gaussians associated with their diffusion, we can then find the diffused profile. And that's exactly what we're doing here. So that equation at the bottom of slide 30 shows that the concentration in this infinite source case can be given by the sum of all those Gaussians.

So we go for the sum from i equals 1 to n, where n is some large number, of this delta xi comes from the dose-- remember, there was a Q front of the Gaussian-- times this exponential of x minus xi because we're sliding this the position of this particular thin slice along the x-axis. That exponential squared over 4 Dt. So we're summing up all these Gaussians at the bottom of slide 30.

So, in fact, analytically, or in an exact sense, the solution which satisfies Fick's second law is written down at the top of slide 31. The concentration is actually equal to concentration C prime over 2 times the quantity in square brackets 1 minus the error function of the argument x over 2 times the square root of Dt.

And we can write this as C sub s times the complementary error function of x over 2 times the square root of Dt, where the second equation and third gives you the definition of what we mean by the error function. Error function of z is just equal to 2 over the square root of pi times the integral from 0 to z of this exponential-- it's integral of a Gaussian, basically.

So the error function is the integral of the Gaussian. The complementary error function is 1 minus that. And then these error functions and complementary error functions have been tabulated. So in that sense, it can be calculated analytically. So we know that the complementary error function is what the shape of this profile can be calculated according to.

So let's look at slide number 32, which, again, the error function solutions are made up of a sum of Gaussian delta function solutions. And what you see is that here in this plot, the initial profile is shown in the dashed line in green and that you have the subsequent profiles are time t equals t0 in black, 4t0, in blue, and 9t0 in red. And that the dose beyond x equals 0 continues to increase with annealing time in this infinite source sort of solution.

So let's go on to slide 33. We can take this as another special case, in fact, by just looking at the plot in the upper part of the slide 33. What we see is that at x equals 0, the concentration is actually held fixed. So if we have a situation where we have a constant surface concentration, then, in fact, the solution to the diffusion equation is just the right-hand side of the above figure.

An example of this might be the case where we're doing diffusion from a gas ambient into the solid where the gas concentration above the solid solubility of the dopant. Then, in that case, at the surface of the silicon wafer the concentration of the dopant is fixed at the solid solubility. So it's constant. So if we take just that right-hand solution in the above figure, we can write it down from the previous solution.

Concentration is just C sub s, which is the surface concentration, which is a constant, times the complementary error function. And in fact, you can integrate this equation to find the dose on the right-hand side. And the dose is given by this integral, which can be done integrating from 0 to infinity. And you can do that. And it turns out that dose Q is equal to 2 C sub s over the square root of pi times the square root of Dt. So, again, now we see that this dose or the number of the integral of these curves on the right-hand side is increasing with time according to the square root of Dt. So we're getting a higher and higher dose into the sample.

So let's go on to slide number 34. And here we are graphically comparing on the left and the right-hand side the two different types of classical processes that we talked about in terms of their diffusion profile shapes. On the left is the predeposition case where we have, say, a constant surface concentration, assuming the pre-dep was being done by a gas phase in diffusion.

And there are two plots shown on the left. In the upper left is a plot on a linear scale. The lower left is a plot on a semi-log plot. But in either case, what you're looking at is a complementary error function. And you can see that the surface concentration is a constant normalized at C over C s equals 1. And that at the different times, you can see the twice square root of Dt is 0.1, 0.5, and 1 micron.

The area under these curves is increasing. And fact, it's increasing by this factor square root of Dt. So the longer you would do the pre-dep, the more dose you would deposit into this silicon surface. On the right hand side, we see instead their Gaussian profiles . This would be the case for a drive-in, which has a constant dose. And so you see what's happening over time at the shorter time.

We have a certain peak concentration. That peak concentration is then falling or dropping for the second profile. And the profile is broadening, and then it falls again, and the profile broadens further. So that's for Q equals a constant, integral is constant, and the left-hand side is for the surface concentration is a constant. That's just to get your eyes calibrated for complementary error function versus a Gaussian type of solution.

Let's go on to slide number 35 and talk a little bit about dopant diffusion coefficients themselves. And we're just going to talk first about what we call intrinsic dopant diffusion, which happens in the case when the dopant concentration is less than n sub i. So the semiconductor is considered to be intrinsic.

And generally, we can write these intrinsic diffusion coefficients in an Arrhenius-type relationship for the diffusion coefficient is just some constant D 0 written here times exponential minus EA over kT. And this chart, which is taken from your textbook, shows some rough numbers for what D 0 looks like in units of centimeters squared per second and what the activation energy looks like for a couple of different species and some of the pot that are dopants in silicon.

So, for example, if you look at boron, it has an activation energy here of 3.5 electron volts. And the prefactor is about 1. Thing to note is that n sub i-- the intrinsic carrier concentration-- is very large at process temperatures. So, actually, intrinsic diffusion conditions apply under many different conditions or cases. So, for example, at 1,000 degrees C, n sub i is roughly equal to about 7 times 10 to the 18th.

So this diffusion coefficient written here in this simple chart would apply anytime the concentration of that dopant at 1,000 degrees is less than about 7 10 to the 18th. You can use this constant diffusion coefficient. When we get above that-- when the doping concentration is larger than n sub i-- we'll talk about next time there's some interesting Fermi level effects that come into play as the point defect concentrations become modulated by the carrier concentrations themselves of the diffusion species.

So we'd go on to slide number 36. This is just a graphical representation. It's a plot of the diffusivity in centimeters squared per second as a function of 1,000 over T. So it's Arrhenius-type plot. In the upper y-axis, you can read the temperature if you prefer that. And what you can see for these dopants right off the bat looking at them is that there's a pretty large difference or discrepancy between so-called fast diffusers in silicon and the slower ones-- say, the fast diffusers being boron and phosphorus among the common dopants.

Boron, again, the only really available p-type dopant, is relatively fast. It can be up to a factor of 10 or 20 or 30 faster diffusion coefficient than the slower diffusers such as arsenic or antimony. So this gives you a rough idea of, when we talk about doping diffusion, what we're going to have to worry about a little bit more would be the fast diffusers-- say, boron.

The other thing I want to point out with respect to this plot on slide number 36 is that earlier versions of the text had an error in this corresponding figure. And so on the website, we had posted the errata. And you'll be able to see that to make sure that you're using, if you're reading curves off the plot-- reading values off the plot-- that you have the right values. One way to check that, of course, is to back to slide 35 and actually compute directly with a calculator the diffusion coefficients.

So let's go on to slide 37. And I'd like to summarize this introduction to dopant diffusion. We've talked about that the placement of dope regions is critical because it determines many of the characteristics of short-channel MOSFETs. That's why we spent so much time calculating in great detail dopant diffusion, as we'll do over the next three or four lectures.

Turns out there's a design tradeoff between the series resistance of a MOSFET, which means you would like to have deeper source-drains to minimize the series resistance and the short-channel effects, such as the control of Vt, would dictate that you have a shallower source-drain. So this is a fundamental tradeoff. And therefore, the channel doping profile engineering is a way of compromising that design tradeoff. And so channel doping profiles need to be controlled very accurately.

Beyond that motivation, we talked about some simple analytic solutions that the time evolution of a doping profile, if the case is simple, is governed by a fixed loss-- the so-called diffusion equation. There are a couple of cases where there are analytic solutions. We talked about the diffusion of a Gaussian profile with a fixed dose or diffusion of a complementary error function, which we apply for a constant surface concentration. And finally--