JUDY HOYT: This is-- today's lecture is going to be the final lecture on chapter 7. It's not the final lecture on diffusion. We're going to continue talking about diffusion in chapter 8, once we've talked about ion implantation. But this will finish up chapter 7. So hopefully you're finishing reading that chapter at this point. Diffusion is probably the biggest topic we cover in this course in terms of the overall complexity and the depth of the modeling.

So there's a lot of points I want to cover or just try to review from last time. The last lecture was pretty dense. There was a lot of material to cover. So what we talked about last time was that there are these so-called Fermi-level effects, or high-concentration effects, and electric field effects, and that both of those become important when the carrier concentration-- either N over P, depending on how the material's doped-- when either of those carrier concentrations is greater than NI, N sub I, at the diffusion temperature, then these high concentration effects become important.

The basic idea of the model that we have is that the Fermi level, as it moves up and down in the bandgap in response to the doping level, depending on how you have it doped, as you're doped more heavily, it tends to increase the point defect concentration as the doping increases. We know we've seen that in one of your homework problems. You had to calculate that.

And the increase in point defect concentration then increases the diffusivity of a dopant. And this leads to, or can lead to, generally more box-like shape profiles because as the dopant diffuses, as you get out to the edge of that profile, the concentration is dropping. And as the concentration of the doping is dropping, the Fermi level is moving. The total point defect concentration is going down. So you tend to get these very steep fall-offs in your profiles.

So in addition to this, this leads to this concentration-dependent diffusion. We talked about effects like oxidation-enhanced diffusion, which is abbreviated OED, oxidation retarded diffusion, ORD, and the growth or shrinkage of stacking faults. And these phenomenon can be explained by not by looking at Fick's law per se or whatever, but by looking at atomic scale diffusion models.

And in fact, what we talked about last time is that based on the impact on stacking faults, that oxidation, people believe, injects excess silicon interstitials into the bulk, while thermal nitridation, which, again, you may not be as familiar with nitridation, is basically the reaction of the silicon surface with a gas-like ammonia to create silicon nitride.

That process is believed to inject vacancies into the bulk. And based on monitoring these processes, people believe that boron and phosphorus diffuse primarily with interstitials. So they have an F sub I number in the fraction interstitial contribution to diffusion of 1. And antimony diffuses primarily with vacancies. So F sub I is pretty close to zero, maybe 0.02.

So today I wanted to review these atomic scale mechanisms because we kind of breezed through them at the end of the last lecture pretty fast. And then I want to go on to some new material and just give you a brief example, or a brief look at one example, of how the point defect gradient-- we talked about this last time-- that not only does the dopant gradient, D arsenic by DX, but the point defect gradient, for instance, the interstitial gradient, can actually drive diffusion in a way that actually looks like uphill diffusion. So it's kind of a nonFickian phenomenon.

And this leads to a phenomenon in MOSFETs that's called the-- in CMOS [INAUDIBLE] what's called the reverse short-channel effect. And I'll just introduce that. We'll talk about it in subsequent lectures in more detail. And then the second half of the lecture, we're going to talk about profile measurement techniques. How do we measure all these complex profiles? What's the best way to do it?

OK, so let's review a little bit on this atomic scale mechanism. And what I would encourage you in thinking about the atomic scale mechanisms, rather than just looking at this particular, very static lattice and imagining how atoms might move, maybe you sit down with a piece of paper, start out and draw yourself a simple two-dimensional lattice.

Draw silicon atoms everywhere except one point, which you call your vacancy. Draw a blue atom, or a pink or red or whatever, which you call your arsenic doping, or whatever it happens to be. And draw a couple of time steps for yourself. And see if you can figure out how this arsenic atom, this dopant atom, A, and the vacancy together, by staying in close proximity to each other, can move throughout the lattice. Give yourself a couple of time shots. And I think that will help understand how these things move together.

So in this vacancy-assisted model or mechanism, what we do is we say A, where is A dopant-- it could be boron, phosphorus, arsenic, antimony-- we combine that with a vacancy, which is pictured here, an absence of a silicon atom. And it forms a pair, AV. OK, so you say, well, that's fairly simple. What's unique about this?

Well, here A is a substitutional dopant atom. This pair AV is mobile. The substitutional dopant atom A is immobile. So what we're saying is that dopants in silicon do not move. They do not diffuse. This is the assumption, unless they are paired with either a vacancy or interstitial. But if they're just in the lattice and there were no vacancies around, no interstitials around, the dopant won't diffuse.

So this is a hypothesis that we're making. And we use it to basically model all the diffusion of dopants in silicon. So here's the example of being paired with a vacancy. Obviously the dopant here could hop into this vacant site. The vacancy can move up here where the dopant was. Now what happens? Well, the dopant can stay put for a moment. The vacancy can move over here to this lattice site. And a silicon atom can move there.

The vacancy can continue. It can move right here. And this silicon atom would move up there. Now I have the vacancy at this spot down here at the bottom and the dopant atom next to it. Well, now they can exchange places. And the dopant atom has moved here. So it moved from that point to that point. And how it got there was just by a couple of steps of the vacancy hanging around nearby as a pair, essentially. And that's how the dopant atom moved. Without that vacancy, it couldn't have.

Similarly, there are interstitial or interstitial-assisted mechanisms. And we write this as a chemical equation. Substitutional atom A, dopant plus silicon interstitial forms an AI pair. And it's only this AI pair that can move. And this is an example. There are a couple of examples here. One is called the kick-out mechanism where you have a pure interstitial, a silicon atom that's actually in the interstitial spaces. It's an extra atom in the lattice.

And it can come along and it can kick out that dopant atom. And then they exist as a pair. And then the dopant can kick out a silicon atom. And so they kind of move along, you can imagine this pair, always one of them-- either the dopant or the silicon atom-- being off-site, so to speak. That's one mechanism. The other is an interstitial C mechanism. It's a little bit-- the distinction is a little bit different.

The interstitial mechanism is this atom that's here in the interstitial space is presumed to be not bonded, not covalently bonded in any way. In the interstitial C mechanism, the idea is that a dopant and a silicon atom are kind of both bonded and both sharing partial bonds. They're sharing a substitutional site, if you want. There's an extra atom there because there's only supposed to be one atom at that site in the lattice. There's two of them. One of them happens to be silicon. One happens to be a dopant.

And they can go along sharing. This can share, and then the dopant atom can then share with its neighbor and could diffuse perhaps along bond directions. So this is an interstitial or interstitial C assisted mechanism. Again, unless it's paired with one of the point defects, we assume that a substitutional dopant atom otherwise cannot move.

OK. So last time we also talked about this picture. I've shown it several times now, but hopefully you start to become comfortable with it. What this exemplifies is that how we use a process going on at the surface in this field to probe the atomic scale mechanisms. The problem is nobody can really see interstitials. They diffuse too fast. There's no real way of looking at them with a microscope or something like that. That's a problem.

So what we do more often is subject the surface to some kind of a process that we feel either injects interstitials or injects vacancies. And then we see what happens somewhere at a distance in the crystal to infer, or to probe, what we think is going on. Admittedly, it's indirect. But if you build up enough evidence, people believe they can make a case that this is what's probably going on.

So there are two service processes that are very commonly used. First one is oxidation. And it's usually done locally. In order to do an OED experiment, you typically don't put one wafer in a furnace and oxidize it and put a second wafer in a furnace and have it under inert because first of all, the furnace may not be the exact same temperature every run. We know they're not perfect. Temperature may be slightly different. That'll screw up your whole experiment.

So typically, instead you start with a single wafer to do an OED experiment, and you coat part of the wafer-- maybe one dye area or one small area of the chip. You coat it with silicon nitride so it won't oxidize, which is a good barrier to oxidation. And you have down here a buried layer, some distance-- maybe half micron or so. Could be below the surface. A marker layer of some dopant that's going to diffuse.

And you're going to measure the extent that it diffuses. So the dark green here is meant to represent the dopant when it starts out at the beginning of the oxidation. The light green is the edge or the junction as it appears after some time of oxidation. And you can see underneath the region where you were doing locals, where you were locally oxidizing the surface, you get a large spread of the boron layer. And this light green layer is very wide.

Underneath the inert surface-- this is called inert because there is no oxidation taking place-- the boron doesn't diffuse nearly as much. So this is inert diffusion. This is OED, oxidation enhanced, diffusion. And also, schematically, in this cartoon, underneath the area where you're oxidizing, the stacking faults, these defects, are growing because they're adding silicon interstitials to the ends of these defects. And here they're just staying put or maybe even shrinking, but they're certainly not growing.

So the key point is that people see the enhancement of boron diffusion, the oxidation enhancement, under the same conditions that caused stacking faults to grow. And stacking fault growth is interpreted to be injection of interstitials because people from materials point of view believe that the way you make the stacking fault grow is to add interstitials. So that's sort of the key piece of evidence that links. It's sort of like a criminal case. You really can never find the fingerprints, but you have some circumstantial evidence, and you put it all together, and you convict the person on trial. It's kind of like that.

The circumstantial evidence-- and it's reasonably strong-- is that stacking faults are growing when you're oxidizing. Boron is also being enhanced. Aha. Well, interstitials cause stacking faults to grow. Therefore, the interstitials are probably causing boron to diffuse more rapidly. That's kind of how the logic goes.

Similarly, people have found that nitridation-- so that's the reaction of silicon with ammonia-- has exactly the opposite effect. The boron actually diffuses slower. It's retarded, as we say in its diffusion. And the stacking faults actually shrink relative to inert anneals. So people feel that nitridation injects vacancies. And that's how those simple arguments go.

So what we say when we want to get to an atomistic level modeling-- and this is the type of model that when you're modeling the diffusion in supreme, we say that the dopants diffuse with a fraction, F sub I. So F sub I goes from 0 to 1 of interstitial type mechanism. And a fraction F sub V, which is 1 minus FI, of vacancy type mechanism. So we break down its diffusion coefficient into separate terms.

There's one term, which is proportional to F sub I times CI over CI star. So again, F sub I, it goes from 0 to 1. If you believe the dopant only diffuses with interstitials and vacancies are insignificant, you would make F sub I equal to 1, like the case of boron or phosphorus. And then this term becomes negligible. If you believe that it only diffuses with vacancies like antimony, you'd make F sub V close to 1. If you think it's both, well, then you apportion it accordingly.

And how do you figure out F sub I and F sub V, those proportions? Well, you subject the dopant in the silicon to different circumstances of injecting vacancies or injecting interstitials and seeing how much the diffusivity goes up or down under those conditions. And that's what people have done over the years to try to nail down F sub I and F sub V.

So the DA here is the effective diffusivity. Now, it's being measured under any condition, but particularly under conditions where the point defect populations are disturbed or perturbed. So this would be during oxidation or nitridation. DA star. What does that mean? That's the normal equilibrium diffusivity of the dopant, measured under inert conditions. So again, when we're talking about diffusion and we're talking about inert, what we mean is no oxidation, no nitridation.

Inert means nothing that injects excess point defects into the bulk. So by using this equation, you can understand how to model OED. What happens in OED is this term, CI over CI star. The interstitial equilibrium-- the interstitial population relative to what it is in equilibrium goes up by some fraction, some amount. Could be five, ten times higher. That causes the diffusivity to be enhanced by 5 to 10 times.

In addition, because of recombination-- interstitials and vacancies can recombine-- if you pump up the interstitials, the vacancy is going to go down. So this term can actually go down to a certain extent. And that's how the model accounts for OED or for nitridation retardant diffusion because nitrogen does just the opposite. It enhances the CV over CV star and suppresses this term.

And depending on your dopant, your FI and your FV values, it will have more or less of an effect on diffusivity. And that gets modeled in the simulations. I actually pulled out of the literature some data just to show you some of the classic experiments that took place quite some time ago on people trying to determine F sub I and F sub V.

And this is a paper by-- it's a review article. It's quite long. It's many pages by Fahey, Griffin, and Plummer. It's getting older. It was published back in 1989, but that was sort of at the height of people really coalescing all this data on these atomic scale mechanisms. So it's a good review article if you want to understand in more detail than is given in your text about this. Since that time, of course, there have been new discoveries, and we'll talk about those newer discoveries when we talk about TED.

But what people did was-- this is particular data I took from Fahey's article. This is an experiment on nitridation. So you're flowing ammonia at very high temperatures. Ammonia doesn't react at as low temperatures as oxygen does with the surface. So you kind of have to go to pretty high temperatures. That's one disadvantage of nitridation.

So you typically see nitrogen experiments-- 900, 1,000, 1,100-- in order to get reasonable injection. So under nitridation, 1,100 degrees c and there are three different dopants studied here-- phosphorus, arsenic, and antimony. And this axis here shows the time averaged because they do diffusion experiments for a certain amount of time. Here you could do it for half an hour or five hours, 10 hours, 20 hours. And you take the time averaged enhancement, or time averaged diffusivity, that you get averaged over that time interval.

And you divide it by DA star, so the equilibrium, the diffusion coefficient measured right next to it in the stripe right next to it in the inert case. So if I go a couple slides back to slide 3, so the DA star they measured right over here by measuring the amount of diffusion underneath this region, the inert region, and the DA time average they measured over here under either oxidation or nitridation. This particular one is nitridation.

And so if you look at these dopants-- look at antimony here. What you see pretty much at all times, DA over DA star is enhanced under nitridation. So antimony has a strong dependence on the excess vacancy population. And that's how people came to eventually give antimony an F sub V number that's close to 1. Arsenic has some effect, but not very much.

What happens in the case of phosphorus? Actually, we do nitridation and we inject vacancies. Phosphorus diffusion over time is actually slowing down. It's actually retarded? Now, how can that be? Well, let's say phosphorus doesn't diffuse by vacancy mechanism, but it diffuses almost entirely by interstitials. As I inject vacancies you say, well, then it shouldn't have any effect, but it does via recombination. Excess vacancies recombine and they reduce the interstitial population lower than it would be in equilibrium.

And that lower interstitial population lowers the diffusion coefficient. And so for phosphorus, this gives us a hint that F sub I is probably pretty close to one. Then you can do the analogous experiment with these same dopants with oxidation and see how it reacts during injection of interstitials. And between those two experiments, you try to get an estimate of what this F sub I and F sub V value could be. So those are some classic experiments people have done.

And in fact, also on the next slide, slide 6, I've taken a table from that paper that just summarizes at that time what was known at the state of the art of the interface processes at the surface and how they affect the diffusion of different dopants under different conditions, and also how they affect stacking faults.

There are three columns here. The first and the last, we've already talked about. Oxidation injects interstitials, causes vacancies, population to go down, stacking faults grow. The right column in nitridation, interstitial population goes down, vacancies go up, and stacking faults shrink. And for either one of these columns, you can see what happens, say, to phosphorus and boron diffusion.

Now he's sort of broken this out to intrinsic diffusion when n is less than or equal to NI, or extrinsic when it's higher. In either case, when you do oxidation, phosphorus and boron are enhanced. Antimony is a little tricky, actually. There's a little bit of enhancement initially. But overall, the effect is believed to be retarded or slowed down.

Arsenic is a tough one, because again, its F sub I value is going to be close to half, it turns out. It can be enhanced to a certain extent, depending on if it's intrinsic or extrinsic, it can be retarded, so it's a little bit tricky under oxidation. Arsenic also is enhanced under nitrogen. So it's tough. That's why we believe that F sub I and F sub v are somewhat equally weighted, depending on the amount of enhancement. It's enhanced in both cases.

Boron's easier to understand because it's pure. It's enhanced under oxidation, but it's retarded under nitridation. So people have concluded based on a lot of data that the F sub I is close to 1. Now, there's one other column in the middle that we didn't talk about in this class, and I don't think it's mentioned much in your text.

But Fahey talks about it in this article. It's called oxynitridation. It's a little bit trickier. Oxynitridation refers to the fact that if you start with a thin oxide-- so you have a thin oxide on the surface, and then you nitride that. You subject it to a high-temperature ammonia. You are doing something called [INAUDIBLE]. You're growing something called oxynitride. It's not pure nitride. It's not pure silicon dioxide. It's kind of got both silicon oxygen and nitrogen.

And so people believe-- because stacking faults, grow people believe that interstitials are injected and that vacancy population goes down. But it's another marker, another process people use. It's a little bit harder, in some ways, to interpret. So that's a summary of some of their classic data. If I go on to slide 7, now I want to talk about what I think are some very clever experiments.

What we've talked about so far when I've been showing this cartoon, I've been showing talking about one-dimensional. So I've been assuming I'm very far from the interface between the neutral or the inert ambient and the reactive surface. So far I haven't talked about what happens near the interface. But you know as you're injecting the interstitials here due to the oxidation, they don't just diffuse straight down. They diffuse out in a two-dimensional fashion.

They diffuse down, sideways, over to the edge. So they're actually diffusing in different directions. Now, if you're far from this edge, you don't see much effect of the interstitial. They primarily go down. But if you're near the edge here, you're going to see some edge effects. And in fact, depending on how far from this stripe the enhancement of the oxidation occurs will give you some idea of how rapidly the interstitials are diffusing in this lateral direction.

But there's another process besides just the fact that the interstitials are diffusing around. So they're going vertically and laterally. One other process is recombination that we have to consider. So in fact, the interstitial flux into the surface is the difference between the generation rate. There's a certain number of interstitials being generated per unit time. That's G. There's a certain number of interstitials being recombined at the surface per unit time. That's R.

The net flux of interstitials injected is-- you can think of the net flux is if G and R are fluxes, it's G minus R. OK, now, the problem is-- and we know that the generation rate here is proportional to the oxidation rate. But the question is, in your mind, the recombination rate at this point here at this interface is not necessarily the same-- at the active oxidizing interface is not necessarily the same as the recombination rate at the inert interface.

For one thing, this interface is changing. It's continuously reacting and oxidizing. So it's not clear that R should be the same here as it is here in the inert. And in fact, it's not necessarily the same. Depending on the reaction, it can be quite different. So when in supreme, you'll see various parameters. There'll be a diffusivity for a silicon interstitial. You need to know that.

There'll also be a recombination rate. Sometimes they call it K sub S. There'll be a recombination constant associated with this oxidizing interface. And there'll be another recombination constant associated with the inert interface. And they'll also be bulk recombination.

We need to know those three parameters if we're really going to understand what this lateral extent of the oxidation-enhanced diffusion looks like because after all, if I'm getting a lot of surface recombination over here, interstitials injected, basically, they all get sucked into that surface and recombine. So the net excess interstitial population will fall off more rapidly.

And here's an example on the right of a test structure that people came up with to try to study two-dimensional effects. And I think it's kind of a neat test structure. This is in cross-section now, so it's a little tricky. But I'm looking at three different cross-sections. So this is the surface of the sample. This is a phosphorous junction that was initially diffused in. So it has a starting junction depth, say of half a micron, whatever. Phosphorous diffused into lightly doped boron.

And they put stripes on the sample. And the stripes are masked such that the open region where they cut away the nitride, these open regions, these open stripes, are getting smaller and smaller as we go from left to right. So here's a pretty wide open region, narrower, narrower, until it gets to the point where it's a very narrow open region.

So the nice thing is, what you're doing is you're sort of changing the region over which you inject these point defects. And then they did the oxidation. They did the oxidation of phosphorus. And what they see is the junction depth now, which initially was flat because it was diffused originally before they put the stripes down, after they put the stripes down and they do oxidation, right underneath where the oxidation takes place, of course you see a big enhancement of the phosphorus. That's OED.

But interestingly, look at the shape as you go to different opening widths of these open stripes. The overall junction depth kind of reaches the same point here, regardless of the width of the open area. It's changing a little bit in shape. But nevertheless, it reaches the same junction depth.

So what that's saying is that at this point down here, the interstitial concentration or supersaturation at the little tip here is about the same as it is here. It's about the same as it here, as it is here. And so it's not that much perturbed by the presence of, all around it, these interfaces, these inert interfaces. So the recombination probably at the oxidizing interface and at the inert interface, those rates are probably fairly comparable when you're talking about interstitials.

But let's look at the other case. Here's another example. They did an experiment like that, but instead they started with an antimony junction. So this initial starting junction depth looks similar, but it was antimony. And what do they see as they decrease the opening width? In fact, the overall junction depth, even right below the opening, directly below it, actually goes down.

This is for nitridation of antimony. Unlike here, where these tips all stay-- the end of the junction was the same regardless of stripe width. Here it's actually going down. So what's that saying? That's a two-dimensional effect. That's actually saying that directly below this opening where the nitridation is taking place, locally the supersaturation of vacancies is actually smaller than below an opening that's much wider. So these supersaturation of vacancies must be being impacted by recombination that's taking place on either side of that opening.

And in fact, people believe that the vacancy recombination rate here at these inert interfaces is a lot faster than it is at the nitriding interface. And so the vacancy supersaturation level is actually impacted at the center point by what's happening all around it. So it's a two-dimensional diffusion and recombination problem. And there are a number of parameters. There's the diffusivity of the point defect. There's recombination at this surface. There's recombination at the reactant surface. And there's recombination in the bowl.

So it's kind of a neat way of looking at things, even with a one-dimensional test structure-- I mean, all we have is one dimension here that we can measure, essentially, the junction depth. But we can get two-dimensional information by changing the stripe, period.

So actually, if you go on to slide number 8, I also took this from that same paper by Fahey, Griffin, and Plummer. And this is a little more quantitative description where they've actually done diffusion modeling of this two-dimensional diffusion problem, two-dimensional diffusion of the interstitials being injected and recombining, and what effect they would have on a dopant diffusion.

So it's a little bit tricky here, but what you're doing is a local oxidation. Out here on the wings there's nitride, which is hatched region. Underneath the nitride that's inert. That's an inert interface. So that's masked. Now what they did was they're looking at simulations for case A and B. And case A is when the mask opening is fairly wide. So the mask opening in case A goes from here to here. Case B is you imagine the opening in the mask to be very small. So the region that's being oxidized is much narrower.

And they did the simulations for the two cases, A and B. So A is the case, these two curves for a wide opening, and B is for a narrow opening. But there are two different types of curves here. There's the solid and the dashed. In the solid curve, what they assumed in their simulations is that they assumed the case of S value or the point defects were recombining more slowly at the inert interface compared to the oxidizing interface. So they're saying there's not that much recombination over here at the inert interface compared to the oxidizing.

When they do that, when they adjust those surface recombination velocities in that way, what they see is when you change the stripe width from narrow to wide, the opening, you get profiles that look like the solid lines where at the very center of the stripe, the junction depths are almost the same, which is the case, if we go back, just go one slide back to slide 7, it almost looks like this phosphorus case in oxidation.

The junction depths, regardless of the width of the opening, were about the same. And so you can actually fit this shape, this shape right here, to the experimentally measured shape by changing the ratio of the surface recombination velocity at this oxidizing interface to in inert. In the dashed lines, what did they assume? Well, the dashed lines show the case where the surface recombination velocity of interstitials at this oxidizing interface is about the same as the non-oxidizing. So they made them equally.

So if it recombines equally in this interface versus that interface-- in fact, what you see for a narrow stripe is that the overall junction depth is much lower than it is for a wide stripe. So the enhancement is much less. So actually, people use the shapes of these junctions as a function of stripe width to say something to infer about the recombination velocities at this interface versus at the inert interface, just by changing the duty cycle.

So this is how, in supreme 4, if you look at some of those coefficients, this is how they were actually measured. It's kind of a clever experiment. Slide 9 is actually showing you some experimental data, again, just to give you a feel maybe from a different vantage point. It's a little bit hard to see this. But this is a photograph from a microscope. And you got to get used to seeing this. It's been beveled and stained.

So this surface up here that you're looking at is the top surface of the chip. This surface down here is the bevel. So if, actually, I could find a piece of chalk-- there happens to be one here, which is kind of rare. So what they do is, this is the top surface. This region here has been beveled at some very shallow angle. And so the top surface here has the stripes on it. So in the microscope, if you look at the top surface, it looks like this. So these are the regions here where the oxidation-enhanced diffusion took place.

Here's a 50-micron opening, 25, 20, 10, all the way down to 4-micron opening. This phosphorus, this region, it turns out you can chemically etch the silicon surface, and it stains. So you get a different surface appearance under the microscope where it's N type. And so you use. This is a really old-fashioned way of doing it. But beveling and staining the silicon was a way to measure the junction depth because you could actually measure it-- basically by doing this at a shallow enough angle, you can spread this out over a long distance, a distance such that you can see it in an optical microscope.

So this was just to give you an idea of what the junction actually looks like. Under a wide opening, looks like this. Under a narrow opening at 4 microns, it's about the same depth. So this was experimental evidence that people used and then fit the supreme profiles. We go back one shape, they fit to that shape. And they found that the model looked more like the solid line model. So they could actually fit the recombination coefficients based on some of this data.

So it was a relatively-- it was a clever experiment, relatively simple techniques-- oxidizing, patterning, and beveling and staining were used in some of these original experiments, then at high temperatures. OK, so that's to give you an idea of how some of these experiments were done originally.

Now, let's talk about-- I want to talk about a specific example that we kind of whizzed through last time when we were doing towards the end of the class notes. Let's talk about boron diffusion. And we're going to say that boron diffuses based on all the data that people have with only two point defects. Of all the point defects it could have as pairs, these are the two that it prefers.

It prefers a neutrally charged interstitial, silicon interstitial, which we write as I0. And it pairs with that. And so we have a B, a boron minus. Again, that boron substitutional in the lattice is an acceptor. So it has a net negative charge when it's substitutional lattice. It combines with a neutral interstitial and forms a pair, DI super minus. Minus because the net charge on that pair, if you were consider it as a pair, is one negative net electron charge.

So it likes to do that. Or boron might like to pair with a positively charged interstitial, particularly because the concentration of these interstitials tends to go up in high doping concentrations. But even in low doping concentrations, you can see there might be some kind of coulombic interaction. A boron minus might pair with a positively charged interstitial and form a BI pair.

The pair charge of this particular pair now is actually neutral as a pair. So we have two of these simple chemical reactions. And they give rise to fluxes of mobile species. So remember originally, on the left-hand side, the boron is immobile. One of these guys comes along, pairs with it. Now it's mobile on the right-hand side.

And we sum these two, the flux of these two. And we say the total flux of boron in the sample is going to be the flux of the BI minus pair and the BI neutral, uncharged, pair. So that's an example of how supreme might consider this, at least from a chemical equation point of view. So given that, those simple two equations, and all that we know about deviations from Fick's law-- I showed this last time, but again, I think we went through it a little too quickly on slide 11.

I just want to show, again, the overall equation that supreme 4 is solving for boron, assuming it's diffusing with just two species with neutral interstitials and with positive interstitials. This is the actual flux, so-called diffusion equation, that supreme is solving. And it's a far cry from what you would think in a simple case. You would think it would just be DC boron, or partial T, if you want, is equal to a diffusivity times partial concentration of boron with respect to X.

This is what you might think. That's the simple version of boron diffusion when we started this chapter 7. It's just regular old Gaussian type diffusion. That equation looks pretty simple compared to that one up on slide 11. And this is actually what supreme is solving. So where do all these terms come from? Rather than deriving it, let's just sort of examine and see if we can understand based on what we've talked about, what the different effects are that are causing it to have this large term here in curly brackets.

Well, the first thing you can notice is the concentration of boron with respect to time is the D by DX of a flux. That does hold. Everything in curly brackets here is represented by a flux. The question is, why does the flux look so complicated? Well, the first part of the flux, there's a term that depends on DBI star. So what is that? That is inert, low-concentration diffusion of boron that's driven by the boron gradient itself.

So that would be equivalent to this D up here. So that's sort of the simplest part of the equation. But multiplying DBI star, there is this interstitial supersaturation coefficient. CI over CI star appears right here. And again, that's because we're trying to take into account here these non-equilibrium effects. This was sort of an equilibrium diffusion, but of excess silicon interstitials, and that we have pair diffusion. So there's CI over CI star here. And it's gradient is in here, as well.

There are high concentration effects on the dopant diffusivity. So this you recognize this thing in parentheses, it's 1 plus beta P over NI divided by 1 plus beta. That whole thing is Fermi-level effect. That's the high concentration effect. Remember, beta was the ratio of boron diffusing with the I star divided by the ratio diffusing of neutral interstitials.

And so as this kicks in, as P over NI gets large, it's going to bump up the diffusivity, the effect of diffusivity. So if you want, you can think of all these-- there's a lot of these things here upfront, these three terms, as all multiplying the inner diffusivity by some number that's going to pump it up, or maybe pump it down if the CI over Ci star goes down under injection of vacancies. And this last bit here, this partial x of the lin of everything in parentheses, that came-- you covered last week when Maggie was lecturing.

That's the electric field effect, just the fact that besides just the gradient of the concentration gradient that drives diffusion. We also know electric fields. We can have a field-aided term. So that's where this last term is coming from. So you can get a feel for where all these terms are coming from. It's a fairly-- you can't imagine doing this by hand. To solve this equation, obviously you're going to have to do something on the computer, something numerically.

OK, so let's go to slide-- so that kind of gives you an example of the boron case. Now I want to give you an example of the case which is a little bit strange, puzzled people for a long time, but now is understood as an example of the impact of the gradient in the point defect. Not the gradient in the dopant itself, but a gradient in injected point defects. How are they going to change the device dopant profiles?

And we're going to talk about this in much more detail once I talk about ion implant damage. But just want to introduce it at this point to make the point that the gradient in the point defects can drive diffusion, as well. There is something called the reverse short channel effect that puzzled people for quite a bit in the early 90s and ended up being explained, this electrical effect in devices being explained by this boron pair diffusion, boron interstitial pair diffusion.

So before we talk about reverse short channel effect, which sounds really weird, how about what's the regular or the usual short channel effect? We talked about it a little bit several lectures ago. But basically, the usual short channel effect is for a given process, the threshold voltage goes down as you decrease the L. So for a given process on chip, as you look at smaller and smaller devices made on chip and you plot the threshold voltage, what you'll see is it rolls off.

Generally the VT goes down. It becomes easier to turn on the device. And here's a textbook if you want to go through in understanding some of the physics. But basically, what happens is as you bring the source and drain closer, the potential due to the drain actually starts interacting with the potential in the channel, and it starts having an effect.

Ordinarily you'd like that not to be the case. You'd like to have just the gate have the only effect. So the lowering of the potential barrier by the drain voltage is what causes this to go down as you shrink the channel length. And how much that potential barrier can be up raised or lowered is a strong function of the profile of the boron, say in an NFET, underneath the channel.

So in fact, if we go to slide 13-- just, again, I took this from Tower and Ning's book on fundamentals of modern devices for the device physics. Again, you don't have to understand the detail, but just gives you an idea of where this comes from. Imagine here I have-- this is my source, my gate up here, and the drain over on the right. And this distance, 0, is right at the source injection point. And L is the channel length. So that's right at the drain.

And what he's plotted here in this book is the surface potential. So that's the potential at the surface that a carrier would experience as a function of distance along the channel. So here at 0, you're just starting in the source. And it looks something like this. Curve A is for the case of a 6-micron device. Curve B is a 1.25 at, say, half a volt. And curve C, when I put a drain bias of 5 volts, what happens? Actually, the potential, even in the center of the channel, instead of just being the potential of being lowered in the drain, it's actually because it's a short channel the potential is actually lowered here.

And so it's that ability of the drain voltage to impact things that is a short-channel effect, and is partly responsible for the lowering of the VT. So the way people-- when you scale devices shorter, the way people counteract this is they dope underneath the channel. They counter-dope it more heavily with the opposite dopant. So they add more boron, for example. If you were to add more boron, there'd be much less of this effect of this potential being impacted by the drain.

So there's a tendency, as you scale devices smaller and smaller, you'll see the doping in the channel. If you go on the ITR roadmap, every year it goes up. mid 10 to the 17th, 20 to the 18th, mid 10 to the 18th as we shrink devices. So people counteract this short channel, the normal short channel effect, by upping the doping in the channel.

But what people saw was the reverse short channel effect, which was confusing people, is that for short channel lengths on the chip, it was actually found that the threshold voltage actually increased in a certain range of channel length. So as they shrunk the channel length in a certain range, VT actually went up, almost as if the doping in the channel was getting higher in that range. So all these devices are fabricated on the same chip, subject to the same temperatures on the same wafer, same iron implants, everything.

Why would it be the doping would be different in the center of the channel, depending on the channel length? And people were mystified by this for a while. So the reverse short channel effect basically looked something like this. And this is kind of a backwards plot. But if you plot threshold voltage-- now this is on the right axis. Inverse channel length is increasing this way.

But if you want to look at the upper X-axis, it's helpful. The channel length is here on the right going from 0.2 all the way up to 1 micron. Or if you're shrinking the device, you're going from here to here. So as I'm shrinking down here from, say, a 10-micron device down to 1 micron, what was actually happening from 10 to 1 micron? Actually, it's easier to see it if you look at the closed squares from the experiment. The VT was actually going up from about 1.1 volts up to 1.2 or 1.25 volts, which is a significant increase.

So you're shrinking devices from 10 microns or 5 microns down to the third of a micron. VT is going up. That's the opposite of what everyone expected from the old days from the regular short channel effect. So people call this the reverse short channel effect. And in fact, people were able to-- it turns out there's a paper in 1993 at IEDM from Rafferty. They were actually able to explain this based on simulating what they thought the boron profile would be in this nMOSFET in a case where there was transient enhanced diffusion. And we haven't yet talked about TED.

But basically, what they found was that the source drain implants were injecting excess interstitials and setting up fluxes of interstitials that were then driving the boron, which was originally buried, driving it closer to the surface. And if you made the channel shorter, this effect was even greater, basically, because you're bringing in-- the center of the channel is being brought closer as you change the channel, closer to the source drain regions where the excess interstitials were being pumped in.

But in either case, whether it's OED or TED, the point was that they were able to explain based on these diffusion models that more boron, more p-type dopant, was ending up in the channel region of a short device, say of 0.3 micron device, than would be in a 1-micron device. More boron means instead of the short channel effect and VT rolling off-- actually, there was so much extra boron that VT was going up.

So the electrical engineers and the circuit people who would see these weird VT variations with channel length and didn't know about the processing were mystified as to how this could be happening. They just assumed, well, the boron concentration is the same on a short device and a long device. Well, actually, it's not. It can vary. And of course, that means you need these more complex models that haven't taken into account point defect injection from the sides in order to be able to model the boron profile accurately and to get the device VT right.

So that's sort of a classic example of how these complex models impact device performance. And in fact, here's a cartoon explanation on slide 15 where we-- this is a two-dimensional supreme simulation of what's going on in the reverse short channel effect. And what's being shown here is a relatively short device, short channel length. And the different colors are the different contours corresponding to constant doping concentrations.

And if you just for a moment imagine, this is the drain over on the right, you see this little region here is the drain extension. And down here is the deep drain. Here's the source extension and the deep source extending down to about this blue-colored region. These arrows that are emanating from the source and drain, they are interstitial fluxes coming from the surface of the source and drain.

These interstitials, people believed in this particular model, were being injected due to the damage due to the ion implant. And we're going to talk in the next few lectures about ion implantation and how it damages. But just take that with a grain of salt, that you believe that there was a process that introduced a lot of excess interstitials only in these regions, not where the gate was.

So these interstitials were coming in and they were diffusing all around. So these arrows represent interstitial fluxes. And they're recombining at the various interfaces. Now, look at this interstitial flux here that goes like this. This arrow points this direction and goes up towards the surface. So this interstitial flux, this gradient of interstitials where there's a high concentration here and going down at this point.

That gradient of interstitials was actually dragging the boron with it. It was actually dragging the boron with it and moving the peak boron profile from where it originally was more deep in the sample, moving it up towards the surface. So in a short device, it was causing the boron to be higher at the surface. And so this is a two-dimensional representation.

If you take a cut straight through the center, right here at x equals 0 and look along the y direction, in the vertical direction, you can see three different curves on this plot. So this is a plot right through the center, the concentration of boron versus depth into the device. So 0 would be right at the surface by the channel. And there are three different curves here. The red one is for one micron. And you can see it peaks here at a certain depth. 1 micron channel length.

The blue, the dashed blue, is a quarter micron channel length. And the dashed green is a 0.18, even smaller. And indeed, what the model is predicting, as I'm going to shorter and shorter channels, the amount of boron at the surface is going up here by a factor of almost 3. Something like that. 3 to 4. So indeed, if you increase the boron concentration at the surface of a MOSFET by a factor of 3 to 4, the VT is going to go up. It's not going to go down.

And how can the channel length impact this? Well, it's directly through this mechanism. The normal Fickian diffusion, the simple version of Fick's law, there was no way people, when they use a simple models, they can get that to happen as a function of this channel. They had to invoke some other thing to take an original Gaussian-like boron profile, peaked right here, and then, in fact, move it in closer to the surface.

In fact, if you just looked at these diffusion profiles and I didn't say anything about interstitial fluxes, I just gave you these three profiles and I said, OK, look, is that normal Gaussian Fickian diffusion? It can't be. I mean, for normal Fickian diffusion, what happens is as the profile gets broader, sure, it gets broader. Would you get more diffusion when it's driven by the flux of the dopant?

But the peak doesn't change. The position of the peak in normal Fickian diffusion never changes. If you did your homework and you did the Gaussian diffusion, you found, sure, it goes down with time, but it's not like the peak shifts over to the left or the right. It's weird. If it's driven by its own concentration gradient, by definition, the peak doesn't shift.

But here's an example of diffusion where the peak was actually shifting. The only way to explain that is some non-Fickian sort of phenomena. And in fact, people use this pairing, the fact that boron diffuses as a pair with interstitials, and the grades of interstitials was dragging the bond, and so much so that the peak of the barn was moving towards the surface.

So we're going to talk about this in more detail when we do ion implant damage and how much interstitials are injected and all that. But it just gives you a-- I think it's a nice example of how the process engineers got together with a device people who couldn't figure out what the heck was going on and developed a process model that actually could explain pretty well the VT and predict what it should be.

OK, so that's what I want to say now about this diffusion. I'm going to go on in slide 16 and talk about profile measurement techniques. I showed you some really old photographs from Plummer and Fahey in the late 80s. And I talked about beveling and staining. That's one way, but it was cheap and relatively simple, but not totally sophisticated.

But let me give you some more examples. It's really critical-- you can see from the examples I've given you-- to have some method of measuring the dopant diffusion profile as it goes from the surface into depth in one dimension. And actually, if we're going to explain the reverse short channel effect like this, we really need a two-dimensional map of where all the dopants end up. That's not trivial.

One dimension is pretty sophisticated. There are a number of methods. Probably one of the most sophisticated right now that you will use in your research or people use in the fab, or in semiconductor fabrication, is secondary ion mass spectrometry. I think we've talked a little bit about SIMS already. But that's probably the number one technique. And it continues to improve. It's improved dramatically over the last 10 years.

As devices have shrunk, as junction depths have become narrower and narrower, they've found a way to make SIMS higher and higher resolution. SIMS only gives you the physical number of atoms of arsenic or boron per cubic centimeter. It's the physical chemical amount of atoms in the sample at any given point in depth. Sometimes you're interested in knowing not the number of arsenic atoms, but the number of electrons at that point per cubic centimeter. Those are different.

Remember arsenic, you can put a lot of doping in, but not all of it may be electrically active. Or it may be compensated by the presence of boron or another dopant. So there are electrical techniques like spreading resistance. We'll talk about one-dimensional CV and differential vanderpol. These measure the carrier concentration. That's not the same as the dopant concentration.

So 1D though, is reasonably sophisticated. Two dimension is still kind of tough. The methods are still being developed. They've certainly gotten beyond the junction staining methods. But they're generally indirect methods. They're more difficult. And this is an area with a lot of research, and development is taking place just to develop the metrology tools and techniques.

Here's an example of some 2D techniques I'll show you. Cross-section transmission electron microscopy is a way of visualizing junctions with chemical etching. It's like the modern analog of the old Paul Fahey days, look in a microscope after you stain the junction. But it has resolutions that are a factor of 10 to 100 higher resolution.

The scanning probe microscopy, I'll talk about. And the last one is inverse modeling. This is kind of a funny one, but it actually can be very useful in complicated cases. What people do is they take a device. They take all its current voltage characteristics, its capacitance voltage characteristics. They put them all in one giant database and think about it and try to figure out, given all those characteristics, what must be the doping profiles. So it's an inverse technique.

Unfortunately it relies on knowing really detailed electrical models because you're really extracting this all from electrical measurements. So it's not the same as doing SIMS or whatever, but when you get to small dimensions and two dimensions, it is a technique people try to use. OK, so let's look at slide 17 and talk a little bit about SIMS. I've already mentioned that in the past, but now we're going to talk about it specifically for depth profiling of dopants.

There are two modes for SIMS. There's what's called dynamic or static. Dynamic is what you typically use. Dynamic means as the ion beam comes in and hits the surface, it's sputtering away a significant amount. And it continues to go deeper and deeper into the surface as a function of time, at some constant sputter rate, say 5 angstroms per second.

It sputters away the surface. And it looks at the atoms that come off it. They get ionized. It puts them in a mass spec and it tells you at any given depth what's coming off. Static SIMS is a little different. The energy and the angle are adjusted of the primary ion beam such that it doesn't actually sputter very much. It's primarily just taking off what's at the surface. So it's very gentle and low energy.

That's just for looking at just what atoms or molecules are at the surface. But for depth profiling of dopants, typically using dynamic SIMS. How does it work for depth profiling? Well, it's somewhat intuitive, and you should take yourself through these three cartoons. I think you'll have a better understanding. First case, I have a sample of silicon. It has some dopant A up at the top. In some depth, there's a layer of B.

At this interface, let's say there's a spike of X. It's a different dopant or a different contaminant, maybe carbon or oxygen. And down here it's doped with C, whatever these three elements are. So what happens? As I'm sputtering, I'm bringing in my initial ion beam, it starts to create a crater. And it constantly is cratering the sample. And you're looking off-- as a function of time, you're looking at what atoms come off in a spectrometer.

So if you look at the intensity as a function of time, the spectrometer can scan several different masses. Well, lo and behold, it finds for this time period, while it's sputtering through this cap region, it just detects dopant A. When it gets to this interface right here, it's starting to sputter off not only a little bit of dopant A, but a little bit of B, and also X pops in.

And then as you continue to crater down and you get into this point, you're sputtering off B. So you're linearly in time sputtering through the sample. And you collect as a function of time the intensity. So that's what you get out of SIMS. You don't get doping concentration versus depth. You get the intensity. How many ions are coming off per second? It counts per second on a detector as a function of sputter time.

So you have to somehow convert the x-axis of time into depth. Well, that seems obvious on this plot. If you know the sputter rate, you figure out-- let's say you measure at the end. You put it in a deck tack and you measure how deep the hole is, the crater is. And you assume it's linear in time. Well, then I can figure out how many angstroms came off per second. And I can convert time to depth.

And that's a big assumption there because what you're assuming is that the sputtering rate is constant throughout the entire experiment and measurement. If you have different materials-- maybe you're going through oxides, silicon oxide and silicon, sputter rates differ in those materials. So you run into some distortion of profiles. So you have to be very careful. But it's an assumption pretty much that people need to make. So you convert the time axis to depth.

Intensity, that's even trickier because it's just intensity. It has to be calibrated to convert that to atoms per cubic centimeter. So I'll say a few words about how that conversion is done. But let me first show you just a-- I took this off the Charles Evans website, www.cea.com. Again, they're a large commercial company-- they're probably the largest in the world-- international company that for a business does materials analysis.

And one of their big thing that they focus on is dynamic SIMS, primarily-- well, for a lot of industries, primarily for semiconductor and magnetic industries. And this shows you just is a little thing that they like to advertise just to give you an idea of the detection limit, so the number of parts per billion or the percent of an impurity you can measure.

It tends to be related, to a certain extent, to the analytical spot size. So how big is the little spot that you're looking at? And there are lots of different acronyms. And I apologize for all this. If you want to go on to their website, they have each one of these acronyms is defined. In fact, we've talked about some of them. We talked about earlier in the course, we talked about TXRF, total XY fluorescence. It measures about a centimeter. It only measures at the surface. And it's good for measuring in this range, 10 to the 17th atoms per CC.

The most sense-- as you go down to get to parts per trillion, the only technique that can get into that range, and it's only for certain elements, is dynamic SIMS. You see this little bubble is in this range. It has a spot size maybe 10 microns. Maybe closer to 100 microns is more typical these days. Maybe 100-micron spot size. But you can get elemental information down with dynamic SIMS down into the part per billion, and maybe even hundreds of parts per trillion kind of range.

So it's what's used for profiling dopant atoms because remember, dopant atoms can exist at very low concentrations. 10 to the 14th. That could be a typical doping concentrations. And SIMS is the only thing that can measure that right now from a chemical point of view. So the primary ions that we typically use-- I'll just mention there are two primary ones or common ones.

Oxygen is an ion coming in, has come in. People use oxygen because it enhances the secondary ions that come off. It enhances the production of positive. Oxygen tends to strip off electrons. So O2 plus. So it will enhance the ionization of the atoms coming off. Because remember, if the atoms come off as neutrals, you can't mass analyze them in a spectrometer. So you've got to get the atoms off the sample, and you have to hope that they come off ionized.

And the way you increase the ionization yield, so to speak, is you use oxygen to increase the positive ion production. It's good for groups 1 through 3 if you're trying to measure-- and the transition metals in silicon. For instance, for boron, typically if you're profiling for boron, you use O2 plus beam. Cesium plus, on the other hand, is just the opposite. It enhances the negative ion production. So it's used for groups 4 through 7.

And these are good electron acceptors, and they form negative secondary ions. So arsenic, you would typically use cesium primary beam. And they have machines set up for these two different-- so you use a different machine depending on what your profiling yield. Now, the ion yields, when I mean the ion yield-- so that's the number of ionized species that are coming off of some element, say arsenic, compared to the total amount of sputtering of the silicon, that yield.

That depends on the matrix material. So if you have arsenic at the same concentration in oxide and in silicon at the identical concentration, the ionization yield will be different. So it will come off with different intensities. So that's called the matrix effect. So that's a bit of a problem. So the measured intensity has to be measured on a test sample that's calibrated where you know the concentration. And you compare that intensity on the same day to whatever's coming off of your unknown sample. And so it's always calibrated to a known sample.

And that's one of the big drawbacks of SIMS. You need standards. It's not an absolute measurement technique. It's always relative. It's only as good as your standards. And in fact, on slide 21, I have an example of some actual SIMS data and how people quantify. The x-axis we talked about, quantifying is not too bad. You use a deck tack. You assume constant sputter rate. You cross your fingers. And that's how you get the depth scale.

How do I get the y scale, which was originally intensity? I need to convert that to concentration. Well, typically what people do is they take a sample that was ion implanted. And again, we haven't talked about implantation. That's next lecture. Implantation is electrical means of very accurately controlling the integrated dose, the integral under the curve, of an element that you implant into a sample. And that accuracy of that control makes the implanter a very good way of generating SIM standards because the implanter can exactly control the total number of atoms per square centimeter that go into the sample.

And in fact, this was a phosphorus implant at a total dose of 1 times 10 to the 14th phosphorus per square centimeter. So that's given as a known. So once we know that, since we know this area under the curve, we can then convert intensity coming off on that given day to a certain concentration of phosphorus. So every time we want to measure an unknown-- so the right-hand sample was unknown. The left-hand sample, the concentration was known because it was ion implanted. So you use a known dose.

So from this integrated ion implant dose, I can generate a sensitivity factor that enables me to convert from intensity coming off to a phosphorus concentration on the unknown sample. And so this is what I get, for example, on an unknown sample. But you always need a standard which you trust in order to compare it to. And it has to be compared on that same day because the SIMS machines, their calibrations can be changing from day to day.

And it has to be, in fact, a sample hopefully that has roughly the same total amount of peak phosphorus as in your unknown. If it was dramatically different, [? MIP ?] effects. And it should be in the same material. Notice this was phosphorus in silicon in a known sample. This is phosphorus in silicon in an unknown sample. You wouldn't want to use phosphorus and silicon dioxide to calibrate this because the matrix effect would kill you.

OK, so that's an example of a practical idea of how the SIMS actually works. These are some considerations. People are very concerned about the depth resolution. What do I mean? Actually, let me go back for a minute for the depth resolution-- is let's say this actual sample is really a box. Let's say it's really a box-like profile. What SIMS does to it is it smears it out a little. The edges are not perfectly straight up and down. They have a little bit of exponential decay on the front and the back.

How much of that is real? Is the phosphorus really decaying exponentially? Or is it really a box-like profile? What effects determine how well I can resolve the profile in depth? So these are some considerations. The depth resolution depends on the element you're profiling-- phosphorous or boron or whatever-- and the matrix that it's in, because after all, the impact conditions-- remember, these primary cesium ions are coming in. They're hitting the phosphorous or the arsenic, and they're imparting energy to it.

So of course, they can knock that arsenic or phosphorus in a little deeper. And of course they could smear out a profile. So you try to change it so you can minimize the amount of ion damage or ion smearing. For sputter depth, the deeper you go into that and the deeper you make your crater, the more the bottom of the crater gets roughened just by the sputtering process as a random process. When you have a rough bottom to your crater, well, you're pulling atoms off from slightly different depths at that point in space.

So the deeper you sputter, the worse the depth resolution for SIMS. So if you want to measure a sharp profile, try to put it near the surface, and you'll get much more box-like profile than putting it 1 or 2 microns in where the sputtering process has roughened the bottom of your crater. So what people have improved these days to try to get sputtering processes that are as smooth as possible by rotating the sample, and rastering the beam, and trying to make it really a nice, flat, perfectly shaped crater bottom. So they've done a lot of techniques.

Also, these days they have ultra low-energy SIMS has been developed where the incoming beam has such a low energy, it doesn't perturb the profiles too much. And that's an important low-energy SIMS that's been developed in the last 10 years. OK, so that's chemical measurements. How about electrical? Well, remember I talked about beveling and staining? Well, you don't have to just bevel and stain. You can actually bevel, just like I talked about before. You create this surface that's beveled at a certain angle.

And instead of staining, you take two little probes and you measure the resistance between those two probes. You flow a current and measure the voltage drop as a function of lateral distance along the bevel. As you move along the bevel laterally, of course you're also moving in depth. So you're essentially spreading out by the cosine theta. You're spreading out-- or the sine theta. You're spreading out that profile.

So you can move laterally. And literally, it takes the probes on a machine, moves it laterally by certain steps. And you can convert this, then, to a resistivity, or resistance plot, as a function of depth. And you convert resistivity into carrier concentration. So this measures the doping concentration versus depth, not the dopant. Doping refers to the electrons or holes.

So it's complementary to SIMS, which provides the chemical information. The depth resolution is not nearly as good. I mean, with spreading resistance, you're lucky to measure a junction that's 1,000 angstroms deep. Typically with SIMS, you can measure 100-angstrom deep junctions. And the big problem is, again, it needs some standards. You need to know how to convert how from resistivity to carrier concentration. So you need to know these curves. There's some famous curves on the Silicon website.

This company Silicon does spreading resistance for commercial purposes. You can go onto there, and they show you their calibration curves. But if you have a material other than silicon, single crystal-- let's say you have polysilicon-- the relationship between carrier concentration N and the dopant concentration is not very well understood. Or if you have silicon germanium, there are no standards. So spreading resistance is OK, but it does have some limitations.

I won't go through this in any great detail. This is shown on slide 25. I'm referring you to the-- I pulled this off the Silicon website. But in fact, even in the most perfect case, you can try to correct the data for artifacts by solving Poisson's equation to account for things like space charge layers where you have an N and a P region meeting and things like that. There's a lot of literature on this.

But you can see it's measuring junctions that are pretty deep, microns deep, not hundreds of angstroms. But it can be used to measure the well profile and things like that. Slide 26 actually has kind of an interesting technique. It's not quantitative, but if you want to talk about 2D, it can work. This is a cross-section TEM. So you're taking a sample from a silicon MOSFET and you've cut it and cross-sectioned it. And you've made that thickness of that specimen only about 2,000 angstroms thick. There's special ways of cutting through and cross-sectioning it.

And then you send electrons through and you look at their diffraction patterns. So you can actually image the gate, the gate oxide, the silicon. And interestingly, what people have done is you take the specimen after you've cross-sectioned it and you dump it in some acid. And it turns out the acid etches much more rapidly regions that are very heavily N-type. So when you etch that region, you change the thickness of the sample and you change the transparency of the sample to electron beams.

So where it's very heavily doped, it appears very light. The electrons don't-- they go right through. So the contrast on this gives you good qualitative information. And people believe that it delineates this line, this dark line, corresponds to a concentration of about 10 to the 19th. So it gives you an idea of where the edge of the source and drain might be, but it's qualitative. And it's very time-consuming.

There's another 2D technique, again, based on beveling. Again, we'll go back to the same old bevel idea. Here's a MOSFET with a source and drain. People actually take an atomic force microscope with AFM cantilever, little tip. And they measure where they are across the sample. And they measure at each point, a CV curve. So it's like very locally doing a capacitance voltage measurement where the tip corresponds to the metal point of the CV.

And the backside contact corresponds to the back. And you're actually measuring-- you know from CV-- remember, we talked about CV on dots. If I put a dot over a uniformly doped sample, you can extract from the CV the local doping concentration. Well, they're doing it, but here on a very small scale with a very small AFM tip. So this is scanning capacitance microscopy.

It's very tricky, though, the spatial resolution issues, limited by the probe size. The tip is only so small. And then it has fringing electric fields. So the actual area of the capacitor that you're creating with the tip is somewhat uncertain. It has to be modeled with sophisticated ENM modeling. But just to give you an idea, again, this is the website if you want to go there, if you're interested.

Basically how it works on an N-type CV curve, basically it takes DC by DV, the derivative of the capacitor with respect to the voltage. And you can relate that locally to the carrier concentration. Very similar to what you do in 1D CV, but now over a surface. And in fact, here's a beveled junction where this region corresponds to phosphorus.

This is the scanning capacitance microscopy image. So this is very high doped. This is the junction region in p-type silicon near the edge of a mask. So here's a mask. There's no phosphorus over here. It's lightly doped. So it's tricky. It's still under development, but it is kind of a popular topic for modern metrology.

So let me just summarize about techniques for profile measurement. We talked about SIMS, the most popular. It does very good 1D profiles in depth, excellent as the best sensitivity of any technique to the dopant concentration. Excellent depth resolution. Methods are still under development trying to improve it. The very near surface region is troublesome, but there have been recent improvements.

You have to watch out for matrix effects. If you're profiling a dopant in oxide or nitride or silicon, the ion yield varies dramatically, it has to be calibrated. Spreading resistance is only generally mostly one-dimensional, although people are trying to do it two-dimensional. But it measures carriers, so the active electrons and holes. Pretty good sensitivity. Depth resolution is not great.

And it's hard to do shallow junctions. And you need to do some electromagnetic modeling to really understand it. These newer techniques are kind of exciting. These two-dimensional scanning capacitance and scanning resistance microscopy with using small probes are very interesting. They do rely on beveling, but there's a lot of advanced models for the process itself that are being developed in R&D today to try to come up with a better way to get quantitative two-dimensional dopant profile measurements.

So summarizing on what we've talked about so far on dopant diffusion, we said that they diffuse by interacting with point defects, vacancies, and interstitials. The diffusivity is proportional to the concentration of those point defects. These point defect concentrations go up exponentially as I increase the temperature. And so the diffusivity goes up exponentially.

They can also be changed by things other than temperature-- the local Fermi level-- the local doping concentration, that is. Ion implant damage, as we'll see in the next chapter, can change the point defect concentration. Surface processes like oxidation and nitrogen change it. All of these affect the effect of diffusivity. So the dopant diffusivity can vary in space. It can vary in time. All of that means that we cannot calculate accurate profiles in Silicon devices by hand. We're pretty much have to monitor all that by doing numerical solutions.

There's been a lot of progress in the last 10 or 15 years on getting physically based models for dopant diffusion that will actually help you predict electrical behavior. These simulators-- and we'll talk about it more when we give a lecture on supreme 4. They allow you to fully couple the diffusion of the point defects. So you solve for the diffusion of the interstitials and vacancies, and you solve diffusion of the dopants at the same time.

The problem with all these models is there's a lot of parameters. And so any parameters that you don't know, you can get beautiful profiles, but all the parameters need to be calibrated. So that's about all I have for today. And if you're handing in your homework 3, please bring it up front to this folder.