Started with a couple of announcements. There are two handouts for today. There are the lecture notes for today, which are this handout 25. And you also have an additional handout 26, which we're going to go through. It's a little calculation example I was hoping we could go through during today's lecture and have you do some calculations. So it's not a homework problem. We're going to do that during the lecture.

Let's see. As far as homeworks, I don't have any to pass out quite yet. The TA is still working on those. And she's out today. But I do have the clipboard, which I will have up front here if you want to sign up for your final project. We started circulating that last time where you can-- I'm asking people to put your name down.

And even if you don't know your topic, if you can check whether you want to do a written report or an oral presentation to the class, that would be helpful. And then once you know your topic, that would be good to fill that in, as well. All right. So I'll leave that up here for now. And you can come up towards the end and sign up.

OK, so today's lecture is actually going to be the final lecture on ion implantation and transit enhanced diffusion. Let me review here on handout 25, the first page-- review what we've talked about so far. We've talked about ion implanted profiles. And we said you can model them very simply as a Gaussian, more accurately as a Pearson Ford or dual Pearson distribution. I'll show you some examples of real implants today.

Or they can be simulated by numerical techniques such as Monte Carlo. Last time we talked a lot about the damage modeling. And we introduced something called the plus end model where n is a small number, about 1. And this is the model for residual damage that says that there's roughly n excess silicon interstitials injected per primary iron where n is a small number on the order of 1.

Then last time we talked about how these excess interstitials cluster very, very quickly into defects called 311 defects. And later, these defects then dissolve during annealing. And their evaporation rate is really what determines the kinetics of transient enhanced diffusion, determines how long TED lasts, and also determines the magnitude of TED.

In fact, we had a clustering slash evaporation model that we talked about last time. And that was used to explain the time and temperature dependence, and to a certain extent, the dose of energy dependence of TED in a rough way. So what I wanted to cover this time is to actually give you some real examples of ion implantation profiles, data, and a little bit of simulations. I want to spend a few minutes-- maybe 10, 15 minutes if we have time in class.

I know a lot of people are sleeping in because the Red Sox won last night the World Series. But for those of you who are here, that's what the handout number 26 is all about. We'll calculate together a little simple calculation on transit enhanced diffusion. And then I want to finish the lecture and spend the rest of the time on the effect of TED on devices. Now that we have these models for TED, we can go back and revisit the reverse short channel effect and talk about that in more detail.

OK, so let's go on to slide number 2. And what I'm showing here, these are some actual SUPREM core outputs, output files, modeling ion implant profiles, and just to show you a practical example of how you might use SUPREM-IV along with a dual Pearson. A dual Pearson means you take two Pearson four distributions and you add them together. And you can use this in a quasi empirical way to simulate the ion channeling tail.

So there are four different plots here, four different panels. Look at the first one marked A. These are all for the same condition in terms. of energy. This is all boron, BF2, implant. And the energy, the primary energy, is 65 KEV. And then what we're looking at in each panel is a different dose. So before we get started with looking at those, let me just make a note. The total atomic mass of a molecule of BF2 is 49. You just add up two fluorines and a boron. You get mass 49.

The mass of boron is 11. So as I mentioned last time, people sometimes do a BF2 to get a lower energy or more easily modifies the silicon. For example, look at this little simple calculation. 65 kilovolt BF2 is equivalent to 65 times 11 over 49. So you assume that the energy is partitioned according to the ratio of the masses for a boron implant.

So if I do 65 KEV BF2, it's the same as if I implanted the boron atom itself at 15 KEV. So especially in the old days when it was very, very difficult to build ion implanters that could ion implant at low energies, now they have implanters that can get you down to 1, 2, 3, KEV. But the beam current tends to suffer when you go that low.

So instead of doing very low energy boron, people would do a higher energy molecule, BF2. So you'll often see BF2 implants. And so now let's go ahead and look at these BF2 implants. And the couple of things being shown here-- the plus signs or the symbols are the actual data. And these data was taken from Tasha's work back in 1989. These are actual sims profiles.

And the lines, the smooth lines, are SUPREM simulations. So for example, in part A, or plot number A, we have a dose that's very high of 5e15. And so what that means is with such a high dose of BF2, throughout most of the implant, of the time of the implant, you're really going to be amortized. You're going to be implanting into an amorphous crystalline because it's going to amort-- the BF2 will amorphous the silicon. So this is the profile that you get.

Notice it looks pretty much like a standard Pearson 4. And there's this little small exponential tail. But that tail doesn't really become obvious or evident until you get about one, two, about three orders of magnitude down from the peak. And then you start to see the tail. So the way this profile is constructed is what dual Pearson means. It's two Pearson force added together. So there's one Pearson four here, which is with this dashed line that's labeled as the amorphous profile.

And then there's a second Pearson four here, which has also, unfortunately, a dashed line, but you can see it has a slightly different slope. And that's labeled the channel profile. Those two Pearson fours are added up with a certain ratio. This ratio here, in this case, is 0.969. And you get the total profile. And you can see the simulation, obviously if you pick the right number for the ratio, the simulation then fits the entire profile, including this region here at the near surface and the amorphous-- and the tail, rather, the channel tail.

Let's go to a lower dose, say 1.5e15. Now you do see-- again, we see two Pearsons, this one here and the one that's associated with the channel profile. But now the channel profile is a higher fraction of the total. And you see this exponential tail coming in a little bit more obviously. And if you go down to panel number C or plot number C, the dose is even lower, 5e14. Not that much amorphization going on. So you have a large, very prominent channel tail. So the second Pearson is really dominating.

And finally, at a low dose of 2e13, there is no amorphization. So the implant that really dominates in the dual Pearson is primarily that associated with the channel profile. And the ratio is actually 0, so it's just the channel one. So that's how dual Pearson works. You can see it is two Pearson fours added together, one of which takes care of the channel profile, and the other which takes care of the amorphous. And you add them together depending on a certain number called the ratio. So it's quasi empirical.

So that was for BF2. Let me just show you some simulations. The different symbols here are not data, so I apologize. These are actually different symbols represent different SUPREM simulations. Just to give you an idea when you do a simulation, depending on which model you choose, you'll get slightly different profiles. And it's up to you to figure out by comparing to data or by looking at the literature which one of these is closest to the truth.

For example, just as a comparison, if you look at the open circles here, this profile marked Gaussian-- it's a little bit hard because there's a lot of symbols. But there's an open circle profile, and it's quite symmetric, as it has to be, because it's Gaussian. It has no channel tail. If you look at the open triangles, it's the Pearson four. A single Pearson looks a lot like the Gaussian, not much difference. A little bit skewed.

The dual Pearson is these open boxes. And it does a little bit better job, you would think-- well, we don't know. Actually we haven't seen the real data here. It has the channel tail incorporated in it. And look at the Monte Carlo. Now presumably, the Monte Carlo has the most physics built in into the simulation. And it looks a little noisy because, of course, we only follow a certain number of ions.

But in fact, the dual Pearson looks reasonably close to the Monte Carlo. The Monte Carlo is attempting here to include a little bit of the ion channeling. So you get different models give you different answers in SUPREM, and you have to figure out which one suits your particular application best.

So now what I want to do is I want to take a few minutes here in class if we go on to slide number 4 in the handout. And this is also the first page of your handout number 26 if you want to look at that. They're identical. And what I was hoping you could do is get together as a group. There's not enough people here to really split up into individual groups because everyone, again, sleeping in because Boston won-- is to go through this and make some simple calculations for the next 5 or 10 minutes together on how you would do the simple calculation.

Let me just read through it before we start. So we have an engineer wants to form a shallow boron dope source drain for an advanced technology. And the question, the manager is wondering whether to buy a batch furnace or to use a rapid thermal anneal. And the furnace anneal they are considering for this implant would be 800 degrees for one hour. So that's a relatively long time at low temperature.

Or should they use a rapid thermal processing machine, or RTA, at 1,050 for one second? And the implant that they want to activate is boron, and it's 20 kilovolts, relatively low energy. 20 kV, and a 5e14 dose. So what we're asked to do here in A is to make a rough estimate using a square root of DT estimate of how far the dopants move during an 800 degree one-hour anneal versus 1,050 for one second.

Again, it's not necessarily Gaussian diffusion, but you can calculate the root DT. And this is without any TED effects. And this is pretty simple because I've printed at the bottom of the slide the intrinsic diffusivities under equilibrium conditions. D at 800 is this number. And the diffusivity of boron at 1,050 is this number. So we have those numbers written right there. So part A is pretty easy.

Now, part B, what we want to try to do is include the effects of TED. So we want to use the charts that we handed out last time. And in fact, they're in handout number 26. If you didn't bring your old handouts, I've included them again in handout number 26. Use these charts to figure out the expected enhancement of the diffusivity. Remember we said the diffusivity is now the equilibrium, D star, times CI over CI star. That's the effective diffusivity. And how long Ted lasts to calculate the real square root of DT, including TED effects.

So these are the two things I'd like you to try to do. And I'm hoping you can work as a team on this. Maybe get together with a couple of folks who are near you. Somebody here hopefully has a calculator if you need one to do this. I'm not so interested in exact numbers. I just want to get an idea of how the answers play out.

So why don't you-- I'm going to give you 5 or 10 minutes to get together and work on that. And we'll stop the lecture now, and then we'll come back and we'll see who has an answer that looks reasonable, and we'll talk it through. OK? So we can stop the recording at this point, as well, because it won't be that interesting to record you punching your calculators.

1,300. What can you say about the first 1,300 seconds? How high is the diffusivity during the first 1,300 seconds? It's going to be 7,000 times than it is in the next 1,300 seconds or 2,000 seconds. So who cares, in some sense about, the next 2,000? Now, if this instead of being 3,600 seconds was orders of magnitude more seconds, yeah, then at some point you actually have to take into account the normal diffusion.

So you don't completely ignore the actual time. The actual time that the thing is in the furnace is 3,600 seconds. It's just that for on the order of almost a third of that, it's enhanced diffusivity by 7,000. So we can ignore the rest of the time. But you want to keep that in the back of your mind. So then were you able to calculate either a DT enhanced or the square of DT enhanced? Either one. What'd you get for that?

AUDIENCE: DT 1.06.

Is that the unit?

JUDY HOYT: The DT you got-- OK, I ended up somehow with a 3.5 times to the 10 to the -10 centimeters squared. But doesn't mean-- oh, I multiplied this. So if we multiply the ordinary diffusivity times that, I got 7,000 times 3.7 times 10 to the -17. So the actual diffusivity I got was 2.6 times 10 to the -13 centimeters squared per second during the transit, during the time when the diffusivity was enhanced.

And then multiply that by 13. Oh, my numbers-- I had 1,333 here. Maybe I was carrying a few more significant digits.

AUDIENCE: Well the numbers should be in the root.

JUDY HOYT: Oh, root DT. OK. So for the square root you got what? 18? Right. 10 to -5. So in angstroms, to put it on the same-- that's 1,860, just to put it in the same units. So not even close in the intrinsic diffusion. So really, the TV clearly dominates in that case. And there's quite a bit. That's quite a bit of motion. Now that, you can easily see on a sims profile, and that could definitely affect your device performance junction depth, a difference of that much.

So how about a 1,050? Exact same formulation applies. But the nice thing is-- how about what can we say about CI over CI star at 1,050? It's a lot lower, right? I got about 550. So the enhancement of the diffusivity is only 550 times larger. That's the diffusivity-- and how long does it last? And the TED at 1,050-- well, if you go back to your magic curve, phosphorus, it doesn't look like it's going to last very long.

Again, you have to multiply by 5 and by 0.08 over 0.06, the RP ratio. So the amount of time it looks like TED lasts under this dose condition, I got about 0.67 seconds. 2/3. 2/3 of a second. Basically, 2/3 of the anneal. Remember, the total anneal time was one second. So again, the last one third of the anneal, I'm going to ignore just because this is still a large enhancement factor. It's 550. So I'll ignore it. If it had been 1,050 for hours, then obviously we can't ignore the normal diffusion.

OK. So then in that case, the root DT at 1,050, I ended up with something about 420 angstroms. Is that close to what you guys got? 4.2 times 10 to -6. Again, putting it back all into angstroms. So again, we see this kind of interesting anomalous, and maybe initially somewhat non-intuitive, result.

Here we have a higher temperature. Admittedly a much shorter time. It's only a second. We can get a lot less root DT, a lot less motion or broadening, if we do a 1,050 rapid thermal anneal compared to putting in the furnace. You might say, well, 800 is really low. It should be a safe temperature. It's only a couple of angstroms motion. But in fact, you're much better off, for this implant, doing-- according to this calculation of TED, you're much better off doing a 1,050 anneal for a fraction or for a second.

So that's just to give you a feel for where the numbers come from and how the whole thing works. I think once you work through an example like that, you have a much better feel for TED. OK. Good. Well, that seemed like everybody had a good handle on that. So let's go back to the regular handout, handout 25.

And I think you've got a good feel for how simple calculations and how powerful it is to see you as those two or three charts. With two or three charts, you can say a lot about-- a rough back-of-the-envelope calculation for TED. Anything more sophisticated than what we just did in the last 10 minutes, you probably should be using SUPREM at that point, or some simulator, I should say. All right, let's go on.

So just to remind you ourselves, where those charts just came from-- if I'm on the slide number 5 now on the handout, they came from this observation that was made at Bell Labs and confirmed by a lot of different workers that the time scale of TED was the same as the time scale of the shrinkage of these now famous 311 defects, and that at low temperatures, the 311s ones hang around a lot longer than at high temperatures. And of course, that's why TED lasts a lot longer at low temperatures.

OK, so just to remind ourselves and from the little hand calculation we just did, this is exactly representative of what we just did in class here. The general picture of TED that we showed last time, we have a certain enhancement factor-- CI over CI star max. Again, that's a ballpark. That's the maximum enhancement. So we assume it's just a constant enhancement throughout the entire period of the steady state while the 311s ones are decaying. And then we suddenly say there's a rapid exponential decay right after some period of time called tau enhanced that we just calculated in our example.

So that's the simple model. The critical parameters that determine TED? Well, the amount of TED we just said was the supersaturation level, which we write as I over I star or CI over CI star. And we know that's a function of temperature. And this is the functional dependence, or you can read it right off that plot. And the duration of the steady-state condition, which is the so-called tau enhanced time-- now, that tau enhanced depends linearly on dose. And doses have a wide range. So this is the key.

It depends linearly on q. So if you implant 20 times or 100 times more dose, you get 20 or 100 times more interstitials. So it'll last that much longer. So notice the dose unintuitively somewhat. The dose doesn't determine CI over CI star. The dose determines how long the thing lasts, how long the transient lasts. And there is a certain small dependence on RP. And the energy dependence comes from the dependent-- in the equation, the energy dependence is represented through the dependency on RP.

So that's the general picture. And I think that example helps us understand it. Now, let's look-- anything a little more complicated, I said we should be using a simulator. So the next couple of slides here on slide 7, starting with slide 7, I'm going to show you some SUPREM-IV examples using more and more sophisticated models. So the first model, I'm going to use SUPREM-IV using a Gaussian implant assumption, which of course is not very sophisticated.

And I'm going to be annealing this arsenic implant at 1,000 degrees C for different times. And if you look at the open boxes, that's my initial ion implanted arsenic. And what's being shown here is a series of different curves, different symbols for different times. And what the model that was used in these particular simulations is called the Fermi model in SUPREM. When you do your next homework, there'll be different models. You have to invoke the one here is called Fermi.

As the name suggests, it takes into account the Fermi level, and therefore the concentration dependence of the diffusivity. It does not take into account any TED. So when you tell it to use PD method equals Fermi, you're not taking into account TED. So this is just normal ordinary diffusion. And the junction depth proceeds according to a square root DT type of behavior, as you would expect. It does take into account the concentration dependence. And you see the box-like profile as a result.

Now we go on to a little more sophisticated model. This is that same implant, 34 KEV 4014. But now this is on slide number 8. What we're doing is we're using a Monte Carlo implant model that changes the as implanted slightly. But the reason we use that is because the Monte Carlo model in SUPREM keeps track of the damage because not only does it track the incoming ions, but every time a silicon gets displaced, it can keep track of that. So it can keep track of the damage.

And then we're using a model in SUPREM called the fully coupled or the full coupled model. It takes into account these following factors-- the impact of the Fermi level. It also takes into account point defect injection and 311 cluster dissolution. It has a model there for 311s and their kinetics. And it also, because it's fully coupled, defect gradients can drive diffusion.

So if there's a two-dimensional situation like in the reverse short channel effect, actually the gradient in the defect can drive a diffusion. But this is what you would see if you do this fully coupled model. Look at-- these first four curves all lie straight on top of each other. 30 seconds to 120 seconds. They're all the same. Now, why is that?

Well, I mean, essentially, it's because of TED because whether you're 30 seconds or 120 seconds, there's some portion of tau enhance. There's a certain enhance time. And during that period, CI over CI star is so large because of the amount of 311s that end up being formed that it doesn't matter whether you're doing seconds or 120.

The diffusion is dominated by whatever happens during the transient when the 311s ones are evaporating. And then after that, if you only-- if you go to longer times, so say 600 seconds, 1,200, and 1,800 seconds, do you start to see normal diffusion being large enough? Because the time is large enough that normal diffusion now overtakes the TED and you can actually see that.

And so this is an example of why people didn't discover TED until they had rapid thermal annealers because in a furnace, you can't anneal for very short. So you could never access anneal times that were this short. And you never would discover that, in fact, TED was going on. So a signature of TED, if you do an experiment on annealing and implant, a very clear signature-- and if you do sims on different times, is you see very short times. They all look the same profile.

That's an immediate signature. Aha, there's some kind of TED going on. Some very short transient is dominating all the diffusion. And then you'll go to long enough times that you'll start to see normal diffusion taking place. And these are SUPREM simulations. There's no actual data here the different symbols are not data. I apologize. It's for different-- the symbols are for the different simulations.

OK, so that was a couple of examples. Let's go on to slide 9. I want to give-- we're going to go back now. Now that we've talked about TED and we understand the models in much more detail, I want to remind ourselves why we want to limit diffusion and then talk about some particular device impacts of TED. Well, we already know we want to keep junctions shallow. These junctions here, XJ, either in the deep-source drain and particularly in the shallow-source drain extension has to be kept shallow in order to do gate length scaling.

We need steeper lateral junction. So in this direction, in the L direction, we need the junctions to be very steep in order to lower the series resistance at a given effective channel length. We need a very small under diffusion under the gate to reduce the overlap capacitance. If this underneath the gate, this diffuses too far under the gate then you'll have a lot of overlap capacitance. And that's going to slow down the circuit speed.

We talked about needing a retrograde well in order to get better mobility while controlling short channel effect. We need this retrograde well in depth. We need to retrograde the profile and decrease it towards the surface. And there's these fancy little angled halo implants. These halos here are shown in the bright red. These are implants that are done at an angle to the gate in order to put the dope just right where we want it.

And if we're using this angled implant to obtain the right profile, we'd like to not have that diffuse all over the device because then it kind of-- we don't get the shape that we want. So all of these reasons, we need to limit diffusion. So I want to talk about, now that we know TED and go back to this reverse short channel effect that I introduced three or four lectures ago, and I think we'll have a better understanding at this time.

If you want to learn a little bit more about it, there's an article by Rafferty in IED of '93 or Crowder at IBM of '95 where this was first explained. And what people were trying to explain is the device physicists in the early 90s were all finding if they-- and the circuit guys-- if they plotted their VT a function of the gate length of the transistors on a chip or on a wafer.

What they expect is the normal short channel effect. VT is fairly constant, and you get to short channels, and the VT is supposed to roll off due to well-known electrostatics. What they found and said is the so-called reverse short channel effect. It's reversed to one's expectations. In fact, the VT, the threshold voltage on these devices, was going up as you made them shorter. And it was going up and peaking quite a bit. And then eventually, the normal short channel effect would take place.

So this so-called reverse short channel effect was quite bothersome in the early days because people didn't understand when they scaled to these channel lengths why VT would be going up. How could it be? Well, we talked qualitatively about the reasoning for this several lectures ago. Let me review the qualitative, and then we'll do the more quantitative. This is an actual simulation from SUPREM of what's going on in the reverse short channel effect.

So we have our gate. You see the side wall spacers. This little region under here is the source extension. This is the drain extension and the deep source and the deep drain. And what's happened is that we have implanted these source drain regions and their extensions. We've implanted, say, with arsenic. And we've generated a flux. In fact, these arrows are supposed to indicate the interstitial fluxes. So these are silicon interstitial atoms that are diffusing.

And in fact, and they're then recombining. They can recombine the bulk or they can recombine in the surface. And because of the way the recombination goes at the surface, they end up creating a flux that goes like this. And now that flux and the gradient of the interstitials is so sharp that it actually can drive uphill diffusion of the boron. And you can see that uphill diffusion. You can't see it in this, but if I take a cut right through the center here at X equals 0 and plot it versus depth, so boron concentration versus depth at the center of the channel.

If you have a 1-micron device on the chip it looks like this. If you have a 0.18-micron device on the chip, it looks like the green curve. And what do you see? In the green curve, well, for one thing, the surface concentration of boron is a lot higher than it is in the 1-micron device. So that explains to the circuit designer, oh, I have a higher concentration of boron on the surface. That makes my VT higher. So that explains the VT effect.

How it got to be so high, and how did-- look at even the peak of the boron profile in the 0.18 micron device. Even the peak move towards the surface. Again, that's totally non-Fickian diffusion. If I give you a Gaussian diffusion and you diffuse it in your calculator, the peak stays put. It just goes down. Peak doesn't suddenly move over to the left or the right. That's not Gaussian.

Well, what's pulling this peak over is the fact that there's these interstitials that have a gradient. And they drag with them because remember, boron likes to diffuse with interstitials. It's diffusing as a pair because of the fully coupled diffusion. They drag with it the boron peak. So this is the qualitative explanation, at least, of what's going on in the reverse short channel effect.

And now let's go through a little bit more carefully the articles by Rafferty and Crowder. And you can see what it is that they-- how they explain this in a little more detail. These are some of the key process steps that influence the channel profile. So the reverse short channel effect ends up all up about being what determines the channel profile. So step number one, we know we do an ion implant and it has some shape. This is supposed to represent, if you turn your head sideways, the shape of on implant.

So we do an implant and a couple of energies. It peaks here, and then there's another peak down here. That's what we expect it to look like in the as-implanted case. Step number two in making the device. Well, we grow a gate oxide. All right, there's a certain amount of thermal budget associated with the gate oxide-- 800 degrees, whatever. It'll diffuse a little. OK, we can understand that. That's normal diffusion.

Step number three, we put down a gate and we pattern it. So it has a pattern like that. Usually polysilicon gates are put down at 600, 500-- very low temperatures. Not a whole lot of motion going on. All right, now the step number four is called the lightly doped drain. Actually, we don't use that terminology anymore in devices. We call it the source drain extension. It's a shallow source drain right here. That gets an ion implanted and it gets masked by the gate.

So when you're doing step number four, the sidewall is not there. Step number five hasn't been done yet. So you have this shallow source drain extension. So there you're injecting a certain number of point defects. OK? Now I take that. I've injected these point defects, and now I have to form the sidewall spacer. Depending on what material form the spacer, if I form it of silicon nitride, that goes down at 800 for about an hour. Uh oh. Bad temperature.

800 for about an hour. We just did that calculation. 800 for about an hour, you can get a lot of TED can have a 7,000, perhaps, enhancement in your diffusivity. So you got to watch out for that. That's why nitride spacers are a little tricky. If you do a low-temperature oxide deposit in space-- or you can do that at 400 so it's not so bad. You don't have to worry so much.

All right, so there's a possible indication of problem. And now number six, we do the deep source drain implant. Again, that's probably going to be arsenic. Pretty high dose, 10 to the 15. Not a great thing in terms of TED. I'm introducing a lot of point defects that can then come in here and enhance the boron implant, the boron profile. So I just wanted to give you the order of the steps so you can understand how all these things fit together to determine the channel profile.

And if you want to go a little more sophisticated into Rafferty's IDM article that's being shown here on slide number 13 of your handout, what he calculated here-- is so this is the edge of the gate. This is a two-dimensional plot out of a two-dimensional simulator. And these are contours. And what he's done is he's ion implanted a certain amount of damage or dose into the drain region. And he's calculated something called the time integral of the supersaturation ratio.

So it is essentially-- and he calls it the enhanced time, but it's the integral of CI over CI star, integrated over the anneal. And he's plotted in terms of contours. So these are profiles of damage caused by this shallow source drain implant. And he represents this damage by this integral of CI over CI star, which he calls the enhanced time, so to speak.

And a number here that corresponds to 10 to 6 has the units of seconds. So it's as if you did a 10 to the 6 second anneal, essentially. And that corresponds to an average supersaturation of about a factor of 1,000. If the real time-- if this were really-- the real time is about 20 minutes. This cutoff at the bottom of your 20-minute real time is equivalent to 1,200 seconds. So there's huge amount of enhancement.

But look at these contours look how they go down. You're going from 10 to the 6 here down to 10 to the 5 in this range. So again, I have a large gradient in this CI over CI star. And the gradient is pointing me, pushing me, towards the gate and towards the gate oxide. Don't forget he's saying that the oxide interface under here acts as a sink for recombination of interstitials.

OK, so this is the origin of the reverse short channel effect as he explained it on slide 14. The implant damage from the shallow source drains sets up a retrograde CI over CI star profile under the gate. And the gradient in this profile, so this grad of I over I star-- because again, it's diffusing as a pair-- results in an extra flux in the boron diffusion that wouldn't be there if it were diffusing alone, but because it diffuses as a pair with interstitials.

So that retrograde causes a boron pileup at the interface. And the shorter the channel, the higher the pileup because as you make the channel shorter, you bring in those source drains closer and closer into the center of the channel. So that explains the reverse short channel effect. Why does the boron pileup get bigger and bigger, and therefore the VT goes up and up for a shorter device.

Oops. And in fact, here's an example,. If we go onto slide 15 from his article, this is boron concentration simulated versus depth. Now, he's doing this at the center of the channel. And say a long channel device looks like this. Here's the boron he simulated. It basically has the as-implanted shape. There two implants. There's a low energy implant-- you can see its peak-- and a slightly higher one. So this boron at the center looks sort of like this.

Now, if you take the boron that's the center of a 2-micron long channel, if you go to a 0.45 micron device, it looks like this one. There's a huge pileup at the surface. And the peak is close to the surface. So the surface doping could be three to five times different when you have a very short channel device compared to a long channel device because of the influence of these point defect gradients that get set up.

And so these very different surface dopings can then be used to calculate the VT difference between these devices. And that's exactly what Rafferty did. And this is taken, again, from his article, 1993. He calculated here, based on those profiles, what he thought the threshold voltage should be as a function of 1 over the channel length, or if you want to read the channel length on the top axis for different biases.

And if you look at his calculations here, this smooth, solid line is, including the TED effect, he predicted that the channel length would go-- the VT would go up like this as I go to shorter channel lengths. And the data-- these are the experimental data they obtained from devices at Bell Labs. Indeed, you can see the VT going up. So look at the VT for a 2-micron device is over here. The threshold voltage is about a volt. And for a 0.45 micron device, which is, I think, right about here, the VT is about 1.3 volts.

And that agrees very well. So that roll-up of the VT agreed very well with the simulation when he included the TED. The dashed line is a simulation without the TED. And indeed, you see the normal short channel effect going down like this. So the only way people could really understand the reverse short channel effect was to really understand in detail what was happening with the boron TED and the influence of the damage on each side of the channel on what was going on inside, underneath the gate.

So that's kind of a famous paper on how understanding these process models end up influencing the device model, which ends up influencing the circuit model. So there's a lot-- in simulators, there's been a certain amount of work in the early 90s on how to model reverse short channel effect, what are important parameters.

Well, obviously, the magnitude of the initial implanted damage. Well, you know the dose of the source and drain extensions and the source drain. So that's important. But you need to figure out exactly how much I over I star there is. We did it in our book, our textbook, using a very simple analytic calculation. We calculate the maximum CI over CI star. Remember, that was a ratio of k reverse to k forward in that equation. And we put some estimates down for that, but it's not clear exactly how accurate.

So depending on your simulator, the way they calculate it will be slightly different. So these are different clustering factors in this simulator. They have a parameter called a cluster factor that will adjust, essentially, the CI over CI star max. And depending on how that is, that CI over CI star max, you'll get different amounts of reverse short channel effects.

So I apologize here, the vertical axis here is VT. That got cut off somehow. So this is the threshold voltage. And this is the gate length calculated. And you can see in this particular simulator-- this is not SUPREM. This is a silvaco simulator. But for a cluster factor of 0, they didn't predict any reverse short channel effect. When they allow clustering and they allow these 311s to come in, that you do see a roll-up, a reverse short channel effect.

And depending on the magnitude of the cluster factor, reverse short channel effect can be more or less prominent. So there will be usually-- depending on the simulator you use, there'll be a couple of parameters one can tweak to effect both the initial implant damage and the recombination rate.

This is a two-dimensional effect. So not only how much damage and how many 311s do I end up with, and how many interstitial-- what's the interstitial concentration that's important, but how those fluxes diffuse and how they recombine at that oxide interface will determine the actual gradient of the interstitials. So the k sub s factor is also important, the recombination rate of interstitials at the gate oxide interface in the channel.

And that's a parameter that one can adjust. Hopefully it's fairly well-known, but by adjusting that parameter, you can change these curves, as well. Annealing temperature is important, as you know. Again, this is from that silvaco simulator, as well. Again, VT versus gate length. And the different color curves here are for different annealing temperatures.

And as you might imagine, low temperatures give you more reverse short channel effect because we know at low temperatures, TED is more prominent. The CI over CI star hangs around longer. The time of the enhancement is worse. So here at 850 in the red, you see more a little bit more of a roll-up than if you were to anneal at very high temperatures. In this particular anneal, I don't know what the dose was. There wasn't much reverse short channel effect.

So lower temperatures tend to change that. And that's just because of the CI over CI star term, as well as the tau enhanced term, depend on temperature. And in fact, the next slide in your handout, slide number 19, is just that same plot that we've used to do the example a few minutes ago to remind you that CI over CI star is increasing as we go lower in temperature. And that's why you see more of a reverse short channel effect as you lower the temperature.

So not only does the VT change, but just to give you an idea of other device parameters that can change depending on the channel profile, that boron pileup at the surface underneath the gate can decrease the channel mobility. Channel mobility, at very high concentrations of channel dopant, you get more scattering of the electrons, more scattering of the carrier. So the mobility can go down.

And in fact, this is a plot of a calculation of the mobility that Rafferty made in that article here on slide 20. So he's calculated the mobility as a function of the channel length for two different doses. This EXT dose-- EXT stands for the source drain extension. So that's the dose that gets implanted right next to the gate.

Right after you cut and define the gate poly, you implant the source drain extensions. And here this is for a 3e12, a relatively low dose. Not that many interstitials are injected. The boron profile looks pretty much as implanted, and you don't get much pileup. So the mobility stays pretty high until you get to very short channels.

If you do an extension dose here that's about three times that-- say about 8e12-- what happens is you get this reverse short channel effect to get the boron being drawn to the surface. A very high amount of boron in the channel means you have a lot of ionized acceptors. And the electrons feel those ionized acceptors. They get scattered by coulombic scattering. And in fact, the mobility then can go down.

So for a higher extension dose, different profile causes pile up and you get a lower mobility. Lower mobility can mean you can end up lowering your current drive compared to what it should have been. And that affects the overall circuit speed. So again, just subtle changes in the channel profile which have nothing to do with how I implanted the channel. It's how I implanted the neighboring regions. The source and drain end up impacting not only the VT, but the mobility of the device, and therefore the current drive.

Well, this was a neat paper. Now, several years later, 1995, Scott Crowder came up with an interesting idea. Knowing what he knew about 311s and what we in our class, he said, OK, well, 311s, they anneal out a lot faster at high temperature. So let's say I have to do an 800-degree step because I'm going to have to put down those nitride spacers. Maybe I can still do that if before I do it, I do 1,000-degree short anneal.

So let's do a very short anneal, one second at 1,000, evaporate all those 311s, and get rid of them. Then I can put the wafer in the furnace and put down my 800-degree nitride. All the TED will be over with. So I use a high temperature to cause the 311s to evaporate, get rid of all those excess interstitials in a very short time. I should have less of the reverse short channel effect, less total interstitials. And that's what, in fact, we showed.

This starred region is when he did the high temperature rapid thermal anneal first before he did the 800-degree C, longer time anneal. And he saw he had less roll-up, less reverse short channel effect. The open squares represent the case where he did the 800-degree C anneal in the furnace first. And then did the 1,000. And you see a lot of reverse short channel effect. Interesting.

Exact same amount of time in the furnace and in the RTA in both wafers. He just changed the order of the operations. So this is kind of interesting that what we know about 311 defects and how they anneal, it's important for us to think about the order of the steps in which we make a MOSFET where we insert the anneals. Now, you don't always have a choice. You have to cut the gate before you do the source drain extension so it's self-aligned and all that.

But this was kind of a clever experiment just to show that the process order can have a big effect because of things like 311s. And again, this is just a reminder. We already saw this several times on slide 22 of the amount of time. And what he did was this 1,000-degree, 1-second anneal. He popped it up. And so he can get-- basically within a second or so, he can get rid of all the 311s. The tau enhance is only a few seconds.

And then he could go on later, and without any 311s around, put it in the furnace, or with very few, and go down to 800. So that was his proposal. Another thing that he showed in that paper-- this is from that same paper by Scott Crowder, IEDM of 1995. He did an interesting comparison. He also compared devices made by various similar processes on bulk, so on regular Czochralski wafers, to the similar device made on an SOI wafer.

Remember, we said there's this technology called SOI where you can have single-crystal silicon layer on insulator. Very high quality material. And you can make MOSFETs in that material. And what he did was now he found, when he did the source and drain, implants-- indeed, you of course, even in silicon insulator, you do inject interstitials up here.

But interestingly, remember, the interstitials tend to get injected both down, and then they go up. But ones that went down were going up here now against an interface between oxide and silicon. And it turns out that interface is a very good sync. There's a lot of recombination that can take place at this interface. K sub s is fairly large between oxide and single-crystal silicon.

So a lot of interstitials can be sunk or can be absorbed there, whereas in the bulk, we get these interstitial fluxes that come in and they don't get absorbed because there's no oxide down there. You can end up getting profiles that look like-- fluxes that look like these arrows, driving the boron to the surface.

So what he showed was, in fact, on an SOI wafer, you don't get as much pileup of boron underneath the gate. Anc on the SOI wafer, this plot-- the y-axis didn't get showed up here, but the y-axis is the threshold voltage, VT. Again, the roll-up of the VT is a lot less on an SOI wafer, subject to the same kind of annealing, compared to this bulk wafer shown by the solid line, which had a lot more reverse short channel effect. Interesting idea. Use SOI.

Use the property of the fact that there's a sink here for interstitials to sink out a lot of the interstitials that you implanted and get less motion of the channel doping. So this tells us right away, though, if we're doing a process in SOI-- and we can do the exact same process. We can get quite a different result in insulator in SOI compared to bulk because the channel doping profiles will not be the same. And so the simulators need to be able to simulate these effects.

And another thing I want to talk about is I just want to remind you what's the usual order in which we form the channel, the gate, and the source drain in a MOSFET, and we are doing clever things with the order of things. This is sort of a cartoon in PowerPoint. And so this is supposed to be my wafer. These green regions here on the left and right are going to represent the shallow trench isolation. So green is my STI, which is my isolation region.

And then typically, after you do shallow trench, you implant what they call SSR, super steep retrograde. It's just, in implant regions, if you're making an MOSFET, you have usually a shallow implant up here, or more lightly doping near the surface. And you have the peak of the implant a little bit deeper. So this could be a boron implant. So that boron implant usually goes in pretty early in the process.

And it's that boron implant that's going to determine your channel profile, and therefore your VT and things like that. So it usually goes in fairly early, right about here. After that, it sees the thermal budget of the gate oxide growth. That could be typically 800 degrees. So it's got to go through that diffusion. And then you make the gate. The gate is usually a very low-temperature process, and it's just etching. So that's no motion of the boron there.

Then you implant the shallow source drains and you use the gate as a mask. So now I'm doing implants of arsenic. And you're introducing a little bit of a certain number of point defects here on the left and the right of the channel. Now I put it in the furnace and put on spacers, these green spacers. If they're oxide, I do it at a low temperature. LTO goes down at 400. It's not usually a problem.

If they're silicon nitride spacers, typically nitride LPCVD goes down at 800. So I could get a fair amount of diffusion at 800 of TED, especially because I have the implant damage introduced from the shallow source and drain. So watch out for nitride spacers. 800 for an hour to make these spacers could really-- as we saw in our example, can really cause a lot of motion at that boron.

Then we do the deep source drains using the spacers as the mask, now, again, self-reliant. And then usually right after that, there's a final thermal anneal. So one idea people have had is, well, don't put the P type SSR implant in at the beginning. Why don't you put it at the end of the process? And a very radical idea is to put it in even after the gate has already been formed.

This is not being done in production, but it was a neat idea that people had in research. OK, you have to deal with all these point defects. Well, don't put the boron in until you've already annealed out a lot of those point defects. So here is an alternate process to give you an idea of how device engineers were trying to get around TED, to a certain extent, on slide 25, and also process integration scheme for forming a MOSFET.

And this was published back in 1998 by Philips Corporation called channel profile engineering, 0.1 micron MOSFETs, by doing through-the-gate ion implantation. So they were proposing a flow that goes like this. A conventional flow is on the right. Conventional usual sequence to make a MOSFET. As you form the P well, you implant the channel boron ion profile all the way up here. Then like we just said, we do gate ox. We form the gate. We make spacers. We form the source drain extensions.

Oh, I'm sorry, that's some oxide deposition. Here's the sidewall spacer. Here's one of the killers, the nitride dip at 800. A lot of TED can happen there. Then the deep source drains. And then we do the RTA. So what they were saying is instead of putting the channel implants in here, where they can diffuse during this nitride spacer step, take the channel implant and put it in towards the end after you already have the nitride spacers in.

The thing that's weird about that, though-- think about it. Now I have a topography that looks like this. Now I'm going to implant the channel. I have to make sure I give it enough energy so boron can get through the gate. So you have to calculate that energy. And in fact, your born profile now is going to look sort of like this. It's going to have this shape to it. It's going to be a little deeper here where the gate doesn't exist in the source and drain. It'll be a little shallower here.

That might be OK from a electrical point of view, but it is kind of strange. But the advantage they have is that it doesn't go in-- the only thermal step it sees is the last high-temperature step. It never sees all the TED would happen during the sidewall spacer at 800. So that was an idea they had.

One thing you might think, though-- what do you think about-- if you're an electrical engineer, what do you think about ion planting through a gate oxide? It's a little scary because that gate oxide might only be 20 angstroms thick, right? Are you going to implant a high energy ion through? It's not clear what damage takes place right at the interface between the oxide and the silicon, and interface states and things.

So although this was a neat process, I don't think it was ever accepted in production. I'm not sure people thought it was reliable enough. But they did show-- on the next page, they did show in their IDM article that they could achieve a dramatic improvement over the boron control. This is dopant boron concentration versus depth. And the black line is the reference device.

So that's the device that went through the ordinary flow where they put the boron in at the start of the process. And it goes through everything. It sees the sidewall spacer, nitride depth, all that. Lots of TED. The boron is essentially almost flat. You don't even see much of the implant, the initial implant, whereas when they did through-the-gate as implanted and after processing-- so look at after processing here-- you can see there wasn't that much diffusion at all.

So they were really able to control much better because it didn't see any of the TED the nitride. All it saw was the last high-temperature, 1,000-degree step. But as I say, for reliability reasons, I don't know if it was ever really accepted. They did show that they could reduce the reverse short channel effect. This is now on the slide 27, VT as a function of gate length.

And they have different things here. Here's a reference process, the black. That's when you put the boron in at the very beginning, the standard process flow. You see the roll-up, the reverse short channel effect. And through the gate is TGI. And they had several different doses. TGI, you see no roll-up in VT. The VT is very flat until it goes to the conventional short channel effect sort of effects. So they were able to eliminate the roll-up because they had better control over the profile. They essentially eliminated a lot of the TED in the boron profile.

I just want to mention before we finish chapter 8 some new diffusion modeling issues that are in the literature right now that are people working on today, conditions where we have very high dose being implanted. The 311 model that we talked about is good in intermediate dose regime, but doesn't really work for very high doses. So we need to model the type of damage that takes place in a very high doses. It's being investigated by people today.

Very wide energy range. There are people doing implants less than a kilovolt today. There are some really crazy people trying to do implants very shallow, and some very deep ones, even greater than a megavolt. The physics of stopping and the physics of how to model those implants is not really all that well-known in these two ranges.

We talked about pre-amorphization the substrate. Prior to introducing boron we said, well, hit it with silicon at a high enough dose that you can amorphize it, and then you can avoid channeling. But what kind of damage is produced by a pre-amorphization by an amorphous implant and what effect exactly it has is not all that well understood.

There's a whole bunch of new annealing techniques that have come out, something called a spike anneal where you take a RTA machine, you zap up the lamps really fast, and you zap them down immediately. And the wafer never even spends time at any one temperature. It just sort of goes up and down. You're accessing sub-1 second time regime. Exactly what happens during those ramps is not completely understood in spike annealing.

Laser annealing, where we take a laser and we scan it across and we only heat the area for a nanosecond. Again, the kinetics of defect evolution in the nanosecond regime has not really been very well understood. So that's a very hot topic these days. So process conditions are changing from what they used to be? New mechanisms. Well, people already know 311s. That's kind of been beaten to death, I would say, at this point. A lot of work has been done.

But end of range dislocation loops, which we never completely get rid of. And these are very important. If you do an amortizing implant, you can't avoid them. Their effect on diffusion has never been completely understood. The clustering of dopants with interstitials, which may restore a defect and may affect the electrical solubility of the dopant. How many electrons or holes you end up getting has not really been understood. So that's a new topic.

And at interfaces, what the recombination rates are at interfaces. As we put more and different materials, we're putting high K into the problem now. What happens when you have a high K interface? How do point defects recombine at an interface between a high K and silicon as opposed to SiO2? A lot of interesting new research topics that's not covered.

And I won't go through the summary in any great detail. I think we've gone through all this. I encourage you to read through chapter 8 carefully. And I think we're at a stage now-- next time I'm going to talk about SUPREM-IV in detail-- where we have enough tools that we can really understand to a reasonably high level how people put processes together to make devices and to minimize doping diffusion. OK, that's it. If you haven't signed up yet on the clipboard, I've got it up front. I'd be happy for you to sign up.