MITOCW | Lec 2 | MIT 2.830J Control of Manufacturing Processes, S08

DUANE BONING: The first problem set is out, and I should let you know that Hayden is watching out for you, because we actually trimmed off one problem from past problem sets. So this is a little bit of a revamp from the past, and slightly shorter. So the trade-off is maybe asking slightly fewer questions, but then emphasizing a little bit more the discussive responses from you on the problem set. As we talked about last time, we'd really like a little bit of textual description from you about the problem and your thought processes on those.

OK, good-- so let's start with today. If you recall, Tuesday was the broad overview, a little bit of the outline of what we're going to be seeing throughout the course of the semester. Today I'm going to focus in a little bit on semiconductor process technology, but with a real emphasis on variation in semiconductor processes. And the basic agenda here is give you a very quick-- I think it's a four-slide process summary of a typical process sequence. It's going to be a bit stylized.

Another reading assignment that is coming, but I have not quite issued yet in order to give you some room to work on problem set 1, will be chapter 2 of May and Spanos, which is a fairly nice condensed, but good enough depth and detail description of semiconductor process technology. So it is an excellent 50-page, 30 to-- well, I guess chapter 2 is about something like 50 pages-- nice introduction to and some details on semiconductor processing.

But what I want to do is set a little bit of a baseline to make sure everybody has a course picture of what is involved in making integrated circuits, since that is going to be one of the processes that we use to illustrate lots of things. And so what I'm going to talk a little bit about after going-- flying through the whole process sequence is some of the types of variation that arise in microfabrication, with emphasis on parametric variation.

And then we'll talk a little bit about temporal variation-- run to run, wafer to wafer, if you will-- and then spatial variations. And semiconductor processing is an interesting one, because there are nesting of replication of the products within additional geometric structures. So we'll talk about that a little bit. And then what we'll be doing mixed in with the discussion of these types of variation is trying to give you a little bit of a preview, some examples drawn from both the literature and some of the research that my group has been doing over the last 10 years or so.

We'll try to give you a preview of some of the kinds of manufacturing control techniques that we'll be diving in much more detail throughout the term. So we'll see a little bit early on about basic statistical analysis, trying to decide, is there something varying here or not-- basic detection and analysis of that. Then we'll talk a little bit about process modeling, especially from an empirical point of view, and a little bit of the boundary between physical modeling and empirical modeling.

I'll give you a little example of that. And then we'll talk about using those kinds of models for process optimization and some of the kinds of strategies that arise to have robust designs-- and then lastly, a very simple example of feedback control in a unit process in semiconductor fabrication.

So the following is from a book. It's a little bit dated. About 10 years old. It's actually an interesting book, because it's about statistical case studies in industrial process improvement, where almost all of the cases are really drawn from the semiconductor industry. So before they dive into those statistical methods, they themselves give a very quick overview of the semiconductor fabrication process.

So I like this example or this-- I think there's three or four slides that have a sequence-- a little bit because it's fairly simplified and abstracted, so we can go through it quickly. But the real reason I love this is there are mistakes in it. There are errors, things that don't necessarily make physical sense or don't quite-- well, a lot of this is, of course, simplification of the process to get it down to something easy, but even in that-- with that caveat, there's things that are just-- don't quite line up.

So your challenge is to try to point out to me, as we go through it, what can't be right-- even if you don't have a lot of semiconductor background. So let's see if you can catch these. Actually, let's see if I can skip ahead briefly. I'm skipping to the end of the sequence. This is step 25. What we're showing here is the fabrication sequence of building up thin films on the surface of the wafer, as well as changing the material properties of parts of the surface of the wafer, scoped in or zoomed in to trying to build a single transistor.

Now, there are people that build and sell single transistors, but really we're talking integrated circuits. At the ICC Conference earlier this week, Intel announced their 2 billion transistor microprocessor. So that's the level of integration that's out there now. We're going to focus on one. Clearly, when we talk a little bit later, part of the challenge is to build a product with 2 billion of these things and have them all work.

So this is a wonderful industry in terms of exquisite manufacturing control required to have all 2 billion things work on your circuit. But to get a picture of what's really going on, we're going to focus in on a single transistor. And there's a picture of it geometrically, and over on the right here is a very abstracted picture of the function of a single MOSFET transistor.

And the basic idea is that this can act as an electronic switch under the application of different voltage conditions so that one can control the flow of current in an on/off kind of fashion, and therefore perform certain logic functions, depending on the presence or absence of voltage on components of the structure.

The real action is, right down here in the center of the device, in the center of the transistor, you essentially have a source and a drain. These are electrically conductive, and these little blue things are, in this picture, aluminum paths or metal paths-- metal connections, if you will-- that make contact to the semiconductor-- the silicon semiconductor layer underneath.

These red regions have been very highly doped-- and we'll see the process for doing that-- that make them also much more metallic in nature. They are highly conductive. So this is how we get electrical access down to the semiconductor. The interesting part is the silicon that has been only lightly doped-- and that's what I think these little dots are meant to indicate-- to control the degree at which it will conduct under different amounts of potential applied to it from the control of a gate.

This blue-- I guess it's green on this picture-- this is the gate electrode. And eventually, it's typically made out of polysilicon, and eventually contacted also by metal so that we can apply positive or negative voltage to the gate. And if we apply in a typical NMOS transistor enough of a positive voltage, what we do is we attract lots and lots of negative charges to counteract the positive charges we're creating on the surface of the gate.

We create lots of-- or attract or generate an inversion with lots and lots of electrons, negative charge down here in this very, very thin surface layer of the silicon. So the semiconducting silicon can be made to conduct under certain voltage or bias conditions, and now current can flow from the gate-- or underneath the gate from the source to the drain.

So if we apply a positive voltage, we turn the thing on, and current can flow. Alternatively, if we apply either an opposite voltage, or no voltage in a typical MOSFET for logic operations, the conductivity of the silicon layer will be so low that current is blocked, and the thing is off.

So the basic idea here is we need to control carefully the conditions and properties of the silicon layer, and then we need to set up the rest of the geometry to be able to have access to the-- and form the structures for the source drain and the gate. So what we're going to see in this sequence is how we start from some raw material, build up the thin film layers, and differentiate both the geometry and change those properties.

So in this picture, we're starting out here with the silicon in white. We add or create a thin layer of silicon dioxide in this picture. Now, there's lots of different processes-- many of these are discussed in more detail in May and Spanos-- for creating silicon dioxide. One is deposit on the surface. The other is to actually, quote, "grow" a silicon dioxide layer.

And that is, we essentially subject the wafer to an oxygen carrying ambient, either O2 or a H2O vapor kind of ambient. And the oxygen will react with the silicon, forming SiO2, and we have a very nice stable silicon dioxide layer. It turns out that that simple process of being able to grow a very high-quality silicon dioxide layer was kind of the crux process that enabled integrated circuits in some fashion.

And we'll talk about it a little bit later, but basically, we're going to need insulating layers between conducting layers-- one conducting layer being the silicon itself. The others-- when we build multiple levels of wiring, we need insulators between them. But in particular, for the formation of that transistor, the creation of a very good high-quality, very stable, well-controlled oxide insulating layer between the gate electrode and the silicon electrode was crucial.

That forms, basically, a metal gate oxide semiconductor, or MOS transistor structure. The ability to create a MOS capacitor and then contact it on the sides, where you can invert the silicon layer was a big challenge. The first transistors were, in fact, bipolar transistors.

Shortly after the invention of the bipolar transistor, Shockley went off on his own, and also invented the MOS transistor. In the very early versions, a different semiconductor material was used-- different element than silicon. Anybody know? What were the first transistors built out of?

AUDIENCE: Germanium--

DUANE BONING: Germanium-- why aren't all of our transistors built out of germanium now? We'll come back to that. Some of them are starting to be again. The problem is, if you oxidize germanium, you get an insulating layer. The problem is it's extremely soft and unstable.

And in fact, if you oxidize the surface of germanium, you can wipe away the oxide. It is not stable and manufacturable. And so it was really this ability right here to be able to grow a very hard, high-quality silicon dioxide that was crucial for being able to build up for both the conversion over to silicon and for being able to build up complex structures.

OK, so we've got this thin silicon dioxide layer here that has been grown. Now what we want to do is start building the region where we're actually going to have the active device. So we're going to go through a sequence here that gets repeated again and again this is referred to as the photolithographic sequence.

And so what we start with is depositing a thin layer of photoresist, shown in red here, this is typically spun on. So this is a polymer type of material. It's in a solvent so that one can drop the material on, and then spin the wafer very rapidly in order to get the controlled film thickness that one wants for the photoresist.

And this is-- interacts with light in one two different ways. You can have positive or negative photoresist. What's shown in the next picture is a typical-- let's see. Is this is one positive or negative? This one's a positive photoresist, because we're going to basically transform or imprint, if you will, the mask layer here in the same fashion with the underlying photoresist. The basic idea of the photoresist is it contains photoactive compounds that underexposure to light will change the polymer linking of the photoresist. Yeah-- question?

AUDIENCE: But I understand the [INAUDIBLE] sorry.

DUANE BONING: Yes.

AUDIENCE: I understand the [INAUDIBLE] as the [INAUDIBLE] the black part.

DUANE BONING: Right.

AUDIENCE: [INAUDIBLE] black part. [INAUDIBLE]

DUANE BONING: That's exactly right.

AUDIENCE: --change the chemical structure.

DUANE BONING: That's right. The goal is to take some design, from some CAD system or whatever, that enables us to differentiate the surface. The design is used to make a physical-- well, in most cases, a physical glass plate, if you will, that has metal blocking areas-- the chromium-- that are intended to block the transmission of the light source in some regions, and allow it in others. So that's exactly right.

AUDIENCE: [INAUDIBLE] are not exposed.

DUANE BONING: Yeah. So this is actually a cutaway. This is going on for-- if this is only a couple of microns wide, the entire wafer-- or the entire chip, depending on how the photolithography-- it goes on forever. So this is just a cutaway.

AUDIENCE: No, but I mean the middle is exposed.

DUANE BONING: Yes, yes.

AUDIENCE: OK, so we can wash away the exposed part?

DUANE BONING: Right--

AUDIENCE: But it's not solidifying the--

DUANE BONING: No. OK, so that's the two types of photoresist activity. So in this case, with the positive photoresist, what happens is the light energy from the exposure system will break down some of the polymer chains in the-- in these photoactive polymer chains in the photoresist, weakening them.

We then go in and do a develop process, where a chemical solvent is used that attacks the weakened polymer chains and rinses it away. So we have the spin on resist, we do the exposure to light through the mask, then we do a develop, rinsing away the resist. And in the positive case, the pattern on the mask is transferred directly, if you will, into the pattern of the photoresist.

There is also, in fact more commonly used, a negative photoresist that actually works in the opposite way, which is anywhere where the light exposes, it actually strengthens the chemical bonding of the photoactive compound in the polymer, and makes it more resistant to dissolution in the developer-- in which case, I would rinse away everything that did not get exposed and leave the pattern which did get exposed.

So actually, there's some degree of design choice in picking a positive and negative photoresist, depending on some manufacturing control issues. It turns out, if you're developing away material, you might, in one case or the other, be able to actually overdevelop or overexpose slightly and control slightly the line width of these features a little bit better in one photoresist case than the other, depending on what you want to do.

OK, so so far, this sequence of three steps has simply transferred the mask pattern into a photoresist pattern, but the photoresist itself only temporary. Its purpose is to act as a mask to allow us to differentiate more aggressive physical transformations of the remaining wafer.

And in particular, in the following case, as shown here in step 5, this is saying what we want to do now is do another etch of the silicon dioxide layer, in this case using a plasma-- a gas plasma etch. I guess that's what the little lightning storm here is. So this is typically done in a very low pressure vacuum equipment, where we have some feed gas. We apply RF energy to dissociate the species in the gas.

Many of those species then become very highly chemically reactive, and they interact with the surface of the wafer and etch away particular species. So one would design the plasma process to preferentially etch away the silicon dioxide, while not attacking, or only very, very slowly attacking the protective photoresist layer. So what this is allowing us to do is essentially vertically etch down into the silicon dioxide layer.

Now, in this picture, what they're trying to do here is show that we are partially etching some silicon dioxide layer away, apparently leaving a small remaining level or layer of silicon dioxide. So you can see that shadow down in there? What they're trying to show in this case is they're leaving this thin silicon dioxide layer that is going to be that insulator between the gate electrode and the silicon that we talked about in the final structure.

So typically, this might be, oh-- what you might have is close to modern technology. You might have a 1/2 a micron in the-- of silicon dioxide that you want-- fairly thick silicon dioxide-- out here to act as a very, very big insulating layer separating your metal electrodes from other components.

But this silicon dioxide layer that's used between the gate and the semiconductor needs to be extremely thin. And the reason is the closer I can get my two plates of that capacitor, the more control-- well, the more capacitance I have, but the more control I have all of the silicon surface. So the harder I can turn the device on, the more current I can flow per unit area, and so on.

So in this process, apparently we're etching from about 1/2 micron-- 5,000 angstroms-- down to-- we might be at 100 angstrom. Or below remaining gate oxide layer is, in a modern process, really what you want. Everybody happy with that?

AUDIENCE: So what is the tolerance?

DUANE BONING: Hooray-- how would you do that? Yeah. What is the tolerance? Well, first off, to answer your question, you want that to be 100 angstroms, plus or minus-- well, in fact, some of the kinds of control that they quote are plus or minus 5 angstroms, which is less than an atomic layer in some cases.

What they actually mean is, on average, over certain kinds of areas. But how would you actually achieve that with the manufacturing process? How are you going to etch down through 5,000 angstroms and stop 1,000 angstroms, or 100 angstroms before you hit the silicon?

Very carefully-- no, the answer is you're not. You're not. There are some additional reasons why one wouldn't want to do it, but at present, we don't have the observability to be able to know when we're within 100 angstroms. Second, there's no way you can be that uniform across the whole wafer.

What if, on one part of the wafer, instead of 5,000 angstroms, it was 5,001 angstroms? By the time I get down to know etching 4,900 angstroms everywhere, that little 1 out of 5,000 controllability has been trans-- or variability, which is not bad. That'd be great if I got whatever that percentage is-- 0.02% or whatever.

That's amazing. But it would magnify by the time I etch down, even if I could etch uniformally. There's just no way, most importantly, to be able to know when you're within 100 angstroms and be able to stop controllably at that level. So what do you think they really do?

AUDIENCE: Etch all the way through, then grow another layer--

DUANE BONING: That's exactly right. We can controllably etch all the way through using ideas of etch selectivity. The same idea I mentioned of-- we can form the-- or choose an etch chemistry that allows us to etch silicon dioxide and not etch photoresist. We can also use to etch silicon dioxide but not etch silicon, can have very good selectivity so that one can stop controllably on the surface of the silicon.

Now everybody knows exactly where they are across every place in the wafer. One can even over-etch a little bit. You don't even have to try to set up a timed etch. If I was etching at 1,000 angstroms a minute and I started with 5,000 angstroms, I might etch for five minutes. But to be sure that I've got the same degree of uniformity everywhere, I might add for an extra 10 seconds. That way, I'm sure that I'm clearing everywhere across the wafer, even if there's spatial non-uniformity in my etch, or today's etching's slightly slower than yesterday's.

Then I can go back in and, with different process conditions-- lower temperatures, more pure chemicals, or slightly different-- primarily different temperature, or even using alternative process technologies that involve very, very short time oxidations-- rapid thermal oxidations-- I can now controllably grow up 1,000 angstroms, or 100 angstroms, or whatever I need.

So that's the first mistake. I remember that one. I think there's another one someplace. OK. So still taking this at face value, we had the photolithographic sequence imprinted onto the photoresist. We used that then to controllability change the permanent structure, the silicon dioxide. And then we remove using another chemical solvent, or in many cases, another plasma etch, a plasma ash that removes the temporary photoresist.

Then we go on. We've done that sequence. We've now deposited or grown, and then spatially differentiated to be able to create one level of the structure. And now we're going to repeat that kind of idea multiple times, and then mix in a few additional transformation steps. Next step here as we're depositing a-- our polysilicon layer. Again, this will be fairly highly doped either subsequent to the deposition or during the deposition.

It is a semiconducting layer, but if one dops it-- that is, adds elemental impurities of a particular type to play with the band structure of the silicon-- one can make that behave in a fairly metallic fashion. And in the MOS transistor, that polysilicon is very highly doped, and is a conductor under all conditions.

So we've deposited that. What geometry do we want? We want this thin gate electrode, so we're going to do a photoresist-- or photolithographic step again to do that-- deposit photoresist, expose. In this case, here, again, another positive photoresist case-- so we're blocking this structure. We develop and dissolve away the exposed photoresist, and then we're doing a plasma etch again to-- plasma etch to form the gate electrode, and then remove or ash away the photoresist.

Now, this structure right here is very, very critical. This is the gate electrode. Again, the current is going to flow in this direction along this length underneath that gate electrode in the silicon. And this is referred to as the channel length. So the channel is the structure that we're going to create underneath the silicon.

And the length of that physical electrode is very, very, very critical in determining many of the most important characteristics of the transistor. You can think of it as-- and we're actually getting closer in the dimensions that we're talking about now-- that the time of flight of an electron underneath that electrode will depend on the width or the length that it needs to transit. We're getting close to these velocity limited transport regimes.

And so from one point of view, the ability or the speed of the transistor is critically dependent on variations in that channel length. This is often referred to as the CD of-- or one of the CDs of the process, critical dimensions of the process. And in modern technology, in fact, that is perhaps the smallest lateral dimension that is determinative of the generation of the technology and the performance and characteristics of the technology.

So if we talk about a 65-nanometer technology, they're talking typically about that channel length, that minimum feature size being 65 nanometers. In fact, often, tricks are played where, photolithographically, using the exposure, that's the dimension that they can form. But you might actually over-etch and shrink that, get that to even a smaller dimension than you could photolithographically define in order to boost the speed and the controllability of the device. So this is very critical lithography step and sequence.

Going on, the next thing we need to do is start to form the source and drain regions. We need to differentiate the silicon underneath. And what we're going to do here is use another process, an ion implantation process, I believe, in this-- yes, I think that's what their little picture there is-- where we actually implant those [INAUDIBLE] atoms. We shoot them with essentially a high-energy electron or ion gun into the silicon in a very controlled fashion.

The guys who are going to Varian here in Gloucester, Massachusetts-- Varian sells ion implanters. That's their business. They're-- I don't know-- 75% or something of the market now. They're dominant, and they sell the equipment for ion implantation. There's also other competing ion implant companies, but most of them are also in Massachusetts. So Massachusetts is the ion implant capital of the world.

The basic idea there, again, is we're highly doping in a controllable fashion just these regions. Those regions we don't want to dope underneath the channel-- we want to keep that fairly lightly doped so that we can invert it for electrical action. And in this case, we're getting something often referred to as a self-aligned process, whereby we don't have to selectively shoot the ions just in here.

We can blanket the whole wafer, and use these thick oxide and use the polysilicon as the mask itself to prevent the ions from reaching the underlying silicon regions. So instead of having to put down another photoresist layer and try to open up just exactly those regions, and worry about aligning of the physical mask to match up exactly with the existing structures on the wafer, we can take advantage of the underlying structures to achieve that spatial alignment. Yeah?

AUDIENCE: Can you elaborate a little bit on the process of placing this photoresist? Seems like you're going to have one [INAUDIBLE] put another. How do you make sure everything lines up the way you need it to?

DUANE BONING: This is a wonderful manufacturing control challenge. Overlay and alignment are-- almost more so than the physical dimension that can be imaged are the manufacturing control problems. There's more description in May and Spanos. But essentially, over time, the ability to optically line up the existing structures-- typically, you might actually use alignment marks that you build up in parallel to these transistors to aid in the optical-- automatic optical recognition of spatial locations on the wafer.

It's a huge challenge to then line that up with the physical mask plate, such that you minimize offsets. There's also exquisite environmental control in the equipment that's used for this. Imagine that I don't need to line up just for one transistor-- I got to make sure across maybe a 1 centimeter squared or 2 centimeter squared chip that all of my transistors are lining up.

And in fact, in older technologies, I had to make sure the whole wafer was lined up, because I had a big glass plate that I was using in almost a one-for-one transformation, so everything across a whole wafer had to be lined up. And if the temperature of the glass plate and out of the wafer were off by a degree, the thermal expansion would upset the alignment, for example.

So over time, you've seen-- both to achieve smaller dimension, but also to achieve better control-- a number of [INAUDIBLE] driver for the evolution of the photolithographic equipment. One step was photolithographic reduction from, say, a 10 to 1 or a 5 to 1 glass plate down to a smaller chip or smaller wafer.

The second was chip-to-chip stepping. So in fact, the photolithography equipment's typically called now a stepper, because it's imaging one field, maybe one chip. There may be more than one chip in each field, but I'm stepping through in order to get the control and be able to achieve the alignment on one chip at a time.

And in the most modern equipment, in fact, they don't even image the whole chip at one time. They're imaging slices of that so that they can do all kinds of other on the fly alignment control, as well as leveling, because turns out the depth of focus is a big constraint. And if I have a focal plane mismatch, where I'm in focus on the left side of the chip and not on the right, then I don't get the dimensional control I need.

So it's really amazing innovations that are exactly driven by the manufacturing control necessary. So there's a little bit more description in here talking about some of that equipment as well. In fact, I'm not really talking here much about the processing equipment. I'm trying to get us through at least the process sequence, but the equipment itself is fascinating. It's really fascinating.

OK, we're almost there with this device, because now we're primarily just repeating these same kinds of structures. What we're going to do now is another photolithography step. Here, essentially what we want to do with this yellow layer is build up an insulating layer. So I think a new layer of silicon dioxide is deposited.

In this case, it's not thermally grown through that chemical reaction of oxide. And silicon-- because I've already got polysilicon, I've got-- well, I guess it actually might work in this case. I've already got a very thick layer of silicon in this case, we use typically a chemical vapor deposition process-- again, in vacuum equipment. But we decompose silicon dioxide carrying molecules in a gas plasma.

So there will be a more complex molecule typically coming in, and we decompose that, and basically, it snows on the surface of the wafer and builds up a silicon dioxide layer. So we deposit that silicon dioxide, and what we're going to do is now open up holes within that using a photolithography step. You can see here's the mask, where we want to open up a hole; down, where we'll be able to make contact to the gate and make contact to the source and drain.

So we go through again-- photolithography, put our resist down, expose it, develop it, plasma etch-- again, controllably etching through the oxide, stopping on the silicon layer. Remove the photoresist, and now we put down a metal layer. Now, typically, what's done for metal layers is some form of physical vapor deposition, where-- one example is a metal evaporator where the wafer will be positioned some distance away from a metal source.

The metal source-- aluminum or other materials-- will be heated up, either directly electrically or under sputtering kinds of activity, such that either individual molecules or small agglomerations of molecules become very excited, and traject through the vacuum, and build up the layer on the surface of the wafer.

So in this case, this is an older technology. They're talking about this being aluminum. I'll show you a little bit later that most of the most advanced chips are now using a copper interconnect, which has other challenges. So the basic idea is now, again, we've covered the whole surface of the wafer, and we selectively go in with a photolithography step and remove where we don't want the metal lines protected, where I'm not going to want to etch.

I'm protecting parts of the aluminum, and then I do a plasma etch or a chemical etch in order to remove the aluminum. And then I remove-- strip away the photoresist, leaving my structure. Now, this is a single level metal, very simple transistor. And building up a several billion transistor chip, or even a 100 transistor chip, we now need to also electrically interconnect all of those devices.

And there's a whole additional sequence of process steps that are used to build multiple levels of metal contact, but conceptually, they're very similar to those that we've seen so far. Here I've got access to this individual transistor. Clearly, laterally, I can have neighboring transistors that I've wired together with the patterned aluminum.

When you get a certain number of transistors, it becomes very hard to have just one level of wiring. You need to be able to have a wire that goes over another wire, and that's where we get to multiple level interconnect structures, where one would deposit another layer of silicon dioxide or other insulator and then build vertical [INAUDIBLE] down from one level to the next to contact multiple streets on level 2 with streets on level 1, and so on, up to 6, 8, or even 10 levels of metal in modern structures.

AUDIENCE: So [INAUDIBLE] are there several transistors above each other or just connectors?

DUANE BONING: Yeah-- good question. Almost all existing semiconductor chips are planar processes, meaning all of the transistors are in just one plane, just at the surface of the silicon. In fact, let me hand this around. This just has a copper interconnect test pattern on it, so-- but this is an 8-inch wafer, and the basic idea is we can get that very nicely grown silicon layer and control of the surface of the silicon layer.

It's actually quite difficult to build up multiple levels of independent and high-quality silicon layers on the wafer. There's an awful lot of research-- pass that around-- for 3D integrated circuits, where one would, in fact, build multiple strata or multiple layers of transistors, and then bond the wafers together to be able to have in a very compact form factor multiple devices in the third dimension.

You get lots of benefits of smaller footprint. At that coarse level of 3D, I would say many several memory chip makers are already doing chip stacking within one packet. So in some of the memory chips, if you were to etch away that-- the black packaging casing around one of those chips, you might find a stacking. But they are not heavily electrically interconnected.

It's not a fully integrated structure. In fact, at the end of the day, they typically have bond wires-- big macroscopic wires coming in the sides of the chip separately at each of the stacking. Now, there's other variants. So for example, many of the thin film transistor, TFT displays will actually have a layer of transistors that are a deposited layer of polysilicon, not quite the same pure crystalline layer. And sometimes those include stacked transistors. So there are some technologies. They're just not quite the same quality or the same high-speed nature.

OK, anybody detect any other obvious mistakes or errors in their pictures? Maybe it was just that one, that one inability to control of the etch down to a 100 angstroms. Now, there's lots of other abstractions in here. And one of the other neat things in this picture is it's showing-- you're building up a fairly complicated geometry using controllable processes.

One of the most important things that you have for control of this lateral differentiation is photolithography. Notice, we were never just depositing directly aluminum where we wanted it. It's always a-- deposit everywhere, because you can do that across-- in parallel across the whole surface of the wafer, and then have to go back and do a subtractive process to typically remove what you want.

There are very, very few steps where you can selectively grow or selectively add material only where you want it. So that's an example of a huge constraint in semiconductor processing. So what I want to do now, with that as the overview of the unit process and how they all stack up to build a more complicated structure-- I want to talk a little bit about of the input-output picture of semiconductor processing.

This is very similar to our generic picture that-- hopefully you guys have started reading that process control overview-- where we've got different inputs to the unit process, including the raw material, the wafer. We've got inputs in terms of the environment that we create around the wafer-- that vacuum process, the chemical [INAUDIBLE], the gas ambient used in plasma etching; and then other control factors on the machine or the facility-- things like thermal control in photolithography-- to be able to ensure that there's not thermal expansion of the components.

And then we get changes in many of these wafers, many of these states. Many of these are desirable. We saw the change in the wafer state in terms of the geometry, as well as things like doping processes that change the conductivity of the silicon layers. But we also often have the side effects changes in the state of the equipment. And this turns out to be a big deal in much of semiconductor control.

So for example, every one of those coating processes occurs in some kind of a chamber. I coat up a film. I coat aluminum on the surface of the wafer. What do you think happens to the rest of the chamber? It gets coated too. So over time, you will often get buildup of material in the rest of the equipment, and there will need to be other steps to go in and clean or reset the equipment state.

So actually, we often are worried not only about the product and what's happening with the product, but in manufacturing control, we're also very worried about what's happening with the equipment. So what we often don't have control over are all these disturbances. There may be variations in the process coming from lots of different sources. Some might be the incoming material. In this case, if there's variability in the starting wafer, or the wafer as processed up to some step, that might interact with the step that we're now performing, and limit our ability to control.

And then similarly, there may be influences of variation coming from the machine, or the facility, or the wafer environment. So what I want to do is talk a little bit about some of those different kinds of variations, and build on what we've seen here. Now, I mentioned this a little bit last time. It used to be 10 years ago, 15 years ago, if we talked about yield-- the percentage of functioning devices that came out of semiconductor fab-- the main worry limiting yield was particle defects.

And you can imagine, we're talking about very, very small features here. And if pieces of extraneous matter were to fall between, say, two metal lines that would cause an electrical short-- or could cause an electrical short-- and now the device does not function. It does not function correctly. In other cases, the particle-- if it happened to appear in the photoresist, it might cause an open in one of these lines. It might prevent-- or cause me to actually regions where I didn't want to.

So not only can I have shorts-- I might have opens between lines. And this defect control was the primary source of yield loss in the fab. That's why, if you go into an IC fab, it's a clean room. Everybody's wearing clean room garments. There's incredible amounts of airflow to filter out and remove all kinds of dust particles or other extraneous matter down to very, very tiny dimensions.

That's really the measures that have been taken in order to get the cleanliness you need in order to-- not eliminate, but get those particles down to a particle density and down to a presence of size that are small enough not to have too big of a yield impact. What I mean by size is imagine that little dust particle is only a quarter of the width of any of the minimum feature sizes on your structure.

Well, that might mean I lose a quarter of the width of one of those metal lines, but at least it still conducts. It might have a slightly higher resistance than I expect, but functionally, I'm still OK. Similarly, if it's too small of a particle to bridge between lines, I'm not going to have a short. So the idea is to control the particles down to a certain feature.

Now, there are some very interesting manufacturing control technologies that are used to characterize the presence and density of these kinds of particles. And what I pictured here is a metal layout, a metal patterned structure-- referred to as a nest structure-- that allows one to build a large number of these snaking lines, and electrically connect between these different lines in order to test if there is a bridging defect between lines, and count the number of those events.

So you build this structure to spatially sample as large an area as you can to try to capture some number of these defects. And you can also use this structure to characterize the size of these defects, especially if they're relatively large. So for example, if I had a big particle that bridged three of these lines, I know, within certain bounds, that I've got a particle between some size interval.

And so that's what's shown over here is a defect count as a function of the size, in microns, of different particles for these large area sampling structures. Now, this turns out to be still a-- an important part of semiconductor fabrication is understanding your defect densities, defect counts, understanding what the typical sizes are, what levels of the process they come in. Are you mostly getting these in aluminum deposition? Are you mostly getting these in other etch processes?

And actually, then feeding that information to design for manufacturability to optimize the layout of the circuit to minimize the likelihood of a particle falling in a bad location. And we'll talk about that, I think, somewhere near the midpoint or second-- last third of the course. We'll talk about yield modeling and some of the technologies that are used for characterizing, statistically, the point defects, particle defects, and some of the strategies used for optimizing or minimizing impact on circuits.

But what is becoming much more important-- and we mentioned this on Tuesday as well-- is parametric control. And in some cases, even these defect controls turn into parametric control problems if what I'm worried about is not so much yes-no functionality, but changes in the resistance of lines because of the presence of defects.

And so much more broadly than defects, we're much more concerned with continuous parameter control of geometry as well as some of these other properties that we saw in the sequence. So for example, that critical dimension control might be the channel length of transistors on my device, and inherently, you're going to have some statistical distribution on channel length. Call that L.

So for example, if I were-- let me start down here-- if I were to sample many nominally identical transistors that were laid out exactly the same, were assumed to behave exactly the same, and I measured some electrical characteristic or some geometric characteristic, I might get a distribution-- drawn here as a Gaussian. It's not always, but very often, that's a great model and I might have some distribution of that across the chip.

And very often, there are inherent physical limitations in the process that give rise to those sources of variation, some of which we'll come back to-- might be systematic or might be random. Now, what's interesting is we have a physical nesting. On the wafer, we have many chips.

Now, if I were to take and measure the same transistor on each chip, I would also get a distribution. I have a chip-to-chip variability in parameters, and often that will come because of different physical causes. It may be we're doing that photolithographic stepping, and every time, I have a little bit different overlay control or offset control.

So the source of this may be-- this distribution may be because of wafer level or across wafer sources of variation that are very different than the sources of physical variation of replication of transistors within the chip. And we also can have wafer-to-wafer variations. Again, if I were to either calculate the average over one wafer, or maybe just sample exactly the same location on every wafer and stack that up across 24 wafers or large numbers of wafers that might be in one lot, processed even relatively close in time to each other, there's also going to be a deviation.

And that might arise because of changes in the equipment state from one run to the next. The first wafer-- it may have a pristine equipment with no build-up on the-- of material on the wafer wall. The next wafer might see a slightly different process-- unintentionally-- than the previous one because of changes or degradation in the equipment state over time affecting that wafer state. That might be an example.

And then finally, in semiconductor fab, we typically are processing a lot of wafers. That may be 24 wafers at a time. And from one lot to the next, these might be processed at substantially different points in time. And there may be another distribution associated with that, and physical sources are causes of that variation that are different than these other spatial or temporal sources.

Some companies-- and there's not a lot of them these days-- have multiple fabs, so I could continue this up. Intel in particular is very famous for making the same product in multiple fabs, in multiple states, and in fact, multiple countries around the world. And they work very, very hard to have matching from one fab to the other, in terms of the equipment and so on, to minimize fab-to-fab deviations or differences in their chips.

They really don't want people saying, I'm only going to buy my Pentium or Itanium chip from the Albuquerque fab, because it's better than the one from Portland or whatever. They really want to minimize that. So you've got extensions up the hierarchy as well.

Now, if we were to actually zoom in on one of those distributions, what's also going to be very interesting is trying to understand what the sources of variation are at each of the spatial scales. And you might have a fairly complex distribution, say, over channel length. And what's often very, very much the case, and very interesting is that total distribution is the compounded effective multiple additional sources that are convolved together.

And a lot of work will be done in designing test procedures, designing process control charts, and whatnot to be able to track individual contributors and separate them out into the component sources of variation. Why do you want to do this? If you can identify individual sources, that's crucial to being able to go and squeeze them down.

You got to know what the source is in order to remove it. What else is the case often is that some of these are not completely random. Many of the sources of variation have a systematic physical cause. And if we understand what is systematic or repeatable every time you run this, you might actually find that the transistor in the upper right corner of the chip is a little bit too wide.

You might then track that back to the photolithographic imperfection that's causing that. Once you know that, you can compensate and counteract that source of variation. A lot of the work is trying to break these down. So let me give you a little bit more example of some of these sources of variation, and the tools and techniques that are evolving to handle them.

Let me zoom into one unit process. We've done actually a lot of research on chemical mechanical polishing in semiconductor fab. And the basic process is you take that wafer, you press it-- or you hold it in a wafer carrier and press it face down onto a rotating table or [INAUDIBLE] to which is affixed an abrasive path. You also drip a wet abrasive slurry-- so it's a very, very fine particle grit-- in a chemically active slurry, and then you rotate both the wafer and the pad, or the table.

And the basic goal of this is to flatten the surface of the wafer, if you go back to this picture, with even one level of metal, look at all of the topography, the differences in heights that are generated, because we're etching down, filling in, and so on. If one were to try to build multiple levels of metal, you have a real problem here.

You keep building up in some regions where you have more metal, and in those regions where you don't, you get thinner and thinner by comparison. That becomes a huge challenge in terms of other processes, like photolithography-- talked about that depth of focus. Now I want a pattern or image at the top of one of these levels, and I've got also these recessed regions-- become very difficult to image, say, a metal 4 layer.

And so CMP has been created to basically abrade away the raised surfaces, get it down to a planar surface. Now, one of the big problems in CMP is the process itself is often not all that stable. That polishing pad, that abrasive path-- what happens if you use sandpaper for a while on a piece of wood? Well, you smooth out the grain or whatever in the piece of wood. What happens to your sandpaper?

It also wears-- fills in. It gets gooky. Basically, it changes its state-- the process and the equipment changes it state over fabrication as well. And if you looked at the normalized removal rate for a 350-wafer sequence, you will often find drift in the removal rate. You might start out with a fairly high removal rate, and over time, it drifts down. This is something like at 20% or 40% kind of drift.

Similarly, if you looked at the wafer-- measured some pattern, some eight-sample point, and calculated a non-uniformity metric-- maybe its standard deviation of those points-- one would find-- maybe you can see it-- looks like there's a slight increase in that non-uniformity over time. The pad is not quite as uniform as well. So this is the kind of temporal wafer-to-wafer variation that you don't want, and would like to understand it. If there's a systematic trend, then maybe there are control approaches that can be used to compensate or correct for that. Yeah?

AUDIENCE: This polishing seems to be a bad idea to me. [INAUDIBLE] a way to getting, like you said, [INAUDIBLE]

DUANE BONING: Yeah. The semiconductor industry really got its start about 1960, and for 30 years, it was all about defects, those particles. Keep them away from the wafer at whatever cost. And then primarily, folks from IBM around 1990 said, yeah, we're going to pour slurries with 100-nanometer particles on it, and then we're going to rub the heck out of the surface of your wafer. And everybody thought they were crazy.

But it turns out that, first off, the control of the process can be very good in terms of not a very, very fine abrading of the surface. So you're not gouging out. You're really doing almost nanometer-by-nanometer scale abrasion. But then you still have all those particles. So a lot of the work was very, very effective post-polish cleaning processes that involve chemical baths, that involve brush cleaning, in fact, that involve hydroponic agitation to levitate the particles off of the surface of the wafer in a clean bath-- that they actually found no appreciable defect-oriented yield loss, once you did those extreme measures to clean up the wafer.

And in fact, the irony ended up being that the CMP processed, by planarizing, actually improved the defectivity. Because many of these particles might be present in layer 3, you've got a lump, but it didn't destroy the action of level 3. But if it's sticking up, by the time you go in to level 4 and you do a metalization over that, well, now you've got a problem.

And the CMP process, which would go in and flatten or planarize between levels, would actually chop the heads off of these defects and reduce their impact for subsequent levels. But it is a fascinating story of the adoption of CMP, because it was so counterintuitive of a process, exactly as you said.

So we'll come back at the end or later to control strategies, for dealing with this-- of the course. Basically, this kind of a drift-- you could start to think, well, I can measure this, and maybe what I would like to do on subsequent runs-- if my removal rate has drifted down, I'd like a strategy where I polish a little bit longer or I press a little bit harder-- these sorts of things.

AUDIENCE: So you only do this polishing at the very end?

DUANE BONING: No. You do it at each level--

AUDIENCE: At each level

DUANE BONING: --each metal level, as well as in some other places. So it's actually used repeatedly in multilevel interconnect. Here's an example of a little test chip to try to characterize some of those spatial variations, and decompose them that we talked about. This is a chip design that has a few thousand ring oscillators. A ring oscillator is a chain of an odd number of inverters, each of which does a logic zero to one or one to zero computation.

And if you organize those in a ring fashion, it oscillates at some high frequency that indicates the speed of the individual transistors in that structure. So a ring oscillator is a wonderful stand-in electronic circuit telling you a lot about the speed of your technology.

What we did is built a tiled architecture, where we've got thousands of these structures; replicated spatially, so now we can measure multiple instances and see if there's a spatial pattern; well as we can also play with some of the design parameters of each of those ring oscillators. And in particular, we can do things like take that silicon gate electrode and split it into three of them.

We can play with the spacing between those electrodes. We can orient those [INAUDIBLE] vertically or horizontally on the layout. We can also look at the effective nearby structures. For example, how dense per unit area do you have transistors or polysilicon gate electrodes? Would that have an effect on the ring oscillator?

And now, if you go out and build this, this is a wonderful kind of data, where you can aggregate or map out the data at these different spatial scales to look for trends. So for example, if we look at the wafer scale variation in the ring oscillator speed, we've got something like 30 chips in the upper half of one of these wafers. And mapped out here and the color is the average ring oscillator speed from high to low.

And you can see a spatial trend from one side of the wafer to the other. This is telling us there is a wafer level non-uniformity, and in fact, that across wafer variation is about 9% in speed of these devices. You can also then start to look-- I guess that's the same thing-- you can also start to look and say, depending on how you designed the transistor, what kind of offsets do you get in the frequency and what kind of variation do you get in the speed of the devices?

And what's interesting here is that many of these ring oscillator designs are-- from the circuit designer perspective, they think of them as identical. But in fabrication, because of neighborhood or nearby structures, in fact, there are slight deviations. So nominally identical transistors may, in fact, have substantial systematic, highly repeatable effects. And if the manufacturing line can capture and characterize those, and send that information to design, design can compensate.

So this is another theme. This design and manufacturing link can be very important. We can also map out identical transistors within one of these chips. And here again, you can see a top-to-bottom variation in the ring oscillator speed. And in this case, that within each individual chip is about 4% in this particular case, which is quite interesting.

Notice we said there's 9% across the whole wafer. I might have 10 chips lined up in a line across the wafer. If I were to say, from center to edge, I have a systematic trend-- or from one edge of the way for the other, the systematic trend of about 10%, and I map that down to 1/10, because I've got 10 ships, that would be about a 1% variation, you might expect, because of wafer level sources on each chip.

But what we find is something like 4% That's telling us there's a different source of physical variation at work within each of the chips. And then you can also look for other systematic effects-- things like the proximity of one [INAUDIBLE] to another can have photolithographic effects. And in particular, that can result in an offset, depending on whether I have 1 micron, or 2 micron, or 2x, or 3x spacing between the devices.

But in this case, the variation across many, many replicates of each of those is about the same. An interesting effect is, if we take that same [INAUDIBLE] that single gate electrode and split it up into two, or split it up into three, and then look at the effect on ring oscillator speed offset and variance, what do you think we're going to see? We didn't know, so you may not know. Yeah?

AUDIENCE: Probably see four variations in the [INAUDIBLE] one because there are more edges.

DUANE BONING: Right, you have more edges. And that's exactly. If we look at the sigma with one finger, two finger, or four finger, we found a proportional increase in the variance, with more opportunities for variation by the fingering. So that was an interesting implication for some design kinds of effects.

OK, so that's an example where design of experiments could be used to come up with test structures, and characterize, and split out some of these sources of variation. Now, I'm not going to go into the full detail on these slides. I put more slides in here than I needed because this is a current area of my research and I get really excited about it. But it helps explain a little bit this multilevel interconnect.

I'll cover the first one very briefly. There's two sequences that I didn't talked about in the earlier flow, because it's a 10-year-old process. Modern integrated circuits-- especially for microprocessor logic, and now even memory-- are using copper interconnect. Copper conducts better than aluminum. You get speed improvements. You get reliability improvements.

It's a big drive to move to copper. But it's a little bit more complicated of a process. The basic process is different than an aluminum, because we do not have an effective plasma etch process for the removal-- selective removal of copper. It's just a bad process in that way. So instead, what we have to do is go through the following sequence.

We go to photolithographically etch in our insulating layer where we want, ultimately, the copper line to reside. Then we fill in copper everywhere. We happen to do that using an electroplating process, and that electroplating process results in nice filling of features of different sizes, but it also overfills. And even more problematic is it overfills to different amounts or different thicknesses depending on the feature size that we filled up. So you get this very complicated topography.

Now, I described CMP as wonderful for flattening and removing and abrading away topography. So the second step is using CMP to remove all this excess copper. And it will polish away. And ideally, you would like to stop exactly on the surface of the oxide and have your nice remaining copper lines where you wanted them.

But it also has a geometric imperfection arising from over-polish, where big fat wide lines-- the polishing pad-- that polyurethane pad-- actually digs down in and over-polishes certain features. And in regions where you've got a lot of copper and very little hard supporting oxide, it actually erodes down in appreciably. So these are some systematic effects. They are manufacturing limitations, and one would like to model and characterize those. And in fact, that's what this waiver was for.

What we came up with was a test structure methodology where you would put lots of features of different sizes on a wafer, run it through the process, measure the results-- so this is a measurement of the electroplated profile for different feature sizes, like those little lines and spaces between the lines of quarter micron, quarter micron, versus say 100 micron and 100 micron lines. And then you could measure the surface topography and see how much overinflating or how much underfill you might have in each of those regions.

Now, this is kind of a neat example for something that we'll be learning about later in the term, which is empirical modeling. This is an empirical process, where we run the wafer through, we take measurements, and then we developed in our first-level model response surface models that are basically power models as a function of line width and line space. So you can empirically fit a model that describes how much overplating height or what the step height might be.

And then one can use that to characterize step height or the amount of bulge as a function of line width and line space, and give some feedback to the designers saying, hey, if you design in these ranges. This is what's actually going to happen, as opposed to what you thought was going to happen. And so that's a little response-- surface model kind of example.

And then one can also build those perhaps into design tools that could predict for a whole chip-- maybe a product chip-- what the electroplated process thickness would be or what the CMP process would be. I'm going to skip these. This is essentially doing the same thing for CMP, but the idea here is this linkage, again, between design and manufacture, where the manufacturing fab can characterize what the actual amount of dishing or erosion is going to be for very complicated product chips.

And that information could get fed back to the designer so that they actually know, in this location, my final copper wire is going to be this thick, not this thick. And that could be very important, because the thickness of a copper wire determines its electrical resistance. It even affects the coupling capacitance between to neighbor wires. So those key electrical parameters-- resistance, capacitance-- that determine interconnect speed depend on that critically.

So that linkage with robust design can be very important. One can also use those kinds of models to play with process parameters, and empirically, as a function of the process conditions, minimize these systematic sources of variation. And one approach that is dominant in a number of industries is a capturing of these dependencies on design features, like line width or line space, and what is a good region to be operating in in these ideas of design rules.

And so in fact, this is some work from LSI Logic looking at the copper interconnect space and saying, OK, what region, what limits, what combinations of line width and line space can you have, and where are the boundaries because of some of these different physical effects? So for example, photolithography says I can't patent smaller than certain dimensions. Dishing limits, that dishing down into wide features, says I can't have features wider than 10 microns. I'll dish them too much. So this is another way of coupling between design and manufacture.

And then the last thing I've already alluded to, but coming back to the CMP problem, we also can develop-- and we'll talk about response surface or empirical models that say, OK, here's what the removal rate is, or the non-uniformity is as a function of some of the process conditions, like the speed of rotation, or the pressure or downforce of rotation.

And once you have a basic control model-- for example, the normalized removal rate is a function of downforce and speed-- I've got some input-output empirical relationship between these-- I can now use this in a run-by-run control kind of feedback situation, where I make a measurement on every wafer that comes out-- what the average removal rate was-- use that together with this empirical model, perhaps do a model update to track the changing state of the equipment and pad as-- it also gets perturb from run to run-- use that do model to say, for the next wafer, change the recipe slightly.

Polish a little bit longer or press a little bit harder to compensate for that known characterized variation source. And so this is just an example. You can imagine that, using these simplified models, one can have some compensation strategy that brings the removal rate back up to target.

And so we'll talk about that in the latter part of the term as well. These are just examples showing, for different numbers of wafers, you have to actually change the speed or the pressure, downforce, but you can get very good control of the output. So to wrap up for today, this was a whirlwind tour through the semiconductor fabrication process, and I hope you have a feel for the myriad different kinds of variations that are out there, and a glimpse forward to some of the techniques that we're going to be developing through the term to deal with this.

So we'll see you on Tuesday. Dave will be giving that lecture and talking about the rest of the processes out there in the world besides semiconductor fabrication, and building up some intuition that we'll use in those processes. So get started on the problem set as well.