MITOCW | Lec 10 | MIT 2.830J Control of Manufacturing Processes, S08

DUANE BONING: OK, so today is a little bit different. Today we're going to talk about yield modeling. And this is unabashedly connected to semiconductor manufacturing-- although I think many of the things I talk about here are more widely applicable, especially to anything that's large area manufacturing.

So for example, a few years ago, I did a roll coding process, rolling out these giant plastic sheets of film material at Kodak. And defect modeling, and the yield of-- on a per area basis was a big deal there as well. Similarly, many of the MEMS processes, thin film processes, yield modeling associated especially with defects is very important.

Many of the other ideas I'll be talking about here in this lecture are also connected to the idea of assemblies or systems that consist of many, many, many, many parts, whether it may be a failure of probability or deviations in individual parts. And part of the question is how those aggregate across the whole system.

So all of the examples here will be pretty much drawn from semiconductor manufacturing, but I think they are more broadly applicable. And perhaps many of the tools actually developed here in semiconductor manufacturing can be used and propagate to other processes. So the material will mostly be drawn from chapter 5, so on-- need to add a note as a reading assignment. This is drawn from [INAUDIBLE] chapter 5.

But I'm also showing some examples from a couple of other papers, and I'll put those papers on the website as well. I realized this morning they're not up yet. One is a paper-- sort of a classic paper by Stapper on integrated circuit yield management, yield analysis-- and then a more recent one-- I guess somewhat more recent-- 2000-- on predictive yield modeling. So both of those will be available on the website.

So we've already talked a little bit about some of the kinds of variations that lead to-- ultimately can lead to failures, or failures in specification. When a deviation in some continuous parameter exceeds some spec limit for normal operation, we can think of those as parametric failures.

So we've talked about things like line width or so on that might lead to either direct functional failure-- meaning it really won't work because the parameter's too far away-- or a more fuzzy failure-- this continuous quality loss in performance, meaning that I've gotten such a deviation, the thing might still turn on or might operate, but it does so with decayed or degraded performance.

In addition to that, we also want to talk about random failures. And these are generally thought of as more uncoordinated-- or uncorrelated random failure in some element. They aren't necessarily due to some continuous parameter, but maybe more of a lumped failure. We'll talk in particular about area dependent kinds of failures. In semiconductor manufacturing, the main source of these are point defects associated with very small particles, dust, or debris that interfere with the operation of an electrical element, generally.

And we previewed some of those in one of the first lectures. So we're going to talk about these kinds of defects. So the key idea in these is many of the area-dependent failures are ones where, depending on the total square area of your circuit, you have more or less opportunity for those kinds of failures. So it becomes an interesting problem in terms of analyzing the probabilities associated with failure for different sized circuits. And we'll talk about those.

So here's an example I pulled out of the [INAUDIBLE] paper on an integrated circuit yield tree-- so looking at, say, 100 ASIC chips that are manufactured, and what the breakdown of those might be in terms of their ultimate fate. And in this particular process, out of the 100 chips, about 30 are ultimately shipped to the customer. So in some sense, you've got a yield of 30%-- not great, but may not be entirely unrealistic.

And then, within that 30 chips shipped to the customer, there's already some parametric variation going on that is essentially a reflection of that kind of quality loss degradation that we talked about before. Typically, chips are tested, and we'll do a breakdown, give you a feel for the kinds of tests that are done.

But at the end, they're often tested for a few key performance parameters-- in particular, speed. And then a thing called speed binning is done, and you can see here, binning down into three different speed categories, where you've got a few that are operating at 400 megahertz that-- presumably, you can sell those chips a little bit more.

350 megahertz you might not have quite the same price premium on, and maybe you have to sell with almost no-- or very limited profit the 300 megahertz chips, or something like that. So there's already still the driver and process control to get as tight control as you can and push the speed limit as much as you can.

But what we want to talk about especially today are some of the sources for the chips and sources of variation and kinds of failures that are affecting the chips that are rejected. And we can break those down here on this chart into these other categories. We've got growth functional fail, unrepairable cache, speed less than 300 megahertz, and all others.

So out of these, first off, the speed less than 300 megahertz-- that's just our cut-off on the speed, and that's probably more of a parametric variation. We can break that down and start to look at what devices, or what components, or what's responsible for the slowness perhaps to improve either design or manufacturing control.

And here, for example, the clock speed perhaps is a little bit too slow, either because of the interconnect-- the interconnect delay might be a little bit too long-- or because the transistors, the active devices are not quite strong enough. Knowing those two could tell you a lot about the sources of variation.

For example, if it's interconnect, it's probably something with your back-end process. Some of your resistances in the interconnect wires might be a little too high, or some of those capacitances-- whereas, if you've got an active device strength failures-- the devices are too slow-- that's likely something to do with, say, channel length, or perhaps gate oxide thicknesses a little too thick. So it can start to lead to good knowledge that can give rise to some improvement efforts.

Now, a little bit more interesting for today are these two other categories-- this gross functional fail and unrepairable cache. The unrepairable cache-- this might be an ASIC chip, and it might have different components or different regions on it. Some of it may be random logic performing particular combinatorial or combination logic functions.

But another component is likely to be embedded memory. And in fact, as we get to larger and larger chips and integration scale, most of that additional area these days is going to memory. So I don't know what percentage of the new 2 billion transistor Intel chip is cache, but it's--

AUDIENCE: 90%--

DUANE BONING: 90% is cache-- something you can do with the area that helps with performance. But very interesting issue here is these are among the most dense, most tightly packed and smallest scaled structures-- highly repeated transistor SRAM cells, little memory cells. But there you're already expecting-- and we'll talk about the sources of some of these-- you're already expecting some number of those memory cells to fail, perhaps because of particle-oriented defects.

And so one builds in a certain amount of redundancy into the cache so that one can detect-- or into the memory so one can detect particular failed cells and program in or fold in additional redundant capability. So that's the repair, the direct repair of the cache.

But at some point, depending on where the failure is, you may not-- you may have too many of those failures, or you may fail perhaps even in some of the redundant switching in the circuitry. So you get to a point where you can't repair all of those caches. And so you might start to look then inside and start trying to say, OK, what are the sources there? Some of those you might not know-- these unobservable root causes. Others-- a big chunk maybe due to these points defects landing in particularly damaging locations.

And similarly, if you then look at growth functional fail, you just try to-- you can't even test the memory. The chip simply won't power up or won't operate. One might then also do different kinds of inspection approaches to try to detect what the source of those errors are. And again, random defects are typically a really large part of that.

Systematic defect we'll talk about a little bit later as well. Those might be not quite random point defects, but some other failure that's not quite a parametric failure, but it's something that's affecting an awful lot of the components all together. And a typical example here might be things like overlay between different layers of the process.

So everything's just a little bit off set in terms of the alignment from one layer to the next, causing a substantial yield loss in lots and lots of components together. So what we've bolded here in that big black rectangle are these random defects. And that's one of the key elements I want to talk about here are these point defects associated with particle-oriented problems, and how to model the impact of these basically from a statistical point of view.

But getting back to the sources or the mental picture for these kinds of defects, there's a-- you can be a little bit more careful with the terminology here and talk or differentiate-- talk about the difference between, say, particles and defects. So a particle-- you can think of any kind of foreign matter that might be sitting on the surface or might be embedded in a layer on the chip. Now, some of these might be benign, so a particle is not necessarily a defect.

A defect would be when it affects some functionality of the chip. So a picture here, qualitatively giving you this sense-- we've got three different dust particles on a particular feature on-- lined up with particular features on the mask. And one can easily imagine that perhaps this particle, if it's conductive and those crosshatch areas are also conductive-- or say, metal lines-- that's going to be a problem, potentially.

Actually, if someplace else, these two things are the same wire, maybe it's not a problem. So it's actually kind of interesting. It depends on both the particular layout and on the location of a particle, whether it will lead to degradation or functional failure.

You can also imagine perhaps this structure, this particle might be a defect. In what case would this be a defect? Your initial inclination might be, if I'm just looking at this layer, it's not necessarily bridging from this wire to that wire. So why would that be a problem?

AUDIENCE: [INAUDIBLE] capacitor then [INAUDIBLE]

DUANE BONING: Right. So if it's conductive and these two are conductive, I might have some additional capacity of linking even within that one layer. But I'm trying to give hints, saying one layer here. Yes?

AUDIENCE: This direction.

DUANE BONING: Yes, in the z-direction-- remember also, we've got build up of many, many layers. Remember that stacking of the interconnect layer. And even within the device layer, there's many-- so it can propagate potentially bridge or short or cause deviations in the next layer of processing as well. How about this last one here? Is that a problem? I'm seeing maybe, yes, yes, probably. Why do you think that's a problem?

AUDIENCE: [INAUDIBLE] the resistance on the [INAUDIBLE]?

DUANE BONING: Well, if it were conductive, it might not reduce the resistance-- or increase the resistance.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right. Yeah. So a couple of examples there is certainly, if it's non-conductive, what you've got now is increased resistance in that segment right there, which can also cascade to reliability problems. A well-known problem is migration, where basically, the electron flux of high current flowing very, very high current density up in the 10 to the fifth to 10 to the sixth per centimeter squared kind of amps centimeter squared can actually cause the metal atoms to move.

So they will migrate with the current flow-- or with the electronic flow, and you'll get-- you can very often get voiding, especially in these particular locations, where the wire gets thinner and thinner, and ultimately, in fact, may be an open. But again, maybe it's OK. Maybe you've got enough latitude and you're not really pushing current through there.

So one side message here is you never want particles. You always want to minimize the number of particles and the opportunity for failure. But the other message, of course, is how bad they are kind of depend on particulars of your circuit and your specifications. So we'll actually talk a little bit about some of the tools that have evolved to be able to analyze some of those sorts of things.

So I've talked a lot about opens and-- or I guess short circuits here. You might lead to an open circuit, in terms of losing a conductive path, but there's also a lot of other kinds of failures that-- where we're dust particles or other defects might impact the device operation so that you get failure. It could be, for example in an active device, where you've got perturbation of transistor parameters, not necessarily just a short or an open.

So what is yield? Very qualitatively, yield is the percentage of parts meeting some specification-- set of specification. But what we often do in IC yield terminology is look at different points in the process-- flow or different points of testing-- and we'll differentiate the yield in some cases that way. And the other is we'll sometimes break down what kind of yield losses or what size of a bucket we're talking about for thinking about yield.

By size of a bucket, what we've got in semiconductor manufacturing is a-- and many other manufacturing process is a very hierarchical-- spatially as well as temperately-- structure. What I mean by that is we've got within the fab many different lots of wafers. Each lot maybe is 25 wafers being processed together. Within each lot, I've got-- of those 25 wafers, I pull out one wafer and I've got, what, 50 to thousands of chips on it.

So I can talk about yield in terms of, well, what fraction of lots make it through the line? It's possible I might scrap a whole lot of wafers. Or what fraction of the wafers within the lot make it through? I might have a dropped wafer, and you might break it, or other kinds of large-scale mechanical failure along the line.

You're not going to scrap the whole lot if one wafer breaks, but-- so the actual percentage of wafers that make it to the end of the line. And that's just, mechanically, have the wafers made it to the end? Then you can start looking and saying, OK, what fraction of those wafers appear to be coarsely are grossly within spec?

And then similarly, you start looking at the die and say, which of the die are likely to be able to function? What percentage of die yield do I have? Before I go and I invest the two hours of intense burn in and testing for each of those chips, and packaging of each of those chips, I might then also want to do some on-chip electrical testing.

And then, once it's packaged, I might do a full functional testing at speed to try to determine which fraction of those are working. So we've got some of these different terminologies here. So wafer yield is just-- actually, let me go to the next picture. This graphically defines some of these different yields.

We've drawn the wafer fab itself, the sequence of the processing steps shown up above. And very often, inline tests are being performed all the time-- not necessarily on every piece of equipment or test of the wafer after every individual fab step, but one will be making measurements at various points along the flow to check on the status of the wafer, as well as check on the status of the equipment.

One might also have some amount of real-time measurements actually on the equipment itself. And a whole additional problem is correlating equipment measurements to what's happening on the wafer-- that's especially useful for debug, but it tends not to be used directly for yield calculations. So it is possible that, as you're coming along here, you may scrap out wafers based on some of the inline tests.

Then what's often referred to as the back end-- although it's kind of confusing, because front end and back end can mean different things, depending on who you're talking to in an IC fab. If you're within the IC fab, often they'll talk front end processing as the transistor formation and back end as the interconnect formation.

But once you emerge out of the fab itself into the testing, that's also referred to as the back end. Once the wafer has finished its fabrication, and is now going both in the testing and then dicing up into individual chips for packaging, that's also referred to as the back end. So we've got some wafer fab. We've got wafer yield that maybe those wafers that make it through the very coarse electrical test.

Then you do a more detailed functional test on each die, and as pictured here, the die yield or functional yield would be those fraction that are making it through. Those set of a little bit more elaborate functional tests-- now you go ahead and package up those-- just those chips that are successfully passing those tests, and then you might do a binning or parametric test on all of the chips.

There is another additional failure point that's really important. And this distinction between yield and reliability is a little bit fuzzy. We've got die yield as being-- the full parametric die yield as those chips that meet all the specifications. You ship them to the customer. They go into parts. And three months later, they fail in the field. That's also a yield loss, in some sense.

It's more described as a reliability loss, but that's that field loss can be really bad. You really want to avoid that, because the costs associated typically with dealing with in-field failures is very high. What's interesting is, very often-- and I'm not going to talk too much about it here, but what's very often the case is there is a relationship between reliability failures and yield loss sources back in the fab.

The intuition is not that hard to imagine. We even talked about it back on in this picture. If I have a low yielding process because, on a critical metal layer, I've got lots of point defects leading to these kinds of problems, it might survive through the test, but that problem of, say, migration might be more prone to occur in the field. So in general, actually, yield problems in the fab can actually be a great warning signal that you may have ultimate reliability failures.

So what we want to do is try to get a handle on ways to model and understand the yield loss, and be able to make some predictions not just for-- based on historical data for product A, but also make some predictions for product B on your line-- what you might expect the yield loss to be.

What's interesting here is we've done so much with the Gaussian distribution-- this is a great case where the normal distribution is generally not the operative distribution-- that the probability functions of the binomial and Poisson statistics are typically more at work with random kinds of point failures. So we'll review that just real briefly.

And then what I think is really interesting is the spatial nature of some of these kinds of defect processes, this area dependence. So we want to talk about how-- what some of the basic modeling approaches are for area-dependent failures. So earlier in the semester, we already talked about the binomial distribution. I think this may be the same slide, or almost the same slide. Remember, the binomial distribution is kind of a nice one dealing with just this notion of success or failure.

So if I have a point defect that leads to success or failure with some probability p, and I've got now lots of opportunities for that failure-- n trials for that particular failure or success-- then we can count up or associate the probability of some number of successes x using a binomial distribution.

This very often is, in fact, probably the most important underlying function-- it and its approximation on the next slide-- for thinking about ways to aggregate when I've got multiple opportunities or multiple structures, and I want to estimate, given my probability, that any one-- say, any one chip on a wafer is bad, assuming they were uncorrelated, and just due to random defects, what is the probability that, in a wafer with 100 chips on it, I would have 95, 96, 9-- or better number of chips actually coming out?

What's my probability associated with the yield of at least 95% on that? And that falls out directly from a binomial distribution. You could add up then the probabilities of F, with x being 95, 96, 97, 98, 99, 100, and there you go. Now, the Poisson distribution we also talked about, and this is a good one when we have in particular very large numbers of opportunities for failure, but the failure probability for any one of those occurring is exceptionally small.

So we'll talk about this, especially when we're talking about, say, the tens of thousands or millions of devices or structures within an individual chip. Typically, binomial, the probabilities of failure for any one chip are fairly large, and the numbers of chips on the wafer are fairly moderate, so you can directly use the binomial. But when we start to talk about very, very small probabilities of failure, the Poisson ends up being very, very interesting.

And this is the case also where you start to not just think about discrete failure opportunities, but it's more useful to think about a failure rate lambda. So for example, what-- if you have a particular failure rate per unit area, just as with queuing systems, you have an opportunity for arrival per unit time.

Now think, I've got the opportunity for an arrival of a defect per unit area. How does the probability then, for different areas, give rise to the different probabilities of certain numbers of successes or failures? So this ends up being very useful for defect-oriented, area-dependent oriented modeling, very often.

It's also great for any other case where you've got just a large number of discrete failures. And a typical one that we'll talk about are things like via yield failures or contact yield failures. These are the little electrical connections from one layer to the next in the wiring, or from the wiring down to the transistor level. You can imagine any one metal layer may have millions to perhaps, in some layers, billions of these.

The opportunity for failure of any one of those is exceptionally small, but you've got a heck of a lot of them. And so then you might want to ask, what's the probability of having perfect operation within that? So I saw a hand somewhere. Was there question?

AUDIENCE: No, you answered [INAUDIBLE]

DUANE BONING: Great, great-- OK, so we're going to be using those. Here is the via example. We could use the binomial distribution. Well, first, let me give you a couple of definitions here. So we're looking, say, at one particular metal layer, and the probability of failure for any one via is exceptionally small. We'll call that p sub v, probability of failure for that.

However, again, we have n opportunities or n vias in each layer of the chip. So you might want to ask the question, well, what's the probability that I have one via failure, then I have 10-- I have failures in some range? Or ultimately, you want to really ask the question, what's the probability I have zero via failures, so I don't have any wiring problems on that chip? One could go ahead and directly use the binomial distribution.

Alternatively, what we can start to think of as, what is a failure rate or the average number of total via failures, lambda v, for those vias on a layer? So here, we're essentially-- what the heck happened? That's supposed to be an equals. So this failure rate is simply the product of the opportunities in the individual failure, which gives you per chip now this failure rate. What is the average number of failures-- via failures per chip for that layer?

And now you can use the Poisson distribution, again, because the conditions of very small P, very large n. And just plugging in, we've got this expression here. And again, what I said is what we're really interested in is the probability that the whole chip is good, that none of these via failures are catastrophic. None of them occur. And so I'm really looking for the probability that x equals 0-- I have zero via failures.

With an average number of failures of three per chip, I'd like to know, well, what's the likelihood then-- what's my probability that the whole chip is good? It's not zero. It's not 100%, because on average, I've got three defects, or three via failures per chip. But I've got some out there in a tail that are still going to be good. And so even with non-zero failure rates-- fact, it's hard to imagine that you have a full zero failure rate-- you can still have good chips. Now, of course, you'd like lambda to be perhaps less than 1 on average.

But now we can basically use Poisson statistics to aggregate and calculate, given individual failure likelihoods or failure rates, what the probabilities are that the whole assembly works. Question here--

AUDIENCE: [INAUDIBLE]

DUANE BONING: Right. Right. In fact, what this has already done is multiplied by the area, and so the area multiplication here was per chip. So you ultimately get to some per unit unitless. So lambda is unitless in this case. We'll see other examples when we do some other area dependencies, where you might be looking within the chip, and actually explicitly adding in or calculating some area, and then multiplying the failure per unit area times the area that you're sensitive to to get to a lambda-like parameter. OK?

This is just a little example-- I'm actually not going to go through it-- that's just working through the two cases for the binomial and Poisson distributions-- particularly in the case when n is large and pv as small. So this is just looking at the particular binomial distribution when x is 0, or the Poisson distribution for x equals 0 for no failures in the two cases, and just showing that, for small lambda or for small pv, they both go to the same approximate result.

So we have our simplest yield model. We have the binomial distribution, which you might use, for example-- aggregating chip yield. We've got via kinds of individual failure models-- so per component or a failure rate, and how to aggregate those on a per unit or per area basis.

But I do want to get, actually, to exactly the question you just asked. How do we get our minds around the situation when I've got those little dust particles falling on some area, and I'm trying to understand the area dependence of the circuit? And so what we're going to do is actually want to build a yield model that's a little bit more broken out, that explicitly allows us to make predictions based on the area of the circuit, the area of opportunity for these failures, and a defect density or knowledge about the number of defects on average per unit area that we are likely to have.

And the reason is, if you think about it, the chip gets bigger and bigger. It's got larger area. And if it's-- only takes one defect to fail, the larger it becomes, the more likely that chip is to fail. So one key driver in this that interacts a little bit with design is, how big can I make the chip without incurring undue yield loss, just because I'm going to have some likelihood of defects per unit area?

So we want to understand that interplay. We'll start with just overall area, but quickly get to this notion of a critical area on the chip, which is really just that area where the defect has to fall or a particle has to fall in order for it to actually be a defect, and cause an electrical open, or a short, or some other fault, some other failure in the circuit.

So how might we go about modeling these? Well, first, to help with the mental model here, with spatial defects, we're going to make, in the simplest yield model, a few assumptions. And then I'll show you, over the course of time, some of the improved versions of these defect-oriented models that have arrived that account for a little bit more-- or additional effects or relax a few of these assumptions.

So what I've pictured here-- [INAUDIBLE] it does show up-- is a wafer with some number of chips on it-- I don't know-- 100, 150 different chips, and a splattering of a few little red particles. These actually are defects. Each one of these red particles falls into place that causes a failure, some kind of a short. They're big enough that they actually sort things out.

And you can start to see, essentially, an assumption here is that each one of these defects corresponds to killing one chip in this simple model. Some other assumptions are they are, in fact, randomly distributed by Poisson kinds of statistics. They're also randomly spatially distributed. Knowing where one defect is tells you nothing about where another defect is. So they are spatially uncorrelated.

So those are some of the initial assumptions, and under those assumptions, what has been observed is a very interesting or natural relationship between the density, d0-- the average number per unit area of defects-- in this case, I've got, what, 1, 2, 3, 4-- eight defects here per the total unit area of the wafer, and the number of or percentage of chips that fail, depending on the area of each chip.

And it's pretty obvious, especially if I go to extremes. What if the area of my chip were the area of the entire wafer? I had one chip per wafer. If I had eight defects on average per wafer, that means pretty much every wafer, every time, I'm going to for sure have most likely at least one defect, and my yield's going to be extremely low.

At some point, though, my chip size gets small enough that this assumption of every defect killing only one chip is a very good one. And then I saturate out to basically a relationship that's very close to just being determined by counting the number of defects I have per unit area. And what was done-- and this is either in the Stapper paper or referred to another paper from Stapper-- is very early on, this dependence on the chip area and that percentage functioning was observed, and it was observed to be exponential.

So this is empirical observation that gives credence to this notion of the Poisson statistics are really what's at work, that exponential dependence on area. And what he basically found is that, as the chip area in square millimeters went up, the yield went down. So when the chip area was small enough-- very close to 100% yield.

And then-- this is on a log scale-- notice, this is on a long scale. So there appeared to be roughly a-- on the log scale, a linear decrease in yield as the wafer area-- or excuse me-- as the chip area increased-- so that kind of an exponential dependence.

And so very early on, the first model that was really used was a Poisson defect model that basically treated each defect as a point-- said, again, these same assumptions. Each defect results in a fault, and these things are spatially uncorrelated. So then what you can really do is start to say, for any circuit, any chip, with some critical area a sub c-- so that's the area within the chip that's sensitive to the falling of these particles-- maybe ac is equal to the whole chip area, maybe not-- and some defect density, then the yield is simply exponential-- e to the minus ac times d0.

And recognize, that ac times d0-- that gives rise to something like a lambda parameter, a failure rate kind of parameter. Now, he did a little bit of further breakdown here, which I'm not going to go too much into. This actually distinguishes between a given circuit and then the whole chip, which might have n circuits on it.

Looking individually at each critical circuit on the chip, you could say for that particular circuit-- the wiring pattern, say, of that particular circuit for a metal layer-- what is the critical area a sub c for that layer? And you can get the yield statistics for metal layer 3 for circuit-- maybe it's the adder circuit in the upper left corner of the device-- or the chip.

And then, if you have n circuits, each with a critical area A sub C, for all of them to work, you've just got a multiplicative probability so that your yield is a multiplicative yield factor for all of those individual circuits. So you can read this as your yield for an individual circuit, just to the n-th power.

And so what they're doing here is just simply aggregating and saying the total critical area might be across all of your circuits. If each one of them had equal area a sub c, the total area would be just the product n times ac. Or you might do a summation, might simply add up all of the different critical areas.

Now, an expansion on this starts to pull in a little bit more statistics. And in particular, one of the really interesting statistics is an observation that not every wafer observes the same defect density-- that there, in fact, is a probability density function associated with the defect density. Some wafers might see larger numbers of defect per unit area. Other wafers may see fewer.

In the natural operation-- you've got the fab as clean as you can make it-- there's still a range of different defect densities you expect on any one wafer. So the first extension is to characterize the probability density function associated with defect density-- just number of defects per unit area. And now you can integrate up what your expected yield is, accounting for the fact that I've got a whole range of, or a whole PDF for different defect densities.

And so the first extension here is referred to as the Murphy yield model, discussed, again, in [INAUDIBLE]. And all we do is we have, for any given d, we have our Poisson yield model, and then I'm simply averaging that over my PDF. So I'm integrating that over all possible defect densities.

Now, we can get back to the Poisson yield model, and now we actually recognize that that's the special case when we assume there's only one defect density, and it applies to every wafer. That is, our PDF, our f of d, is just a delta function. All of the defects are at d0. So we can recover and get back to our Poisson yield model.

But what's interesting now is, depending on the statistics associated with defect densities, I might end up with different final yield formulas. And so a number of different-- whoops-- yield distributions, PDF associated with defect density have been explored, and then some empirical fits done to data-- yield data to try to see which matched a little bit better.

And what's nice is there are at least a few PDFs that, if you plug them into that integral, you're going to have a closed form solution. So for example, if you have a uniform probability density function associated with defect density, that yields or gives rise to this uniform yield formula, which is no longer exponential-- or just an exponential.

It's also got a 1 minus the-- this exponential in a scaling factor-- can also do it for a triangular distribution. You get a squared version. If I plug in a Gaussian, we already know that an integral over a Gaussian is kind of nasty-- doesn't have a closed form solution. So it's not directly integrable. One can certainly do it numerically, and things like the phi function does that.

The Murphy yield model was done back when people really wanted closed form kinds of solutions. Oh, I thought I had a picture. Here we go. So here's a comparison of some of these different PDFs that have been examined. Again, Poisson assumes everything is at a d0-- should be a d0 there-- a uniform distribution. I might have defect densities across that whole range, or some triangular distribution where you might, in fact, restrict it in some additional form related to some d0-- or an exponentially decaying defect density function.

What was interesting is, if you go back to the literature, Seeds proposed that-- based on some experimental data, that an exponential defect density distribution appeared to make sense. First off, the qualitative reason was its vanishingly small likelihood that you've got lots and lots of defects, because if you do, your overall process yield is not going to be very good, and so you would have done process correction or process development to remove that.

But as the defect density gets smaller and smaller, a good manufacturing process-- that's where you want to be. You want to have much, much higher likelihood of small numbers of defects per unit area than high ones. So this is not really a statement about physics. It's a statement about manufacturing operation that drives a particular kind of shape of defect density distributions.

It says all of the-- or a huge amount of energy is put in to driving down the defect density distribution. And what that should lead to is something like an exponential falloff, as pictured here. You would expect and hope that your manufacturing process would have a much higher likelihood of a small number of defects per unit area.

And what's nice is, when you plug that in, you can get a closed form expression that's-- that, for the exponential defect density, is very nice and simple. Question--

AUDIENCE: Does that really make sense, though, compared to, say, using half a Gaussian instead with, say, v centered at 0 and just cropping half of it. It kind of seems as you approach 0, it actually becomes more difficult to remove those last couple of defects, rather than going exponentially up that curve, that it kind of flattens off.

DUANE BONING: Yeah. So the question is, what really is the defect density distribution? And does this make sense? Especially with the singularity, as the defect density goes to 0, what's the relative probability? Might you model this with a Gaussian or a half Gaussian? There's all kinds of arguments.

And it's actually difficult to get enough data to really nail down the distribution. Think of how many wafers, if you will-- to get a very careful description of defect density per unit area-- on average you might need. It's hard to fully get the amount of data that you need. So you're really getting a few data points in here that you're trying to get at least the right trend with.

And so it actually doesn't matter too critically, as long as you've got the basic essence of the shape. And I will show you at the end a few of the kinds of test structures that are used to try to approximate or get at these defect density distributions. And in fact, what is very often done, just to give you a little bit of a peek-- people might use the exponential with a fit to just a couple of points or a couple of parameters.

OK, so that's basically the Seeds model, and that often is used. But I want to return to a couple of other further extended models and give you a little bit of a feel for them, because the arguments about defect density distributions continue. But also, reassessing or looking back at some of the other assumptions that I've mentioned, arguments about those also exist.

And one of the most important ones is this notion of no spatial correlation in your defect locations. Rather than show this, let me show this first. So we assumed the picture over on the left, that all of your defects were randomly distributed across the wafer. What's very often observed in practice is that these defects tend to cluster near each other.

And maybe there's some process going on in your chamber that occasionally splattering particles, accelerating particles in some direction. And so those may naturally send multiple particles all together, and they may very often tend to cluster together. So that is very interesting.

If, instead of each and every particle being spatially distributed and causing a fault on an individual chip, now, well, I've got multiple particles all falling and perhaps causing defects on these two chips, but now that assumption that every single defect is causing its own unique kill event is no longer really true. You really can't keep killing the same chip and causing additional yield loss.

So if you've got clustering of your defects, in fact, you may be in better shape than you would have assumed per the count of defects over on the left. And a distribution that has an additional parameter in it-- this alpha parameter-- that gives a defect density distribution with an extra degree of freedom that you can play with this shape-- not necessarily even Gaussian.

But some other amounts of skewness away from that exponential is a negative binomial or gamma probability distribution that gives rise to this negative binomial model. And so empirically, there's this additional alpha parameter that lets one tweak, or fit your data to tweak the defect density distribution.

So if you wanted something that was a little bit more like a Gaussian or a half Gaussian, but maybe behaved a little bit differently right near your low defect density, you've got that opportunity. And what's nice about it is it actually correlates or relates to this notion of spatial clustering of your defects. So there's a reasonable physical explanation for these situations. Yes, question--

AUDIENCE: [INAUDIBLE]

DUANE BONING: I'm sorry. Say that again.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Oh, well, if they're doing the d0 Murphy model, it's probably-- they might be just using the delta function-- simple kind of a model. But you'd actually have to probe a little bit, because they might also have a clustering parameter, and really, then what's going on is something like this.

So where your d0 is in here, there is still a scaling factor to this distribution. You can see the d0 in here. So they might be using, in fact, a negative binomial yield model. In fact, I think that, right now, this is a dominant model that is used, with clustering accounted for. And so the d0 is still your average-- it's your central scaling or average on this distribution.

AUDIENCE: [INAUDIBLE] 10 years or 5 years, then that [INAUDIBLE] changing [INAUDIBLE]

DUANE BONING: So the question is, for a long lifetime, how do these parameters change? And generally, they do change on your fab-- not necessarily the lifetime of your product so much, because I think of-- your d0 tends to be more a characteristic of your unit process or your integrated process. But as you learn more, you have this yield learning, where you hope you drive your defect density down with time-- drive your particle size down.

So you do continue to improve the process. In fact, in some of the yield projections for product you might run in your fab in a year, you might also include some projections on what you think d0 will improve to, based on historical trends over time.

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yes, yes-- typically, perhaps more with the d0-- probably less projections on alpha. Alpha tends to be this clustering, which has two limit that I'll talk about-- low clustering and very highly, tightly clustered. I don't think that's assumed to change that much with time. But the d0 is the main thing that goes down with improved processing.

AUDIENCE: [INAUDIBLE]

DUANE BONING: They should, whether they're drawn to actually integrate out to 1 or not. But they are all still defect [INAUDIBLE] probability density functions. So we do have this alpha clustering parameter. What's nice is, amazingly, you plug that PDF into the integral with an exponential Poisson kernel in e to the minus acd, you get a closed form yield formula here, as shown at the bottom, which has the alpha clustering parameter in it.

And we can take two limits. One limit is the large alpha limit, which is very little clustering. So think of maybe alpha as the distance between individual defects, and as that gets large, you don't have any clustering. And that limit converges to the Poisson model.

And in the very small alpha limit, with very, very strong clustering, that actually converges in the limit for alpha going to 0 to the Seeds model that we saw earlier, which was the pure exponential. So you see, as alpha gets smaller and smaller, this approach is more and more the exponential defect density model.

Turns out that, generally, people are fitting based on experimental data, their D0. And they're also fitting, empirically, alpha. And alpha tends to be related both to the clustering, but also a little bit to the sensitivity of your type of circuit to clustering. So it's not purely-- if I did this just on blanket wafers and looked at the clustering, that may actually not tell me what is going to happen for different kinds of product.

So you actually might end up with different components of or different products, whether it be a memory product or a microprocessor product, or different components on a big multi-product that has a lot of memory cache on it, and also has the combinational logic on it. You might have slightly different yield model components or slightly different alphas for those two different cases, and you would fit those.

AUDIENCE: [INAUDIBLE] is actually designed using different alpha parameters--

DUANE BONING: Interesting--

AUDIENCE: [INAUDIBLE]

DUANE BONING: Yeah. So the observation, if you didn't hear that in Singapore, was that, in practice, with those memory redundancy schemes, those also affect alpha, and so it's an empirical fitting process with different kinds of redundancy to see how that affects alpha-- and what your ultimate yield would be based on that. OK, so so far, we've talked about a probability density associated with the number of defects per unit area.

We can also think about another probability function, another statistical relationship. So far, we've talked about every defect abstractly as being infinitesimally small. So one defect doesn't cover 20 different chips, right? It's just infinitesimally small. But what if there is an area dependence to the size-- or a probability associated with the size of those defects?

That could interact very importantly with some of those original shorting and open physics that we talked about earlier. So for example, if I have wiring lines like this, and I'm really worried about either open or short, and my defect is substantially smaller than either the spacing or the width of the line, it can fall almost anywhere and not cause at least an immediate failure-- might still be a reliability or a parametric resistance change that I'd be worried about-- whereas, if the defect were much larger, it can fall almost anywhere on my circuit and cause either an open or a short.

So the effect of particles of different sizes can be very different, and so I might also want to characterize the size distribution of particles. And the interaction of those science distributions with the particular feature sizes on my circuit is going to be very important. So it interacts with this notion then of a critical area.

Let's see if I've got a better picture. Nope, this is pretty much it. There is a formal notion of critical error-- area for any particular defect size that you can actually analyze for your particular layout and say, which area-- which fraction of the area on that layer does the center of the particle have to fall in order for it to cause either an open or a short?

Let me try to erase this a little bit. So for example, the critical area perhaps associated with the smallest particle may be 0. It can fall anywhere and not cause a problem. This particle is perhaps a more interesting one, in that maybe it's exactly equal to the size of-- actually, let's do an example [INAUDIBLE]

Let's do an example where I've got something that's, say, equal to-- or just slightly larger than the spacing size. But in some places, I've got wires where-- and spaces that are smaller than that, such as pictured here, but I've also got other places-- let's say, I have now another wire up here, where the spacing is larger than the particle size.

Now this same particle can fall right there and not cause a short. So you can calculate across your entire particular layout what is the band where the center of the particle has to fall to cause either an open or a short and some up that area of sensitivity for failure for each of the layers. So there's this interaction between a critical area per particle size, and then a distribution associated with the particle sizes that are very important to also characterize.

And here are some examples-- again, going way back-- for defect size distributions. These are back characterized in mils. Anybody know what a mil is? Thousandth of an inch, or about 25 microns-- so these are giant, giant particles. We have driven down defect sizes a bit.

But what is very interesting is the same trend has continued to be observed-- that there is generally believed to be something close to an exponential dependence in defect size, not just in the number of defects as well. And that exponential-- or power law kind of dependence is very often used in modeling the size distribution for defects.

Now, there's a couple of parameters-- this n and this p, which, again, end up being technology dependent and generally fitting parameters to the data. So now, if we take in that notion of a distribution in defect sizes and the probability associated with that, one can form an aggregate sort of approximate parameter. And so sometimes that's also the d0 that is quoted.

So it's not only averaged in terms of numbers per area, but it's also kind of a boiled down approximate parameter giving you a sense of the basic central moment of the size distribution as well. But if you really wanted to do careful defect modeling, you actually need to know the p parameters and the n parameters. You would like to have that full defect size distribution at hand.

Here are some examples. The typical ranges of that exponent-- I guess it's a p exponent in the previous slide-- is two, three, four. And it may depend on the defect failure mode that you're looking at-- so for example, extra metal or short versus missing metal and opened. They may have slightly different sensitivity to that defect size as well.

AUDIENCE: [INAUDIBLE] more distributions [INAUDIBLE]

DUANE BONING: This is basically this power law. 1/x to the p is the assumed--

AUDIENCE: That's [INAUDIBLE]

DUANE BONING: Yes.

AUDIENCE: But for [INAUDIBLE]

DUANE BONING: For the d0, the d0 is actually still not counting-- adding in the defect density. That will come in in the other distribution. Oh, OK, I did have a slide. I think I've already explained this, but this is talking again about the critical area for different size dependencies. And what's interesting is then you can also start to produce a plot of this critical area versus defect size.

And what's really important here is very intuitive trends that, for larger defects, I've got larger critical area. More part of my chip is sensitive to it. But once it gets smaller than a certain dimension, when my defects tend to be smaller than my minimum feature size on the device, I start to be less sensitive to immediate failures. So if they're a fraction-- your particles are a fraction of your minimum dimension. And that's good, because you probably have lots of defects of those sizes. It's very hard to get rid of those.

Now, what you can do is start to put these together. You aggregate these. If you have a defect size distribution that goes as one of these 1/x to the p's, you can start to try to empirically fit that to your data. One thing that happens with these distributions, of course, is the same thing you were worried about with the exponential. When x gets really small, that goes sky high-- 1 over a very small number.

And so what people basically do is they're mostly worried about the defect sizes larger than their minimum feature size. And then they'll basically truncate it either as a constant or, in fact, as a linear drop-off, once you get down below some minimum size. I should have drawn that near x0 instead.

So there is a size at which it's hard to either detect these, and you don't care about them, and so you're not really trying to model that part of this distribution. In one of the first lectures, I gave you a little bit of a preview and an example here of how you might measure these defect size distributions. Characterization test vehicles, especially early in process development, might be used to try to characterize the capability of the process in terms of these-- both your defect density, but also your defect sizes.

And imagine now that I've got a whole array of these nested metal lines that I can make electrical measurements or connections to. And if I have now a defect that's kind of small in general-- or usually-- it might be even so small that I rarely get short, but I might get some amount of resistance change in these lines-- versus other defects that are so big that they start to bridge, on average, two or three lines.

You can start to build up some electrical measure of the likelihood of having-- or the relative counts of defects of different sizes. So here's an empirical-- this is, again, from a 2003 paper. You can start to see that 1/x to the p empirical relationship, and it gives you a sense for how you can actually measure those.

So now you can have a more careful definition of the critical area, where you have your defect size distribution. I might use dsd to indicate that. So there's our defect size distribution, that 1/x to the p-- again, really only worrying about it above some x0. And then you also have this probability of failure, which folds in this notion of critical area.

You can actually look at your layout and say, if I've got a defect of this size, this is my probability of failure integrated across my whole chip. And now you can aggregate the two of those into a net total critical area. So all I'm doing is saying there is a probability of failure associated with defects of different sizes. The defect is really small. Again, I have that critical area plot, where I'm down in here, and it's very small.

But empirically, for my particular circuit, I may have different critical area dependence curves, where I aggregate the total pof. And then I look at the product of those two, integrate it up, and that gives me a good aggregate sense-- which really says where I'm mostly worried is in this range here, where I've got defects that are close to my feature size. The small ones aren't going to kill me. The bigger ones are going to kill me. Really big ones are not going to kill me, because there's not that many big ones. So really, it tells you the area to worry about.

OK, I'm going to skip over most of this. I just want you to get a feel for this notion of critical area. And this is an area where there's a lot of design automation tools, where it can actually look at your particular layout and start to do those kinds of drawings that I showed you for critical area, if you will, and shade in-- maybe a little hard to see, but you can shade in and say, what is the gray critical area for that particular circuit?

Where am I sensitive to the likelihood of a short for defects of a particular size? Where am I sensitive? Where would a-- the center of a particle have to fall in order to cause an open? Or where might I have to have a particle fall that would cause a short between two different layers?

And you can actually do these calculations for your layout, and then do design modifications that would come back in and say, oh, if that's my critical area where I might be susceptible to a break, let's make those lines a little bit wider. If I make those lines wider, I've improved my yield, because now, if the particle falls in those lines, I'm not as likely to actually have an open failure.

So there is yield improvement strategies that go with these notions of critical area, and the probabilities associated with them. The last notion is simply you can integrate these up and get aggregate notions of overall yield. This is also described in [INAUDIBLE] Just reminding you-- there's also other kinds of yield detractors, where you might have gross yield losses, and so you might have a global factor y0 that's associated with alignment errors or other kinds of gross factors.

And then this last is a little example that you can read about in [INAUDIBLE], which is basically simply saying, in practice, your chip yield is all so aggregated that it doesn't really tell you what's gone wrong. So very often, you want to slice yield into your different layers-- your different process layers, or slice them into different functional blocks-- maybe the memory or cache block, a logic block-- and basically look at where you are most sensitive to yield loss.

For example, that 95% yield loss factor or via 2 inside of your SRAM might be where you're losing most of your yield. There's quite a bit of development of these test chips that have those things like those nested via or nested snake structures in them so that one can characterize defectivity distributions, as well as sensitivity of different kinds of circuits to those failures.

So that's a whirlwind tour there of semiconductor yield, these notions of not just functional yield that we saw last time, but also this-- or parametric yield, but defect yield as well. And you'll have a little bit of fun playing around with some of these notions of area-dependent yield, which I think is really kind of the cool idea, the important idea in yield modeling for semiconductors.

OK, so we'll see you again on Thursday. Thursday is the quiz. I think here, you had also posted for office hours. I think you have those--

AUDIENCE: Yes.

DUANE BONING: --tomorrow as well.

AUDIENCE: [INAUDIBLE] to 6:00.

DUANE BONING: Tomorrow, 5:00 to 6:00, if you want-- any last questions before the quiz-- Hayden's available for those as well. So thanks. We'll see you on Thursday.