PROFESSOR: OK. So let's get started. We want to first review Wald's equality a little bit. Wald's equality is very tricky thing. If you think you understand it, you will go along. And at some point, you will be using it and you will say, I don't understand what this says. And that will happen for a long time. It still happens to me occasionally.

What happens is you work with it for longer and longer times. The periods when it becomes confusing become rarer. And the expected time to straighten it out becomes smaller. It is a strange kind of result.

So we started out with a stopping trial definition. J is a stopping trial for a sequence of random variables. If it's a random variable and it has the property that for each n greater than or equal to 1, the indicator random variable indicator of J equals n is a function of X1 to X sub n. In other words, the decision of whether to stop at time n is a function of 1 up to n.

A possibly effective stopping trial is the same, except that might be a defective random variable. And the reason you want to have possibly defective random variables is that before you start analyzing something that might be a stopping rule, you generally have no way of knowing whether that actually is a stopping rule or whether it's a detective stopping rule. So you might as well just say it's a defective stopping rule to start with and show that it's not.

Then from that, we went on to Wald's equality which added the condition that X sub n that this is based on is a sequence of IID random variables. To be a stopping trial, you don't need IID random variables. You don't need any restriction at all, other than the fact that you can make a decision on when to stop based on what you've already seen. That's the only condition there.

Wald's equality is based on this extra condition that the random variables are IID. Each of them, with some mean, X bar. If J is a stopping trial and as the expected value of J is less than infinity, then the sum S sub J at the stopping trial, J, satisfies this relationship here.

And remember we proved that last time. And the key to proving it, the hard part of it-- well, it's not hard if you see it. But what's difficult is when you start looking at how you find the expected value of S sub J that's equal to a sum over n of X sub n times the indicator function of J being greater than or equal to n.

Now the condition here is that this indicator at J equals n is a function of these quantities here. When you add the IID quantity down here, what you find then is the indicator function for J greater than or equal to n. It's also the indicator function for J 1 minus the indicator function for J less than n. It's independent of X1 up to X sub blah, blah, blah.

That indicator random variable is then a function only of X1 up to X sub n minus 1, because it's 1 minus the indicator of J less than or equal to n minus 1. J less than n, which means that X1 up to X n minus 1 is independent of Xn. And X1 up to X n minus 1 is what determines this indicator function J less than n. It's the J less than n which is important. So we got that Wald's equality.

What I want to do today to start off with is to do the elementary renewal theorem, which is a strange result. Wald's equality, you can use it to determine the expected value of N of t. Now why do we want to determine the expected value of N of t?

We already have shown in great generality that there's a very nice, with probability, one type limit theorem associated with N of t over t. We know that N of t over t approaches t with probability 1. Namely, all the sample functions, except a set of probability 0, all approach this same constant, 1 over X bar. So it seems that we know everything we want to know about the expected value of N of t, about N of t, and everything else. But no.

And there are two reasons why you want to do something more than that. One of

them is that very often you want to know how N of t over t approaches the limit of 1 over X bar. The other is that sometimes you really are interested in, what is N of t at some finite value of t? N of t is a random variable, some finite value of t. So you can't evaluate it. But at least you would like to know what its expected value is. You might want to know what it's variance is too.

But people who study renewal theory have spent an enormous amount of time on trying to find the expected value of N of t. It's the basic problem that people work on all the time. The elementary renewal theorem is something which says a little more for finite times. And it actually says the expected value of N of t over t and the limit is equal to 1 over X bar.

Sounds like much less than what we've done before. And perhaps this is because this was a computational thing that people could do without computers, before computers came along. And all the work on renewal theory went on basically from 1900 until about 1970. People didn't have any computers to compute things.

And all the mathematicians who were interested in this field really loved to compute ugly things. And they computed ugly things all the time. And the expected value of N of t was one of the ugly things that they could really do a lot with. So it probably has more importance in this field than it deserves. But it does have some importance.

Anyway, what we want to do is to get some idea of why this elementary renewal theorem is true. Later on we will study how to actually find the expected value of N of t. N of t is the number of arrivals that have occurred up to and including t. So N of t plus 1 is the number of the first arrival after t since-- that should be a t there-- the expected value of N of t is finite, the expected value J is also finite.

In other words, we're defining J to be N of t plus 1. We want to show the J is actually a stopping trial. And therefore we can use Wald's equality on it.

Let's stop and think for a minute. Why don't we just define a stopping trial to be the number of arrivals that have occurred up until time t? What's the matter with that? If I let you observe a sequence of these IID random variables, the first inter-arrival

interval, the second inter-arrival interval, the third inter-arrival interval, and so forth, and I say stop when you have the last dozen that's less than t.

The question is, is that a stopping trial? You have to be able to determine that just from what you've seen up until the present. But when you see that last arrival, there's no way you know that it's the last arrival less than t unless you see the arrival after that, also to see whether that's greater than t. So the only way you can define a stopping trial of this form is to define the stopping trial not at N of t, but of N of t plus 1.

So now we want to apply Wald's equality to the expected value of the time at which N of t plus 1 arises. Do you see now why we want to call this a stopping trial instead of a stopping time? If you call that a stopping time, you want to know what you were doing at this point. Because the stopping trial says you're looking at the first arrival that occurs after time t. Then you're looking at the time at which that arrival occurs.

Wald's equality says the expected value of that time is equal to the expected value of the inter-arrival time times the expected value of J. J is a stopping trial. The expected value of J is just-- J is N of t plus 1. So that's expected value of N of t plus 1 times X bar.

That's simple. It's simple until you try to recreate that argument. I would suggest to all of you that sometime later today you go back and try to recreate this argument. And if you can do it, then you understand it. If you can't, think about it a little more. Because it is deceptively simple.

So that was a relationship we had. Wald's equality then is relating two unknown quantities. It's relating the expected value of N of t with the expected value of the first arrival time after t. That's one of the troubles with Wald's equality. It relates two things that, neither of which you know in many cases.

There are many situations in which you do know one or the other. We talked about one last time. But in most of them, you don't know either of them. So it doesn't help you immediately. You have to find something else.

So what we're going to do here is say, well we can't find expected value of N of t from this. But we can at least bound it. And we can bound it by saying that S sub N of t plus 1. This is the time of the first arrival after t. The first arrival after t has to be at a time greater than or equal to t. So expected value of S sub N of t plus 1 also is greater than or equal to t. So expected value of N of t, solving this equation is expected value of S sub N of t plus 1 over X bar minus 1.

And then you lower bound expected value of this first arrival after t by t. So the expected value of N of t has to be greater than t over X bar minus 1. Well this seems to be useful, because if what you're interested in is the expected value of N of t over t. And that's what we're interested in for the elementary renewal theorem.

You're saying that that's greater than 1 over X bar minus 1 over t. And when t gets large, that 1 over t term gets very small. What that means is you have a lower bound on expected value of N of t, which is going to 1 over X bar.

The elementary renewal theorem then says, let X bar be the mean inter-renewal. Then the limit of t approaches infinity of the expected value of N of t over t is equal to 1 over X bar. Very weak result. At least it seems to be a very weak result.

But we need an upper bound on the expected value of N of t. We don't have any upper bound on this quantity here. Because you remember these-- Actually S sub N of t plus 1 minus t is this residual life at time t. And we found that the residual life at time t can be pretty badly behaved. So we're in trouble. So what do we do?

Well the argument is done carefully in the notes. It's very discouraging to me, because this should be a simple result. And here this argument is incredibly tricky. So we're working very hard to do something that's not-- Well, it seems like it ought to be elementary. It is an elementary result in terms of its usefulness. It's a very hard result in terms of trying to prove it.

So the way you prove this is first you truncate the random variable X. Namely instead of having a random variable with this arbitrary, non-negative probability distribution, what you do is you look at another random variable, which has the

same distribution up until some large value B. And then you truncate it at that point. So the distribution function looks like--

Here's f of x, however it might look. This is 1 here. This is supposed to be going up towards 1. And what you do is you truncate it at some point B. So this is a truncated version which stops at B. And this here is the actual Fx of x.

So first you truncate the random variable. Then you prove the elementary renewal theorem for that truncated random variable, which is easy. Because at that point, the expected value of S sub N of t plus 1 is less than or equal to t plus B.

And then the next thing you have to do is monkey around with the difference between X bar and X tilde bar. And you have to go through a bunch of strange inequalities. You have to then let B approach infinity as t is approaching infinity in just the right way. And when you get all of done doing this, bingo. You get the elementary renewal theorem.

You see why I don't want to go through that in class. And you see, I hope, why probably most of you will not want to go through it in detail either. Because if you want to learn how to do truncation arguments, this is probably not the best one to use as a way of figuring out how to do that.

If you want to learn how to use truncation arguments, do the Weak law of large numbers where you don't have a variance. It's done back in chapter one. It's a nice relatively simple argument where you truncate a random variable. This one is just peculiar.

We do all of that with stopping trials as we've defined them. Very often, and particularly when you get to queuing theory, you want to define stopping trials in a more general way. And here's a more general definition. A generalized stopping trial for a sequence of pairs of random variables, X1, V1, X2, V2, X3, V3, is a positive integer random variable such as for each n greater than or equal to 1, the indicator function, which tells you whether to stop or not, is a function of X1, V1, X2, V2, up to Xn, Vn.

In other words, you don't have to decide whether to stop or not in terms of these random variables X1 to Xn that you're interested in. You have this other set of random variables too. Typically in queuing situations, these other random variables that you're interested in are service times. And the X's are arrival times. And these things are going on in parallel with each other.

And still, you can do the same arguments about stopping trials by looking at just the past history of the input random variables and these other random variables, whatever they happen to be. Wald's equality then says that the expected value of S sub n is equal to X bar times the expected value J, where Sn is the sum of the X's. And Wald's equality holds by exactly the same proof as before. There's nothing new.

You just look at the proof and you say, OK. All these V's are involved there also. You can replace each of Vi in this definition with a whole vector of random variable. So you can make this as general as you want to make it.

In fact, a lot of people prove Wald's equality by skipping all of this stuff about stopping trials. And what they say is that the rule for stopping at time J is independent of all of the future arrivals. And that lets you prove the theorem immediately.

It doesn't give you any clue at all as to when that holds, so that in fact you're taking Wald's equality and making it essentially a truism. And you're avoiding the real problem, which is to know when you have a stopping rule and when you don't have a stopping rule. But anyway, we can generalize it in this way.

And now we want to use this to look at a fairly general queueing situation. We talked about the G/G/M queue before. Let's just talk about the G/G/1 queue at this point. This is a queue with general independent IID and her arrival times, IID service times. So what you wind up with is at time 0, you assume that there's an arrival at time 0. You don't count it as part of the arrival renewal process.

So this arrival is here. That arrival goes into service. There's some service time V

sub 0. At some inter-arrival time, X1, another arrival comes in. In this particular sample function here, the second arrival comes in before this 0 arrival that we didn't account finishes service.

And I'm sorry. I should have-- It might make better sense to call this X1 and call this V1. It would make all the notation simpler. Or call this X0 and call this V0. But unfortunately that's not consistent with any of the notation that anybody uses for talking about renewal processes. Because you want these X's to be the elements of a renewal process. You want the V's to be IID random variables. And this, unfortunately, is the way it has to be.

So these arrivals come in. This is the inter-arrival time before-- This is S1 here, which is the time of the first arrival after 0. This is S2, which is the time of the first arrival after time 0. And when you count the arrival at time 0 in, this function here is N of t plus 1. So this is N of t plus 1 going along here. This is the arrival process plus 1, where now we've counted this arrival at time 0.

You have these services going on. V0 for this particular sample function is the time of the service of this 0 arrival. V1 is the time of the service of this X1-- well, it's the service time of the arrival that has come in at this point. This is the service time of the arrival at time 0. This is the service time of the arrival at this time. This as the service time of the arrival at this time and so forth.

I want to make sure that you understand this model here before we go on. Because if you don't understand this, you won't understand anything we do after this.

If you look at this now, you see a departure process which is going on also. These departure times are determined partly as a sum of the number of customers to have departed by time t. And this example here is just the sum of V0 plus V1 plus V2.

So it's almost the same as this renewal process here, except here you're talking about a renewal process of dealing with the departure times. This changes any time that the queue empties out. Because when the queue empties out, the server isn't doing anything for a while. So suddenly, the number of departures slows up. And in

some sense, you reset things here.

What I want to do in this argument today is to explain much more carefully why you can actually restart things when this new arrival occurs. So I want to say that, in fact, this is a renewal time for this entire queuing process. These are renewal times for the arrivals to the queuing process. This thing here, the next arrival to an empty server and so forth, are all, in fact, renewals of the entire queuing system. We want to understand why that is. We want to use this idea of stopping trials to understand this.

So look at the first arrival to start a new busy period. Well, think of that as a generalized stopping trial. Why is it generalized stopping trial? Well the sequence of paired random variables we want to look at is X1, V0, X2, V1, X3, V2, and so forth. And stopping at J equals 3 is actually a function of this first pair, the second pair, and the third pair, which is the restriction we need for a stopping trial.

So the stopping trial stop when the first arrival comes to an empty system is in fact a function of the arrivals and as the departure intervals up until that time. Wald's equality holds because, in fact, every time you get a new arrival, it's independent of all the old arrivals. It's independent of all the old service times. It's independent of all the new service times also, sort of. But you have to be more careful with that. But in fact, the service times are all IID. So that works too.

So Wald's equality holds here. But that's not what we're interested in here. What we're interested in is trying to show that you do get renewals at these times here. So let's try to argue why that is. Did I--

AUDIENCE: Question? Do V's have to be independent somehow?

PROFESSOR: What?

AUDIENCE: Do the V's have to be independent somehow?

PROFESSOR: Yes. Yes.

AUDIENCE: So they have to be independent from one another?

PROFESSOR: Yes.

AUDIENCE: It was the case here, you said.

PROFESSOR: Yes. That's what you mean by a G/G/1 queue. A G/G/1 queue is--

AUDIENCE: You have a service time and the arrival.

PROFESSOR: Service time.

AUDIENCE: They're both independent.

PROFESSOR: They're both independent, yes.

AUDIENCE: And that makes the departure?

PROFESSOR: Right. And in fact, the first G says that the arrivals are IID, have an arbitrary distribution. The second G says that the service times are IID, but have an arbitrary distribution. And then the 1, our the M, or whatever, says how many servers there are.

AUDIENCE: And that makes the departures independent?

PROFESSOR: Yeah. So you can really have much more general queuing situations than this. And you often do. Yes?

AUDIENCE: With the departure process, it's not a renewal?

PROFESSOR: What?

AUDIENCE: The departures.

PROFESSOR: The departure process is not a renewal process, because every once in while, the server stops doing anything. What?

AUDIENCE: That's how I got confused actually. Because every once in a while, the server somehow--

PROFESSOR: Every once in a while, the server stops working. But V1, V2, V3 are defined as the service time of the first customer. Well actually V0, the service time of the zeroth customer. V1 is the service time of the first customer. Namely the time from when the customer enters service to when the customer's finished with service. And the reason why these things--

If you looked at the sequence V0, V1, V2, V3, you could call that a renewal process. But it wouldn't be a renewal process of any interest to you, because it would be a renewal process where the renewals occur at V0, at V0 plus V1, at V0 plus V1 plus V2. And those might not be the time when these customers are finishing service. Because there are these idle times to take into account also.

If you didn't have the ideal times to take into account, you wouldn't have to think about queuing theory at all. Because all you'd have is two separate and independent renewal processes. So this is where all of the gimmickry in all of queuing theory comes from.

But what I wanted to come back and say that I didn't say very well before. In the notes, the assumption is that these pairs here are IID. What you really would like is something a little more general than that, which says that each Xi is independent of all of the earlier pairs. So that that's giving you a little bit of flexibility for how these other random variables impact the situation. And in fact, they can be anything at all, just so long as you have this condition that each Xi is independent of all these past things here.

So we said that Wald's equality holds. But the other thing, which is what we really want, is the new arrivals. Namely the arrival after S3, which comes at S3 plus X3. That's X sub J plus 1, X sub J plus 2, and so forth. And the service times, in this case V sub J, is the service time that occurs after this. V sub J plus 1, and so forth, are all independent of the old arrivals and the old service times.

So in fact, this is actually saying that everything that happens after this J's arrival is independent of everything that happens before. And when I say everything, you have to be careful about that, because we're talking about all the inter-arrival times

that occur after this. And we're talking about all of the service times that occur after this. We're not talking about things like S sub n plus 1, which is the time of the next arrival, because that's this arrival time plus that next inter-arrival time.

So each of these intervals are in fact independent of each other. And what we've done here is use this idea of a stopping trial not to define when we stop this whole process, but to define when a renewal occurs. In other words, we're using it to define when the old inter-renewal period ends and a new one begins. This can be called a stopping trial, if in fact you stop the whole process at that point and then don't do anything further.

So that's what's nice about this idea of stopping trials. They let you not only do Wald's equality. But they also let you actually see why it is that these arrivals that start after this point are independent of the arrivals and departures before that.

So let me just try to say that again. The stopping rule is the index of the first arrival in a new busy period. The arrivals and departures in the new busy period are independent and identically distributed to those in the old. Thus the intervals between new busy periods form a renewal process. And that's the thing that we're going to use in everything else we do here.

So we have one renewal process, which is embedded in another renewable process. The outside renewal process of the arrivals are what you start out with. These embedded renewals are, in fact, functions of both the arrival process and the service distribution. So that's a more complicated thing. But at least you now know that that's an arrival process.

It says that you can analyze any one of these queuing systems or any considerably broader type of queueing system, which has the same property of having renewals every once in a while, by looking at what happens within a renewal period. Then using what happens within a renewal period and applying one of these limit theorems on the embedded renewals to get some time average overall time. So that's what we're really up to here.

This same analysis applies to G/G/m. Applies to lots of queueing systems. You have to look at a queuing system and see whether it applies. To get some idea of how you can see that without going through this whole analysis, suppose that you decided to try to call this point here, at which the system first became empty the end of a renewal period.

Why couldn't you do that? Why is this not a renewal period? Well, is it a stopping trial? It certainly isn't a stopping trial for the X's. Because at this time when this last departure has left, you have no idea how long this new inter-arrival period that's going on here is going to be. That has some arbitrary distribution. All we know at this point is that it's longer than this. And if you had a busy period end at some point close to the beginning of this inter-arrival period, you have a different distribution from when it occurs close to the end of an inter-arrival period.

Well, because these inter-arrivals are not memoryless. They have memory which says their conditional distribution depends on how long they've been running. So this is the only sensible place you can define a renewal process here. So let's go on.

There is something called Little's theorem, which was invented by John Little not all that long ago. And Little's theorem is curious, because you can view it as being either trivial or non-trivial. I will try to convince you that it's trivial and also convince you that it's non-trivial. And what I mean by that is that it's trivial in terms of trying to use it someplace. It's then nontrivial to try to justify that you can actually do that. But you get the idea of doing it in many, many places.

We're going to assume an arrival at time 0, like we've been doing before. Service process can be almost anything. But let's assume a G/G/1 queue to be specific. The system empties out eventually with probability 1. And assume that it restarts on the next arrival.

We've seen that intervals between restarting form a renewal process for the G/G/1 queue. And we've argued that it forms a renewal process for an even broader class of queues. That's normally. Who can tell me when you don't get a renewal process when you have a G/G/1 queue? We made an assumption here without talking at all

about it. And what's that assumption? What?

AUDIENCE: Starting with an empty queue.

PROFESSOR: Yes. Every once in a while, you get an empty queue. If you have a system where service times are greater than inter-arrival times, what's going to happen is that arrivals keep coming in, eventually building up more and more. The queue gets longer and longer. And you never have any renewals occurring. So one thing we didn't say here is we're talking about G/G/1 queues and more general queues, and the more important case where the queue sometimes empty out. We want to use this to analyze when they actually do empty out.

So here's the picture. The same kind of picture as before of an arrival process. And every time we're talking about these kinds of things, these queuing situations, we will distinguish the arrival process, which is a renewal process. But when we talk about the renewal process, we'll talk about this renewal process which starts on arrivals to an empty system. So here's a renewal. Here's a renewal, and so forth.

Suppose we look at the difference between A of t and D of t. A of t is the number of arrivals that have come in up until time t, counting this arrival at time 0. D of t is the number of departures that have occurred. What is the difference?

The difference, L of t, is the number of customers in the system at time t. And one of the things you'd like to be able to analyze in a queuing system is, what's the distribution? What do you know about how busy the queue is? If you build storage for 100 customers, you would like to know what's the likelihood that that storage will fill up. And you have to drop customers on the floor or do whatever you do to customers when the queue is full.

So L of t is clearly something we want to look at. It's the number of customers in the system at any time. Not the number of customers waiting in queue, but the number of customers both to waiting in queue and being served currently.

We can view this as a generalized renewal reward function at this point, because this number in the system at any given time t, is a function what happens within this

inter-renewal period here. It's a function of these arrivals within this inter-renewal period and these departures within this inter-renewal period. So we can use renewal reward theory. The total reward with an inner renewal period is the integral of L of tau over that period. This is the way we found--

Every time we looked at renewal reward theory, we looked at the reward within one inter-renewal period. And then we tried to look at what happened when we started to look at many, many inter-renewal periods. So we want to look at L of tau integrated over one renewal period. And then we want to go on to see how we apply renewal theory over that. So in each inter-renewal period, the aggregate number and queue number in the system integrated over the duration of that inter-renewal period is this integral of L of tau.

Here I've stated Little's theorem really. Let's go back and see where that comes from. If I want to look at the integral of L of t over this period of time, what is it? I can view it as W1 plus W2 plus W3. That's the easy way to integrate this function.

You start out here. You're integrating up to here. 1 times D tau. And you have 2 times D tau. Then 1, then 2, and then 1 again. But the easy way to integrate it is just say the integral is W1 plus W2 times this height, plus W3. For those of you who have studied Lebesgue integration, it's the idea of a Lebesgue integral as opposed to a Riemann integral. If you haven't studied it, don't worry about it. It's just a trivial idea that you can integrate this way as well as integrating this way.

That's the crux of Little's theorem. The crux of Little's theorem is this equality between the integral of L of t and the sum of the waiting times of each customer. W1 is the waiting time of customer-- Well it's the waiting time of customer 0 in this case. W2 is the waiting time of customer 1, from the time it comes in, it's waiting in queue until this time. Then it starts service. Next customer comes in here. Waits in queue. Finally starts service.

Then the system is empty. And you integrate L of t from here up to here. You sum the W's. And that's what I did in the next slide.

And I can integrate over as many inter-renewal periods as you want to integrate over. And what's going to happen? I'm adding up these individual inter-renewal periods. I'm adding up this integral over each of them. I'm also adding up the sum of the waiting times over that larger interval. And this equality still occurs.

Any time I integrate over an integer number of inter-renewal periods, I have this equality between the integral of L of t and the sum of W sub i's. One of the things that we observe from renewal theory is when you start taking limits between 0 and infinity, one inter-renewal period doesn't make any difference.

When we go on through the mathematics there, we've done it carefully. We've upper bound it. And we've lower bound it. And we've then shown that it doesn't make any difference. But at this point, you know that it doesn't make any difference.

So that you know that when you take the time average, the time average number in the system is equal to the limit as t goes to infinity of 1 over t times i equals 1 up to the number of arrivals up until time t times W sub i. When you take that limit, it's W sub i over A of t times the limit of A of t over t. Both of these go to a limit with probability 1, because this is just a strong law of large numbers again. This is just the renewal theorem again.

And that limit here is the time average of W. It's a little awkward calling this a time average. Because what this really is is a sample path average. That's a sample path average which exists with probability 1. This is a sample path every over any old sample path you want to look at of the time average over all of the customers that come in. You add up all the customers that come in and you divide by the number of customers. And that gives you the time average delay that a customer experiences.

So what Little's theorem says is that the time average of the expected number of customers in the system at a given time is equal to lambda, which is the arrival rate-- A of t over t, is arrival rate-- times the expected delay that each customer experiences. That's a very useful relationship. Like all of the things that come from the Wald equality, it only gives you a relationship between two things that you don't

know. But one relationship between two things you don't know is better than no relationships between two things you don't know.

Here we have this one useful relationship. We can understand intuitively what's going on here if we look this diagram again. If you look at the diagram again, if you expand this whole diagram in t, if you measure it in hours instead of measuring it in milliseconds, for example. This whole thing spreads out enormously. And what happens? The W's all get multiplied by a very large amount. A of tau stays the same. Lambda, the expected arrival rate, becomes very much larger. Arrivals per hour rather than arrivals per millisecond.

So you see that at least as far as scaling is scaling, there's no way of relating this to this without having something there, which is something per unit time. This is time. This is a time average.

So we have Little's theorem. Useful in many, many situations. It's useful as an approximation in many places where it doesn't hold.

But the right way to look at this, this is really an accounting identity. There was this first thing we brought out, that this integral of L of t over 1 inter-renewal period was equal to the sum of the waiting times over that one renewal period. All of the mathematics and all of the stuff we've been struggling with have all been concerned with showing that when you go to the limit as t goes to infinity, all of these things make sense. Question is not whether L equals lambda W, but whether these quantities exist as sensible time averages or is limiting ensemble averages.

What we're dealing with now is time averages. But you can do all of these same things with limiting ensemble averages. And hopefully, you get the same answer when you do it. And you have to go through a lot of analysis in order to show that these are the same. But the idea that L is equal to lambda W is, in some sense, much more simple and fundamental than that.

Last thing I want to do is talk about the Pollaczek-Khinchin formula for M/G/1 queues. The thing that you get here is you improve on Little by one step, because

now if you have arrivals that are coming in a Poisson way-- Yes?

AUDIENCE: I'm sorry. I just have a question. Is there an easy way to tell when we have the condition we need with probability 1 that restarts? That the system empties out with probability-- is there some kind of condition like the expectation of the probable [INAUDIBLE]?

PROFESSOR: Well certainly one thing that you want is that the expected time between arrivals has to be less than the expected service time if you only have one server. And if you have n servers, then you want the similar situation that the expected inter-arrival arrival time of arriving customers is less than or equal to the expected service time divided by n. Because you have n servers. Now the expected arrival time is less than or equal to n times the expected service time for each server, because you have n servers working at a time.

You have to be careful on that one. Because you can dream up situations where you can have half the servers busy all the time. The system never empties out. And you always keep oscillating between half of the servers full and all of the servers full. You can even control systems and you sometimes want to control systems so that you have about half the servers operating at all time.

If you have half the servers operating at all times, then you have a fair amount of leeway before the queue fills up. You also have a fair amount of leeway before the queue empties out. And you have all these servers that you're paying $100 an hour to and they're not doing anything. So you'd like to keep, somehow, sitting in the middle. You would like to keep servers busy doing other things when they're not busy serving customers, to put it another way.

AUDIENCE: So there is no way to obviously tell when the [INAUDIBLE]?

PROFESSOR: There is no way in general. For particular simple kinds of queueing systems, yes you can tell. We've been going through some of that. For the M/G/1 queue, it's very easy to tell, because there's a nice formula.

And you'll see from the simple formula that it's not as simple as it looks. Because

first I'll tell you what the formula is. Suppose that X1, X2, as usual, are IID exponential arrivals at rate lambda. So the arrival process is Poisson. That's what the M there means, memoryless arrivals. V1, V2, and so forth are going to be IID service times.

We're going to assume they have first and second moments. First moment of the service time, we'll call it V bar. The second moment of the service time, we'll call it V squared bar. What the Pollaczek-Khinchin formula says is that the expected waiting time in queue, namely before you get into service, expected waiting time in queue is lambda times V squared bar divided by 2 times 1 minus rho, where rho is the service factor, which is the arrival rate times the expected service time.

This quantity here, as the simplest answer to your question. If the service factor, which is the arrival rate times the expected service time, if that's bigger than 1, you're in trouble. If that's less than 1, normally you're not in trouble.

This formula says it's something worse than that. It says that as the service factor, if the duty factor is less than 1, this quantity here is positive. But if the second moment of the service time is infinite, you're still going to spend an infinite amount of time waiting in the queue.

And that's unfortunate. You might expect this by now after talking about residual life on all of these things. But it's still is unpleasant to see it.

When you go from the expected number in the queue to the expected number in the system, what is it? Well, the expected number in the system over time is going to be expected number in queue plus the expected service time. Yes. A customer comes into the system. He waits in the queue for a while. Then he gets served.

This is his expected service time. This his expected time waiting for service. The expected number in this system, how do you get that? We get that from this by using Little's relationship.

So the expected number in the system is lambda squared times V squared bar over 2 times 1 minus rho. You multiply this times lambda. So you get lambda squared

from this lambda. From the V bar, you get lambda V bar, which is the duty factor. The expected number sitting in the queue is just going to be this quantity. I won't derive that for you. You can work it out.

If you look at some examples of this, before we do try to derive it, expected number in a queue for an M/D/1 system, namely Poisson arrivals, deterministic service time. When you have deterministic service time, the second moment of the service time is just the first moment squared. Every arrival takes you exactly the same amount of time to be served so that V bar is that service time. V bar squared is a second moment of the service time.

So we just get the formula we had before. I hope it's the formula we had before. Lambda times V bar squared. Oh yes, rho is lambda times V bar. So we have lambda times V bar squared over 2 times 1 minus rho.

You look at exponential inter-arrivals, suddenly you get twice as much wait in queue. It is worse to have exponential service times than it is to have Fx deterministic service times. If you look at this strange kind of distribution we talked about last time, a very, very heavy tailed distribution in a sense, binary distribution where V is equal to epsilon with probability 1 minus epsilon. And V is equal to 1 over epsilon with probability epsilon. Then if you work it out, the expected W sub q is equal to rho over epsilon times 1 minus rho.

This is the same kind of behavior we had before. You have an enormous residual life. And therefore, the time that you wait for a customer in service to get finished is very, very large. It's what you notice all the time if you're waiting in some kind of system where many of the customers get very rapid service. And every once in while, there's a disaster.

A good example of this is at an airline, when you've missed your plane or something. And you go up to a booth there. And many customers, it's just a matter of printing them out a ticket. They're done in 20 or 30 seconds. Every once in while, there's a customer who has enormous problems. And it takes an hour for them to be served. That's exactly this situation here.

This says that the expected waiting time in q, before you even get up to get your own service has this factor of 1 over epsilon in it. This is a nice kind formula, because, in fact, it separates two different kinds of things. It has this one factor of 1 over 1 minus rho, which is a typical factor that you get. Because if you have a server which is overwhelmed by too many customers, if the average service time is very, very close to the average inter-arrival time, then you're going to have relatively large build ups of queues.

Every once in while, it'll empty out. But it will be rare. And you're going to have a large amount of delay. This term here is a separate kind of term. And this has to do simply with a service time distribution that says that bad service time distributions affect you in exactly the same way, no matter what the duty factor is. So there's this nice separation between these two things, which is just the way it is.

Why does the wait in q go up with-- that should be V squared bar. I'm sorry. Look at the time average weight, the expected R of t, for the customer in service to finish service. And picture of that is down here.

If you look at this over time, what you see is at time 0, there's a customer that just came in. The residual time until it gets finished, residual life, is this triangle that we're used to. Then at this point, the next customer starts to be served. That time is there. And the third customer starts to be served.

The difference between this and a renewal process we looked at before, is that when we're looking at this residual life in a queueing system, every once in while the queue empties out. And then there's a period when nothing happens. And then suddenly again we start these triangles here.

So there's a question as to how you evaluate this time average. Very difficult question, isn't it? If you had to do this in a quiz, how would you do it? Would you throw up your arms? Or would you say, well let me just figure out what this has to be. I have these triangles here. I have this idle time here. How much idle time is there? How much triangle time is there?

The triangles are these one half V sub i's that we're used to from before. And then this empty time here is a 1 minus rho term. It's 1 minus the duty factor of the system. And that's the amount of time over a very long period of time, the fraction of time, that the system is empty. That fraction has to be 1 minus rho, because the system is busy with fraction rho. It's empty with fraction 1 minus rho. So you can write down the answer to this without really having much of a justification for it.

So now we want to figure out why this is what it is. This is all right. I want to take the limit as tau goes to infinity of this integral here over time. And what that is going to be is 1 over tau. And what this is is an analytic way of finding that factor of 1 minus rho that I really didn't talk about before.

So this integral here, if I look at it out to A of tau, is going to be 1 over tau times the sum of all of these triangles here up to the number of customers which have arrived up until time tau. So I have all of these out to time tau, which is going to be out here now. So I sum all of these up.

And now I ask the question, what is that sum going to be and the limit as tau goes to infinity? Well, we know how to evaluate that kind of thing, because we've done it four or five times already. The limit as tau goes to infinity of 1 over tau times the summation from i equals 1 to A of tau Vi squared divided by 2.

You can write this as limit of A of tau over tau times 1 over A of tau times the sum from i equals 1 to A of tau Vi squared. Now as tau increases, this A of tau is going to increase. These are IID random variables. So what we're doing is increasing the sample average of these IID random variables. So this quantity goes to what? It goes to the expected value of V squared by the strong law of large numbers.

And what does this quantity go to? A of tau over tau, that's the number of arrivals per unit time. And that goes to lambda. Is that what I got there? I forgot this 2 here. So that goes 1/2 expected value of V squared. So this is the careful way of doing it, where you bring all of this enormous power to bear for something that you could intuit without using any of the power.

And what is W sub q then? What is the average time we spend in the queue? Well you have two components of time you spend in the queue. If you arrive when the server is busy, you're going to have to wait this residual lifetime, which is the expected value of r, which is residual life.

The other thing you're going to have to wait for is for every customer in the queue to be served. Now how long does that take? Well here you have to be quite careful, because you have a number of customers that have been built up in the queue. Each customer that's in the queue has a service time associated with it.

Think of a customer. When it comes in, it's given a little packet with its service time in it. So you look at the sum of all those terms. The question is, is the expected number in the queue independent of the service time of the customers in the queue? And how do you reason that out?

And at this point, you really don't want to reason it out in terms of writing a lot of equations. You want to just think about it. How does a queueing system work, if you have a first come first served queuing system?

Customers come in. Each of them has a certain amount of time to be served, which it's holding in a little bag at its side, but which nobody else sees. And these are independent, random variables. But when it comes in, there are a certain number of customers in the queue in front of it. So the number of customers that are in the queue before it each have their own service times sitting in their own little pockets.

So the amount of time that it takes for me to be served is this sum of all of the previous times. And all these previous times are independent, random variables. And each of them are independent of how many customers are in the queue at the time they arrive.

So what I'm trying to argue is that the number of customers in the queue at time t are independent of the service time of each of those customers, because the service times are not apparent to the system until a person getting into service takes it out of his pocket and looks at what his service time is. And that's unknown to

the system until the service starts. So when you take N q bar independent of V bar0--

Incidentally, if you read the notes, this is explained in a slightly more mathematical than customers having little pockets which they read. But it doesn't help a whole lot.

But anyway, when you go through that argument, you then get this expression. And then you add a little Little. Namely a little Little's theorem, which says that N sub q is equal to lambda times W sub q. So you can take this over on the other side. And this gives you lambda V squared over 2, which is this term. And at the same time, you get 1 over-- Excuse me. Lambda V squared over 2 is this term. And W q minus N q V bar is 1 minus is W q times 1--

I'll do it at the board. Probably I'll see it. But W sub q is equal to lambda V squared over 2. That's the expected value of r plus lambda W sub Q bar. So we take this over on this side. And then we get W q bar equals lambda V squared over 2 times 1 over 1 minus lambda. 1 minus lambda V bar. Where did the V bar come from? Beats me. Oh, came from this quantity there.

Let's patch up our lambda V bar times W sub q bar. So now we get lambda V bar. That is all I had to say today. And so have any questions?