From one point of view we have two types of operational systems; made-to-stock and make-to-order. In made-to-stock, based on the forecast that we have for demand, we produce the product in the absence of customer. From the time that the customer may ask for this product we can benefit from two things, economy of scale. We can produce in economic batches, and we can protect ourselves against changes in customer demand using safety inventory. So if the average customer demand is 100 per day, I can have 150 units available. So if the demand is more than average, I will not encounter shortage. I will not be out of stock. Sometimes I need to exactly know not based on my forecast exactly know what the customer wants and sometimes the customer should be there, and I also should be able to store the product and keep it for the time when customer demand comes in. When you go to a physician, he or she really doesn’t know what you need, so they can’t produce something and put it there. Sometimes we have can it. I can have a hotel room for July 6, but I can’t store it if you don’t come on July 6. That product is gone. I cannot store it for July 7 or July 8. Banks, restaurants, airlines, hotels, hospitals cannot follow made-to-stock. They should make-to-order so they cannot take advantage of economy of scale. They cannot produce in large batches, and they cannot protect themselves against changes in customer demand while using the safety inventory. If demand goes up they must have safety capacity. They should have a facility which its capacity is more than average demand. So safety capacity we need it in make-to-order. Safety inventory we have it in made-to-stock. Safety capacity is much more expensive than safety inventory. And it’s why, for example, when you go to an emergency room – and I hope you never go there – but if someone goes there, a couple of weeks later they get a bill for an amount they cannot imagine. And that is the reason, or at least one of the reasons that some services are so expensive.

Here is an example of a make-to-order process in service industry, in call center. Calls come in. If the capacity of the system is already full, you hear a busy signal, but if it is not full, then you go into a waiting line, and you should wait over there. Until a processor is available, the server is available, then that server can serve you and answer calls relieving the system. For simplicity suppose there is one processor, and suppose it takes 5 minutes on average to serve a customer, then inter-arrival time of customers, inter-arrival time of customers should be greater than 5 minutes, because if it is 4 minutes, then every 4 minutes a customer comes in, and it takes 5 minutes for the server to serve the customer, and the waiting line will go up and up and up. The system will explode. The inter-arrival time should be always greater than service time. For example, it should be 8 minutes. Every 8 minutes a customer comes. Every 5 minutes the server can serve a customer; however, even in this case we still have waiting line. So while the capacity is greater than demand, we still experience waiting line? Yes. We said customer on average arrive every 8 minutes, so 60 divided by 8 is equal to 7 and a half. In each hour we have 7.5 customers coming. Server can serve customer in 5 minutes, so capacity is 12. Demand is 7.5 per hour, but still we may observe waiting line. Why? Because on average it is 8 minutes. On average it is 5 minutes. On average 7.5 customers show per hour. On average we can handle 12 customers per hour, but maybe in one hour instead of 7.5 15 customers show up and in the next hour no one shows up. An average is still 7.5, but in the first hour you see a waiting line. On average it may take the physician 5 minutes to examine a patient, but for some patients it may take 30 minutes. While the capacity must be always greater than demand, but we still observe waiting line due to variability. The time of arrival of an order is not known ahead of time. We exactly don’t know. It is a random value. We only know its average and standard deviation. The time of next telephone call is not exactly known. The time of arrival of the next car into a gas station is not exactly known. We know the average, and we know the standard deviation. The processing time, the service time: the terms process time, service time, server, processor are interchangeable terms. The service time is not exactly known. We know the average, and we know the standard deviation. 5 minutes average service time. 8 minutes average inter-arrival time. 7.5 customers per hour is the demand. 12 customers per hour is the capacity. Capacity more than demand, but still we observe waiting.

There is an article, psychology of waiting lines. This slide summarizes the basics of that article. You can just read these items.

We said variability and interval and variability and service time will lead to some idle times for resources, and will lead to waiting time for customers. In waiting line analysis, we measure two matrix of interest: how long does it take for a flow unit to pass the waiting line, and to pass the total system. System is composed of waiting line and processor. So we have a waiting line, and we have a processor. How long does it take for a flow unit to go from the start of the waiting line to the end of the waiting line, and how long does it take to go through the processor? And also we are interested in average number of flow units in the line and in the processor. Let’s see on average how long a flow unit spends in the processor and how many flow units are in the processor on average.

Next few slides are very important. I explain very simple concepts. If you just listen to me and get the concept, the subject is extremely easy. I introduce notation Tp, T for time and p for processing. Tp processing time. For example, it takes the assistant of the director of the MBA program at CSUN an average of 5 minutes to process your request. You have a request of something. On average it takes her 5 minutes to do the job, from the time that you go into her room until the time that you come out, that is processing time. You go into a gas station. From the time that you stop in front of the pump until the time that you leave the pump on average it takes, for example, 3 minutes. It is processing time. On average this 5 minutes for the assistant of the director of MBA program to handle the job. Sometimes it may take 30 seconds, sometimes it may take 20 minutes, on average 5 minutes per flow unit per student. At the gas station on average 3 minutes, sometimes just one minute, sometimes maybe 10 minutes, but an average 3 minutes, on average it takes 3 minutes from the time that the car stops in front of the pump until the time that it leaves the pump. There is a processor. This is the assistant of the director of the MBA program. These are students coming in. They should wait if she is busy. On average it takes 5 minutes to go from this point on average it takes 5 minutes to serve a customer, a flow unit. Rp is processing late, so it takes 5 minutes to serve a customer, to serve a flow unit. What is the processing rate if I ask you how many customers are served per hour. If I ask you how many flow units are served or processed per hour, you would simply tell me 60 minutes in one hour. Each flow unit takes 5 minutes to be processed; therefore, processing rate is 12 flow units, 12 customers, 12 cars, 12 students per hour. I don’t like you to answer the question this way. I like to make it a little more difficult. Because I want to generalize the concept. I don’t want to stay in the framework of minute and hour. I would like to extend it to any unit of time. At the same time, I like to exercise a little bit of the Little’s Law. So let me ask you once again, processing time is 5 minutes, average processing time is 5 minutes. What is processing rate? Rp, processing rate for any time unit is 1/Tp. If it takes you 5 minutes to process a flow unit, how many flow units per minute can you process? Obviously 1/5 flow units per minute. So if Tp is stated in terms of minutes, Rp will also be stated in terms of minute. Now if you want to make it per hour, you have Rp equal to 1/Tp, is equal to 1/5 per minute. If you want to make it per hour, you just multiply this one by 60. Then you have per hour. 60 divided by 5 is equal to 12 per hour. The same thing that you found at the beginning by dividing 60 by 5. Rp is equal to 1/Tp. I also like to prove this formula using the Little’s Law. Let’s go back to the assistant of the director of MBA program. Suppose a large number of students are waiting in the waiting room. Therefore, the processor always is busy. As soon as she is done with one student, the student goes out and another moves in. The processor is always busy. Always there is one student with the processor; therefore, average inventory is equal to what? 1. On average, it takes each student 5 minutes to go through the process. It takes 5 minutes for the processor to handle each flow unit, therefore T is equal to 5 minutes. So what is Rp? What is the throughput in this situation when there are large numbers of students waiting and the processor is never idle, in that case always there is one person in the system, average T is 5 minutes. R times T is equal to I. Rp times Tp is equal to Ip. Ip is equal to 1. Tp is equal to 5. 5Rp is equal to 1. Rp is 1 over 5 or in general if the system is always busy, it goes Ip is always 1 TpRp is equal to 1 and Rp is equal to 1/Tp. So from now on we know this formula. We know it intuitively, and we know it also based on the Little’s Law. If Tp is 5, Rp is 1/5, per minute is 60 or 12 per hour.

How let’s look at the relationship between Rp and Tp when we have more than one processor. Here we have only one processor. Tp is the processing time. Rp is processing rate. If there is only one processor, then the relationship between Rp and Tp is Rp is equal to 1 over Tp. But what if there are more than one processor? There are c processors? 2 processors? 3 processors? 4 processors? What if there are c processors? We realize that processing rate is equal to 1 divided by processing time. If processing time is 5 minutes, and if I have one processor, processing rate is 1 over 5. Now if I have two processors, processing rate of the other one is also 1 over 5. They both together have a processing rate of 2 over 5. In general if I have c processors, Rp is equal to c divided by Tp. However, each customer, each flow unit still would be with processor for Tp units of time. Let me exercise the relationship between Rp and Tp using the Little’s Law. Suppose we have enough flow units in the waiting line such that no process is empty even for a fraction of a second. They are always full, always one flow unit is here. One flow unit is here. One flow unit is here. In this case I have 3 flow units always with the processors. In general I have c flow units always with the processors if I have c processors. Right? It takes each flow unit Tp times to go from start of the process to the end of the process. According to the Little’s Law, R times T is equal to I. Rp times Tp is equal to I, but because the system is always full; therefore, I is equal to c, Rp times Tp is equal to c; therefore, Rp is equal to c over Tp. And this is a general relationship between processing rate, processing time, and the number of processors. Again, each flow unit will spend Tp units of time with the server or processor, and the processing rate is Rp equal to c divided by Tp. So if Tp is 5 minutes and we have 3 servers, processing rate of each of the servers is 1/5 per minute, and processing rate of all servers is 3 divided by 5 per minute. If you want to make it per hour, you just multiply 3/5 by 12, and that would be 36. If you do have the minimal difficulty with the very simple concept that you just discussed, just go back, listen to it, and go back to the next part.

Now I will repeat the same thing for inter-arrival time. Inter-arrival time is the time between arrival of two consecutive customers. If average processing rate is 5 minutes, average inter-arrival time must be greater than 5 minutes, because if it is less than 5 minutes then every 4 minutes,, for example, one student comes to the office, and on average it takes 5 minutes for the assistant director to handle this student. Therefore, gradually the waiting room gets crowded and crowded and crowded until the structure of the building will go under. Inter-arrival time should always be greater than processing time, and that would lead to arrival rate to be less than processing rate. If arrival rate is greater than processing rate, we cannot handle the flow units. If processing rate is 12 students per hour, we cannot have 15 students per hour coming in because then each hour 3 flow units will be added to the waiting room, and the waiting room will finally become filled with students. Ta, inter-arrival rate of the flow units. On average every 10 minutes one student arrives, and it takes 5 minutes for the processor to handle the student. So we have enough capacity. Ra is customer arrival rate. Now we should be very comfortable with the relationship between Ta and Ra. Every Ta units of time one flow unit arrives, therefore, 1/Ta flow units will arrive per unit of time. Every 10 minutes one student arrives, therefore, 1/10 students will arrive per minute. If you don’t know how many per hour, I multiply by 60 and that would be 6 students per hour. 1/10 students per minute. Ra is equal to 1/Ta.

Optional, we know Tp. We know Rp. We know Ta. We know Ra. These are related to processor. These are related to the flow units, the customer, the products. Suppose these processors can handle 100 units per hour, and suppose 110 flow units arrive per hour. What would be the output? 110 per hour come. We can handle 100. Output would be 100. Gradually we would have more and more flow units waiting here. The system will explode. If Ra is greater than Rp, output is Rp. Waiting line will go up and up. The system cannot reach steady state, so we cannot have Ra greater than Rp. If we have, output would be equal to Rp. Waiting line will go to infinity. System will collapse. Therefore, the only possibility for a system to work is Ra less than or equal to Rp. So if Ra comes in and capacity here is Rp, and Rp is greater than or equal to Ra, output would be Ra; therefore, throughput to a system is equal to min of Ra and Rp. And R would be equal to Ra. Throughput of the system, Ra must be less than or equal to Rp, and we will later show it can’t be equal to Rp because in that case we will have infinite number of people, infinite number of flow units waiting in the line. Incoming rate must be less than processing rate. Demand must be less than capacity, and throughput is equal to minimum of arrival rate and processing rate and minimum of these is always arrival rate, and the difference is referred to as safety capacity. Safety capacity protected against increasing demand. When demand goes up we don’t want to have a real long rating line in order to protect ourselves against a situation we need some safety capacity, which is the difference between processing rate and arrival. The difference between capacity and average.

Let’s go back and take a look at the whole process. What we are interested in is the flow time and the number of flow units within the boundary of the system. The system is composed of an inflow buffer and a processor or a set of processors; therefore, the flow time is composed of the time that the flow unit spends in the inflow buffer, and we will discuss it later, and the time that the flow unit, the customer, the product spend with the processor, with the server. And it is Tp, and clearly we know what Tp is. We are also interested in knowing the average number of flow units within the system which is equal to the average number of flow units within the system, which is equal to the average number of flow units in the inflow buffer, and the average number of flow units within the processor. We clearly know what Tp is. Later on I will give you a formula for the number of customers in the inflow buffer, and then we can compute Ti based on Ii. But let’s first focus on Ip, because to understand the number of flow units within the server, we really don’t need a formula. It is common sense. Suppose customers arrive on average every 10 minutes, so Ta is equal to 10 minutes. And suppose on average it takes 5 minutes to serve the customer. Therefore, a customer arrives. It takes 5 minutes, and then the processor becomes idle. For 5 minutes the processor is with one customer, for the rest is with no one, and then it will go to the next 10 minutes on average. Every 10 minutes the processor is busy for 5 minutes and idle for 5 minutes.

Any problem with that? No. Let’s look at Ra. If Ta is 10 minutes, Ra is 1/10 per minute or we can multiply it by 60 or 6 customers per hour. Rp is equal to 1/5 per minute, and if I multiply by 60, it is 12 customers per hour. R is equal to minimum of 6 and 12. So R is equal to 6, and Rp is equal to 12 in this case. The processor is busy for 50 percent of the time and idle for 50 percent of the time. 50 percent of the time with one customer, 50 percent of the time with no customer, on average, half a customer.

There is one other way to compute that half a customer. R is equal to 6. Rp is equal to 12. If I divide R by Rp, I will get .5. The average number of customers with this processor.

Let’s define utilization. Utilization is inflow rate divided by processing rate, or throughput divided by process capacity. Utilization must always be less than or equal to 1. We will show it’s not possible to be equal to 1. Safety capacity is the difference between processing capacity and throughput. So utilization is a relative measure. Safety capacity is an absolute measure. In our example, R was 6 per hour. Rp was 12 per hour, therefore, utilization is .5. If you remember, .5 was also the number of customers with the processor on average. And safety capacity is 12 minus 6 equal to 6. Even a single server and the utilization of .3, how many flow units are in the server? Utilization of .3 means the processor is busy 30 percent of the time and is idle 70 percent of the time. R divided by Rp is .3. What we expect from the processor compared to the capacity of the processor is .3. The demand for the server is 30 percent of its capacity. So the processor is only busy 30 percent of the time, and because each processor can only handle one customer at a time, 20 percent of the time the processor is with one flow unit, and 70 percent of the time with 0 flow units. On average the processor is with .3 flow units; therefore, average inventory in the server is equal to utilization. Ip is equal to 1 times U is equal to U.

Now let’s increase the number of servers and see what would happen. Suppose we have two servers, and suppose utilization is equal to .5. That means each server is busy 50 percent of the time and is idle 50 percent of the time. So each server is with one customer 50 percent of the time and with 0 customers 50 percent of the time. Each server is with .5 customers. Utilization is .5. U is equal to .5. Ip is equal to what? Ip is the number of customers, the number of flow units with the processors. .5 customers on average is with this processor. .5 customer on average with the second processor totally one customer is there. How did we get it? Ip is equal to cU. Let’s have another look at the system. Now, we clearly know what is Tp. We know what is R. We know what is Rp. Because we know these two, we can compute utilization which is R/Rp. If I have c processors, Ip is equal to cU. Without going through any formula, just based on common sense, we are able to compute Tp and Ip. Later on we will talk about Ti and Ii, and as soon as we are done with them, we can compute T and I.

Processing time, Tp, for example, is 5 minutes. Number of servers for example c equals 3. Tp is also the waiting time in the server no matter one server or 2 servers or 3 servers or 1 million servers, on average each flow unit, each product, each customer should spend Tp units of time with the processor. Tp does not depend on the number of processors. Processing rate is 1/Tp for one server and c divided by Tp for c servers. In this example I have 3 servers, Tp is 5, Rp is 3/5 flow units per unit of time, which is a minute in this case. If I multiplied by 60, then it is 36 flow units per hour. Suppose in this example utilization is .8. Servers are busy 80 percent of the time and idle 20 percent of the time in this example. The number of flow units with all servers we already have shown it. Ip is equal to cU. U is 8. C is 3. Therefore, Ip is 2.4.

Can we compute R? Given these examples, given these numbers, can we compute R? Given these numbers – so as soon as you give me Tp and c, I can compute Rp. That’s simple. Then if you give me U, I can compute Ip. So I have Rp, which is equal to 3/5 per minute, and I have Ip which is equal to 2.4. This Rp was 3/5 per minute or 36 per hour. Can you compute R? Of course. U is equal to R divided by Rp; therefore, R is equal to U times Rp. U is equal to .8. Rp is equal to 36 per hour; therefore, R is equal to 28.8.

We can do all of these computations without looking at any formula.