This is my second talk on waiting line. In previous talk, we discussed what is happening in the processor. Now we discuss what is happening in the buffers. This is a general representation of a waiting line. A simple waiting line, formed by one or more parallel processor. So we don’t discuss a sequence of processors, just one or more than one parallel processors. Customers or products arrive here; they will go to the empty processor, R process, and leave the system. In this system, customers or products will arrive rate of R per time period. R per minute, R per hour, R per day, R per week. That is our throughput. This throughput goes through the buffer, then goes through the processors, and then leaves the system. R per time period comes in, R per time period goes out because the system is in a stable situation. So one parameter is R. Throughput per time period. The other parameter that we know it well is processing time; effective processing time. The time it takes on average to process one flow unit. One patient, one customer, one visitor, one product. Tp, effective time of processing one flow unit. If I know Tp, and if I know the number of processors, then Rp in the case of one processor is one over Tp and in the case of c processors, Rp is equal to c divided by Tp. And this c is small c not capital c. and then Ip is quite easy to computer because in the spirit of the Little’s Law, R which is throughput, times T which is Tp in this case, is equal to I which is Ip in this case. And therefore if I know TP, and if I know R, then I can compute IP. We already know these components of the system. All I do need to know are these two components. And in the spirit of the Little’s Law, if I know one of them, then because I also know R, then I can know the other one. If I know Ip, and if I know Ii, Ip is the number of units in the processor, or processors, and Ii is the number of flow units in the buffer. If I know these two, then I can add then up, then I can know I in the total system. Also, if I know Tp, and if I know Ti, then I can add them up, and I can compute T. T is the total time in the system. I is the total number of flow units in the system. Both in the buffer and in the processors. Ti is the time in the buffer, Ii number of flow units in the buffer. Tp, the time in the processor, Ip, the total number of flow units in all processors. according to the little law, for the total system, I is equal to RT and R is the throughput. The number of units which come in, and is equal to the number of units going out. I can write the Little’s Law here. Again, R which is throughput throughout the system which is equal to I divided by T for the buffer. Ii is equal to R times T. and regarding the processors, Ip is equal to R times Tp. Please note that this is not Rp, this is R, because R comes in and R goes out. The capacity is Rp, but the throughout is R, and therefore because throughput is here R, it is R, it is R, it is R, therefore here I have R equal to I divided by T, here I have R equal to Ii divided by Ti, here I have R equal to Ip divided by TP. From out previous discussion, we know utilization is equal to R divided by Rp. We have already leaned Rp is equal to c divided by Tp. And here, we learn that R is equal to Ip divided by Tp. So I have R and I have Rp and according to the definition of the utilization, utilization is equal to R divided by Rp. Instead of R I can use Ip or Tp. And instead of Rp I can use c divided by Tp. And if I put them into this equation I will get u is equal to Ip divided by c. So utilization which was defined as throughput divided by capacity can be also stated in terms of the number of flow units in all processors divided by the number of processors. And this is intuitively clear; we didn’t need mathematical proof that utilization is also equal to number of flow units in the processor divided by the number of processors. There are two important characteristics is all waiting lines which have profound impact on the number of flow units in the waiting line. The first one is utilization. As utilization goes up, the number of people in the buffer go up. The other one is variability. If every four minutes, exactly every four minutes, one customer arrives, one flow unit arrives, and it takes exactly three minutes for that flow unit to pass the processor, then we never observe the waiting line. Even if this is four minutes, if we have a situation like this, where utilization is one hundred percent, that is the maxim possible utilization, but there is absolutely no variability, neither in inter-arrival time nor flow time, then we never see flow units in the buffer. Buffer is always empty. However if processing time is one minute and inter-arrival time is still four minutes, but there is significant variability in inter-arrival time, then we still have waiting line. Suppose two people arrive exactly at the same time, and the processor is empty. One of them goes in, it takes one minute, but the other one, should wait for one minute, until the processor gets empty. Therefore we can have one hundred percent utilization only in the case there is no variably. But if there is variability, we can have waiting line even in the case where processing time is much less than inter-arrival time. It is possible that the process is idle because no customer is here. And we may have the processor busy and customers are waiting here. As I explain earlier in a waiting line analysis, we are interested in knowing the number of people in the total system, in the buffer and in the processor. In the total system, in the waiting line, in the buffer, and in the processor. As I explained earlier, if we know R, and we know Tp, we can easily computer Ip. Therefore all we need to compute is Ti and Ii. And if I know one of them because I know R, then I can computer the other one. We have an approximation formula for Ii and we do have also exact formula for some specific cases of waiting line. But in this course, we will mostly use the approximation formula. Let’s once again discuss the impact of utilization and variability on the number of people in the waiting line or the time spent in the system or in the waiting. This is zero utilization, this is one hundred percent utilization, and this is the time that the flow unit spends in total system. If there is no waiting line, on average, each flow unit will spend Tp time in the system because they come, no one is in the waiting in. There are one or more processors empty and available. They go into one of them, and on average it takes Tp units of time to enter the processor and exit. However, if there are people in the waiting line, then the time in the system is greater that Tp. If utilization is low, it may be a little grater that Tp, but if utilization is high, it may be significantly greater than Tp. If variability is low, and utilization is low, it will be a little bit greater that Tp, but if utilization is high, and variability is also high, it will be more. If variability is even higher, then it may be like this. Our measures of effectiveness, as we explained, are the numbers of people in the system and the time they spend in the system. And that is as we discussed affected by two units: utilization, the higher the utilization, the longer the waiting line and waiting time. Variability, the higher the variability, in inter-arrival time, and in processing time, the longer the waiting time, and waiting time in the waiting line. We said utilization is equal to R divided by Rp, so if utilization is higher, either R is high or Rp is low. Either inflow rate is high, or processing capacity is low. Processing capacity itself is c divided by Tp. If Rp is small, then it is either because c is small or it is because Tp is large. We first try to look at Tp and make Tp smaller by improving the methods by automation, by training, by better management. Decreasing Tp is much cheaper than increasing c. Increasing c means if we have two machines and two operators, we should make it three machines, and three operators. Decreasing Tp means if I am doing this job in five minutes, I can train the worker, I can improve the method of doing the job, I can automate a portion of the job, I can have a better management to make sure always input raw materials is available, and there is no absentees, then Tp may go down. The second driver of the process performance is variability. Variability in the inter-arrival time and variability in the processing time. In the simplest case, variability is measured by stand deviation or variance. And variance is the square of stand deviation. Standard deviation is not enough to understand the extent of variability. Does the standard deviation of twenty represent more variability or standard deviation of one fifty, we don’t know. Does standard deviation for an average of eighty represent more variability, or a standard deviation of one fifty for an average of one thousand? Twenty with respect to eighty. Twenty percent. One fifty divided by one thousand, fifteen percent. Therefore with respect to the mean, this twenty is greater than this one fifty; stand deviation divided by age is referred to as coefficient of variation. Then we have coefficient of variation on inter-arrival time, and coefficient of variation of processing time. We refer to coefficient of variation of inter-arrival time as Ca, and coefficient of variation of processing time as Cp, capital C. The number of servers we show it by small C, coefficient of variation, we show it by capital C. Capital C small a for coefficient of variation of inter-arrival time, and capital C small p for coefficient of variation of processing time. Now we present the approximation formula for the number of flow units in the buffer, the waiting line. Ii, the number of flow units in the buffer is equal to utilization to the power of two times one plus the number of servers divided by one minus the utilization. And we refer to it as utilization affect or u-part. We explained earlier that when utilization goes up, number of people in waiting line goes up. And here we can see as utilization goes up, as utilization gets close to one, the denominator gets closer to zero, and something divided by something close to zero gets larger and larger. Utilization affect is multiplied by variability affect. And that is squared coefficient of variability of inter-arrival time plus square of coefficient of variability of processing time divided by two. And we refer to it as variability affect or v-part. Please note that if there is no variability in inter-arrival time or processing time, then this number is zero. And when this number is zero, u can become even one, and still we could have empty waiting line. However, in presence of variability, even a small value of utilization can lead to some people in waiting line. Ca and Cp are coefficients of variation and they are standard deviation divided by mean for inter-arrival time and processing time. As I explained, utilization effect or u-part states that the queue length increases rapidly as utilization approaches one. It will not increase linear. It will not increase in a decreasing order. It will increase rapidly, as utilization gets close to one. We cannot have hundred percent utilization. Let it alone we cannot have one hundred ten percent utilization, one hundred twenty percent utilization. Therefore if someone tells you that we are working at one hundred twenty percent of capacity, there are two possibilities. One is that the person does not know what he’s talking about, second, he does not know what the capacity is. And then we have variability effect which says that wailing line length increases as variability increases in inter-arrival time and processing time increases. Again, while throughput is always less than capacity we cannot have throughput equal to capacity in the presence of variability. We cannot have utilization equal to one in presence of variability because queue length goes to infinity while capacity always is not fully utilized still in the presence of variability we will see flow units in the buffer, flow units in the waiting line. Suppose on average, six people per hour come into our system, and suppose on average Tp is equal to five minutes, six people per hour arrive, we can handle 12 people per hour. But suppose in the first thirty minutes, because we have variability, no one shows up. Fifty percent of this capacity flies out. It’s purged. Suppose the other six come together, five of them should stay in the waiting line, then four should stay in the waiting line, three, two, one should stay in the waiting line. And we will have people in the waiting line while utilization is R divided by Rp equal to fifty percent. Capacity is much higher than throughput; still we have flow units in the buffer, flow units in the waiting line. Let’s see what is coefficient of variation for a couple of well-known distribution function. We know if Tp is average processing time, Rp, capacity, or processing rate is equal to number of servers divided by Tp. We know if Ta is inter-arrival time, the time between arrival of two consecutive flow units, then Ra, which is arrival rate, is equal to one over T. Suppose Sp is standard deviation of processing time and Sa is standard deviation of inter-arrival time. Our processor, or, our processors, has a processing time of Tp, and standard deviation of Sp. Our inter-arrival times have an average of Ta, and standard deviation of Sa. For a general distribution with exception of Poisson, exponential, and constant for a general distribution, if processing time, average processing time, is TP, and inter-arrival time is Ta. So if processing time does not have a specific distribution such as exponential, Poisson, or constant, and if its average is TP, and its standard deviation is Sp, then its coefficient of variation, is Sp divided by Tp. The same is for inter-arrival time. If inter-arrival time is a general distribution, and by general distribution we mean it is not Poisson, not exponential, and it is not constant, and it has an average of Ta and standard deviation of Sa, then for inter-arrival time, coefficient of variation is Sa divided by T. However, if the number of units arriving or the number of units being processed for a specific period of time has Poisson distribution, then in Poisson distribution, mean and standard deviation are equal. So if average processing time is Tp, standard deviation of processing time is also Tp, and therefore coefficient of variation is Tp divided by Tp which is one. The same is true for inter-arrival. If the number of units that arrive over a specific period of time follow Poisson distribution, then the inter time has a standard deviation which is equal to the average. Ta and Sa are equal, they are both the same, and therefore coefficient of variation is one. If the inter-arrival time or processing time follows exponential distribution, then again, standard deviation of the distribution is equal to its mean. Standard deviation of the distribution is equal to its mean, equal to its average. And therefore, standard deviation of p divided Tp or stand deviation of a divided Ta. If processing time follows exponential distribution, then Sp divided Tp is equal to one, and if inter-arrival time follows exponential distribution, then Sa divided Ta is equal to one. Finally if the distribution is constant, that means there is no variability; absolutely no variability. And when there is no variability, standard deviation is equal to zero, and therefore coefficient of variation is equal to zero divided by something and that is equal to zero. In general, in a problem, if you don’t have any additional information regarding processing time or inter-arrival time, coefficients of variation are computed this way. If the number of units which are processed over a period of time follow Poisson distribution, then coefficient of variation is equal to one. If number of units that arrive follows Poisson distribution, then coefficient of variation of inter-arrival time is equal to one. The same for exponential distribution. If processing time follows exponential distribution, coefficient of variation, is one. If inter-arrival time follows exponential distribution, coefficient of variation is equal to one. For constant or deterministic, coefficient of variation is equal to zero. You may have been surprised why I show Poisson and exponential, using a notation M. Because general it is clear. I show it to you as notation G that means we do not have specific information about type of distribution. I mean we know that is not exponential, it is not Poisson, it is not constant and we show it, because general we show it by G. Constant or deterministic, we show it by D. But by Poisson and exponential are shown by M. Let me first tell you something. If the number of units which are processed or arrive over a period of time follows Poisson distribution, then the time between two consecutive units either arriving to the system or going out of the system follow exponential distribution. So exponential shows the time between two events and Poisson shows the number of events over a specific time period. That is the relationship between Poisson distribution and exponential distribution. And hopefully you remember them from you statistics course. But that doesn’t resolve our issue why we show Poisson and exponential by M. We show them by M because they are related to a process called Markovian process or Markov process and Markov was a mathematician, a Russian mathematician.