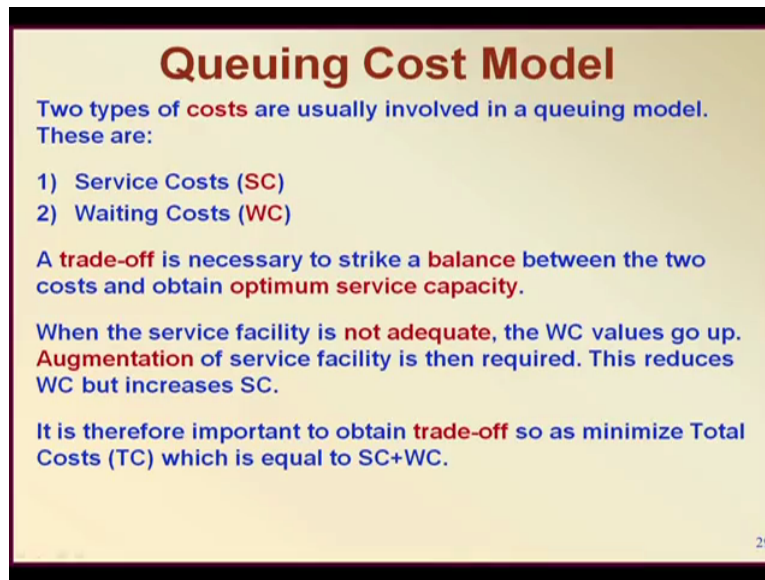


Course on Decision Modeling
Professor Biswajit Mahanty
Department of Industrial and Systems Engineering
Indian Institute of Technology Kharagpur
Module 04
Lecture No. 20
Queuing Cost, Priority and Networking Models

(Refer Slide Time: 00:33)



Queuing Cost Model

Two types of **costs** are usually involved in a queuing model. These are:

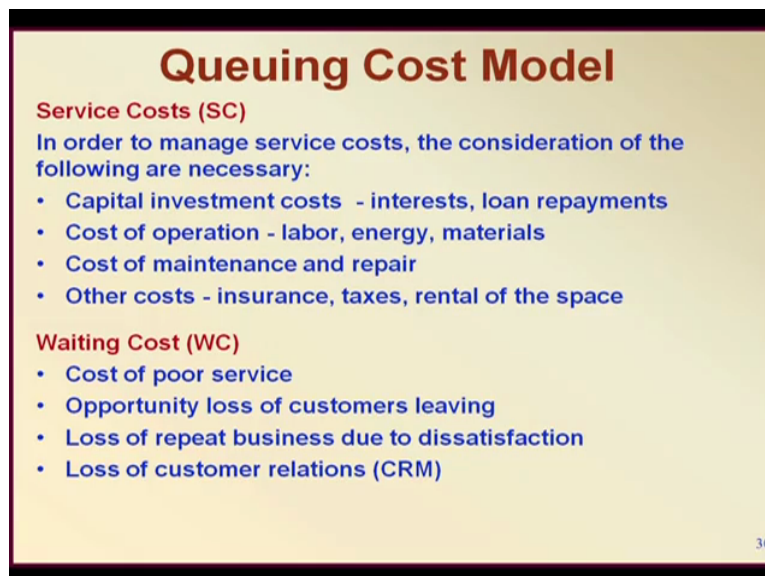
- 1) Service Costs (**SC**)
- 2) Waiting Costs (**WC**)

A **trade-off** is necessary to strike a **balance** between the two costs and obtain **optimum service capacity**.

When the service facility is **not adequate**, the WC values go up. **Augmentation** of service facility is then required. This reduces WC but increases SC.

It is therefore important to obtain **trade-off** so as minimize Total Costs (TC) which is equal to SC+WC.

29



Queuing Cost Model

Service Costs (SC)

In order to manage service costs, the consideration of the following are necessary:

- Capital investment costs - interests, loan repayments
- Cost of operation - labor, energy, materials
- Cost of maintenance and repair
- Other costs - insurance, taxes, rental of the space

Waiting Cost (WC)

- Cost of poor service
- Opportunity loss of customers leaving
- Loss of repeat business due to dissatisfaction
- Loss of customer relations (CRM)

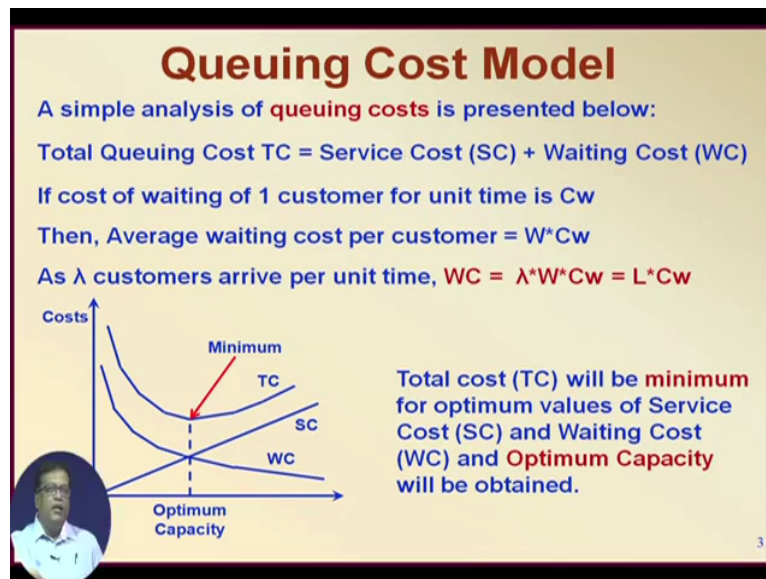
30

In this particular session we are going to discuss the queuing cost, the priority and the network models and conclude the queuing system that we have discussed so far. Look here as we discussed in our previous class that there are 2 kinds of queuing cost, the service cost and waiting cost and a trade-off is necessary between the two. So what are some components of service cost, as you can see it could be capital investment costs, the cost of operation; cost of

labor, energy, materials, cost of maintenance and repair and other cost like insurance, taxes, rentals of the space and so on.

Waiting cost on the other side may be poor service cost, opportunity loss, loss of repeat business or loss of customer relations all this could be contributing to waiting cost. You see real analyses will be little complicated and one really needs to look at all of these factors to really come up with a particular queuing setup.

(Refer Slide Time: 01:48)



But once the queuing setup is already there and steady-state of operation is found at that time you know a simple analyses of queuing constant can be really obtained where just look at this calculations, the total cost will be service cost plus waiting cost and cost of waiting. Let us look at this simple thing, supposed cost of waiting of one customer per unit time is C_w and average waiting cost because W is the waiting time per customer, so what will the average waiting cost per customer is W into C_w .

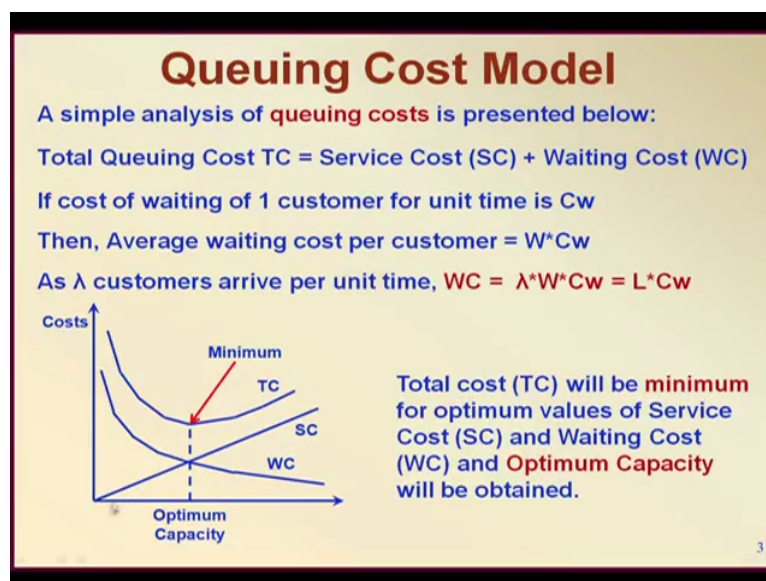
As λ customers are arriving per unit time really WC the waiting cost is equal to λ times W times C_w but from Little's formula λW equal to L , so what is the total expression for waiting cost L into C_w , C_w is cost of waiting for one customer per unit time, right. One point we remember here although most of the time we shall take L or W which is for the system, but certain kind of problem could be there where more relevant cost can be on the basis of WQ rather than on LQ , is it alright?

Say suppose you are really going to a movie and when you are going to a movie then the waiting for entering the hall getting the ticket etcetera, is it alright? Supposing the service

here is actually seeing the movie and the rest of the time is waiting, so if that is so then that particular thing when you are actually getting a service that is not really a cost as far as the customer's concern, what is really is more important is the waiting.

But on the other hand suppose a customer is going to get a tool then it does not matter whether you know you spend time on the queue or more time getting the tool really it is the total time that is important. So based on the given situation what is more important W or WQ based that also has to be remembered and waiting cost really can be calculated on that basis.

(Refer Slide Time: 04:18)



So if you really look at this graph that service cost on the other hand really depends on the number of service facilities or number of servers and usually it is a straight-line type of thing that this more capacity, more service capacity more will be service cost. And more serving capacity is given the waiting cost reduces, if you add the 2 cost you get what is known as the total cost and the total cost figure will look something like this, right.

So there will be a minimum total costs somewhere usually it will be at the crossing of SC and WC and that is where we will get the optimum capacity, is it alright? So we have to really look at that as you increase the service you get more service cost but less waiting cost at a given a point of time you know both will be you know the sum of the 2 will be minimum and that is where we will get what is known as optimum capacity.

(Refer Slide Time: 05:28)

Queuing Cost Model Example

Service mechanics come to take spares at a shop at 6/hour on the average. Waiting for them costs Rs. 8/- per hour. A Shop attendant's wage is Rs. 5/- per hour. There is only one counter.

The service rates are: 1 attendant: 8/hour, 2 attendants: 12/hour, 3 attendants: 16/hour.

With usual M/M/1 assumptions, find how many attendants to choose.

32

So let us look at an example, service mechanics come to take spare at a shop at 6 per hour on the average. Waiting for them cost rupees 8 per hour, shop attendants wage is rupees 5 per hour and there is only one counter. The services rates are 1 attendant 8 hour 8 per hour, 2 attendant 12 per hour and 3 attendants 16 per hour, with usual MM 1 assumptions find how many attendants to choose.

(Refer Slide Time: 06:18)

Queuing Cost Model Example

Service mechanics come to take spares at a shop at 6/hour on the average. Waiting for them costs Rs. 8/- per hour. A Shop attendant's wage is Rs. 5/- per hour. There is only one counter. The service rates are: 1 attendant: 8/hour, 2 attendants: 12/hour, 3 attendants: 16/hour. With usual M/M/1 assumptions, find how many attendants to choose.

Answer: First Compute ρ and W.

No. of Attendants (N)	Arrival Rate per hour (λ)	Service Rate per hour (μ)	Utilization Factor $\rho = \lambda/\mu$	Waiting time (W) in hour $W = \rho/((1 - \rho)*\lambda)$
1	6	8	$6/8 = 3/4$	$(3/4)/(6*1/4) = 1/2$
2	6	12	$6/12 = 1/2$	$(1/2)/(6*1/2) = 1/6$
3	6	16	$6/16 = 3/8$	$(3/8)/(6*5/8) = 1/10$

33

Look here you know if you choose only one attendant your service will be less, but if you choose more attendant then your service will be better but cost will be more. So really if you look at this theories and analysis if the number of attendant is 1, arrival rate is 6, service rate

is 8, 2, 6 and 12, 3, 6 and 16, right. So based on that the utilisation factor are different and waiting time which is rho by 1 minus rho that is L divided by lambda that is to be W will be less, also if 1 attendant waiting time is half you know half hour, 2 attendant waiting time is 1/6th hour and 3 attendant waiting time is 1/10th hour.

(Refer Slide Time: 07:27)

Queuing Cost Model Example

Service mechanics come to take spares at a shop at 6/hour on the average. Waiting for them costs Rs. 8/- per hour. A Shop attendant's wage is Rs. 5/- per hour. There is only one counter. The service rates are: 1 attendant: 8/hour, 2 attendants: 12/hour, 3 attendants: 16/hour. With usual M/M/1 assumptions, find how many attendants to choose.

Answer:

No. of Attendants (N)	Arrival Rate per hour (λ)	Service Rate per hour (μ)	Utilization Factor $\rho = \lambda / \mu$	Waiting time (W) in hour $W = \rho / ((1 - \rho) * \lambda)$	Service Cost (Rs.) $SC = N * 8hr * 5$	Waiting Cost (Rs.) $WC = W * 8hr * \lambda * 8$	Total Cost $TC = SC + WC$
1	6	8	3/4	1/2	$1 * 8 * 5 = 40$	$(1/2) * 8 * 6 * 8 = 192$	232
2	6	12	1/2	1/6	$2 * 8 * 5 = 80$	$(1/6) * 8 * 6 * 8 = 64$	144 Lowest
3	6	16	3/8	1/10	$3 * 8 * 5 = 120$	$(1/10) * 8 * 6 * 8 = 38.4$	158.4

Hence, We may choose 2 attendants – total cost is Rs. 144.

34

Now question is whether you want your people to wait for half-hour or 10 minutes or really even less like 6 minutes, right. What is the level of service that you should provide? So it requires a calculation of cost, so what is the cost of waiting in terms of W but then how many customers are arriving that also is to be seen.

So here look at those details, what is a service cost? The service cost is if you put one attendant assuming an 8 hours day for an 8 hours day you know there will be 5 rupees per hour that is shop attendant wage, so it comes to 40 rupees, if you put 2 attendant 80 rupees, 3 attendant 120 rupees. But what is the waiting cost? Waiting cost as you know how many customers arrival in in a given day that is lambda into 8 hours, right lambda into 8 hours that many customers are arriving.

So what is the waiting time? Waiting time is W. So W into lambda into 8 hours into 8, why 8? Because waiting cost is Rupees 8 per hour that will be the total waiting cost, right. So W into 8 into lambda, because lambda is number of arrivals per hour in 8 hours so many peoples have arrived multiplied by W that is the total waiting time of people and cost 8 per hour. So for one attendant it is 192 rupees, 2 attendant 64 rupees and 3 attendant 38.4 rupees. So if you

add them all then you get then total cost comes out to be 232, 144 and 158.4 and since 144 is the lowest amongst them, so we may choose 2 attendants total cost is 144, right.

(Refer Slide Time: 09:11)

SCET
I.T.KGP

Queuing Costs

Service Cost = No. of attendants * hours in a day
* hourly rate of an attendant.

Waiting Cost = Waiting time per customer (w) *
No. of customers arriving in a day
($\lambda * 8$) * Waiting cost per hour
per customer.

So how have we really done the queuing cost? Let us look at very carefully queuing cost, service cost equal to number of attendants multiplied by hours in a day multiplied by hourly rate of an attendant and waiting cost equal to waiting time per customer it is W into number of customers arriving in a day that is lambda star 8, if 8 hours a day star waiting cost per hour per customer, right. So this is how the service cost and waiting cost are calculated and when you add them you will get the total cost and that particular combination have to be chosen for which the waiting cost is minimum.

(Refer Slide Time: 11:05)


Another Queuing Cost Example

Machines fail at 4 per hour and the cost of non-productive machine is Rs. 9 per hour. A fast repairman charges Rs. 6 per hour and repairs at 7 per hour. A slow repairman charges Rs. 3 per hour but repairs at 5 per hour. With usual M/M/1 assumptions, find which repairman to hire.

Answer:

Repairman	Arrival Rate per hour (λ)	Service Rate per hour (μ)	Utilization Factor $\rho = \lambda/\mu$	Average No. of machines in repair (L) $L = \rho/(1-\rho)$	M/c Idle Cost (Rs.) $IC = L \cdot 8 \text{ hr} \cdot \text{hourly cost}$	Repairman Cost (Rs.) $RC = 8 \text{ hr} \cdot \text{hourly cost}$	Total Cost $TC = SC + WC$
Fast	4	7	4/7	$(4/7)/(3/7) = 4/3$	$(4/3) \cdot 8 \cdot 9 = 96$	$8 \cdot 6 = 48$	144 Lower
Slow	4	5	4/5	$(4/5)/(1/5) = 4$	$4 \cdot 8 \cdot 9 = 288$	$8 \cdot 3 = 24$	312

We may choose the Fast Repairman – total cost is Rs. 144.

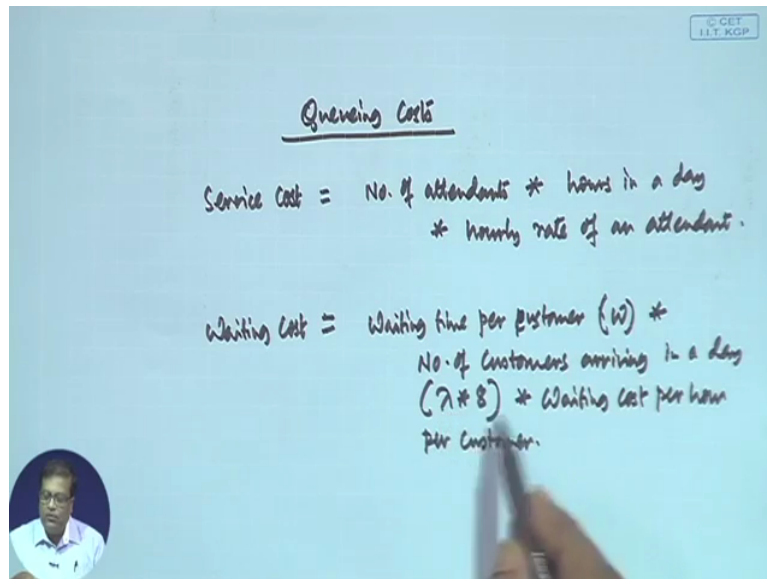


36

Another example; machines fail at 4 per hour and cost of non-productive machine is 9 per hour. A fast repairman charges is rupees 6 per hour and repairs at 7 per hour. A slow repairman charges rupees 3 per hour but repairs at 5 per hour, with usual MM 1 assumptions find which repairman to hire. So if you look at the arrival rate is 4 and 7 for fast repairman, for the slow repairman the arrival rate is 4 and service rate is 5 per hour.

So what are the utilisation factor ρ is λ/μ and slope 4 by 5, so average numbers of machines in repair L which comes out to be 4 by 3 and for slow it becomes 4. So machines idle cost is a L into 8 hours, because L is the number and L into 8 hours, so if you see we have what has really done since λW equal to L you know it is replaced by L.

(Refer Slide Time: 12:19)



So because you know look at this formula here we have multiplied W into lambda into number of hours in a day that is 8, but lambda W is L you could have also taken L into 8 hours that is also another way of doing it or even if you do not multiply it by 8 hours if you simply leave it as L then remember service cost also to be compute per hour so both ways possible.

(Refer Slide Time: 12:41)

Another Queuing Cost Example

Machines fail at 4 per hour and the cost of non-productive machine is Rs. 9 per hour. A fast repairman charges Rs. 6 per hour and repairs at 7 per hour. A slow repairman charges Rs. 3 per hour but repairs at 5 per hour. With usual M/M/1 assumptions, find which repairman to hire.

Answer:

Repairman	Arrival Rate per hour (λ)	Service Rate per hour (μ)	Utilization Factor $\rho = \lambda/\mu$	Average No. of machines in repair (L) $L = \rho/(1-\rho)$	M/c Idle Cost (Rs.) $IC = L \cdot 8 \text{ hr} \cdot \text{hourly cost}$	Repairman Cost (Rs.) $RC = 8 \text{ hr} \cdot \text{hourly cost}$	Total Cost TC = SC+WC
Fast	4	7	4/7	$(4/7)/(3/7) = 4/3$	$(4/3) \cdot 8 \cdot 9 = 96$	$8 \cdot 6 = 48$	144 Lower
Slow	4	5	4/5	$(4/5)/(1/5) = 4$	$4 \cdot 8 \cdot 9 = 288$	$8 \cdot 3 = 24$	312

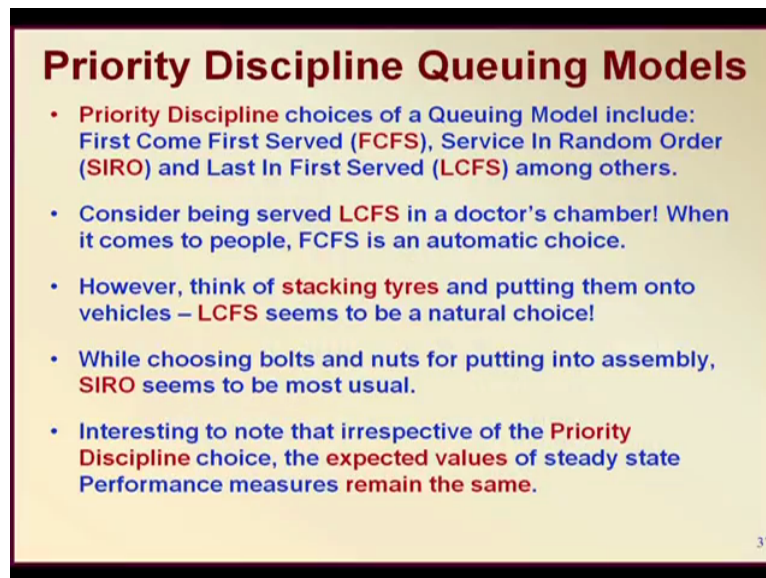
Hence, We may choose the Fast Repairman – total cost is Rs. 144.

36

Anyway machine idle cost L into 8 hours so it comes out to be 96, 4 by 3 is the average number of machines in repair into 8 into 9 and slow comes out to be 288 and this is the repairman cost for the fast and the slow. So total cost comes out to be 144 here for the fast

repairman and for the slow repairman 312, since this is lower we choose the fast repairman. So for such examples you must remember that more better facilities requires better more service cost and however waiting cost will be lower, but on the other hand if you go for something like slow repairman, the repairman cost is low but ideal cost is high, so trade-off is necessary.

(Refer Slide Time: 13:47)



Priority Discipline Queuing Models

- **Priority Discipline** choices of a Queuing Model include: **First Come First Served (FCFS)**, **Service In Random Order (SIRO)** and **Last In First Served (LCFS)** among others.
- Consider being served **LCFS** in a doctor's chamber! When it comes to people, **FCFS** is an automatic choice.
- However, think of **stacking tyres** and putting them onto vehicles – **LCFS** seems to be a natural choice!
- While choosing bolts and nuts for putting into assembly, **SIRO** seems to be most usual.
- Interesting to note that irrespective of the **Priority Discipline** choice, the **expected values of steady state Performance measures** remain the same.

37

That is about queuing cost, a very important some observations let us make on because you know we really do not have time and scope to do justice to very important topic like priority discipline queuing models, something like the network queuing models, but since they are very important topic let us at least mention the major features of them. The priority discipline choice of a queuing model it could be first-come first-served, service in the random order and last in first-served, right.

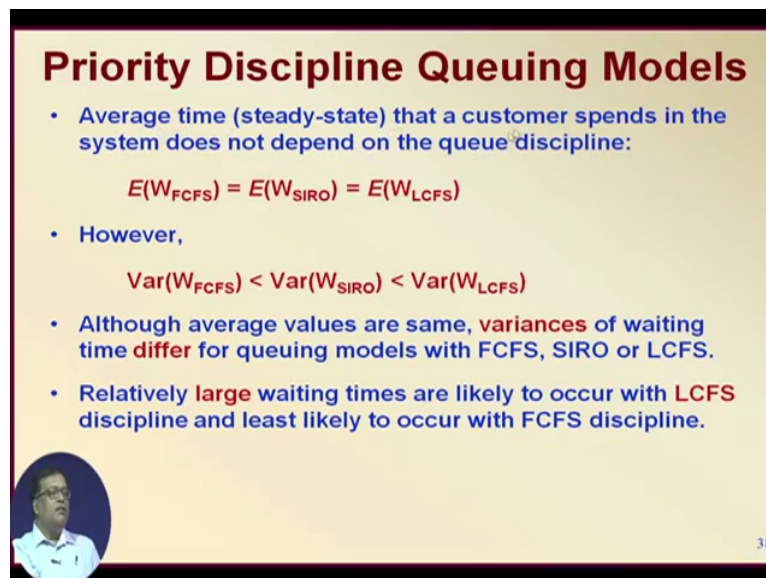
So some examples supposing you know you are going to a doctor and you are a patient and you are waiting and doctor is following last come first-served, how will you feel? You will find that you have come in 8 o'clock in the morning and waiting, someone comes at 12, maybe 4 hours you are waiting and that person is given service earlier, right. So it looks very odd and therefore when it comes to people usually the last come first-served and other choices are not automatic, FCFS first-come service is the automatic choice.

But think of staking tyres, know one tyre is put above another tyre right and those tyres will be now used in let us say earth moving vehicles for tyre replacement. Now you see obviously are tyre that is at the lowest must have come first and tyre that is on top must have come last,

so which tyre will you put into your vehicle, obviously the tyre on top which has come last, so last in first out or last come first-served becomes an automatic choice wherever such staking is involved, so it could be those type of things.


And also think suppose you are putting bolts and nuts into an assembly process of cars and those bolts and nuts are really in a box and somebody come in and replacing them from time to time. In this kind of situation obviously we will choose a bolt and nut randomly from the box and because they are all of the same size and put it into your regular assembly process right, it is an process where you are using the same type of bolts and nuts and for a repetitive process, so obviously we choose those bolts and nuts in a random order, so that is an example of service in random order.

(Refer Slide Time: 16:50)



Priority Discipline Queuing Models

- Average time (steady-state) that a customer spends in the system does not depend on the queue discipline:
$$E(W_{FCFS}) = E(W_{SIRO}) = E(W_{LCFS})$$
- However,
$$\text{Var}(W_{FCFS}) < \text{Var}(W_{SIRO}) < \text{Var}(W_{LCFS})$$
- Although average values are same, **variances of waiting time differ** for queuing models with FCFS, SIRO or LCFS.
- Relatively **large** waiting times are likely to occur with **LCFS** discipline and least likely to occur with **FCFS** discipline.

 38

It is interesting to know irrespective of the priority discipline choice the expected value of steady-state performance measures remains the same. This I told many times but let us look at formula that average time steady-state that a customer spends in the system does not depends on the queue discipline. That means expected value or waiting time for FCFS-SIRO or LCFS they are all equal, that is why I said the for usual queues where the arrival or service pattern does not depends on the queue discipline right.

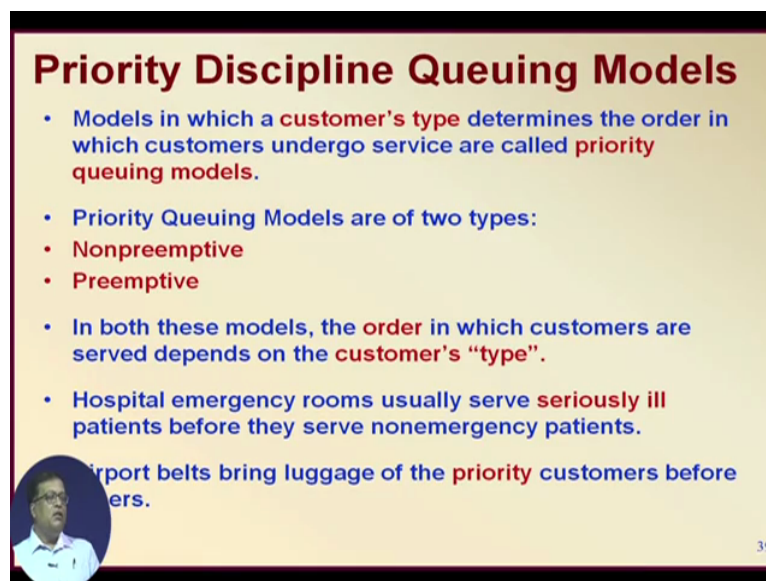
But if the arrival or service pattern changes the queue discipline, for example the SPT rule, what happens in SPT rule? Supposing there are several jobs they have different processing time and you take that job which has got the shortest processing time and process it fast. So

basically the processing is dependent on arrival because you have already resorted them and sending them in these order.

Here this definitely the queue discipline will affect the system performance, but that first-come first-served and this kind of rules which is coming from a different process they really do not change the steady-state parameters values. But do they not change different in their variances, look at the variances the variances will be the LCFS will be the highest followed by that of SIRO followed by that of FCFS, why? Because in first-come first-served everyone waits at a given amount of time, but in the last come first-served someone waits very less, someone waits for a very long time.

So therefore their variances are very different, right, so therefore although average values are same variances are differing. Relatively large waiting times are likely to occur with last come first-served discipline and least likely to occur with first-come first-served discipline, alright this point has to be remembered.

(Refer Slide Time: 19:04)



Priority Discipline Queuing Models

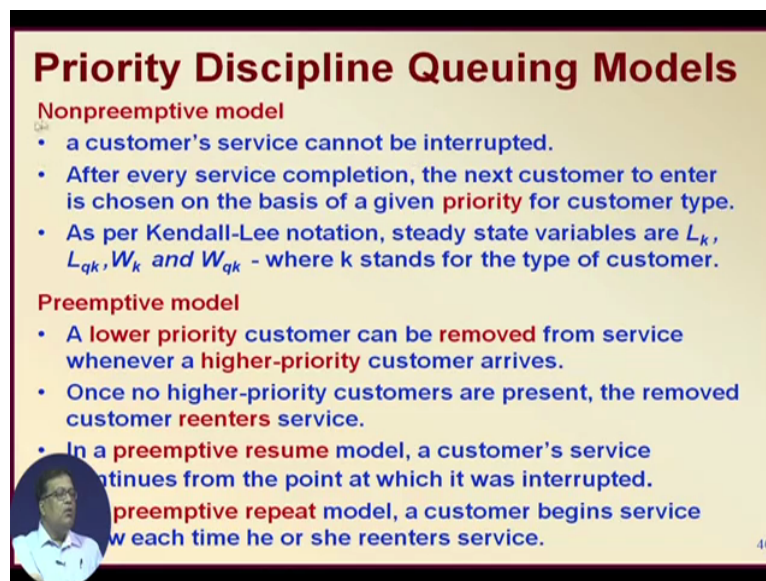
- Models in which a customer's type determines the order in which customers undergo service are called priority queuing models.
- Priority Queuing Models are of two types:
 - Nonpreemptive
 - Preemptive
- In both these models, the order in which customers are served depends on the customer's "type".
- Hospital emergency rooms usually serve seriously ill patients before they serve nonemergency patients.
- Airport belts bring luggage of the priority customers before others.

39

Now priority discipline queuing models you know models in which customers type determines the order in which customers undergo service they are called priority queuing models. You see what happens here the customers are separate, there are some customers there are priority customers, there are some other customers they are non-priority customers. Maybe there could be several classifications, maybe the customers in 3-4 types, one is low priority, medium priority, high-priority, very high priority.

Think of an emergency service in hospital, there are several customers mean patients those patients who are seriously ill, obviously you have to attend them first. Someone who has a minor injury may he or she can wait a little bit more, so that means someone who has major injury, seriously ill conditions should be given priority, right and what happens those customers are serviced earlier. We see that airport belts bring languages of priority customers before the others luggages are brought.

(Refer Slide Time: 20:34)



Priority Discipline Queuing Models

Nonpreemptive model

- a customer's service cannot be interrupted.
- After every service completion, the next customer to enter is chosen on the basis of a given **priority** for customer type.
- As per Kendall-Lee notation, steady state variables are L_k , L_{qk} , W_k and W_{qk} - where k stands for the type of customer.

Preemptive model

- A **lower priority** customer can be **removed** from service whenever a **higher-priority** customer arrives.
- Once no higher-priority customers are present, the removed customer **reenters** service.
- In a **preemptive resume** model, a customer's service continues from the point at which it was interrupted.
- In a **preemptive repeat** model, a customer begins service **afresh** each time he or she reenters service.

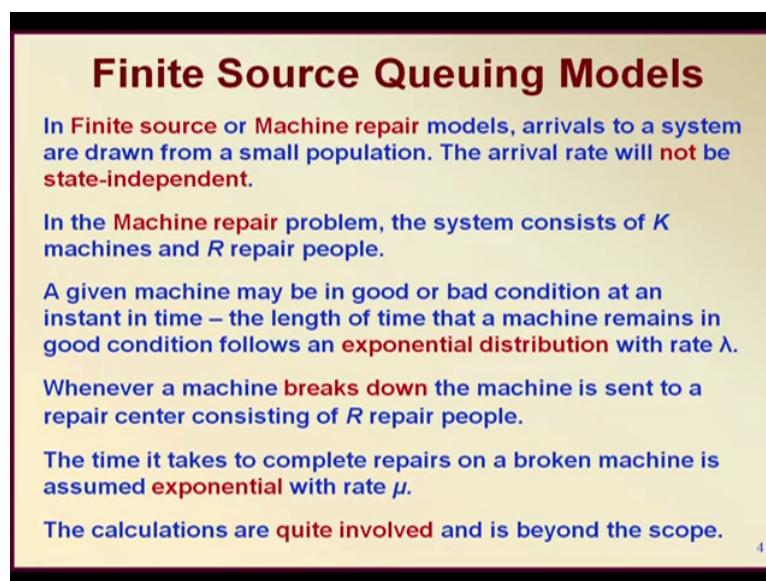
40

So again in this kind of priority queuing models there are 2 types, one is called non-preemptive and other is called preemptive. What is their difference? In the non-preemptive model a customer service cannot be interrupted. Suppose a low priority customer is already in the service and a high priority customer is arrived, will you interrupt the service of the low priority customer? In non-preemptive type it does not happen, right. So under that kind of situations the steady-state variables are L_k , L_{qk} , W_k and W_{qk} where k stands for the type of customers.

Basically what it says that for such kind of systems the different what is called type of priority customers will have different set of steady-state parameters values. In pre-emptive models a lower priority customers can be removed from service whenever a high priority customer arrives, so that is the pre-emptive model. What happens in a pre-emptive model? That a low priority customer is in service and a high priority customer suddenly arrives, so the low priority customers goes back and his service again begins when higher priority customers are not available.

So again 2 types, one is the resume type and other is repeat type. Resume means the customer service starts exactly at that point where it was interrupted, in a repeat service the service is repeated, right. Obviously it also depends on the type of service, there are certain types of service which cannot be interrupted, there are certain kind of service which can be interrupted but has to be redone from the beginning, alright. So these are the different kinds of models all these models have their specific equations specifically given by different work will be available but they are quite involved so we are not going into the mathematical treatment of this models.

(Refer Slide Time: 22:45)



Finite Source Queuing Models

- In **Finite source** or **Machine repair** models, arrivals to a system are drawn from a small population. The arrival rate will **not** be **state-independent**.
- In the **Machine repair** problem, the system consists of K machines and R repair people.
- A given machine may be in good or bad condition at an instant in time – the length of time that a machine remains in good condition follows an **exponential distribution** with rate λ .
- Whenever a machine **breaks down** the machine is sent to a repair center consisting of R repair people.
- The time it takes to complete repairs on a broken machine is assumed **exponential** with rate μ .
- The calculations are **quite involved** and is beyond the scope.

41

Another example could be the finite source queuing models. You see the example could be one of machine repair, what happens in machine repairs there are final numbers of machines, the number of machines are finite so obviously the calling populations in this case or the input population is not infinite, what will happen then? The Markovian assumption or any other assumption, a distribution assumption will not be possible exactly.

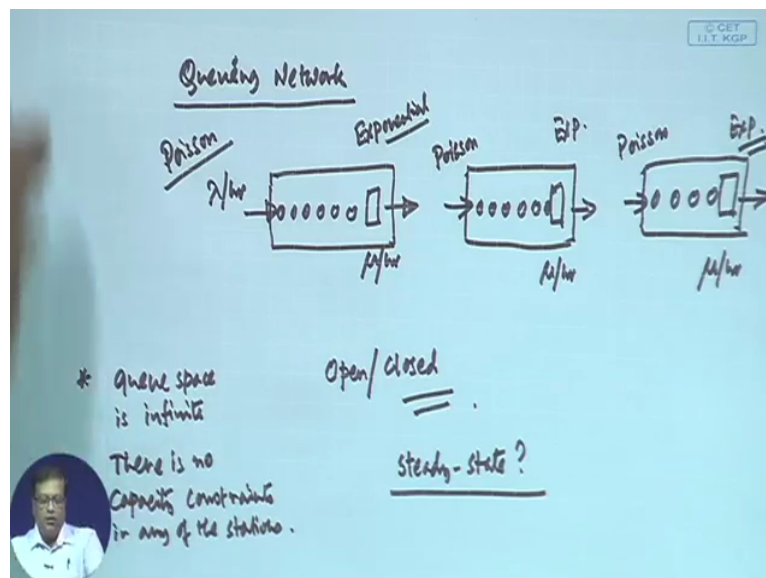
So what should we do? We should try to look at it from that yes the service distribution is exponential, but the arriving population distribution should take care about the calling populations, is it alright. So when you do that then we find that calculation becomes quite involved and you know we can only set certain roles that is whenever a machine breaks down the machine sent to represent consisting of R repair people, not that it cannot be done but it is beyond the scope of this particular discussions.

(Refer Slide Time: 23:58)

Queues in Series – Queuing Networks

- In a **Queuing Network**, a customer needs service in **multiple stages** – one after another. One such system is the **k-stage series queuing system**.
- For **Poisson arrival** to a series queuing system and **exponential service times** and **infinite queue space** at each stage, it can be shown that **interarrival times** for each stage are **exponential**.
- It is required that **enough capacities** are available at each stage for above result to hold.
- For the **entire system**, **L** can be found by adding expected number of customers present at each stage. **W** can be found from Little's formula **$L = \lambda W$** .

42



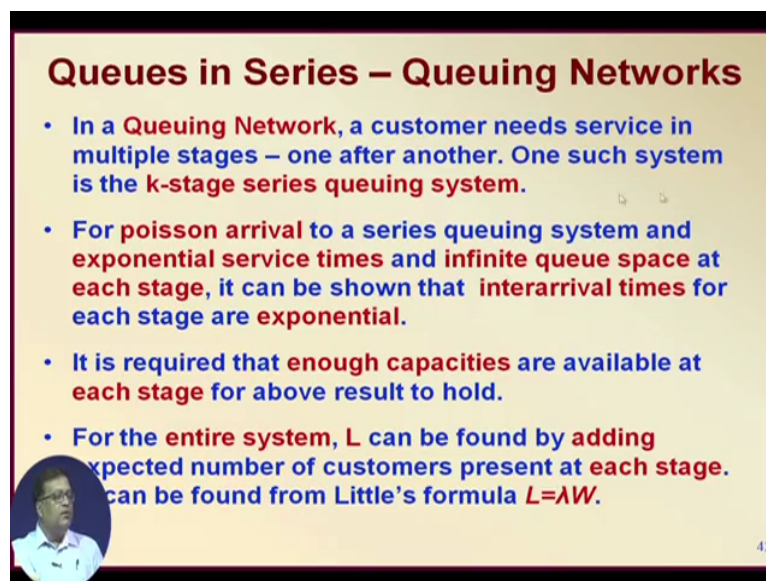
Now before we conclude our queuing thing let us make a note on the queues in series or particularly the queuing network. See queuing network a very interesting thing let us observe, let us think of a queuing network particularly let us talk about open network. Supposing this is a queue system in this queue system there is a service, there is a queue space there is an arriving customer and there is a service customer. This is a λ per hour and the service is μ per hour, but then it does not end here, the customers re-enters another service or let us say another service, for simplicity let us assume the service rate in all of these are μ per hour.

Now this is a queuing network an open queuing network, it is not closed. So what happens? There could be open versus closed queuing network. Now when this happens can there be a

steady-state? What are some conditions? Can there be a steady-state? What happens if the queue space is not infinite? See if the queue space is not infinite and 2nd consideration there is no right so there is no capacity constraints in any of the stations.

So what happens if queue space is not a finite or infinite or capacity constraints maybe disrupt service at any of the stations then the entire process get disrupted and steady-state is not possible. But suppose steady-state is achievable that means the queue space all are infinite and you know there are no capacity constraints at any individual stations then very interestingly you can see that if this is a Poisson process, the arrival is poisson or inter-arrival time is exponential and servers is exponential, right then exponential service is exponential at each station then all the input rates will be all Poisson, right and they all will be equal to the lambda value. So there is a theorem called Burke's theorem, so you know this kind of things are really possible.

(Refer Slide Time: 28:04)



Queues in Series – Queuing Networks

- In a **Queuing Network**, a customer needs service in **multiple stages** – one after another. One such system is the **k-stage series queuing system**.
- For **poisson arrival** to a series queuing system and **exponential service times** and **infinite queue space** at each stage, it can be shown that **interarrival times** for each stage are **exponential**.
- It is required that **enough capacities** are available at each stage for above result to hold.
- For the **entire system**, **L** can be found by adding **expected number of customers present at each stage**. **L** can be found from Little's formula **$L = \lambda W$** .

42

So exactly same thing you can see here, the customer needs service in multiple stages for Poisson arrival to a series queuing system and exponential service time and infinite queue space at each stage it can be shown that inter-arrival times for each stage are exponential. It is required that enough capacities are available at each stage for above result to hold, right. For the entire system L can be found out by adding expected numbers of customers present at each stage. So that means you can analyze every stage individually and then finally add them all up to find the total what you call number in the system that means the total network and then W can be obtained for the network as L equal to lambda W, right.

(Refer Slide Time: 28:53)

Conclusions

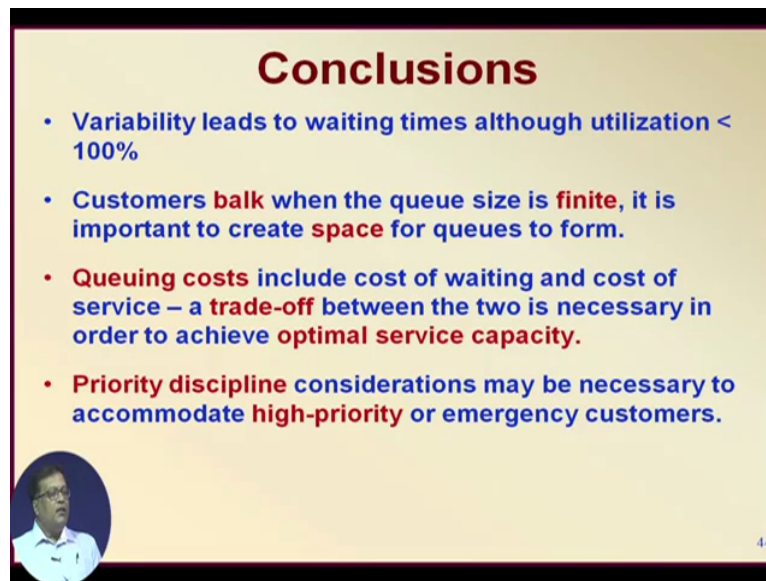
- **Variability in arrival as well as service leads to queues being formed although service rate is higher than arrival rate.**
- **When arrival rate is higher than service rate, no steady state is reached, and steady state analysis is not possible.**
- **In order to improve queuing model performance, it is necessary to identify and eliminate the variability in the process.**
- **Some measures are pooling of resources, resorting to self-service and so on.**

43

So let us conclude some of the very important thing we have seen while discussing the queuing system. The first one is variability, the variability in arriving as well as service leads to queues being formed although service rate is higher than arrival rate. The service rate if it is not higher than arrival rate then no steady-state will be reached right, because transient the system will remain in the transient stage and you cannot really get a steady-state. All the analysis that we have presented in our queuing system they are all true for the steady-state behavior and not at transients. In fact those birth and death rate diagrams can only be drawn when the system has achieved the steady-state.

And primary requirement, the total number of service should be greater than total number of arrival that is a first requirement. Now in order to improve the queuing model performance it is necessary to identify and eliminate the variability in the process, how? First of all you know one measure could be pooling of resources, another measure could be resorting to self-service, right. We have seen how the W really changes with pooling of resources and with self-service.

(Refer Slide Time: 30:27)



Conclusions

- **Variability** leads to waiting times although utilization < 100%
- Customers **balk** when the queue size is **finite**, it is important to create **space** for queues to form.
- **Queuing costs** include cost of waiting and cost of service – a **trade-off** between the two is necessary in order to achieve **optimal service capacity**.
- **Priority discipline** considerations may be necessary to accommodate **high-priority** or emergency customers.

Then variability also leads to waiting times although utilisation could be less than 100 percent, right. Sometimes what happens that utilisation is less than 100 percent but still people are waiting, why? Because there is the queue is formed because of variability's in the arrival and the service process. So if the variability can be removed, if we can service time deterministic we have seen improvements are possible.

Sometimes what happens customers balk when the queue size is finite, it is therefore important to create space for queues to form, if that can be done the customers balking will be reduced and you know the service stations will take more customers. Then there is queuing cost which includes cost of waiting and cost of service, a trade-off between the 2 is necessary in order to achieve optimal service capacity. And finally priority discipline consideration may be necessary to accommodate high priority for emergency customers, right.

And before I end let me just say that most of the real world queues they formed in the queuing networks, right, because they are in happening in one queuing system to another queuing system and the service is a continued process. And many times we really get the arrival process and the service process which is not Markovian. So it becomes very difficult really to analyze in a you know by the kind of equations that we got, so what we should do then? The answer really lies in stimulation and that will be our next concept to discuss, right. So thank you very much.