Chapter 45

Queuing Analysis

45.1 Introduction

One of the major issues in the analysis of any traffic system is the analysis of delay. Delay is a more subtle concept. It may be defined as the difference between the actual travel time on a given segment and some ideal travel time of that segment. This raises the question as to what is the ideal travel time. In practice, the ideal travel time chosen will depend on the situation; in general, however, there are two particular travel times that seem best suited as benchmarks for comparison with the actual performance of the system. These are the travel time under free flow conditions and travel time at capacity.

Most recent research has found that for highway systems, there is comparatively little difference between these two speeds. That being the case, the analysis of delay normally focuses on delay that results when demand exceeds its capacity; such delay is known as queuing delay, and may be studied by means of queuing theory. This theory involves the analysis of what is known as a queuing system, which is composed of a server; a stream of customers, who demand service; and a queue, or line of customers waiting to be served.

45.2 Queuing System

Figure 45:1 shows a schematic diagram illustrating the concept of a queuing system. Various components are discussed below.

45.2.1 Input parameters

- Mean arrival rate
- Mean service rate
- The number of servers

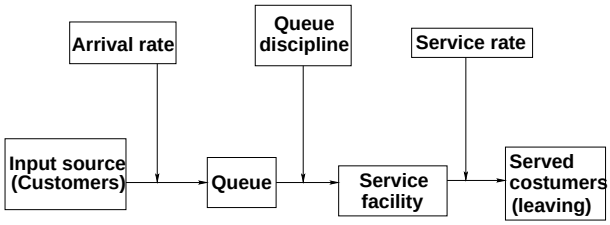

Figure 45:1: Components of a basic queuing system

- Queue discipline

These are explained in the following sections.

Mean Arrival rate (λ)

It is rate at which customers arrive at a service facility. It is expressed in flow (customers/hr or vehicles/hour in transportation scenario) or time headway (seconds/customer or seconds/vehicle in transportation scenario). If inter arrival time that is time headway (h) is known, the arrival rate can be found out from the equation:

$$\lambda = \frac{3600}{h} \tag{45.1}$$

Mean arrival rate can be specified as a deterministic distribution or probabilistic distribution and sometimes demand or input are substituted for arrival.

Mean arrival rate (μ)

It is the rate at which customers (vehicles in transportation scenario) depart from a transportation facility. It is expressed in flow (customers/hr or vehicles/hour in transportation scenario) or time headway (seconds/customer or seconds/vehicle in transportation scenario). If inter service time that is time headway (h) is known, the service rate can be found out from the equation:

$$\mu = \frac{3600}{h} \tag{45.2}$$

Number of servers

The number of servers that are being utilized should be specified and in the manner they work that is they work as parallel servers or series servers has to be specified.

Queue discipline

Queue discipline is a parameter that explains how the customers arrive at a service facility. The various types of queue disciplines are

1. First in first out (FIFO)
2. First in last out (FILO)
3. Served in random order (SIRO)
4. Priority scheduling
5. Processor (or Time) Sharing

1. **First in first out (FIFO):** If the customers are served in the order of their arrival, then this is known as the first-come, first-served (FCFS) service discipline. Prepaid taxi queue at airports where a taxi is engaged on a first-come, first-served basis is an example of this discipline.
2. **First in last out (FILO):** Sometimes, the customers are serviced in the reverse order of their entry so that the ones who join the last are served first. For example, assume that letters to be typed, or order forms to be processed accumulate in a pile, each new addition being put on the top of them. The typist or the clerk might process these letters or orders by taking each new task from the top of the pile. Thus, a just arriving task would be the next to be serviced provided that no fresh task arrives before it is picked up. Similarly, the people who join an elevator first are the last ones to leave it.
3. **Served in random order (SIRO):** Under this rule customers are selected for service at random, irrespective of their arrivals in the service system. In this every customer in the queue is equally likely to be selected. The time of arrival of the customers is, therefore, of no relevance in such a case.
4. **Priority Service:** Under this rule customers are grouped in priority classes on the basis of some attributes such as service time or urgency or according to some identifiable characteristic, and FIFO rule is used within each class to provide service. Treatment of VIPs in preference to other patients in a hospital is an example of priority service.
5. **Processor (or Time) Sharing:** The server is switched between all the queues for a predefined slice of time (quantum time) in a round-robin manner. Each queue head is served for that specific time. It doesn't matter if the service is complete for a customer or

not. If not then it'll be served in it's next turn. This is used to avoid the server time killed by customer for the external activities (e.g. Preparing for payment or filling half-filled form).

45.3 System performance measures

The following notation assumes that the system is in a steady-state condition (At a given time t):

1. Utilization factor $\rho = \frac{\lambda}{\mu}$
2. Pn = probability of exactly n customers in queuing system (waiting + service).
3. L= expected(avg) number of customers in queuing system. [sometimes denoted as Ls]
4. Lq=expected (avg) queue length (excludes customers being served) or no of Customers.
5. W = Expected waiting time in system (includes service time) for each individual customer or time a customer spends in the system. [sometimes denoted as Ws]
6. Wq = waiting time in queue (excludes service time) for each individual customer or Expected time a customer spends in a queue

45.3.1 Relationships between L, W, Lq and Wq:

Assume that λ_n is a constant λ for all n. It has been proved that in a steady-state queuing process, (λ may be considered as avg):

1. $L = \lambda W$
2. $Lq = \lambda Wq$
3. $W = Wq + \frac{1}{\lambda}$

45.3.2 Queuing Patterns

A variety of queuing patterns can be encountered and a classification of these patterns is proposed in this section. The classification scheme is based on how the arrival and service rates vary over time. In the following figures the top two graphs are drawn taking time as independent variable and volume of vehicles as dependant variable and the bottom two graphs are drawn taking time as independent variable and cumulative volume of vehicles as dependant variable.

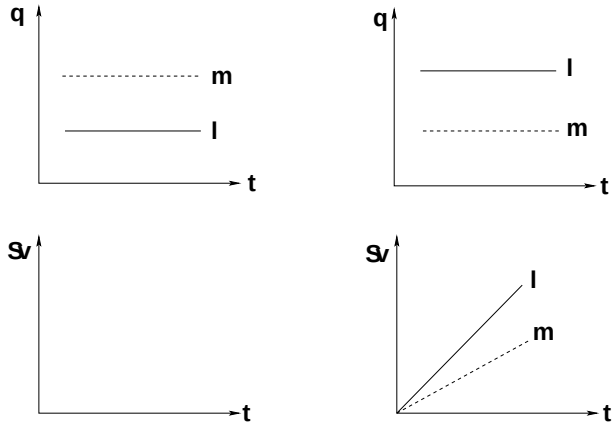

Figure 45:2: Constant arrival and service rates

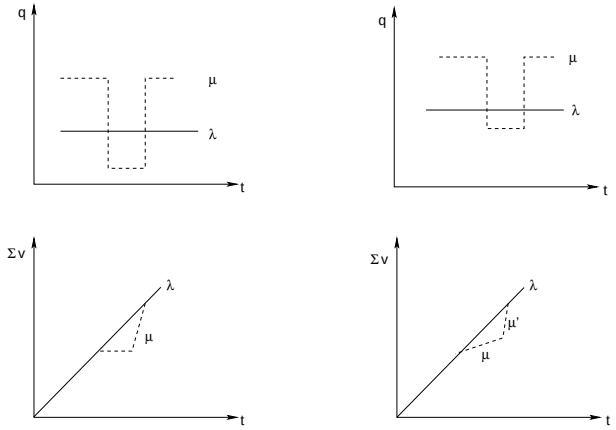

Figure 45:3: Constant arrival rate and varying service rate

45.3.3 Constant arrival and service rates

In the left hand part of the Fig.45:2 arrival rate is less than service rate so no queuing is encountered and in the right hand part of the figure the arrival rate is higher than service rate, the queue has a never ending growth with a queue length equal to the product of time and the difference between the arrival and service rates.

45.3.4 Constant arrival rate and varying service rate

In the left hand of Fig. 45:3 the arrival rate is constant over time while the service rates vary over time. It should be noted that the service rate must be less than the arrival rate for some periods of time but greater than the arrival rate for other periods of time.

One of the examples of the left hand part of the figure is a signalized intersection and that

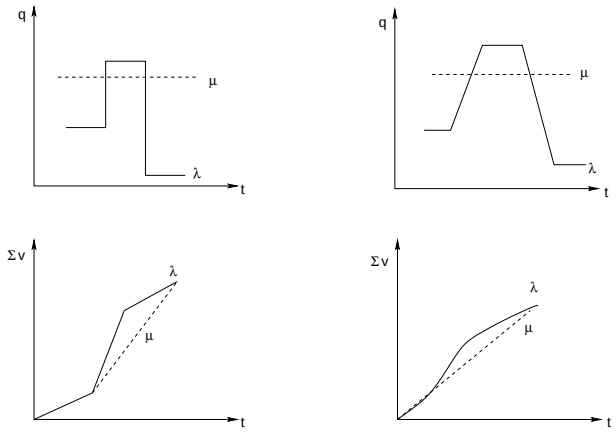

Figure 45:4: Varying arrival rate and constant service rate

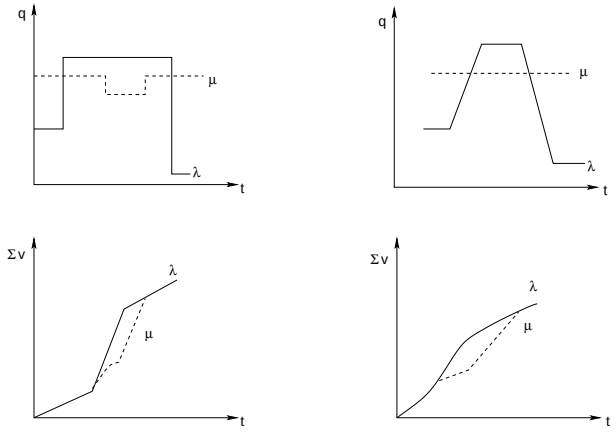

Figure 45:5: Varying arrival and service rates

of the right hand side part of the figure is an incident or an accident on the roads which causes a reduction in the service rate.

45.3.5 Varying arrival rate and constant service rate

In the left part of Fig. 45:4 the arrival rate vary over time but service rate is constant. Both the left and right parts are examples of traffic variation over a day on a facility but the left hand side one is an approximation to make formulations and calculations simpler and the right hand side one considers all the transition periods during changes in arrival rates.

45.3.6 Varying arrival and service rates

In the Fig.45:5 the arrival rate follows a square wave type and service rate follows inverted square wave type. The diagrams on the right side are an extension of the first one with transitional periods during changes in the arrival and service rates. These are more complex to analyzed using analytical methods so simulation is often employed particularly when sensitivity parameter is to be investigated.

45.4 Queuing models

There are various kinds of queuing models. These queuing models have a set of defined characteristics like some arrival and service distribution, queue discipline, etc. The queuing models are represented by using a notation which is discussed in the following section of queue notation.

45.4.1 M/M/1 model

In this model the arrival times and service rates follow Markovian distribution or exponential distribution which are probabilistic distributions, so this is an example of stochastic process. In this model there is only one server. The important results of this model are:

1. Average number of customers in the system $= L = \frac{\rho}{1-\rho}$
2. Average number of customers in the system $= L_q = \frac{\rho^2}{1-\rho}$
3. Expected waiting time in the system $W = \frac{L}{\lambda} = (1/\lambda)\frac{\lambda}{\mu-\lambda} = \frac{1}{\mu-\lambda}$
4. Expected waiting time in the queue $W_q = \frac{L_q}{\lambda} = \frac{1}{\lambda} \times \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\lambda}{\mu(\mu-\lambda)}$

Numerical example

Vehicles arrive at a toll booth at an average rate of 300 per hour. Average waiting time at the toll booth is 10s per vehicle. If both arrivals and departures are exponentially distributed, what is the average number of vehicles in the system, average queue length, the average delay per vehicle, the average time a vehicle is in the system?

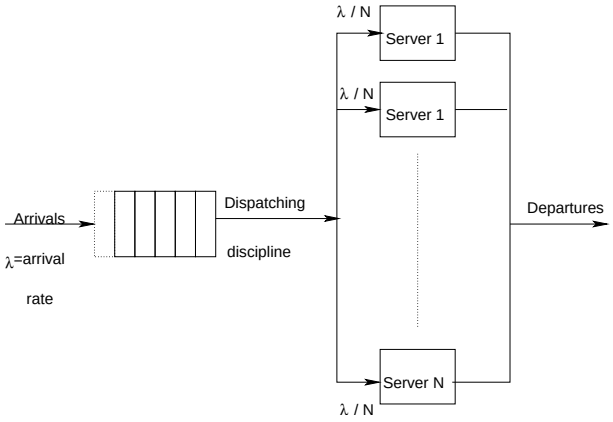

Figure 45:6: Multi-server model

Solution Mean arrival rate $\lambda = 300$ vehicles/hr. Mean service rate $\mu = \frac{3600}{10}$ vehicles/hr. Utilization factor = traffic intensity = $\rho = \frac{\lambda}{\mu} = \frac{300}{360} = 0.833$. Percent of time the toll booth will be idle = P(0) = P(X=0) = $\rho^0(1-\rho) = (0.833)^0(1-0.833) = 0.139(60min)$=8.34 min. Average number of vehicles in the system = $E[X] = \frac{\rho}{1-\rho}$=4.98. Average number of vehicles in the queue =$E[L_q] = \frac{\rho^2}{1-\rho}$= 4.01. Average a vehicle spend in the system =$E[T] = \frac{1}{\mu-\lambda}$= 0.016 hr = 0.96 min = 57.6 sec. Average time a vehicle spends in the queue =$E[T_q] = \frac{\lambda}{\mu(\mu-\lambda)}$= 0.013hr = 0.83 min = 50 sec.

45.4.2 M/M/N model

The difference between the earlier model and this model is the number of servers. This is a multi -server model with N number of servers whereas the earlier one was single server model. The assumptions stated in M/M/1 model are also assumed here. Here μ is the average service rate for N identical service counters in parallel. For x=0

$$P(0) = \left[\sum_{x=0}^{N-1}\left(\frac{\rho^x}{x!} + \frac{\rho^N}{(N-1)!(N-\rho)}\right)\right]^{-1} \tag{45.3}$$

The probability of x number of customers in the system is given by P(x). For $1 \leq x \leq N$

$$P(x) = \frac{\rho^x}{x!} * P(0) \tag{45.4}$$

For $x > N$

$$P(x) = \frac{\rho^x}{N!N^{x-N}} * P(0) \tag{45.5}$$

The average number of customers in the system is

$$E[X] = \rho + [\frac{\rho^{N+1}}{(N-1)!(N-\rho)^2}]P(0) \tag{45.6}$$

The average queue length

$$E[L_q] = [\frac{\rho^{N+1}}{(N-1)!(N-\rho)^2}]P(0) \tag{45.7}$$

The expected time in the system

$$E[T] = \frac{E[X]}{\lambda} \tag{45.8}$$

The expected time in the queue

$$E[T_q] = \frac{E[L_q]}{\lambda} \tag{45.9}$$

45.4.3 Numerical example

Consider the earlier problem as a multi-server problem with two servers in parallel.

Solution Average arrival rate = λ = 300 vehicles/hr. Average service rate = $\mu = \frac{3600}{10}$ vehicles/hr. Utilization factor = traffic intensity = $\rho = \frac{\lambda}{\mu} = \frac{300}{360} = 0.833$.

$$\begin{aligned} P(0) &= \left[\sum_{x=0}^{N-1}\left(\frac{\rho^x}{x!} + \frac{\rho^N}{(N-1)!(N-\rho)}\right)\right]^{-1} \\ &= 0.92(60) = 55.2min \end{aligned}$$

Average number of vehicles in the system is = L = $E[X] = \rho + [\frac{\rho^{N+1}}{(N-1)!(N-\rho)^2}]P(0) = 1.22$. The average number of customers in the queue = $L_q = E[L_q] = [\frac{\rho^{N+1}}{(N-1)!(N-\rho)^2}]P(0)$= 0.387. Expected time in the system =$W = \frac{E[X]}{\lambda}$= 0.004 hr = 14 sec. The expected time in the queue =$W_q = \frac{L_q}{\lambda}$= 0.00129 hr = 4.64 sec.

45.4.4 Multiple single servers' model

In this model there are N numbers of identical independent parallel servers which receive customers from a same source but in different parallel queues (Compare to M/M/N model. It has only one queue) each one receiving customers at a rate of $\frac{\lambda}{N}$. Fig. 45:7 shows how a typical multiple single servers' model looks like.

45.4.5 Numerical example

Consider the problem 1 as a multiple single server's model with two servers which work independently with each one receiving half the arrival rate that is 150 vehicles/hr.

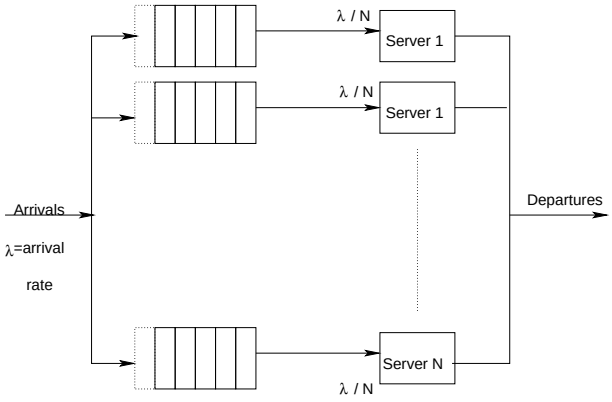

Figure 45:7: Multiple single server

	M/M/1 model	M/M/2 model	Multiple single server model
Idle time of toll booths(minutes)	8.34	55.2	35.04
Number of vehicles in the system(units)	4.98	1.22	0.712
Number of vehicles in the queue(units)	4.01	0.387	0.296
Average waiting time in system(seconds)	57.6	14	17.14
Average waiting time in queue(seconds)	50	4.64	8.05

Solution Mean arrival rate = $\lambda = 150$ vehicles/hr. Mean service rate $=\mu = \frac{3600}{10}$ vehicles/hr. Utilization factor = traffic intensity = $\rho = \frac{\lambda}{\mu} = \frac{150}{360} = 0.416$. The percent of time the toll booth will be idle = P(0) = P(X=0) = $(0.416)^0(1 - 0.416) = 0.584(60min)$=35.04 min. The average number of vehicles in the system = $E[X] = \frac{\rho}{1-\rho}$= 0.712. The average number of vehicles in the queue $=L_q = \frac{\rho^2}{1-\rho}$= 0.296. The average a vehicle spend in the system $=E[T] = W = \frac{1}{\mu-\lambda}$= 0.0047 hr = 0.285 min = 17.14 sec. The average time a vehicle spends in the queue $=E[T_q] = W_q = \frac{\lambda}{\mu(\mu-\lambda)}$= 0.0022hr = 0.13 min = 8.05 sec

Comparison of the three models

From the Table 1 by providing 2 servers the queue length reduced from 4.01 to 0.387 and the average waiting time of the vehicles came down from 50 sec to 4.64 sec, but at the expense of having either one or both of the toll booths idle 92% of the time as compared to 13.9% of the time for the single-server situation. Thus there exists a trade-off between the customers' convenience and the cost of running the system.

45.4.6 D/D/N model

In this model the arrival and service rates are deterministic that is the arrival and service times of each vehicle are known.

Assumptions

1. Customers are assumed to be patient.
2. System is assumed to have unlimited capacity.
3. Users arrive from an unlimited source.
4. The queue discipline is assumed to be first in first out.

45.4.7 Numerical example

Morning peak traffic upstream of a toll booth is given in the table 2. The toll plaza consists of three booths, each of which can handle an average of one vehicle every 8 seconds. Determine the maximum queue, the longest delay to an individual vehicle.

Time period	10 min volume
7.00-7.10	200
7.10-7.20	400
7.20-7.30	500
7.30-7.40	250
7.40-7.50	200
7.50-8.00	150

Time period	10 min flow (3)	Cum. rate(4)	Service service(5)	Cumulative =(3)-(4)	Queue (6)	Delay
7.00-7.10	200	200	200	200	0	0
7.10-7.20	400	600	225	425	175	7.78
7.20-7.30	500	1100	225	650	450	20.00
7.30-7.40	250	1350	225	875	475	21.11
7.40-7.50	200	1550	225	1100	450	20.00
7.50-8.00	150	1700	225	1325	375	16.67

Solution The arrival volume is given in the table. Service rate is given as 8 seconds per vehicle. This implies for 10 min, 75 vehicles can be served by each server. It is given there are 3 servers. Hence 225 vehicles can be served by 3 servers in 10 min. In the first 10 min only 200 vehicles arrive which are served so the service rate for rest 50 min is 225 veh/10 min as there is a queue for the rest period. The solution to the problem is showed in the table 3 following. The cumulative arrivals and services are calculated in columns 3 and 5. Queue length at the end of any 10 min interval is got by simply subtracting column 5 from column 3 and is recorded in column 6. Maximum of the column 6 is maximum queue length for the study period which is 300 vehicles. The service rate has been found out as 225 vehicles per hour. From proportioning we get the time required for each queue length to be served and as 475 vehicles is the max queue length, the max delay is corresponding to this queue. Therefore max delay is 21.11 min.

45.5 Conclusions

The queuing models often assume infinite numbers of customers, infinite queue capacity, or no bounds on inter-arrival or service times, when it is quite apparent that these bounds must exist in reality. Often, although the bounds do exist, they can be safely ignored because the differences between the real-world and theory is not statistically significant, as the probability that such boundary situations might occur is remote compared to the expected normal situation. Furthermore, several studies show the robustness of queuing models outside their assumptions. In other cases the theoretical solution may either prove intractable or insufficiently informative to be useful. Alternative means of analysis have thus been devised in order to provide some insight into problems that do not fall under the scope of queuing theory, although they are often scenario-specific because they generally consist of computer simulations or analysis of experimental data.

45.6 References

1. James H Banks. *Introduction to transportation engineering*. Tata Mc-Graw Hill, 2004.

2. Frederick S. Hillier and Gerald J. Lieberman. *Operations Research*. CBS publishers, 2019.

3. Adolf D. May. *Fundamentals of Traffic Flow*. Prentice - Hall, Inc. Englewood Cliff New Jersey 07632, second edition, 1990.

4. C S Papacostas. *Transportation engineering and planning by Papacostas. C. S, 3rd edition, Prentice-Hall of India in 2001.* Prentice-Hall of India, 2001.