
QUT

SCHOOL OF MECHANICAL, MANUFACTURING & MEDICAL ENGINEERING

MEN170: SYSTEMS MODELLING AND SIMULATION

7. SIMPLE QUEUING MODELS:

7.1 INTRODUCTION: A queuing system consists of one or more servers that provide service of some sort to arriving customers. Customers who arrive to find all servers busy generally join one or more queues (lines) in front of the servers, hence the name **queuing systems**. There are several everyday examples that can be described as queuing systems, such as bank-teller service, computer systems, manufacturing systems, maintenance systems, communications systems and so on.

Components of a Queuing System: A queuing system is characterised by three components:

- Arrival process
- Service mechanism
- Queue discipline.

Arrival Process

Arrivals may originate from one or several sources referred to as the **calling population**. The calling population can be limited or 'unlimited'. An example of a limited calling population may be that of a fixed number of machines that fail randomly. The arrival process consists of describing how customers arrive to the system. If A_i is the inter-arrival time between the arrivals of the (i-1)th and ith customers, we shall denote the mean (or expected) inter-arrival time by $E(A)$ and call it (λ) ; $= 1/(E(A))$ the arrival frequency.

Service Mechanism

The service mechanism of a queuing system is specified by the number of servers (denoted by s), each server having its own queue or a common queue and the probability

distribution of customer's service time. let S_i be the service time of the i th customer, we shall denote the mean service time of a customer by $E(S)$ and $\mu = 1/(E(S))$ the service rate of a server.

Queue Discipline

Discipline of a queuing system means the rule that a server uses to choose the next customer from the queue (if any) when the server completes the service of the current customer. Commonly used queue disciplines are:

FIFO - Customers are served on a first-in first-out basis.

LIFO - Customers are served in a last-in first-out manner.

Priority - Customers are served in order of their importance on the basis of their service requirements.

Measures of Performance for Queuing Systems:

There are many possible measures of performance for queuing systems. Only some of these will be discussed here.

Let,

D_i be the delay in queue of the i th customer

W_i be the waiting time in the system of the i th customer = $D_i + S_i$

$Q(t)$ be the number of customers in queue at time t

$L(t)$ be the number of customers in the system at time $t = Q(t) + \text{No. of customers being served at } t$

Then the measures,

$$d = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^{i=n} D_i}{n} \quad \text{and}$$

$$w = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^{i=n} W_i}{n}$$

(if they exist) are called the **steady state average delay** and the **steady state average waiting time in the system**. Similarly, the measures,

$$Q = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T Q(t) \cdot dt \quad \text{and}$$

$$L = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T L(t) \cdot dt$$

(if they exist) are called the **steady state time average number in queue** and the **steady state time average number in the system**. Among the most general and useful results of a queuing system are the **conservation equations**:

$$Q = (\lambda) d \text{ and } L = (\lambda) w$$

These equations hold for every queuing system for which d and w exist. Another equation of considerable practical value is given by,

$$w = d + E(S)$$

Other performance measures are:

- the probability that any delay will occur.
- the probability that the total delay will be greater than some pre-determined value
- that probability that all service facilities will be idle.
- the expected idle time of the total facility.
- the probability of turn-aways, due to insufficient waiting accommodation.

7.2: Notation for Queues.

Since all queues are characterised by arrival, service and queue and its discipline, the queue system is usually described in shorten form by using these characteristics. The general notation is:

[A/B/s]:{d/e/f}

Where,

- A = Probability distribution of the arrivals
- B = Probability distribution of the departures
- s = Number of servers (channels)
- d = The capacity of the queue(s)
- e = The size of the calling population
- f = Queue ranking rule (Ordering of the queue)

There are some special notation that has been developed for various probability distributions describing the arrivals and departures. Some examples are,

- M = Arrival or departure distribution that is a Poisson process**
- E = Erlang distribution**
- G = General distribution**
- GI = General independent distribution.**

Thus for example, the [M/M/1]:{infinity/infinity/FCFS} system is one where the arrivals and departures are a Poisson distribution with a single server, infinite queue length, calling population infinite and the queue discipline is FCFS. This is the simplest queue system that can be studied mathematically. This queue system is also simply referred to as the M/M/1 queue.

7.3 Single Channel Queuing Theory

7.3.1: [M/M/1]:{//FCFS} Queue System.

A Arrival Time Distribution.

The simple model assumes that the number of arrivals occurring within a given interval of time t , follows a **Poisson distribution**, with parameter $(\lambda)t$. This parameter $(\lambda)t$ is the average number of arrivals in time t which is also the variance of the distribution. If n denotes the number of arrivals within a time interval t , then the probability function $p(n)$ is given by,

$$P(n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad n = 0, 1, 2, \dots$$

The arrival process is called **Poisson input**.

The probability of no(zero) arrival in the interval $[0,t]$ is,

$$\Pr(\text{zero arrival in } [0,t]) = e^{-\lambda t} = p(0)$$

also,

$$P(\text{zero arrival in } [0,t]) = P(\text{next arrival occurs after } t)$$

$$= P(\text{time bet. two successive arrivals exceeds } t)$$

From this it can be shown that the probability density function of the **inter-arrival** times is given by,

$$e^{-\lambda t} \text{ for } t \geq 0$$

This called the **negative exponential distribution with parameter λ** or simply exponential distribution. The mean inter-arrival time and standard deviation of this distribution are both $1/(\lambda)$ where, (λ) is the arrival rate.

NOTE: At first glance this distribution seems unrealistic. But it turns out that this is an extremely robust distribution and approximates closely a large number of arrival and breakdown patterns in practice.

B. Property of Stationarity and Lack of Memory.

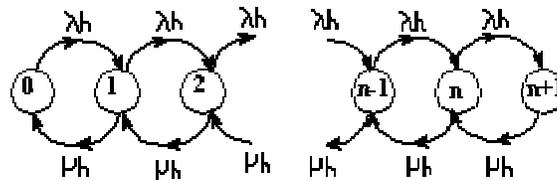
A Poisson input implies that arrivals are independent of one another or the state of the system. The probability of an arrival in any interval of time h **does not** depend on the starting point of the arrival or on the specific history of arrivals preceding it, but depends only on the length h . Thus the queuing systems with Poisson input can be considered as Markovian process. (The reason for using M in the notation)

C. Analysis of the System.

In this case both the inter-arrival times and the service times are assumed to be negative exponential distribution with parameters (λ) and μ . As with Markovian process, we are interested only in the long-run behaviour of the system. i.e. **steady state or statistical equilibrium state**. It is obvious that if the arrival rate is higher than the service rate the system will be blocked. Hence, we consider only the analysis of the system where the arrival rate is less than the service rate.

At any moment in time, the state of the queuing system can be completely described by the number of units in the system. Thus the state of the process can assume values $0, 1, 2, \dots$ (0 means none in the queue and the service is idle) Unlike Markov process, here the change of state can occur at any time. However the process will approach a steady state which is independent of the starting position or state.

Let the steady state probabilities are denoted by P_n , $n = 0, 1, 2, 3, \dots$, where n refers to the number in the system. P_n is the probability that there are n units in the system. By considering a very small interval of time h , the transition diagram for this system can be seen as:



If h is sufficiently small, no more than **one** arrival can occur and no more than **one** service completion can occur in that time. Also the probability of observing a service completion and an arrival in time h is $\mu(\lambda) \cdot h^2$ which is very small (approximately zero) and is neglected. Thus only the following four events are possible:

1. There are n units and 1 arrival occurs in h
2. There are n units and 1 service is completed in h
3. There are $n-1$ units and 1 arrival occurs in h
4. There are $n+1$ units and 1 service is completed in h

For $n > 1$, (because of steady state condition)

$\Pr(\text{being in state } n \text{ and leaving it}) = \Pr(\text{being in other states and entering state } n)$

$= \Pr(\text{being in state } n-1 \text{ or } n+1 \text{ and entering state } n)$

Thus,

$$P_n(\lambda) * h + P_n * \mu h = P_{n-1}(\lambda) * h + P_{n+1} * \mu h$$

This equation is called **steady state balance equation**.

For $n = 0$, only events 1 and 4 are possible,

$$P_0(\lambda) * h = P_1 * \mu h$$

Therefore,

$$P_1 = \frac{\lambda}{\mu} P_0 \quad P_n = \frac{\lambda}{\mu} P_{n-1}$$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

P_0 can be determined by using the fact that the sum of the steady state probabilities must be 1. Therefore,

$$P_0 + P_1 + P_2 + \dots + P_n + P_{n+1} + \dots = 1$$

$$P_0 + P_0 \left[\frac{\lambda}{\mu}\right] + P_0 \left[\frac{\lambda}{\mu}\right]^2 + \dots + P_0 \left[\frac{\lambda}{\mu}\right]^n + P_0 \left[\frac{\lambda}{\mu}\right]^{n+1} + \dots = 1$$

$$P_0 [1 + \rho + \rho^2 + \dots + \rho^n + \rho^{n+1} + \dots] = 1 \quad \rho = \frac{\lambda}{\mu}$$

This is the sum of a geometric series. Therefore,

$$P_0 \left[\frac{1 - \rho^{n+1}}{1 - \rho} \right] = 1 \quad \text{as } n \rightarrow \infty$$

Since $\rho < 1$,

$$P_0 = (1 - \rho) = \left(1 - \frac{\lambda}{\mu}\right)$$

The term $\rho = (\lambda) / \mu$ is called **utilisation factor or traffic intensity**. This is also equal to the probability that the service is busy, referred to as Pr(busy period).

Performance measures

The average number of units in the system L can be found from

$L = \text{Sum of } [n \cdot P_n] \text{ for } n= 1 \text{ to (infinity)}$.

$$L = \frac{\lambda}{(\mu - \lambda)} = \frac{\rho}{(1 - \rho)} \quad \text{where} \quad \rho = \frac{\lambda}{\mu}$$

The average number in the queue is

$$Q = L - (1 - P_0)$$

Sum of $[(n-1) \cdot P_n]$ for $n=1$ to (infinity) .

$$Q = \frac{\lambda^2}{[\mu \cdot (\mu - \lambda)]} = \frac{\rho^2}{(1 - \rho)}$$

The average waiting time in the system (time in the system) can be obtained from,

$$w = \frac{L}{\lambda} = \frac{1}{(\mu - \lambda)} \quad \text{and}$$

$$d = w - \frac{1}{\mu} = \frac{\lambda}{[\mu \cdot (\mu - \lambda)]}$$

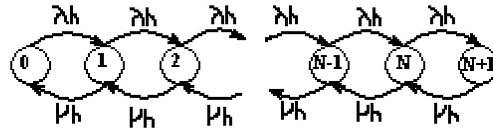
An Example: A firm operates a 10-ton truck on a job contracting basis. The job requests is Poisson distributed with a mean request rate of 1.4 per day. The average service time is 4 hours and is exponentially distributed. Determine all the performance measures and the probability that there are more than two jobs in the system.

7.3.2 [M/M/1] : {N//FCFS} System (Limited queue length system)

If the queue length is limited to N, then some customers (jobs) will be lost. The maximum number in the system can only be (N+1). Thus the transition diagram will have (N+2) states as shown.

The steady state balance equation is the same as before except the first and the last one.

The relationship $P_n = \rho^n \cdot P_0$ holds true. $\rho = (\lambda) / \mu$



As before, from the fact that the sum of all steady state probabilities is 1, we can obtain P_0 . Thus

$$\begin{aligned}
 P_0 + P_1 + P_2 + \dots + P_{N-1} + P_N + P_{N+1} &= 1 \\
 P_0 [1 + \rho + \rho^2 + \dots + \rho^{N-1} + \rho^N + \rho^{N+1}] &= 1 \\
 P_0 &= \frac{[1 - \rho]}{[1 - \rho^{N+2}]} \\
 P_n &= (\rho^n) \frac{[1 - \rho]}{[1 - \rho^{N+2}]}
 \end{aligned}$$

It is also evident that the average number in the system is not the same as before. It is given by,

$$L = \sum_{n=1}^{N+1} n \cdot P_n \quad (\text{Finite number of terms})$$

The above equations hold for $(\lambda) < \mu$. or $\rho < 1$

The average queue length is,

$$Q = \sum_{n=1}^{N+1} (n - 1) \cdot P_n$$

The average waiting time in the system is

$$w = L / (\lambda \cdot (1 - P_{N+1}))$$

$(1 - P_{N+1})$ is required as we know that there are no more than $N+1$ units will be in the system at any time because of the limitation of queue length to N .

It can be seen from the above that, limiting the queue length has the following consequences:

- Average idle time will increase
- Average queue length will decrease

- Average waiting time will decrease
- A portion of the customers will be lost.

EXAMPLE: In the above example suppose the number in the **queue** is limited to 2.
(Refer pp313 - 337 of Introduction to Operations Research Techniques by Daellenbach & George

[Continue with second part?](#)

[Back to MEN170 Contents Page](#)