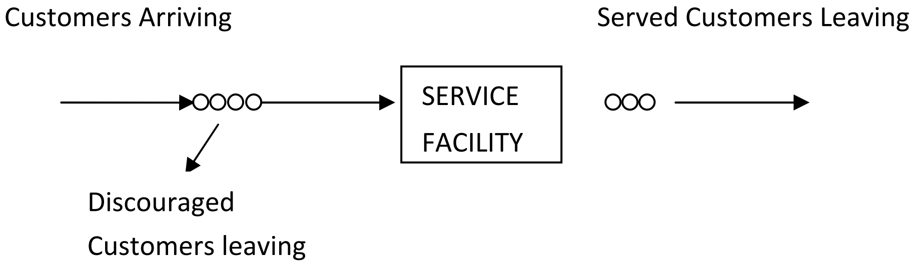
Contents

1 Basics of Queuing

Queuing process: Queuing process is a class of random processes describing phenomena of queue formation.



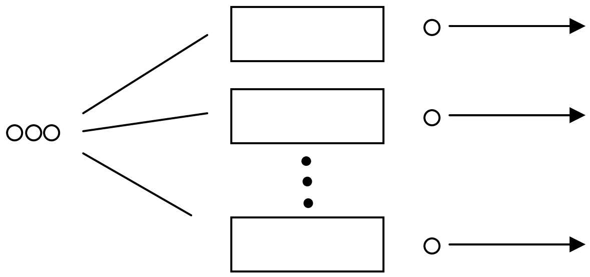
A Typical Queuing Process

Queueing theory: is the mathematical study of waiting lines (or *queues*). The theory enables mathematical analysis of several related processes, including arriving at the (back of the) queue, waiting in the queue (essentially a storage process), and being served by the server(s) at the front of the queue.

1.1 Characteristics of a queuing process

The following are the six basic characteristics of a queuing process:

1. Arrival pattern of customers: In queuing the arrival process is usually stochastic. As a result it is necessary to determine the probability distribution of the interarrival times (times between successive customer arrivals) as well. Also customers can arrive in individually or simultaneously (batch or bulk arrivals).
2. Service pattern of customers: As in arrivals, a probability distribution is needed for describing the sequence of customer service time. Service may also be single or batch. The service process may depend on the number of customers waiting in queue for service. In this it is called state dependent service.
3. Queue discipline: Queue discipline refers to the manner in which customers are selected for service when a queue has formed. The default is FCFS i.e. is first come first served. Some others are LCFS (last come first served), RSS (random service selection) i.e. selection for service in random order independent of the time of arrival and there are other priority systems where customers are given priorities upon entering the system, ones with higher priority are selected first.
4. System capacity: A queuing system can be finite or infinite. In certain queuing process there is a limitation on the length of the queue i.e. customers are not allowed to enter if the queue has reached a certain length. These are called finite queuing systems. If there is no restriction on the length of the queue then it is called an infinite queuing system.
5. Number of service channels: A queuing system can be single or a multiserver system. In a multiserver queuing system there are several parallel servers running to serve a single line.



A Multiserver Queuing system

6. Number of service stages: A queuing system may have only a single stage of service. But as an example of a multistage queuing system consider the physical examination procedure, where each patient proceeds through various stages of medical examination, like throat check up, eye test, blood test etc.

Kendall's notation

A queuing process can be described using a notation which uses series of symbols and slashes such as A/B/X/Y/Z, where

A: indicates the interarrival time distribution

B: indicates the service time distribution

X: the number of parallel servers

Y: the restriction on system capacity

Z: the queue discipline

Standard symbols for the characteristics A and B

M	**Exponential**
D	**Deterministic**
EK	**Erlang type k**
HK	**Mixture of k exponentials**
PH	**Phase type**
G	**General**

1.2 Poisson Process and Exponential Distribution

The most common stochastic models assume that the arrival rate and service rate follow a poisson distribution.

The Poisson process is a counting process {N(t),t≥0}, where N(t) denotes the total number of arrivals up to time t with N(0)=0 and which satisfies the following three assumptions:

i. The probability that an arrival occurs between time t and t+Δt is equal to λΔt+o(Δt), where λ is a constant independent of N(t) and

$$\lim_{\Delta t\to 0}\frac{o(\Delta t)}{\Delta t}= 0$$

ii. Pr{more than one arrival between t and t+Δt }= o(Δt)

iii. The number of arrivals in nonoverlapping intervals is independent, i.e. the process has independent increments.

Stationary increments: One of the most important properties of Poisson process is that the number of occurrences in intervals of equal width are identically distributed. In particular, for t>s, the difference N(t) – N(s) is identically distributed as N(t+h) – N(s+h).

Now, if the arrival process is Poisson then it can be easily shown that the associated random variable defined as the time between successive arrivals (interarrival time) follows the exponential distribution. Likewise if the interarrival times are independent and have the same exponential distribution , then the arrival rate follows a Poisson distribution.

The above theory also holds for service rate and service times.

Markovian property of the exponential distribution: A stochastic process has the Markov property if the conditional probability distribution of future states of the process depends only upon the present state; that is, given the present, the future does not depend on the past. For service times this property states that the probability that a customer currently in service has t units of remaining service is independent of how long it has already been in service. Thus we have,

$$\Pr\{T \leq t_1 | T \geq t_0\} = \Pr\{0 \leq T \leq t_1 - t_0\}$$

Stationary process (or **strict(ly) stationary process**) is a stochastic process whose joint probability distribution does not change when shifted in time or space i.e.

If X_t is a stochastic process then X_t is said to be stationary if, for all k, for all τ, and for all, $t_1, t_2, \ldots t_k$

$$F_{X_{t_1},\ldots,X_{t_k}}(x_{t_1},\ldots,x_{t_k}) = F_{X_{t_1+\tau},\ldots,X_{t_k+\tau}}(x_{t_1},\ldots,x_{t_k}).$$

1.3 Markov process

Markov process, named after the Russian mathematician Andrey Markov, is a time-varying random phenomenon for which a specific property the Markov property holds.

Markov Chain: If we assume that the state space, I, is discrete, then the Markov process is known as a Markov Chain

DTMC(Discrete Time Markov Chain): If the parametric space , T, is also discrete, then the Markov chain is known as a discrete time Markov chain. In this case we let T= {0,1,2,...}. For a DTMC the Markov propert can be stated as

$$P\,(X_n=i_n | X_0=i_0, X_1=i_1, \ldots, X_{n-1}=i_{n-1}) = P\,(X_n=i_n | X_{n-1}=i_{n-1}), \quad i_0, i_1, \ldots, i_n \in I$$

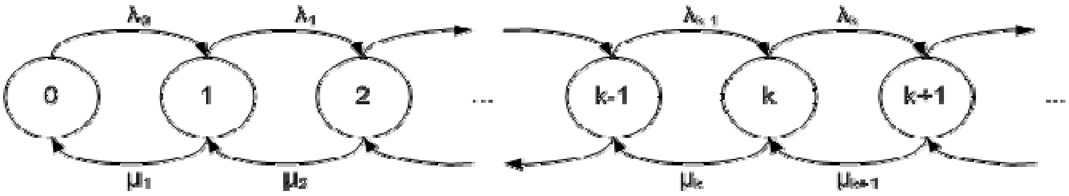
CTMC(Continuous Time Markov Chain): If the parametric space, T, is continuous, then the Markov chain is called a continuous time Markov chain. In this case we let T= [0,∞). For a CTMC the Markov property can be stated as

$$P\,(X(t)=x | X(t_n)=x_n, X(t_{n-1})=x_{n-1}, \ldots, X(t_0)=x_0) = P\,(X(t)=x \mid X(t_n)=x_n)$$

1.4 Birth-death process

The **birth-death process** is a special case of CTMC where the states represent the current size of a population and where the transitions are limited to births and deaths. Birth-death processes have many applications in demography, queueing theory, performance engineering, or in biology.

When a birth occurs, the process goes from state n to n+1. When a death occurs, the process goes from state n to state n-1. The process is specified by birth rates $\{\lambda_i\}_{i=0..\infty}$ and death rates $\{\mu_i\}_{i=1..\infty}$.



Examples

A **pure birth process** is a birth-death process where $\mu_i = 0$ for all $i \geq 0$.

A **pure death process** is a birth-death process where $\lambda_i = 0$ for all $i \geq 0$.

A (homogeneous) Poisson process is a pure birth process where $\lambda_i = \lambda$ for all $i \geq 0$

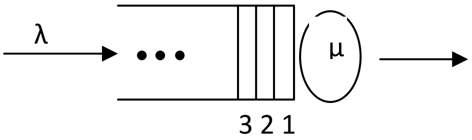
M/M/1 model and M/M/c model, both used in queueing theory, are birth-death processes used to describe customers in an infinite queue.

2 Markovian Queuing Models

2.1 M/M/1/∞

The M/M/1 queuing system is described as a queuing model where:

- arrivals are a Poisson process i.e. interarrival time is exponentially distributed;
- service time is exponentially distributed;
- there is one server;
- the length of queue in which arriving users wait before being served is infinite;
- the population of users (i.e. the pool of users) available to join the system is infinite



A simple M/M/1 queue with arrival rate λ and service rate μ

2.1.1 Steady State Distribution

Let p_n represents the probability mass function of a discrete random variable denoting the number of customers in the system in long run

$$p_n = (1-\rho)\,\rho^n \qquad , \rho<1$$

where,

$\rho = \lambda/\mu$ represents the traffic intensity of the system. For a stable system the intensity ρ must be less than 1.

It can be seen above that the steady state probabilities for an M/M/1 queue follows the geometric distribution with parameter (1- ρ)

Measures of Effectiveness

Measure	Expression
Average number of customers in the system(L_s)	$\rho/(1-\rho)$
Average number of customers in the Queue(L_q)	$\rho^2/(1-\rho)$
Expected waiting time in system(W)	$1/(\mu-\lambda)$
Expected waiting time in queue(W_q)	$\rho/(\mu-\lambda)$
Utilization	P

2.1.2 Transient solution

The transient probabilities $p_n(t)=\Pr\{X(t)=n\}$ for an M/M/1 queue are given by

$$p_n(t) = e^{-(\lambda+\mu)t}\left[\frac{\lambda^{(n-i)/2}}{\mu} I_{n-i}\left(2\sqrt{\lambda\mu}t\right) + \frac{\lambda^{(n-i-1)/2}}{\mu} I_{n+i+1}\left(2\sqrt{\lambda\mu}t\right) + \left(1-\frac{\lambda}{\mu}\right)\frac{\lambda^n}{\mu}\sum_{j=n+i+2}^{\infty}\left(\frac{\lambda}{\mu}\right)^{-j/2} I_j\left(2\sqrt{\lambda\mu}t\right)\right]$$

for all n≥0, where

$$I_n(y) = \sum_{k=0}^{\infty}\frac{(y/2)^{n+2k}}{k!\,(n+k)!} \qquad (n > -1)$$

is the infinite series for the modified Bessel function of the first kind.

2.2 M/M/1/N

This system is a type of M/M/1/∞ queue with at most N(+ve integer) customers allowed in the system.

2.2.1 Steady State Distribution

The state probabilities in equilibrium are given by:

$$P_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{N+1}} & , n = 0,1,\ldots,N, \ \rho \neq 1 \\ \dfrac{1}{N+1} & , \qquad \rho = 1 \end{cases}$$

Measures of Effectiveness

Average number of customers in the system(L_s)	$\begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{N+1}{1-\rho^{N+1}}\rho^{N+1} & \rho \neq 1 \\ \dfrac{N}{2} & \rho = 1 \end{cases}$
Average number of customers in queue(L_q)	$L_s - (\lambda/\mu)$
Expected waiting time in system(W)	L_s/ λ
Expected waiting time in queue(Wq)	L_q/ λ
Utilization	ρ
Blocking Probability(**P_B**)	$\begin{cases} \dfrac{(1-\rho)\rho^N}{1-\rho^{N+1}} & \rho \neq 1 \\ \dfrac{1}{N+1} & \rho = 1 \end{cases}$
Throughput	$\rho\,(1 - P_B)$

2.3 M/M/c/∞

This system is a multiserver model in which there are c servers and each server has an independent and identically distributed exponential service time distribution, with the arrival process again assumed to be Poisson.

2.3.1 Steady State Distribution

For this model the steady state probabilities are given by:

$$P_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & , 1 \leq n \leq c \\ \dfrac{1}{c^{n-c}c!}\left(\dfrac{\lambda}{\mu}\right)^n P_0 & , c \leq n \end{cases}$$

where,

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{(r)^n}{n!} + \frac{r^c}{c!(1-\rho)}\right]^{-1} \qquad , \rho<1$$

ρ=λ/cμ, r= ρ=λ/μ

Measures of Effectiveness

Measure	Expression
Average number of customers in the system(L_s)	$r + \left(\frac{r^c \rho}{c!(1-\rho)^2}\right) P_0$
Average number of customers in the Queue(L_q)	$\left(\frac{r^c \rho}{c!(1-\rho)^2}\right) P_0$
Expected waiting time in system(W)	$\frac{1}{\mu} + \left(\frac{r^c}{c!(c\mu)(1-\rho)^2}\right) P_0$
Expected waiting time in queue(W_q)	$\left(\frac{r^c}{c!(c\mu)(1-\rho)^2}\right) P_0$

2.4 M/M/c/K

2.4.1 Steady state distribution

In this model there is a limit K placed on the number allowed in the system at any time. The steady state system size probabilities are given by:

For this model the steady state probabilities are given by:

$$P_n = \begin{cases} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 & ,1 \le n \le c \\ \frac{1}{c^{n-c} c!}\left(\frac{\lambda}{\mu}\right)^n P_0 & ,c \le n \le K \end{cases}$$

where,

$$P_0 = \begin{cases} \left[\sum_{n=0}^{c-1} \frac{(r)^n}{n!} + \frac{r^c}{c!} \frac{1-\rho^{K-c+1}}{1-\rho}\right]^{-1} & \rho \ne 1 \\ \left[\sum_{n=0}^{c-1} \frac{(r)^n}{n!} + \frac{r^c}{c!}(K-c+1)\right]^{-1} & \rho = 1 \end{cases}$$

where, $\rho = \frac{\lambda}{c\mu}, r = \frac{\lambda}{\mu}$

Measures of Effectiveness

Measure	Expression
Average number of customers in the Queue(L_q)	$\frac{P_0 r^c \rho}{c!(1-\rho)^2}\left[1 - \rho^{K-c+1} - (1-\rho)(K-c+1)\rho^{K-c}\right]$,ρ≠1
Average number of customers in the system(L_s)	$L_q + r(1-P_K)$
Expected waiting time in system(W)	$\frac{L}{\lambda(1-P_K)}$

Expected waiting time in queue(W_q)	$\frac{L_q}{\lambda(1-P_K)}$
Utilization	ρ
Blocking Probability(P_B)	$\frac{1}{c^{K-c}c!}\left(\frac{\lambda}{\mu}\right)^K P_0$
Throughput	$\rho(1-P_B)$

2.5 M/M/c/c

2.5.1 Steady State Distribution

The special case of the truncated queue m/M/c/K for which K=c, i.e. where no line is allowed to form. For this model the steady state probabilities are given by:

$$P_n = \frac{\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n}{\sum_{i=0}^{c}\frac{1}{i!}\left(\frac{\lambda}{\mu}\right)^i} \qquad 0 \le n \le c$$

In case of an M/M/c model we define the following performance measures:

Measure	Expression
Blocking Probability(**P**$_B$)	$\frac{\frac{r^c}{c!}}{\sum_{i=0}^{c}\frac{(r)^i}{i!}}$ for $\rho \ge 0$
Throughput	ρ (1 − P_B)

where , r=(λ/μ)

2.6 Bulk Input

2.6.1 Steady State distribution

This is a queuing model wherein in addition to the assumption that the arrival process forms a Poisson process, we assume that the actual number of customers in any arriving module is a random variable X, which may take on any possible integral value less than ∞ with probability C_x. In this model the batch size can be a constant or a random variable with some distribution. Below we have discussed two cases, size constant or geometrically distributed.

When the batch sizes are distributed geometrically

$$p_n = (1-\rho)[\alpha + (1-\alpha)\rho]^{n-1}[(1-\alpha)\rho] \qquad n > 0$$

where, $\rho = \frac{\lambda E[X]}{\mu} = \frac{\lambda}{\mu(1-\alpha)}$

Measures of effectiveness when batch size is constant(K)

Measure	Expression
Average number of customers in the Queue(L_q)	$\frac{2\rho^2+(K-1)\rho}{2(1-\rho)}$

Average number of customers in the system(L_s)	$\frac{K+1}{2}\frac{\rho}{(1-\rho)}$

Measures of effectiveness when batch size is distributed geometrically with parameter (1-α)

Measure	Expression
Average number of customers in the Queue(L_q)	$\frac{2\rho^2+rE[X^2]}{2(1-\rho)}$
Average number of customers in the system(L_s)	$\frac{\rho+rE[X^2]}{2(1-\rho)}$

where, $\rho = \frac{\lambda E[X]}{\mu} = \frac{\lambda}{\mu(1-\alpha)}$, $E[X^2] = \frac{1+\alpha}{(1-\alpha)^2}$

2.7 Bulk Service

For this model it is assumed that the arrivals occur at a single channel facility as an ordinary Poisson process, they are served FCFS and that these customers are served K at a time. If less than k are in service, new arrivals immediately enter service up to the limit K, and finish with the others regardless of the time into service after the service begins. Also the amount of time required for the service of a batch is exponentially distributed regardless of the fact that the batch is of full size K.

GLOSSARY

Arrival rate (λ): Average rate at which customers arrive to the system. Has units of "customers / time unit".

Blocking Probability: Blocking probability gives the probability of the event that an arrival finds all servers busy and leaves without service.

FIFO: First in, first out (FIFO) queuing is the most basic queue scheduling discipline. In FIFO queuing, all packets are treated equally by placing them into a single queue, and then servicing them in the same order that they were placed into the queue. FIFO queuing is also referred to as First come, first served (FCFS) queuing.

LIFO: The LIFO Queue model supports the last-in, first-out (LIFO) queuing discipline. The entity that departs from the queue at a given time is the most recent arrival.

M/M/1 model: A queuing model with one server and arrival and service time exponentially distributed.

M/M/c model: A queuing model with c servers and arrival and service time exponentially distributed.

M/M/1/N model: A queuing model with 1 server, system capacity N and arrival and service time exponentially distributed.

M/M/c/K model: A queuing model with c servers, system capacity K and arrival and service time exponentially distributed

Orbit: Queuing systems with retrial of the attempts are characterized by the fact that an arrival customer who finds the server occupied is obliged to join a group of blocked customers, called orbit, and reapply after random intervals of time to obtain the service.

Queuing delay: is the time a job waits in queue until it can be executed.

Queuing model: is used to approximate a real queuing situation or system, so the queuing behaviour can be analysed mathematically. Queuing models help measure a number of useful steady state performance measures including the average number in the queue, or the system, average time spent in the queue, or the system, distribution of waiting times, the probability the queue is full, or empty, and the probability of finding the system in a particular state.

Retrial queues: Many queuing systems have the feature that, customers who find all servers busy upon arrival are obliged to leave the service area and come back to the system after a random amount of time

Sample Path: A sample path for the stochastic process $\{X_t: t \in T\}$ is the function on T to the range of the process which assigns to each t the value $X_t(w)$, where w is a previously given fixed point in the domain of the process.

Service Rate (µ): The average rate at which an individual server can serve customers. Has units of "customers / time unit" and is the reciprocal of the average time it takes to serve one customer.

Service Discipline: Queue discipline or service discipline refers to the manner in which customers are selected for service when a queue has formed like FIFO/FCFS, LIFO etc.

Servers: The number of servers available to serve customers entering a queuing system. The number of servers must be a whole number that is greater than or equal to one.

Steady State: The state of the system after it has been in operation for a long time.

System Capacity: The total number of customers that can be in the system, either waiting or being served. Must be a whole number that is greater than or equal to one.

Traffic Intensity: Traffic intensity is a measure of the average occupancy of a server or resource during a specified period of time.

Throughput: is the number of customers served per unit time

Utilization: The proportion of time a server is busy is the server utilization.

Waiting Times: Customer waiting time can be of two types, the time the customer spends in the queue and the total time a customer spends in the system(waiting time in queue + service).