Queuing Theory

Introduction: We will study a class of modes in which customer arrive in some random manner at a service facility. Upon arrival they are made to wait in queue until it is their turn to be served. Once served, they are generally assumed to leave the system. For such models we will be interested in determining among other things, such quantities as the average number of customers in the system (or in the queue) and average time a customer spends in the system (or spends waiting in the queue).

Preliminaries: Some fundamental quantities of interest for queueing models are

L = the average number of customer in the system;
L_Q = the average number of customers waiting in queue;
W = the average amount of time a customer spends in the system;
W_Q = the average amount of time a customer spends waiting in queue.

L = (lambda) * W, Suppose you are waiting for 10 sec, on average 5 customers/second ARRIVE korche.
Then: no. of customers in the System will be 10 sec * 5 customers/second = 50 customers (L = W * lambda)

Imagine that entering customers are forced to pay money (according to the rule) to the system. We would then have the following basic cost identity:

Average rate at which the system earns = $\lambda_a \times$ average amount an entering customer pays.
Where, λ_a = average arrival rate of entering customers. That is, if $N(t)$ denotes the number of customer arrivals by time t, then $\lambda_a = \lim_{t\to\infty} \frac{N(t)}{t}$.

Supposing that each customer pays \$1 per unit time while in the system yields the so-called Littles's formula,

$$L = \lambda_a W \qquad \cdots \qquad (1)$$

This follows since, under this cost rule, the rate at which the system earns is just the number of customer in the system and the amount a customer pays is just equal to its time in the system.

Similarly, if we suppose that each customer pays \$1 per unit time while in queue, then it yields

$$L_Q = \lambda_a W_Q \qquad \cdots \qquad (2)$$

Steady-State Probabilities: Let, $X(t)$ denote the number of customers in the system at time t and define $P_n, n \geq 0$, by

$$P_n = \lim_{t\to\infty} P\{X(t) = n\}$$

P_n equals the (long-run) proportion of time that the system contains exactly n customers. For example, if $P_0 = 0.3$, then in the long run, the system will be empty of customers for 30 percent of the time.

Two other sets of limiting probabilities are $\{a_n, n \geq 0\}$ and $\{d_n, n \geq 0\}$, where

$a_n =$ proportion of customers that find *n* in the system when they arrive.

$d_n =$ proportion of customers leaving behind *n* in the system when they depart.

Example 1: Consider a queuing model in which all customers have service times equal to 1 and where the times between successive customers are always greater than 1 [for instance, the inter arrival times could be uniformly distributed over (1,2)]. Hence as every arrival finds the system empty and every departure leaves it empty, we have

$$a_0 = d_0 = 1$$

However,

$$P_0 \neq 1$$

as the system is not always empty of customers.

Proposition: In any system in which customers arrive one at a time and are served one at a time

$$a_n = d_n, \qquad n \geq 0$$

Proof: An arrival will see *n* in the system whenever the number in the system goes from *n* to *n* + 1; similarly, a departure will leave behind *n* whenever the number in the system goes from *n* + 1 to *n*. Now in any interval of time *T* the number of transitions from *n* to *n* + 1 must equal to within 1 the number from *n* + 1 to *n*. [For instance, if transitions from 2 to 3 occur 10 times, then 10 times there must have been transition back to 2 from a higher state (namely, 3).] Hence, the rate of transitions from *n* to *n* + 1 equals the rate from *n* + 1 to *n*; or equivalently, the rate at which arrivals find *n* equals the rate at which departures leave *n*. Thus, $a_n = d_n, \qquad n \geq 0$ (proved).

Exponential Models:

Arrival Rate = (Lambda) .... Arrivals are Probabilistic ...sometimes more, sometimes less no. of customers arrive ... So queues may be formed because of fixed service rate
Service Rate = Departure Rate = (Mu) ... This is deterministic, NOT Random!

A Single-Server Exponential Queuing System: Suppose that customers arrive at a single-server service station in accordance with a Poisson process having rate λ. That is, the time between successive arrivals are independent exponential random variables having mean ${}^{1}\!/_{\lambda}$. Each customer upon arrival goes directly into service if the server is free and if not the customer joins the queue. When the server finishes serving a customer, the customer leaves the system and the next customer in line, if there is any, enters service. The successive service times are assumed to be independent exponential random variables having mean ${}^{1}\!/_{\mu}$.

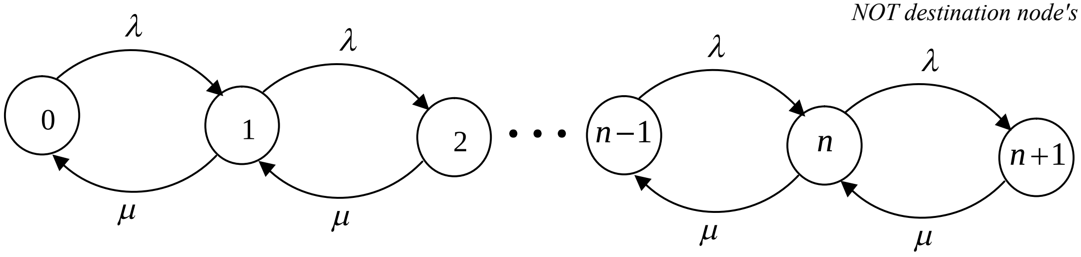
The above is called the *M*/*M*/1 queue. The two *M*s refer to the fact that both the inter arrival and the service distributions are exponential (and thus memoryless, or Markovian) and the 1 to the fact that there is a single server. To analyze it, we shall begin by determining the limiting probabilities P_n for $n = 0,1,\cdots$

We know that, *the rate at which the process enters state n equals the rate at which it leaves state n*. Let us now determine these rates. Consider first state 0. When in state 0, the process can leave only by an arrival as clearly there cannot be a departure when the system is empty. Since the arrival rate is λ and the proportion of time the process is in state 0 is P_0, it follows that the rate at which the process leaves state 0 is λP_0. On the other hand, state 0 can only be reached from state 1 via a departure. That is, if there is a single customer in the system and he completes the service, then the system becomes empty. Since the service rate is μ and the proportion of time that the system has exactly one customer is P_1, it follows that the rate at which the process enters state 0 is μP_1.

Hence, from our rate equality principle we get our first equation,

Rules of Thumb:
i. inflow = outflow in Network flow diagram
ii. edge weight is always multiplied by originating (source) node's P, NOT destination node's P

$$\lambda P_0 = \mu P_1$$



Now consider state 1. The process can leave this state either by an arrival (which occurs at rate λ) or a departure (which occurs at rate μ). Hence, when in state 1, the process will leave this state at a rate of $\lambda + \mu$. Since the proportion of time the process is in state 1 is P_1, the rate at which the process leaves state 1 is $(\lambda + \mu)P_1$. On the other hand, state 1 can be entered either from state 0 via an arrival or from state 2 via a departure. Hence, the rate at which the process enters state 1 is $\lambda P_0 + \mu P_2$. Though the reasoning for other states is similar, we obtain the following set of equations:

State	*Rate at which the process leaves = rate at which it enters*	
0	$\lambda P_0 = \mu P_1$	
$n, n \geq 1$	$(\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$	$\cdots$ (3)

From equation (3), we get

$$P_1 = \frac{\lambda}{\mu} P_0$$

$$P_{n+1} = \frac{\lambda}{\mu} P_n + (P_n - \frac{\lambda}{\mu} P_{n-1}), \qquad n \geq 1$$

Solving in terms of P_0 yields

Putting n = 0, we get $P_1 = \frac{\lambda}{\mu} P_0$

Putting n = 1, we get $P_2 = \frac{\lambda}{\mu} P_1 + \left(P_1 - \frac{\lambda}{\mu} P_0 \right) = \frac{\lambda}{\mu} P_1 = \frac{\lambda}{\mu} \cdot \frac{\lambda}{\mu} P_0 = \left(\frac{\lambda}{\mu} \right)^2 P_0$

Putting n = 2, we get $P_3 = \frac{\lambda}{\mu} P_2 + \left(P_2 - \frac{\lambda}{\mu} P_1 \right) = \frac{\lambda}{\mu} P_2 = \frac{\lambda}{\mu} \cdot \left(\frac{\lambda}{\mu} \right)^2 P_0 = \left(\frac{\lambda}{\mu} \right)^3 P_0$

Putting n = 3, we get $P_4 = \frac{\lambda}{\mu}P_3 + \left(P_3 - \frac{\lambda}{\mu}P_2\right) = \frac{\lambda}{\mu}P_3 = \frac{\lambda}{\mu}\cdot\left(\frac{\lambda}{\mu}\right)^3 P_0 = \left(\frac{\lambda}{\mu}\right)^4 P_0$

$$\vdots$$

Putting n = n, we get $P_{n+1} = \frac{\lambda}{\mu}P_n + \left(P_n - \frac{\lambda}{\mu}P_{n-1}\right) = \frac{\lambda}{\mu}P_n = \frac{\lambda}{\mu}\cdot\left(\frac{\lambda}{\mu}\right)^n P_0 = \left(\frac{\lambda}{\mu}\right)^{n+1} P_0$

To determine P_0 we use the fact that, P_n must sum to 1 and thus

$$1 = \sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n P_0 = \frac{P_0}{1-\lambda/\mu}, \qquad \left[\because 1 + x + x^2 + x^3 + \cdots = \frac{1}{1-x}\right]$$

$$\Rightarrow P_0 = 1 - \frac{\lambda}{\mu}$$

$$\therefore P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \qquad n \geq 1 \qquad \cdots \quad (4)$$

Now let us attempt to express the quantities L, L_Q, W and W_Q in terms of the limiting probabilities P_n. Since P_n is the long-run probability that the system contains exactly n customers, the average number of customers in the system clearly is given by

$$L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \qquad \left[\because E(x) = \sum xP(x)\right]$$

$$= \left(1 - \frac{\lambda}{\mu}\right)\sum_{n=0}^{\infty} n\left(\frac{\lambda}{\mu}\right)^n$$

$$= \left(1 - \frac{\lambda}{\mu}\right)\left[\frac{\lambda/\mu}{\left(1-\lambda/\mu\right)^2}\right], \qquad \left[\because \sum_{n=0}^{\infty} nx^n = x + 2x^2 + 3x^3 + \cdots = \frac{x}{(1-x)^2}\right]$$

$$= \frac{\lambda/\mu}{1-\frac{\lambda}{\mu}} = \frac{\lambda/\mu}{\frac{\mu-\lambda}{\mu}} = \frac{\lambda}{\mu}\cdot\frac{\mu}{\mu-\lambda} = \frac{\lambda}{\mu-\lambda} \qquad \cdots \quad (5)$$

The quantities W, W_Q and L_Q now can be obtained with the help of equations (1) and (2). That is, since $\lambda_a = \lambda$, we have from equation (5) that

$$W = \frac{L}{\lambda} = \frac{\lambda/(\mu-\lambda)}{\lambda} = \frac{\lambda}{\mu-\lambda}\cdot\frac{1}{\lambda} = \frac{1}{\mu-\lambda}$$

$$W_Q = W - E[S] = W - \frac{1}{\mu} = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\mu-\mu+\lambda}{\mu(\mu-\lambda)} = \frac{\lambda}{\mu(\mu-\lambda)}$$

where, $\underline{E}[S]$ = average service time = $\frac{1}{\mu}$ (for exponential distribution).

$$L_Q = \lambda W_Q = \lambda \cdot \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Example: Suppose that customer arrive at a Poisson rate of one per every 12 minutes, and that the service time is exponential at a rate of one service per 8 minutes. What are *L*, *W*, W_Q and L_Q ?

Solution: Since $\lambda = \frac{1}{12},\ \mu = \frac{1}{8}$, we have

$$L = \frac{\lambda}{\mu-\lambda} = \frac{\frac{1}{12}}{\frac{1}{8}-\frac{1}{12}} = \frac{\frac{1}{12}}{\frac{3-2}{24}} = \frac{1}{12} \times \frac{24}{1} = 2 \qquad \text{(Ans.)}$$

$$W = \frac{1}{\mu-\lambda} = \frac{1}{\frac{1}{8}-\frac{1}{12}} = \frac{1}{\frac{3-2}{24}} = 24 \qquad \text{(Ans.)}$$

$$W_Q = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\frac{1}{12}}{\frac{1}{8}\left(\frac{1}{8}-\frac{1}{12}\right)} = \frac{\frac{1}{12}}{\frac{1}{8}\left(\frac{3-2}{24}\right)} = \frac{1}{12} \times \frac{8\times 24}{1} = 16 \qquad \text{(Ans.)}$$

$$L_Q = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\left(\frac{1}{12}\right)^2}{\frac{1}{8}\left(\frac{1}{8}-\frac{1}{12}\right)} = \frac{\left(\frac{1}{12}\right)^2}{\frac{1}{8}\left(\frac{3-2}{24}\right)} = \frac{1}{12\times 12} \times \frac{8\times 24}{1} = \frac{4}{3} \qquad \text{(Ans.)}$$

☺Good Luck☺