Chapter #4

Analysis of the G/M/1 Queue

We consider here the analysis of the G/M/1 queue. This may be considered to be the dual of the M/G/1 queue which has been discussed in detail in Chapter 3. It may be noted that the G/M/1 queue is a queue with a single server and an infinite buffer. In this queue, the service times are exponentially distributed but arrivals may come from any general process, i.e. one where the inter-arrival times are generally distributed with a given distribution. It can be easily seen that this is exactly the opposite of what we had assumed for the M/G/1 queue, i.e. inter-arrival times are exponentially distributed but the service times may have any general distribution.

The G/M/1 queue may also be analysed in a variety of ways. Here we follow an approach based on an appropriately imbedded Markov chain at the customer arrival instants. An alternative approach using the method of supplementary variables may also be used where a two-dimensional state description is used; this would consist of the number in the system and the elapsed time since the last arrival. The approach using the imbedded Markov chain is easier to follow. It may also be easily extended to analyse a general G/M/m queue (i.e. with m servers) as given in [Kle75].

We consider the G/M/1 queue following an FCFS service strategy. Jobs are assumed to arrive with inter-arrival times that are identical, independently distributed random variables with pdf $a(t)$ and cdf $A(t)$. We also assume that the Laplace Transform of $a(t)$ is $L_A(s)$. The mean inter-arrival time is l^{-1} corresponding to a mean arrival rate of l. The service times are exponentially distributed with mean m^{-1} so that the total traffic offered to the queue is $r = l/m$ erlangs. Stability considerations will require that the queue will be at equilibrium only when $r < 1$.

For the imbedded Markov chain analysis of the G/M/1 queue, we consider the imbedded time points to be the arrival instants of jobs to the system. The system state is considered to be the number in the system *immediately before the arrival instants*. For the i^{th} arrival, let n_i be the number in the system just before this arrival. Let s_{i+1} be the number of jobs served between the i^{th} and the $(i+1)^{th}$ arrival. We then get

$$n_{i+1} = n_i + 1 - s_{i+1} \qquad n_i=0,1,....,¥ \qquad s_{i+1}£\ n_i+1 \tag{1}$$

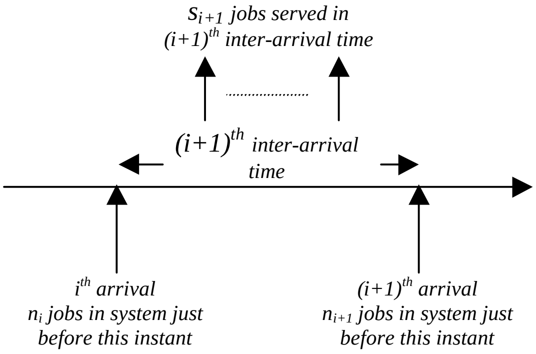
s_{i+1} jobs served in
$(i+1)^{th}$ inter-arrival time
$(i+1)^{th}$ inter-arrival time
i^{th} arrival
n_i jobs in system just before this instant
$(i+1)^{th}$ arrival
n_{i+1} jobs in system just before this instant

Figure 1. Imbedded Markov Chain for the G/M/1 Queue

This has been illustrated in Fig. 1. It may be easily verified from (1), that the $\{n_i\}$ will indeed form a Markov chain. We consider this chain at equilibrium, i.e. with $i ® ¥$. Under these equilibrium conditions, let $n_{i+1}=k$ and $n_i=j$. The one-step transition probability of this Markov chain will then be given by

$$\begin{aligned} p_{jk} &= P\{n_{i+1} = k \mid n_i = j\} \\ p_{jk} &= 0 \quad for \quad k > j+1 \end{aligned} \tag{2}$$

where p_{jk} is the probability that $(j+1-k)$ jobs get served between the consecutive arrival instants. The equilibrium values of these one-step, state transition probabilities are obtained subsequently. However, we note that once these are obtained for a G/M/1 queue at equilibrium, we can use them to obtain the equilibrium state probabilities p_j $j=0,1,.....¥$ of finding j jobs in the system just before an arbitrary arrival instant. This would require solving the set of equations

$$p_k = \sum_{j=0}^{\infty} p_j p_{jk} \qquad k=0,1,........,¥ \tag{3}$$

Note that the p_k's must also satisfy the normalisation condition $\sum_{k=0}^{\infty} p_k = 1$.

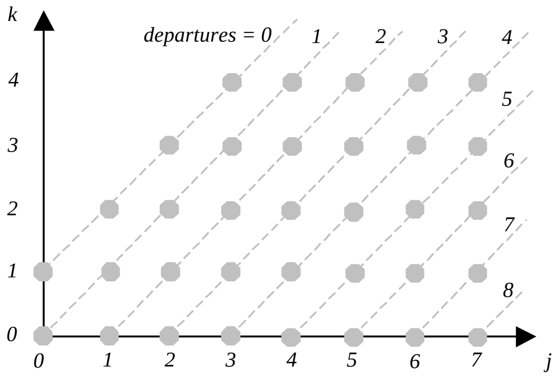

Figure 2. The Allowable State Transitions of the G/M/1 Queue

In Fig. 2, we have shown the state transitions j®k which will be allowed from just before one job arrival instant to just before the subsequent job arrival instant. The points that have not been shown on Fig. 2 are ones for which the corresponding j®k transition will not occur. The dashed lines shown in Fig. 2 correspond to the number of departures between successive arrival instants that must occur in order to get the corresponding j®k transition. Let a_n be the probability of n departures in an inter-arrival time interval, given that the server is busy for that entire interval. We can use this to write the following balance equations.

$$p_0 = \sum_{k=0}^{\infty} a_{k+1} p_k$$

$$p_j = a_0 p_{j-1} + \sum_{k=0}^{\infty} a_{k+1} p_{j+k} \qquad j = 1,.......,\infty \qquad (4)$$

where a_j may be found using

$$a_j = \int_{x=0}^{\infty} \frac{(mx)^j}{j!} e^{-mx} a(x)dx \qquad j = 0,1,.......,\infty \qquad (5)$$

Note that (5) follows from the fact that the service times are exponentially distributed and that, therefore, the number of departures within an inter-arrival time (where the server is always busy) would have a Poisson distribution. From (5), it also follows that the a_j*'s* may also be found as the coefficient of the z^j in the expansion of the Laplace Transform $L_A(m\text{-}mz)$ of the pdf of the inter-arrival times. These values of a_j's may be used to solve the balance equations of (4) to obtain the state probabilities p_j $j=0,1,.....,¥$.

The solution for the equilibrium state probabilities p_j $j=0,1,....,,¥$ just before the job arrival instants are given by

$$p_j = (1-s)s^j \qquad j = 0,1,.......,\infty \tag{6}$$

where s is a unique root of

$$s = L_A(m - ms) \tag{7}$$

We verify this below by direct substitution in the expression for p_j given in (4).

$$(1-s)s^j = a_0(1-s)s^{j-1} + \sum_{k=0}^{\infty} a_{k+1}(1-s)s^{j+k}$$

$$s^j = a_0 s^{j-1} + \sum_{k=1}^{\infty} a_k s^{j+k-1}$$

$$s = a_0 + \sum_{k=1}^{\infty} a_k s^k$$

which will be satisfied if s is a unique root of $L_A(m\text{-}ms)$. It has been shown that if $r<1$, then there will be a unique real solution for s using (7) which will be in the range $0<s<1$. This is the value of s which should be used to obtain the state probabilities given by (6). Note that $s=1$ will always be a solution of (7) since $L_A(0)=1$ will always hold.

The number of jobs in the system at the job arrival instants is geometrically distributed with parameter s. This follows from the state distribution given in (6). It is interesting to note that this holds for the G/M/1 queue regardless of the actual nature of the arrival process, i.e. the distribution of the inter-arrival times. The actual solution procedure would be to use the pdf (or its Laplace Transform) of the inter-arrival time to get a solution s of $s=L_A(m\text{-}ms)$ which lies in the range $0<s<1$. This value can then be used to get the state distribution from (6). It may be noted that this state distribution will be the equilibrium distribution that will be seen if the system is examined just before the arrival instant of a job to the system. This will not hold at any arbitrary time instant as PASTA will not be applicable for a G/M/1 queue. Since the state changes in the actual queue are at the most +1 or -1, Kleinrock's principle may applied to claim that the state distribution of (6) will also be applicable at the departure instants of jobs from the system.

Consider an arrival to a G/M/1 queue. Its waiting time w_q in queue will be zero if it finds the system to be empty on arrival (with probability p_0). If it finds n jobs in the system (including the one currently in service), then it must wait for all of them to get served before its own service can begin. Note that we are assuming an FCFS model here even though the mean results will be the same regardless of the actual discipline being followed. Using the memory less property of the exponential distribution (for the job that is currently in service), we get the mean waiting time in queue, W_q, to be

$$W_q = \sum_{n=1}^{\infty} \frac{n}{m}(1-s)s^{n} = \frac{s}{m(1-s)} \tag{8}$$

Note that a job, which sees n jobs already in the system (including the one in service) when it arrives, will encounter a random waiting time that will be the sum of n independent, exponentially distributed random variables. Using this, it can be shown that the pdf $f_{Wq}(t)$ of the waiting time will be

$$f_{Wq}(t) = (1-s)d(t) + ms(1-s)e^{-m(1-s)t} \qquad t \geq 0 \tag{9}$$

Note that this corresponds to an exponential distribution which has a jump (because of the delta function) of $(1\text{-}s)$ at the origin $t=0$. This may also be used to directly find the mean waiting time in queue to be the same result as in (8).