## The M/G/1 Queueing System

For the M/G/1 queueing system being operated under the FIFO service rule, we derive the expressions of the following quantities in terms of the arrival rate $\lambda$, the mean service time $E[S]$, and the variance of service time $\sigma_S^2$.

- $W$: the average time a randomly arriving customer will spend in the system, which is composed of the waiting time in the queue and the service time.
- $L$: the average number of people in the system that a randomly arriving customer finds, which is composed of the number of people in the queue and the person in service.
- $\rho$: the long run fraction of time the server is busy, which is equivalently the probability that the server is busy at a random point in time.
- $B$: the long run average duration of a server busy period.

Recall that the M/G/1 queueing system is a single server queueing system in which the customer arrival process is Poisson with rate $\lambda$ and the service time, $S$, for each customer follows a general distribution with PDF $f_S(s)$, mean $E[S]$, and variance $\sigma_S^2$. Let us first compute $\rho$ and $B$.

Suppose we have been watching the queueing system for a very very long time, say $T^*$ minutes where $T^*$ is very large. If we recorded the number of minutes the server was busy during this long period of time and then divided it by $T^*$, we would obtain $\rho$, the long run fraction of time the server is busy. During the long period of time $T^*$, we would expect there have been $\lambda T^*$ customers arriving to the queueing system, each of who takes on average $E[S]$ minutes to be served. This means

$$\rho = \frac{\text{number of minutes server is busy}}{T^*} = \frac{\lambda T^* \times E[S]}{T^*} = \lambda E[S]\,.$$

To compute $B$, the long run average length of a server busy period, we again think of observing the system for a long period of time. During this long period of time, there occur a large number, say $N$, of busy periods. Since every busy period is followed by an idle period, we could say that the number of idle periods is $N$ as well. (The difference between the number of busy periods and the number of idle periods would be at most one, and this is negligible compared to $N$.). Since the average length of a busy period is $B$ and the number of busy periods is $N$, the total amount of time the server is busy, over the the long period of time we are observing, is $NB$.

The length of an idle period is, on average, $1/\lambda$ since an idle period occurs when the server is waiting for a customer to arrive after the queue becomes empty. Since the arrival process is Poisson and therefore memoryless, the server will wait a negative exponentially distributed amount of time until the next customer arrives. Therefore if there are $N$ idle periods, the total amount of time the server is idle is $N/\lambda$. The fraction of time the server is busy, $\rho$, can now be computed by

$$\rho = \frac{NB}{(NB + N/\lambda)} = \frac{B}{B + \frac{1}{\lambda}}\,.$$

Solving for $B$ and using $\rho = \lambda E[S]$, we have

$$B = \frac{\frac{\rho}{\lambda}}{1-\rho} = \frac{E[S]}{1 - \lambda E[S]}\,.$$

Let $T$ be a random variable denoting the amount of time that a randomly arriving customer, say I, will spend in the system. Our goal is to compute $W = E[T]$. Note that we can decompose $T$ into the following three random variables:

- $T_1$, the remaining service time of the customer currently in service.
- $T_2$, the time required to serve the customers waiting ahead of me in the queue.
- $T_3$, my service time.

Clearly, $W = E[T_1] + E[T_2] + E[T_3]$. Since the expected service time for each customer is $E[S]$, we have

$$E[T_3] = E[S]\,.$$

To obtain $E[T_2]$, we first compute the conditional expectation of $T_2$ given that there are already $n$ customers in the system when I, a randomly arriving customer, arrive in the system. Since one customer is being served and $n-1$ customers are waiting in the queue,

$$E[T_2 \mid n] = \begin{cases} (n-1)E[S], & n \geq 1, \\ 0, & n = 0. \end{cases}$$

Using the total expectation theorem, we obtain

$$\begin{aligned} E[T_2] &= \sum_n E[T_2 \mid n] P_n = \sum_{n \geq 1} (n-1) E[S] P_n \\ &= E[S] \sum_{n \geq 1} n P_n - E[S] \sum_{n \geq 1} P_n\,. \end{aligned}$$

Note that $\sum_{n\geq 1} nP_n = L$ and $\sum_{n\geq 1} P_n = \rho$. So we have

$$E[T_2] = E[S]L - E[S]\rho\,.$$

Now let us compute $E[T_1]$. Since the service time distribution may not be negative exponential, we should consider the issue of *random incidence*. Note that I, a randomly arriving customer, am more likely to join the queueing system when the duration of the current service is long than when it is short, because long services take up more of the time horizon. The expected remaining service time for the customer in service when I randomly arrive in the system is given by (see Equation (2.66) in the textbook)

$$E[T_1 \mid n] = \begin{cases} \frac{\sigma_S^2 + E[S]^2}{2E[S]} = \frac{\sigma_S^2}{2E[S]} + \frac{E[S]}{2}, & n \geq 1, \\ 0, & n = 0. \end{cases}$$

$E[T_1]$ is then computed by

$$\begin{aligned} E[T_1] &= \sum_n E[T_1 \mid n] P_n \\ &= \sum_{n\geq 1} \left( \frac{\sigma_S^2}{2E[S]} + \frac{E[S]}{2} \right) P_n \\ &= \left( \frac{\sigma_S^2}{2E[S]} + \frac{E[S]}{2} \right) \sum_{n\geq 1} P_n \\ &= \left( \frac{\sigma_S^2}{2E[S]} + \frac{E[S]}{2} \right) \rho\,. \end{aligned}$$

Note that the last equality holds because $\sum_{n\geq 1} P_n$ is the probability of the server being busy, which is $\rho$.

Adding up all the parts we have derived above, we get

$$\begin{aligned} W &= E[T_1] + E[T_2] + E[T_3] \\ &= \left( \frac{\sigma_S^2}{2E[S]} + \frac{E[S]}{2} \right) \rho + E[S]L - E[S]\rho + E[S]\,. \end{aligned}$$

We now have one linear relationship between $L$ and $W$. Combining this with Little's Law, $L = \lambda W$, we have two equations for two unknowns, $L$ and $W$. A little algebra gives us

$$L = \rho + \frac{\rho^2 + \lambda^2 \sigma_S^2}{2(1-\rho)}\,,$$

and $W$ follows.