9 The $M/G/1$ system

In the previous chapters we considered queueing system with Poisson arrivals and exponentially distributed service times. Poisson arrivlas are in many cases a fairly realistic model for the arrival process, but exponential service times are not very common in practice. In many systems the coefficient of variation of the service times will be smaller (or geater) than 1. Therefore it is essential to extend the theory to the case of generally distributed service times. In this chapter we will treat the case of Poisson arrivals and generally distributed (though independent) service times. So we will consider the $M/G/1$ system. Customers are again served in order of arrival.

9.1 The queue length distribution on departure instants

In the $M/G/1$ queue customers arrive one by one according to a Poisson stream with rate λ. The service times have a general distribution with density $f_B(\cdot)$ and mean $E(B)$. For stability we assume that

$$\rho = \lambda \cdot E(B) < 1.$$

The state of the $M/G/1$ queue at time t can be described by the pair (n, x) where n denotes the number of customers in the system and x the service time already received by the customer in service. We thus need a two-dimensional state description. The first dimension is still discrete, but the other one is continuous and this essentially complicates the analysis. However, if we look at the system just after departures, then the state description can be simplified to n only, because $x = 0$ for the new customer (if any) in service. In other words, we are going to look at the *Markov chain embedded on departure instants*; recall that for the $G/M/1$ we studied the Markov chain embedded on arrival instants. A question that immediately arises is whether the distribution of the number of customers in the system on departure instants is useful? The first observation is that, since customers arrive and leave one by one, the queue length distribution on departure instants is exactly the same as the queue length distribution on arrival instants (see section 4.7), and by PASTA, the latter is equal to the equilibrium queue length distribution. Further, we shall see that the departure distribution can be connected to the sojourn time distribution by using a distributional version of Little's law.

To specify the transition probabilities $p_{i,j}$ of the embedded Markov chain we first introduce the probabilities a_n defined as the probability that exactly n customers arrive during a service time. By conditioning on the length of the service time it follows that

$$a_n = \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_B(t) d(t), \qquad n = 0, 1, 2, \ldots \tag{1}$$

Clearly $p_{i,j} = 0$ for all $j < i - 1$ and $p_{i,j}$ for $j \geq i - 1$ gives the probability that exactly $j - i + 1$ customers arrived during the service time of a customer. This holds for $i > 0$. In state 0 a customer leaves behind an empty system and then $p_{0,j}$ gives the probability that during the service time of the next customer exactly j customers arrived. Hence the

matrix P of transition probabilities takes the form

$$P = \begin{pmatrix} a_0 & a_1 & a_2 & a_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ 0 & 0 & 0 & a_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix},$$

The equilibrium probabilities p_i satisfy the equilibrium equations

$$\begin{aligned} p_i &= p_{i+1}a_0 + p_i a_1 + \cdots + p_1 a_i + p_0 a_i \\ &= \sum_{n=0}^{i} p_{i+1-n} a_n + p_0 a_i, \qquad i = 0, 1, \ldots \end{aligned} \tag{2}$$

Note that these equations can be easily solved numerically: starting with $p_0 = 1 - \rho$ (Why?), we can use the above equations to recursively determine $p_1, p_2, \ldots$; namely the above equation can be rewritten as

$$p_{i+1}a_0 = p_i - (p_i a_1 + \cdots + p_1 a_i + p_0 a_i), \qquad i = 0, 1, 2, \ldots,$$

so, once we have determined p_0 up to p_i, we can use this equation to compute p_{i+1}. To solve the equilibrium equations analytically we are going to use generating functions. Let us introduce the probability generating functions

$$P(z) = \sum_{i=0}^{\infty} p_i z^i, \qquad A(z) = \sum_{i=0}^{\infty} a_i z^i,$$

which are defined for all $z \leq 1$. Multiplying (2) by z^i and summing over all i leads to

$$\begin{aligned} P(z) &= \sum_{i=0}^{\infty} \left(\sum_{n=0}^{i} p_{i+1-n} a_n + p_0 a_n \right) z^i \\ &= z^{-1} \sum_{i=0}^{\infty} \sum_{n=0}^{i} p_{i+1-n} z^{i+1-n} a_n z^n + \sum_{i=0}^{\infty} p_0 a_i z^i \\ &= z^{-1} \sum_{n=0}^{\infty} \sum_{i=n}^{\infty} p_{i+1-n} z^{i+1-n} a_n z^n + p_0 A(z) \\ &= z^{-1} \sum_{n=0}^{\infty} a_n z^n \sum_{i=n}^{\infty} p_{i+1-n} z^{i+1-n} + p_0 A(z) \\ &= z^{-1} A(z)(P(z) - p_0) + p_0 A(z). \end{aligned}$$

Hence we find

$$P(z) = \frac{p_0 A(z)(1 - z^{-1})}{1 - z^{-1} A(z)}.$$

Substituting $p_0 = 1-\rho$ (which also follows from the requirement $P(1) = 1$) and multiplying numerator and denominator by $-z$ we obtain

$$P(z) = \frac{(1-\rho)A(z)(1-z)}{A(z)-z}\,. \tag{3}$$

By using (1), the generating function $A(z)$ can be rewritten as

$$\begin{aligned} A(z) &= \sum_{n=0}^{\infty} \int_{t=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} f_B(t) dt z^n \\ &= \int_{t=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\lambda t z)^n}{n!} e^{-\lambda t} f_B(t) dt \\ &= \int_{t=0}^{\infty} e^{-(\lambda-\lambda z)t} f_B(t) dt \\ &= \widetilde{B}(\lambda - \lambda z) \end{aligned} \tag{4}$$

Substitution of (4) into (3) finally yields

$$P(z) = \frac{(1-\rho)\widetilde{B}(\lambda-\lambda z)(1-z)}{\widetilde{B}(\lambda-\lambda z)-z}\,. \tag{5}$$

This formula is one form of the *Pollaczek-Khinchin formula*. Below we will derive a similar formula for the sojourn time. By differentiating formula (5) we can determine the moments of the queue length (see section 1.2). To find its distribution, however, we have to invert formula (5), which usually is very difficult. In the special case that $\widetilde{B}(s)$ is a quotient of polynomials in s, i.e., *a rational function*, then in principle the right-hand side of (5) can be decomposed into partial fractions, the inverse transform of which can be easily determined. The service time has a rational transform for, e.g., mixtures of Erlang distributions or Hyperexponential distributions (see section 1.4). The inversion of (5) is demonstrated below for exponential and Erlang-2 service times.

Example 9.1 *(M/M/1)*
Suppose the service time is exponentially distributed with mean $1/\mu$. Then (see section 1.4.3)

$$\widetilde{B}(s) = \frac{\mu}{\mu+s}\,.$$

Thus

$$P(z) = \frac{(1-\rho)\frac{\mu}{\mu+\lambda-\lambda z}(1-z)}{\frac{\mu}{\mu+\lambda-\lambda z}-z} = \frac{(1-\rho)\mu(1-z)}{\mu-z(\mu+\lambda-\lambda z)} = \frac{(1-\rho)\mu(1-z)}{(\mu-\lambda z)(1-z)} = \frac{1-\rho}{1-\rho z}\,.$$

Hence

$$p_n = (1-\rho)\rho^n, \qquad n = 0, 1, 2, \ldots$$

Example 9.2 *($M/E_2/1$)*
Suppose the service time is Erlang-2 distributed with mean $2/\mu$. Then (see section 1.4.4)

$$\widetilde{B}(s) = \left(\frac{\mu}{\mu+s}\right)^2,$$

so

$$\begin{aligned} P(z) &= \frac{(1-\rho)\left(\frac{\mu}{\mu+\lambda-\lambda z}\right)^2(1-z)}{\left(\frac{\mu}{\mu+\lambda-\lambda z}\right)^2 - z} \\ &= \frac{(1-\rho)\mu^2(1-z)}{\mu^2 - z(\mu+\lambda-\lambda z)^2} \\ &= \frac{(1-\rho)(1-z)}{1-z(1+\rho(1-z)/2)^2} \\ &= \frac{1-\rho}{1-\rho z - \rho^2 z(1-z)/4}. \end{aligned}$$

For $\rho = 1/3$ we then find

$$\begin{aligned} P(z) &= \frac{2/3}{1-z/3-z(1-z)/36} = \frac{24}{36-13z+z^2} \\ &= \frac{24}{(4-z)(9-z)} = \frac{24/5}{4-z} - \frac{24/5}{9-z} = \frac{6/5}{1-z/4} - \frac{8/15}{1-z/9}. \end{aligned}$$

Hence,

$$p_n = \frac{6}{5}\left(\frac{1}{4}\right)^n - \frac{8}{15}\left(\frac{1}{9}\right)^n, \qquad n = 0, 1, 2, \ldots$$

9.2 The sojourn time and the waiting time

We now turn to the calculation of how long a customer spends in the system. We will show that there is a nice relationship between the transforms of the time spent in the system and the departure distribution.

Let us consider a customer arriving at the system in equilibrium. Denote the total time spent in the system for this customer by the random variable S with distribution function $F_S(\cdot)$ and density $f_S(\cdot)$. The probability that our cstomer leaves behind i customers is equal to p_i (since the system is in equilibrium). For a first-come first-served system it is clear that all customers left behind are precisely those who arrived during his stay in the system. Thus we have

$$p_i = \int_{t=0}^{\infty} \frac{(\lambda t)^i}{i!} e^{-\lambda t} f_S(t)dt.$$

Hence, we find similarly to (4) that

$$P(z) = \widetilde{S}(\lambda - \lambda z).$$

This relation is commonly referred to as Little's distributional law. Substitution of this relation into (5) yields

$$\widetilde{S}(\lambda - \lambda z) = \frac{(1-\rho)\widetilde{B}(\lambda - \lambda z)(1-z)}{\widetilde{B}(\lambda - \lambda z) - z}\,.$$

Making the change of variable $s = \lambda - \lambda z$ we finally arrive at

$$\widetilde{S}(s) = \frac{(1-\rho)\widetilde{B}(s)s}{\lambda\widetilde{B}(s) + s - \lambda}\,. \tag{6}$$

This formula is also known as the *Pollaczek-Khinchin formula*. Since S is the sum of W (his waiting time) and B (his service time), where W and B are independent, it follows that

$$\widetilde{S}(s) = \widetilde{W}(s) \cdot \widetilde{B}(s) \tag{7}$$

(since the transform of the sum of two independent random variables is the product of the transforms of these two random variables; see section 1.3). Together with (6) we get

$$\widetilde{W}(s) = \frac{(1-\rho)s}{\lambda\widetilde{B}(s) + s - \lambda}\,, \tag{8}$$

which is the third form of the Pollaczek-Khinchin formula.

Example 9.3 *($M/M/1$)*
For exponential service times with mean $1/\mu$ we have

$$\widetilde{B}(s) = \frac{\mu}{\mu + s}\,.$$

Thus

$$\widetilde{S}(s) = \frac{(1-\rho)\frac{\mu}{\mu+s}s}{\lambda\frac{\mu}{\mu+s} + s - \lambda} = \frac{(1-\rho)\mu s}{\lambda\mu + (s-\lambda)(\mu+s)} = \frac{(1-\rho)\mu s}{(\mu-\lambda)s + s^2} = \frac{\mu(1-\rho)}{\mu(1-\rho) + s}\,.$$

Hence, S is exponentially distributed with parameter $\mu(1-\rho)$, i.e.,

$$F_S(t) = P(S \le t) = 1 - e^{-\mu(1-\rho)t}, \qquad t \ge 0.$$

Example 9.4 *($M/E_2/1$)*
Suppose that $\lambda = 1$ and that the service time is Erlang-2 distributed with mean $1/3$, so

$$\widetilde{B}(s) = \left(\frac{6}{6+s}\right)^2.$$

Then it follows that (verify)

$$F_S(t) = \frac{8}{5}(1 - e^{-3t}) - \frac{3}{5}(1 - e^{-8t}), \qquad t \ge 0.$$

9.3 The mean value approach

The mean sojourn time can of course be calculated from the Laplace-Stieltjes transform (6) by differentiating and substituting $s = 0$ (see section 1.3). In this section we show that the mean sojourn time can also be determined directly (i.e., without using transforms) with the *mean value approach*.

First, let us derive the arrival relation. A newly arriving customer first has to wait for the *residual service time* of the job in service (if there is one) and then continues to wait for the servicing of all customers which are already waiting in the queue on arrival. By PASTA we know that with probability ρ the server is busy on arrival. Let the random variable B denote the service time, R the residual service time and let L^q denote the number of customers waiting in the queue. Hence,

$$E(W) = E(L^q)E(B) + \rho E(R).$$

Furthermore, we get by Little's law (applied to the queue of waiting customers),

$$E(L^q) = \lambda E(W).$$

Combining these two relations, we find for the mean waiting time

$$E(W) = \frac{\rho E(R)}{1-\rho}\,. \tag{9}$$

Formula (9) is commonly referred to as the *Pollaczek-Khinchin mean value formula*. It remains to calculate the mean residual service time. In the following section we will show that

$$E(R) = \frac{E(B^2)}{2E(B)}\,, \tag{10}$$

which may also be written in the form

$$E(R) = \frac{E(B^2)}{2E(B)} = \frac{\sigma_B^2 + E(B)^2}{2E(B)} = \frac{1}{2}(c_B^2 + 1)E(B), \tag{11}$$

where c_B^2 denotes the squared coefficinet of variation of the service time. An important observation is that, clearly, the mean waiting time only depends upon the first two moments of the service time (and not upon its distribution). So, in practice, it is sufficient to know the mean and standard deviation of the service time in order to estimate the mean waiting time. Finally, expressions for $E(S)$ and $E(L)$ easily follow from the relations $E(S) = E(W) + E(B)$ and $E(L) = E(L^q) + \rho$.

Example 9.5 *(Exponential service times)*
For exponential service times we have $c_B^2 = 1$ and hence $E(R) = E(B)$ (memoryless property!). So, in this case the expressions for the mean performance measures simplify to

$$E(W) = \frac{\rho}{1-\rho}E(B), \qquad E(L^q) = \frac{\rho^2}{1-\rho}, \qquad E(S) = \frac{1}{1-\rho}E(B), \qquad E(L) = \frac{\rho}{1-\rho}.$$

Example 9.6 *(Deterministic service times)*
For deterministic service times we have $c_B^2 = 0$ and hence $E(R) = E(B)/2$. In this case we have

$$\begin{aligned} E(W) &= \frac{\rho}{1-\rho}\frac{E(B)}{2}, \qquad E(L^q) = \frac{\rho^2}{2(1-\rho)}, \\ E(S) &= \frac{\rho}{1-\rho}\frac{E(B)}{2} + E(B), \qquad E(L) = \rho + \frac{\rho^2}{2(1-\rho)}. \end{aligned}$$

9.4 The residual service time

Suppose that a customer arrives when the server is busy and denote the total service time of the customer in service by X. Further let $f_X(\cdot)$ denote the density of X. The basic observation to find $f_X(\cdot)$ is that it is more likely that our customer arrives in a long service time than in a short one. So the probability that X is of length x should be proportional to the length x as well as the frequency of such service times, which is $f_B(x)dx$. Thus we may write

$$P(x \le X \le x + dx) = f_X(x)dx = Cxf_B(x)dx,$$

where C is a constant to normalize this density. So

$$C^{-1} = \int_{x=0}^{\infty} xf_B(x)dx = E(B).$$

Hence

$$f_X(x) = \frac{xf_B(x)}{E(B)}\,.$$

Given that our customer arrives in a service time of length x, the arrival instant will be a random point within this service time, i.e., it will be uniformly distributed within the service time interval $(0, x)$. So

$$P(t \le R \le t + dt | X = x) = \frac{dt}{x}, \qquad 0 \le t \le x.$$

Of course, this conditional probability is zero when $t > x$. Thus we have

$$P(t \le R \le t + dt) = f_R(t)dt = \int_{x=t}^{\infty} \frac{dt}{x} f_X(x)dx = \int_{x=t}^{\infty} \frac{f_B(x)}{E(B)}\,dxdt = \frac{1 - F_B(t)}{E(B)}\,dt.$$

This gives the final result

$$f_R(t) = \frac{1 - F_B(t)}{E(B)}\,,$$

from which we immediately obtain, by partial integration,

$$E(R) = \int_{t=0}^{\infty} tf_R(t)dt = \frac{1}{E(B)}\int_{t=0}^{\infty} t(1 - F_B(t))dt = \frac{1}{E(B)}\int_{t=0}^{\infty} \frac{1}{2}t^2 f_B(t)dt = \frac{E(B^2)}{2E(B)}\,.$$

The above computation can be repeated to obtain all moments of R, yielding

$$E(R^n) = \frac{E(B^{n+1})}{(n+1)E(B)}\,.$$

Example 9.7 *(Erlang service times)*
For an Erlang-r service time with mean r/μ we have

$$E(B) = \frac{r}{\mu}\,, \qquad \sigma^2(B) = \frac{r}{\mu^2}\,,$$

so

$$E(B^2) = \sigma^2(B) + (E(B))^2 = \frac{r(1+r)}{\mu^2}\,.$$

Hence

$$E(R) = \frac{1+r}{2\mu}$$