11 The $M/G/1$ system with priorities

In this chapter we analyse queueing models with different types of customers, where one or more types of customers have priority over other types. More precisely we consider an $M/G/1$ queue with r types of customers. The type i customers arrive according to a Poisson stream with rate λ_i, $i = 1, \ldots, r$. The service time and residual service of a type i customer is denoted by B_i and R_i, respectively. The type 1 customers have the highest priority, type 2 customers the second highest priority and so on. We consider two kinds of priorities. For the *non-preemptive* priority rule higher priority customers may not interrupt the service time of a lower priority customer, but they have to wait till the service time of the low priority customer has been completed. For the *preemptive-resume* priority rule interruptions are allowed and after the interruption the service time of the lower priority customer resumes at the point where it was interrupted. In the following two sections we show how the mean waiting times can be found for these two kinds of priorities.

11.1 Non-preemptive priority

The mean waiting time of a type i customer is denoted by $E(W_i)$ and $E(L_i^q)$ is the number of type i customers waiting in the queue. Further define $\rho_i = \lambda_i E(B_i)$. For the highest priority customers it holds that

$$E(W_1) = E(L_1^q)E(B_1) + \sum_{j=1}^{r} \rho_j E(R_j).$$

According to Little's law we have

$$E(L_1^q) = \lambda_1 E(W_1)$$

Combining the two equations yields

$$E(W_1) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{1 - \rho_1}. \tag{1}$$

The determination of the mean waiting time for the lower priority customers is more complicated. Consider type i customers with $i > 1$. The waiting time of a type i customer can be divided in a number of portions. The first portion is the amount of work associated with the customer in service and all customers with the same or higher priority present in the queue upon his arrival. Call this portion X_1. The second portion, say X_2, is the amount of higher priority work arriving during X_1. Subsequently the third portion X_3 is the amount of higher priority work arriving during X_2, and so on. A realization of the waiting time for a type 2 customer is shown in figure 1. The increments of the amount of work are the service times of the arriving type 1 customers.

Hence the mean waiting time is given by

$$E(W_i) = E(X_1 + X_2 + X_3 + \cdots) = \sum_{k=1}^{\infty} E(X_k).$$

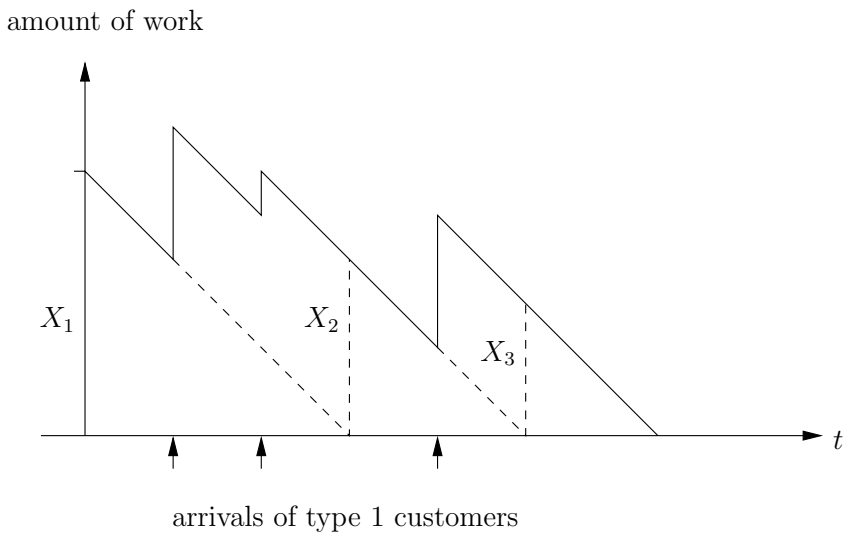

Figure 1: Realization of the waiting time of a type 2 customer

As mentioned above, the first portion of work an arriving type i customer has to wait for is the sum of the service times of all customers with the same or higher priority present in the queue plus the remaining service time of the customer in service. So

$$E(X_1) = \sum_{j=1}^{i} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j).$$

To determine $E(X_{k+1})$ note that X_{k+1} depends on X_k. We therefore condition on the length of X_k. Denote the density of X_k by $f_k(x)$. Then it follows that

$$\begin{aligned} E(X_{k+1}) &= \int_{x=0}^{\infty} E(X_{k+1}|X_k = x) f_k(x) dx \\ &= \int_{x=0}^{\infty} (\lambda_1 x E(B_1) + \cdots + \lambda_{i-1} x E(B_{i-1})) f_k(x) dx \\ &= (\rho_1 + \cdots + \rho_{i-1}) E(X_k). \end{aligned}$$

Repeated application of the relation above yields

$$E(X_{k+1}) = (\rho_1 + \cdots + \rho_{i-1})^k E(X_1), \qquad k = 0, 1, 2, \ldots$$

Hence we find for $i = 2, \ldots, r$

$$E(W_i) = \frac{E(X_1)}{1 - (\rho_1 + \cdots + \rho_{i-1})} = \frac{\sum_{j=1}^{i} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j)}{1 - (\rho_1 + \cdots + \rho_{i-1})}, \tag{2}$$

An intuitive argument (which can be made rigorous) to directly obtain the above equation is by observing that the waiting time of a type i customer is equal to the first portion of

work plus all the higher priority work arriving during his waiting time. So

$$E(W_i) = E(X_1) + \sum_{j=1}^{i-1} \lambda_j E(W_i) E(B_j),$$

from which equation (2) immediately follows. Substitution of Little's law

$$E(L_i^q) = \lambda_i E(W_i)$$

into equation (2) yields

$$\begin{aligned}(1 - (\rho_1 + \ldots + \rho_i))E(W_i) &= \sum_{j=1}^{i-1} E(L_j^q)E(B_j) + \sum_{j=1}^{r} \rho_j E(R_j) \\ &= (1 - (\rho_1 + \ldots + \rho_{i-2}))E(W_{i-1}).\end{aligned}$$

By multiplying both sides of this equality with $1 - (\rho_1 + \ldots + \rho_{i-1})$ we get the simple recursive relation

$$(1 - \sum_{j=1}^{i} \rho_j)(1 - \sum_{j=1}^{i-1} \rho_j)E(W_i) = (1 - \sum_{j=1}^{i-1} \rho_j)(1 - \sum_{j=1}^{i-2} \rho_j)E(W_{i-1}).$$

Repeatedly applying this relation and using (1) finally leads to

$$E(W_i) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{(1 - (\rho_1 + \ldots + \rho_i))(1 - (\rho_1 + \ldots + \rho_{i-1}))}, \qquad i = 1, \cdots, r. \tag{3}$$

The mean sojourn time $E(S_i)$ of a type i customer follows from $E(S_i) = E(W_i) + E(B_i)$, yielding

$$E(S_i) = \frac{\sum_{j=1}^{r} \rho_j E(R_j)}{(1 - (\rho_1 + \ldots + \rho_i))(1 - (\rho_1 + \ldots + \rho_{i-1}))} + E(B_i), \tag{4}$$

for $i = 1, \cdots, r$.

11.2 Preemptive-resume priority

We will show that the results in case the service times may be interrupted easily follow from the ones in the previous section.

Consider a type i customer. For a type i customer there do not exist lower priority customers due to the preemption rule. So we henceforth assume that $\lambda_{i+1} = \cdots = \lambda_r = 0$.

The waiting time of a type i customer can again be divided into portions $X_1, X_2, \ldots$. Now X_1 is equal to the *total* amount of work in the system upon arrival, since we assumed that there are no lower priority customers. Observe that the total amount of work in the system does not depend on the order in which the customers are served. Hence, at each point in time, it is exactly the same as in the system where the customers are served according to the non-preemptive priority rule. So $X_1, X_2, \ldots$, and thus also W_i have the

same distribution as in the system with non-preemptive priorities and, of course, with $\lambda_{i+1} = \cdots = \lambda_r = 0$. From (3) we then obtain

$$E(W_i) = \frac{\sum_{j=1}^{i} \rho_j E(R_j)}{(1-(\rho_1 + \ldots + \rho_i))(1-(\rho_1 + \ldots + \rho_{i-1}))}, \qquad i = 1, \cdots, r.$$

For the mean sojourn time we have to add the service time plus all the interruptions of higher priority customers during the service time. The mean of such a generalized service time can be found along the same lines as (2), yielding

$$\frac{E(B_i)}{1-(\rho_1 + \cdots + \rho_{i-1})}.$$

So the mean sojourn time of a type i customer is given by

$$E(S_i) = \frac{\sum_{j=1}^{i} \rho_j E(R_j)}{(1-(\rho_1 + \ldots + \rho_i))(1-(\rho_1 + \ldots + \rho_{i-1}))} + \frac{E(B_i)}{1-(\rho_1 + \cdots + \rho_{i-1})}, \qquad (5)$$

for $i = 1, \cdots, r$.

11.3 Shortest processing time first

In production systems one often processes jobs according to the shortest processing time first rule (SPTF). The mean production lead time in a single machine system operating according to the SPTF rule can be found using the results in section 11.1.

Consider an $M/G/1$ queue with arrival rate λ and service times B with density $f_B(x)$. Assume that $\rho = \lambda E(B) < 1$. The server works according to the SPTF rule. That is, after a service completion, the next customer to be served is the one with the shortest service time.

Define type x customers as the ones with a service time between x and $x + dx$. The mean waiting time of a type x customer is denoted by $E(W(x))$ and $\rho(x)dx$ is the fraction of time the server helps type x customers, so

$$\rho(x)dx = (\lambda f_B(x)dx)x = \lambda x f_B(x)dx. \qquad (6)$$

From (3) and by observing that the numerator in (3) corresponds to the mean amount of work at the server, which in the present situation is simply given by $\rho E(R)$, we obtain

$$\begin{aligned} E(W(x)) &= \frac{\rho E(R)}{(1-\int_{y=0}^{x} \rho(y)dy)^2} \\ &= \frac{\rho E(R)}{(1-\lambda \int_{y=0}^{x} y f_B(y)dy)^2}. \end{aligned}$$

Hence the mean overall waiting time is given by

$$\begin{aligned} E(W) &= \int_{x=0}^{\infty} E(W(x)) f_B(x)dx \\ &= \rho E(R) \int_{x=0}^{\infty} \frac{f_B(x)dx}{(1-\lambda \int_{y=0}^{x} y f_B(y)dy)^2}. \end{aligned} \qquad (7)$$

In table 1 we compare the SPTF rule with the usual first come first served (FCFS) rule for an $M/M/1$ system with mean service time 1. The mean waiting time for the SPTF rule is given by

$$E(W) = \rho \int_{x=0}^{\infty} \frac{e^{-x}dx}{(1-\rho(1-e^{-x}-xe^{-x}))^2}$$

and for the FCFS rule it satisfies

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

The results in table 1 show that considerable reductions in the mean waiting time are possible.

	$E(W)$	
ρ	FCFS	SPTF
0.5	1	0.713
0.8	4	1.883
0.9	9	3.198
0.95	19	5.265

Table 1: The mean waiting time for FCFS and SPTF in an $M/M/1$ with $E(B) = 1$

11.4 A conservation law

In this section we consider a single-server queue with r types of customers. The type i customers arrive according to a general arrival stream with rate λ_i, $i = 1, \ldots, r$. The mean service time and mean residual service time of a type i customer is denoted by $E(B_i)$ and $E(R_i)$, respectively. Define $\rho_i = \lambda_i E(B_i)$. We assume that

$$\sum_{i=1}^{r} \lambda_i E(B_i) < 1,$$

so that the server can handle the amount of work offered per unit of time. Customers enter service in an order independent of their service times and they may not be interrupted during their service. So, for example, the customers may be served according to FCFS, random or a non-preemptive priority rule. Below we derive a conservation law for the mean waiting times of the r types of customers, which expresses that a weighted sum of these mean waiting times is independent of the service discipline. This implies that an improvement in the mean waiting of one customer type owing to a service discipline will always degrade the mean waiting time of another customer type.

Let $E(V(P))$ and $E(L_i^q(P))$ denote the mean amount of work in the system and the mean number of type i customers waiting in the queue, respectively, for discipline P. The

mean amount of work in the system is given by

$$E(V(P)) = \sum_{i=1}^{r} E(L_i^q(P))E(B_i) + \sum_{i=1}^{r} \rho_i E(R_i). \tag{8}$$

The first sum at the right-hand side is the mean amount of work in the queue, and the second one is the mean amount of work at the server. Clearly the latter does not depend on the discipline P.

The crucial observation is that the amount of work in the system does not depend on the order in which the customers are served. The amount of work decreases with one unit per unit of time independent of the customer being served and when a new customer arrives the amount of work is increased by the service time of the new customer. Hence, the amount of work does not depend on P. Thus from equation (8) and Little's law

$$E(L_i^q) = \lambda_i E(W_i(P)),$$

we obtain the following conservation law for the mean waiting times,

$$\sum_{i=1}^{r} \rho_i E(W_i(P)) = \text{constant with respect to service discipline } P.$$

Below we present two examples where this law is used.

Example 11.1 *($M/G/1$ with FCFS and SPTF)*
The SPTF rule selects customers in a way that is dependent of their service times. Nevertheless, the law above also applies to this rule. The reason is that the SPTF rule can be translated into a non-preemptive rule as explained in section 11.3. Below we check whether for the $M/G/1$ the weighted sum of the mean waiting times for SPTF is indeed the same as for FCFS.

In case the customers are served in order of arrival it holds that

$$\rho E(W) = \frac{\rho^2 E(R)}{1-\rho}\,.$$

When the server works according to the SPTF rule we have (see (6) and (7))

$$\begin{aligned}
\int_{x=0}^{\infty} E(W(x))\rho(x)dx &= \int_{x=0}^{\infty} \frac{\rho E(R)\lambda x f_B(x)dx}{(1-\lambda\int_{y=0}^{x} y f_B(y)dy)^2} \\
&= \left.\frac{\rho E(R)}{1-\lambda\int_{y=0}^{x} y f_B(y)dy}\right|_{x=0}^{\infty} \\
&= \frac{\rho^2 E(R)}{1-\rho}\,,
\end{aligned}$$

which indeed is the same as for the FCFS rule.

Example 11.2 *($M/G/1$ with non-preemptive priority)*
Consider an $M/G/1$ queue with two types of customers. The type 1 customers have non-preemptive priority over the type 2 customers. The mean waiting time for the type 1 customers is given by (1). We now derive the mean waiting time for the type 2 customers by using the conservation law. According to this law it holds that

$$\rho_1 E(W_1) + \rho_2 E(W_2) = C, \tag{9}$$

where C is some constant independent of the service discipline. To determine C consider the FCFS discipline. For FCFS it follows that

$$E(W_1) = E(W_2) = \frac{\rho_1 E(R_1) + \rho_2 E(R_2)}{1 - \rho_1 - \rho_2}\,.$$

Hence,

$$C = (\rho_1 + \rho_2)\,\frac{\rho_1 E(R_1) + \rho_2 E(R_2)}{1 - \rho_1 - \rho_2}\,. \tag{10}$$

By substituting (1) and (10) into equation (9) we retrieve formula (3) for the mean waiting time of the type 2 customers under the non-preemptive priority rule.