5 The General Service Time Queue

5.1 The M/G/1 FIFO queue

This is a queue where customers are served according to the first-in-first out (FIFO) discipline, the arrivals are Poisson (rate $\lambda > 0$), and the successive customer service times are mutually independent with the *same*, *arbitrary*, cumulative distribution function $G(x)$. More precisely, if σ_i and σ_j are the service times of two customers, say customers i and j, $i \neq j$, respectively, then

(1) σ_i and σ_j are independent rvs

(2) $G(x) = P(\sigma_i \leq x) = P(\sigma_j \leq x)$ for all $x \geq 0$.

Let $1/\mu$ be the mean service time, namely, $1/\mu = E[\sigma_i] = \int_0^\infty (1 - G(x))dx$. The service times are further assumed to be independent of the arrival process.

As usual we will set $\rho = \lambda/\mu$.

For this queueing system, the process $(N(t), t \geq 0)$, where $N(t)$ is the number of customers in the queue at time t, is *not* a Markov process. This is because the probabilistic future of $N(t)$ for $t > s$ cannot be determined if one only knows $N(s)$, except if $N(s) = 0$ (consider for instance the case when the service times are all equal to the same constant).

Mean queue-length and mean response time

We assume that the queue is empty at time $t = 0$. Customers are served according to the FIFO service discipline.

Let

- $0 < t_i$ be the arrival time of the ith customer;
- W_i be the waiting time in queue of the i-th customer;
- $\overline{W}$ be the expected waiting time in steady-state ($\overline{W} = \lim_{n\to\infty}(1/n)\sum_{i=1}^n W_i$ when this limit exists);
- $X(t)$ be the number of customers in the *waiting room* at time $t > 0$;
- $R(t)$ be the *residual service time* of the customer in the server at time t, if any. By convention, $R(t) = 0$ if the system is empty at time t;
- σ_i the service time of customer i. Note that $E[\sigma_i] = 1/\mu$.

We will assume by convention that $X(t_i)$ is the number of customers in the *waiting room* just *before* the arrival of the i-th customer. We have

$$
\begin{aligned}
E[W_i] &= E[R(t_i)] + E\left[\sum_{j=i-X(t_i)}^{i-1} \sigma_j\right] \\
&= E[R(t_i)] + \sum_{k=0}^{\infty}\sum_{j=i-k}^{i-1} E[\sigma_j \,|\, X(t_i)=k]\, P(X(t_i)=k) \\
&= E[R(t_i)] + \frac{1}{\mu}\, E[X(t_i)]. \qquad (78)
\end{aligned}
$$

To derive (78) we have used the fact that σ_j is independent of $X(t_i)$ for $j = i - X(t_i), \ldots, i-1$, which implies that $E[\sigma_j \,|\, X(t_i)=k] = 1/\mu$. Indeed, $X(t_i)$ only depends on the service times σ_j for $j = 1, \ldots, i - X(t_i) - 1$ and not on σ_j for $j \geq i - X(t_i)$ since the service discipline is FIFO.

Letting now $i \to \infty$ in (78) yields

$$\overline{W} = \overline{R} + \frac{\overline{X}}{\mu} \qquad (79)$$

with

- $\overline{R} := \lim_{i\to\infty} E[R(t_i)]$ is the mean service time at *arrival epochs* in steady-state, and
- $\overline{X} := \lim_{i\to\infty} E[X(t_i)]$ is the mean number of customers in the waiting room at *arrival epochs* in steady-state.

Because the arrival process is a Poisson process (PASTA property: Poisson Arrivals See Times Averages), we have that

$$
\begin{aligned}
\overline{R} &= \lim_{t\to\infty} \frac{1}{t}\int_0^t R(s)\,ds \qquad (80)\\
\overline{X} &= \lim_{t\to\infty} \frac{1}{t}\int_0^t X(s)\,ds. \qquad (81)
\end{aligned}
$$

We shall not prove these results.

In words, (80) says that the mean residual service times at *arrival* epochs and at *arbitrary* epochs are the same. Similarly, (81) expresses the fact that the mean number of customers at *arrival* epochs and at *arbitrary* epochs are the same.

Example 7. If the arrivals are not Poisson then formulae (80) and (81) are in general not true. Here is an example where (80) is not true: assume that the nth customer arrives at time $t_n = n$ seconds (s) for all $n \geq 1$ and that it requires 0.999s of service (i.e., $\sigma_n = 0.999$s). If the system is empty at time 0, then clearly $R(t_n) = 0$ for all $n \geq 1$ since an incoming customer will always find the system empty, and therefore the left-hand side of (80) is zero; however, since the server is always working in $(n, n + 0.999)$ for all $n \geq 1$ it should be clear that the right-hand side of (80) is $(0.999)^2/2$. ♦

Applying Little's formula to the *waiting room* yields

$$\overline{X} = \lambda\, \overline{W}$$

so that, cf. (79),

$$\overline{W}\,(1-\rho) = \overline{R}. \tag{82}$$

From now on we will assume that $\rho < 1$. Hence, cf. (82),

$$\overline{W} = \frac{\overline{R}}{1-\rho}. \tag{83}$$

The condition $\rho < 1$ is the stability condition of the M/G/1 queue. This condition is again very natural. We will compute $\overline{R}$ under the assumption that the queue empties infinitely often (it can be shown that this occurs with probability 1 if $\rho < 1$). Let C be a time when the queue is empty and define $Y(C)$ to be the number of customers served in $(0, C)$.

We have (hint: display the curve $t \to R(t)$):

$$\begin{aligned}
\overline{R} &= \lim_{C\to\infty} \frac{1}{C} \sum_{n=1}^{Y(C)} \frac{\sigma_i^2}{2} \\
&= \lim_{C\to\infty} \left(\frac{Y(C)}{C}\right) \lim_{C\to\infty} \left(\frac{1}{Y(C)} \sum_{n=1}^{Y(C)} \frac{\sigma_i^2}{2}\right) \\
&= \lambda\, \frac{E[\sigma^2]}{2}
\end{aligned}$$

where $E[\sigma^2]$ is the second-order moment of the service times (i.e., $E[\sigma^2] = E[\sigma_i^2]$ for all $i \geq 1$).

Hence, for $\rho < 1$,

$$\overline{W} = \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)}. \tag{84}$$

This formula is the *Pollaczek-Khinchin* (abbreviated as P-K) formula for the mean waiting time in an M/G/1 queue. Since $\rho = \lambda\, E[\sigma]$, we can rewrite (84) as follows

$$\overline{W} = \frac{\rho}{2(1-\rho)} \cdot \frac{\mathrm{Var}(\sigma) + E[\sigma]^2}{E[\sigma]}.$$

Clearly, higher service time variability yields longer waiting times.

The mean system response time $\overline{T}$ is given by

$$\overline{T} = \frac{1}{\mu} + \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)} \tag{85}$$

and, by Little's formula, the mean number of customers $E[N]$ in the *entire* system (waiting room + server) is given by

$$\overline{N} = \rho + \frac{\lambda^2\, E[\sigma^2]}{2\,(1-\rho)}. \tag{86}$$

Consider the particular case when $P(\sigma_i \leq x) = 1 - \exp(-\mu x)$ for $x \geq 0$, that is, the M/M/1 queue. Since $E[\sigma^2] = 2/\mu^2$, we see from (84) that

$$\overline{W} = \frac{\lambda}{\mu^2\,(1-\rho)} = \frac{\rho}{\mu\,(1-\rho)}$$

which agrees with (77).

It should be emphasized that $\overline{W}$, $\overline{T}$ and $\overline{N}$ now depend upon the *first two moments* ($1/\mu$ and $E[\sigma^2]$) of the service time distribution function (and of course upon the arrival rate). This is in contrast with the M/M/1 queue where these quantities only depend upon the mean of the service time (and upon the arrival rate).

Example 8. *Compare the M/M/1 queue and the M/D/1 queues.*

5.2 The M/G/1 FIFO queue with vacations

This is an M/G/1 FIFO queue in which the server goes on vacation as soon as the queue empties. If at the end of a vacation, the queue is still empty, then the server starts a new vacation. Vacations are independent and identically distributed random variables. More precisely, if V_i and V_j are the ith and jth vacation times , $i \neq j$, then

(1) V_i and V_j are independent rvs

(2) $G(x) = P(V_i \leq x) = P(V_j \leq x)$ for all $x \geq 0$.

Let V be a generic random variable having CDF $G(x)$. Observe that the server is never idle, either it is busy serving customers or it is on vacations.

Arrivals form a Poisson process with rate λ.

For this queueing system, the process $(N(t),\, t \geq 0)$, where $N(t)$ is the number of customers in the queue at time t, is *not* a Markov process, unless service times and vacation times are exponentially distributed.

Mean waiting time

We assume that the queue is empty at time $t = 0$. Customers are served according to the FIFO service discipline. Let

- W_i be the waiting time in queue of the ith customer;
- $\overline{W}$ be the expected waiting time in steady-state ($\overline{W} = \lim_{n\to\infty}(1/n)\sum_{i=1}^{n} W_i$ when this limit exists);
- $X(t)$ be the number of customers in the *waiting room* at time $t > 0$;
- $R(t)$ be the *residual time* of a customer service time if the server is busy, or of a vacation if the server is on vacation at time t;

- σ_i the service time of customer i. Note that $E[\sigma_i] = 1/\mu$;
- V_k the kth vacation time of the server;
- $\rho = \lambda/\mu$ as usual.

Similarly to what has been done in analyzing the M/G/1 FIFO queue, we can derive

$$\overline{W} = \overline{R} + \overline{X}\overline{\sigma}. \tag{87}$$

where

- $\overline{R}$ is the mean residual time at *arrival epochs* in steady-state, and
- $\overline{X}$ is the mean number of customers in the waiting room at *arrival epochs* in steady-state.

Because of the PASTA property, $\overline{R}$ and $\overline{X}$ are also the time averages. Applying Little's formula to the *waiting room* yields

$$\overline{X} = \lambda \overline{W}$$

so that, (87) is rewritten ($\rho < 1$)

$$\overline{W} = \frac{\overline{R}}{1-\rho}. \tag{88}$$

The condition $\rho < 1$ is the stability condition of the M/G/1 queue, with or without vacations. When the server returns from vacation, the number of customers in the waiting room can be very large and we must have $\rho < 1$ to ensure that the queue will eventually empty. Equation (88) is the same as (83) for the M/G/1 FIFO queue, but here the residual time has a different meaning.

We will now compute $\overline{R}$. Let $D(t)$ be the number of customers that have completed their service in the interval $(0, t)$. Let $V(t)$ be the number of complete vacations in $(0, t)$.

We have (hint: display the curve $t \to R(t)$):

$$\begin{aligned}
\overline{R} &= \lim_{t\to\infty} \frac{1}{t}\int_0^t R(u)du = \lim_{t\to\infty}\frac{1}{t}\left[\sum_{i=1}^{D(t)} \frac{\sigma_i^2}{2} + \sum_{k=1}^{V(t)} \frac{V_k^2}{2} + \text{incomplete triangle}\right] \\
&= \lim_{t\to\infty}\left[\frac{D(t)}{t}\frac{1}{D(t)}\sum_{i=1}^{D(t)}\frac{\sigma_i^2}{2} + \frac{V(t)}{t}\frac{1}{V(t)}\sum_{k=1}^{V(t)}\frac{V_k^2}{2}\right] \\
&= \lambda\frac{\mathrm{E}\left[\sigma^2\right]}{2} + \lim_{t\to\infty}\frac{V(t)}{t}\frac{\mathrm{E}\left[V^2\right]}{2}
\end{aligned}$$

where $E[\sigma^2]$ is the second-order moment of the service times and $E[V^2]$ is the second-order moment of the residual times.

We can write the following

$$\begin{aligned}
t &= \sum_{i=1}^{D(t)} \sigma_i + \sum_{k=1}^{V(t)} V_k + \text{residual} \\
1 &= \frac{D(t)}{t}\frac{1}{D(t)}\sum_{i=1}^{D(t)} \sigma_i + \frac{V(t)}{t}\frac{1}{V(t)}\sum_{k=1}^{V(t)} V_k + \frac{\text{residual}}{t}.
\end{aligned}$$

Taking the limit as t goes to ∞, we get

$$1 = \lambda \mathrm{E}\left[\sigma\right] + \lim_{t\to\infty} \frac{V(t)}{t} \mathrm{E}\left[V\right] \quad \Rightarrow \quad \lim_{t\to\infty} \frac{V(t)}{t} = \frac{1-\rho}{\mathrm{E}\left[V\right]} \; .$$

Combining with the expression for $\overline{R}$, we can rewrite (88) as follows

$$\overline{W} = \frac{\lambda\, E[\sigma^2]}{2\,(1-\rho)} + \frac{E[V^2]}{2\mathrm{E}\left[V\right]} \; . \tag{89}$$

The first term in the sum is the waiting time in the M/G/1 FIFO queue without vacations.

The mean system response time $\overline{T}$ is given by

$$\overline{T} = \frac{1}{\mu} + \overline{W} \; .$$