A *queue* is a service line. A single-server queue requires that everyone be served one at a time in the order of arrival. Some examples include an ATM machine, a networked printer, and a fast-lane checkout counter. We shall study a scenario that has the following conditions:

(i) The service time is an exponential distribution with an average time of $M$.
(ii) Arrivals to the queue follow a Poisson distribution with an average of $\lambda$ per unit time. Thus, the average time of new arrivals is $L = 1/\lambda$.
(iii) Arrivals are independent of services.
(iv) So that the queue does not overflow, we assume that the average service time is no more than the average arrival time: $M \leq L$.
(v) The queue has an initial amount of $j$ people in line, including the one being served, and no one leaves the line until his service is complete.

As an example throughout, consider $j = 4$ already queued, $M = 3$ minutes average service time, and $L = 5$ minutes as average time of new arrivals.

**Some Questions**

1. On average, how long must a new arrival wait to be served? What is the distribution of this waiting time?

2. What is the probability that the queue eventually clears? What is the average amount of time needed to clear the queue?

3. What is the average number of people in the queue at any given time $t$?

**Exponential Distribution Background**

Let $Y \sim \text{exp}(\theta)$ be any exponential distribution. Then $f_Y(t) = \frac{1}{\theta} e^{-t/\theta}$ for $t \geq 0$, and $F_Y(t) = P(Y \leq t) = 1 - e^{-t/\theta}$. Moreover, $E[Y] = \theta$. In our context, we will use an exponential distribution to measure the time of the next occurrence.

Suppose we have been waiting some time for an occurrence, but none have happened. So we then re-start the clock and start anew in measuring the time for the next occurrence. What is the distribution now? Does the average waiting time change?

Consider $P(Y > s + r \mid Y > s)$, for $r \geq 0$. In other words, we have already waited $s$ units of time with no occurrence, so we know $Y > s$. We now re-start measuring the time $W$ for the occurrence as a function of time $r \geq 0$. Then,

$$P(W > r) = P(Y > s + r \mid Y > s) = \frac{P(Y > s + r \cap Y > s)}{P(Y > s)}$$

$$= \frac{P(Y > s + r)}{P(Y > s)} = \frac{e^{-(s+r)/\theta}}{e^{-s/\theta}} = e^{-r/\theta}.$$
Thus, $P(W \leq r) = 1 - e^{-r/\theta}$, which is still the cdf of an $\exp(\theta)$ distribution. So it does not matter when we start measuring the time of the occurrence. The remaining time, starting at any point, is an $\exp(\theta)$ distribution with an average wait time of $\theta$.

In particular, when we arrive at a queue, there may already be someone being served. We do not know how long they have been waited on so far. But if we start measuring time anew, the time for this service completion is still an exponential distribution with an average service time of $M$.

**Time of the Next Occurrence**

Let $S \sim \exp(M)$ be the service time, where the average service time is $M$ (minutes). Then $S$ has pdf $f_S(t) = \frac{1}{M}e^{-t/M}$ for $t \geq 0$, and cdf $F_S(t) = P(S \leq t) = 1 - e^{-t/M}$. Then let $A \sim \exp(L)$ be the distribution times of the next new arrival as measured from the time of the previous arrival. For these independent exponential distributions, the minimum exponential distributions, the *minimum* time of service or arrival is still an exponential distribution. Indeed, let $X = \min(S, A)$.

Then,

$$P(X > t) = P(S > t \cap A > t) = P(S > t) \times P(A > t) \text{ by indep.}$$

$$= e^{-t/M} \times e^{-t/L} = e^{-t\left(\frac{1}{M} + \frac{1}{L}\right)}.$$ 

So the cdf of $X = \min(S, A)$ is $P(X \leq t) = 1 - e^{-t\left(\frac{1}{M} + \frac{1}{L}\right)} = 1 - e^{-t/R}$, which is the cdf of an $\exp(R)$ distribution where $R = \frac{1}{\frac{1}{M} + \frac{1}{L}} = \frac{LM}{L + M}$.

Henceforth, an occurrence shall denote either a service completion or an arrival. When the queue is empty, then the next occurrence must be an arrival. If the queue is occupied, then the next occurrence is either a service completion or an arrival. So we can state:

**Theorem 1.** The average time of the next occurrence is $L$ if the queue is empty, but it is $E[\min(S, A)] = \frac{LM}{L + M}$ if the queue is occupied.

**A New Arrival**

Suppose there are already $j$ people in the queue including the one being served, and a new person arrives at the end of the line. What is the distribution of the new arrival’s waiting time to be served and the average wait time?
Upon the new arrival, the first person has already spent some time being served; however, the remaining service time for this person is still an exp($M$) distribution. (The service time of every other person in line is also exp($M$).) Because the sum of independent exponential distributions is a gamma distribution, the wait time $W$ until service of the new arrival is a $\Gamma[M, j]$ distribution with pdf, cdf, and expected value given by

\[ f_W(t) = \frac{t^{j-1} e^{-t/M}}{M^j (j-1)!} \quad \text{for} \quad t \geq 0 \]

\[ F_W(t) = 1 - e^{-t/M} \left( \sum_{k=0}^{j-1} \frac{t^k}{M^k k!} \right) \]

and $E[W] = j \times M$

**Example 1.** Let $j = 4$ be already queued with an average service time of $M = 3$ min.

(a) What is the average wait time for service for a new arrival?
(b) What is the probability that it takes no more than 15 minutes for a new arrival to be served?

**Solution.** (a) A new arrival must wait an average of $4 \times 3 = 12$ minutes to be served.

(b) With these values of $j = 4$ and $M = 3$, the cdf of the wait time is

\[ F_W(t) = 1 - e^{-t/3} \left( \sum_{k=0}^{3} \frac{t^k}{3^k k!} \right) = 1 - e^{-t/3} \left( 1 + \frac{t}{3} + \frac{t^2}{18} + \frac{t^3}{162} \right) \]

Hence, $P(W \leq 15) = F_W(15) = 1 - \frac{118 e^{-5}}{3} = 0.735$.

**Which is More Likely to Occur First?**

While the queue is occupied, an occurrence is either a service completion or a new arrival. Given an occurrence, what is the probability that it is a service completion?

To find this probability, we let $S \sim \text{exp}(M)$ be the service time and let $A \sim \text{exp}(L)$ be the time of the next arrival, where $S$ and $A$ are independent. By independence, the joint pdf of $(S, A)$ is

\[ f(s, a) = f_S(s) f_A(a) = \frac{1}{M} e^{-s/M} \times \frac{1}{L} e^{-a/L} \]

We now need $P(S < A)$ which is the probability that a service completion occurs before the next arrival (measured from any instant in time). Using $s$ as the $x$-axis and $a$ as the $y$-axis, we then have
\[ P(S < A) = \int_0^\infty \int_0^\infty f(s, a) \, da \, ds = \int_0^\infty \frac{1}{M} e^{-s/M} \times \frac{1}{L} e^{-a/L} \, da \, ds \]

\[ = \int_0^\infty \frac{1}{M} e^{-s/M} \left(-e^{-a/L}\right) ds = \int_0^\infty \frac{1}{M} e^{-s/M} \left(e^{-s/L}\right) ds \]

\[ = \frac{1}{M} \int_0^\infty e^{-s(ML)} \frac{s(L+M)}{ML} ds = -\frac{1}{M} \frac{ML}{L+M} e^{-s(ML)} \bigg|_0^\infty = \frac{L}{L+M}. \]

And the probability of a new arrival before a the next service completion (while the queue is occupied) is given by \( P(A < S) = 1 - \frac{L}{L+M} = \frac{M}{L+M} \). (Note: The probability of a service completion and an arrival happening at the exact same instant is \( P(A = S) = 0 \).)

Henceforth, let \( p = P(A < S) = \frac{M}{L+M} \)

(which is the probability of the queue going up with a new arrival)

and let \( q = P(S < A) = \frac{L}{L+M} \)

(which is the probability of the queue going down with a service completion).

For example, with \( M = 3 \) average service time and \( L = 5 \) average arrival time, then

\[ p = \frac{3}{8} \quad \text{and} \quad q = \frac{5}{8}. \]

The next result shows that the average time of the next occurrence does not depend on whether the next occurrence is a service completion or an arrival.

**Theorem 2.** When the queue is occupied, the average time of the next occurrence when the next occurrence is an arrival and the average time of the next occurrence when the next occurrence is a service completion are both \( R = \frac{LM}{L+M} \).

**Proof.** First let \( R = \frac{LM}{L+M} \). The conditional average time until the next occurrence among next occurrences created by service completion is given by
By Theorem 1, the total average time between occurrences, \( E[\min(S, A)] \), is given by

\[
R = E[\min(S, A) | S < A] \times P(S < A) + E[\min(S, A) | A < S] \times P(A < S)
\]

\[
= Rq + E[\min(S, A) | A < S] \times p
\]

thus, \( R(1 - q) = E[\min(S, A) | A < S] \times p \) and \( R = E[\min(S, A) | A < S] \) also.

The Embedded Random Walk

Suppose we begin with \( j \) people already in the queue. An increment or step will be the process of either someone arriving (an upward movement that adds 1 to the queue) or someone’s service being completed (a downward movement that subtracts 1 from the queue). Then the probabilities of moving up 1 unit or down one unit on each step (while the queue is occupied) are

\[
p = \frac{M}{L + M} = P(A < S) \quad \text{and} \quad q = \frac{L}{L + M} = P(S < A).
\]

That is, if a step occurs, then there is probability \( p \) that it was an arrival which adds 1 to the queue (we go up one unit with probability \( p \)), and there is probability \( q \) that there was a service completion (we go down one unit with probability \( q \)).

We now have the scenario of a simple random walk that begins at integer height \( j \) and moves either up or down 1 unit at a time on each step. Because \( M \leq L \), we have \( q \geq p \); so the random walk will tend to drift downward and will drop to any lower boundary with probability 1. In particular, there is probability 1 that it will drop to height 0. Under these conditions, there is probability 1 that the single-server queue will eventually clear and be empty.
Boundary Problem Formulas

Next, we recall some of the basic formulas for the simple random walk that begins at height \(j\) and moves up or down 1 unit at a time, where \(0 < p < 1\) is the probability of up. We assume \(0 \leq j \leq n\).

Probability hit \(n\) before \(0\):

\[
jP^n_0 = \begin{cases} 
\frac{j}{n} & \text{if } p = q \\
\frac{1-(q/p)^j}{1-(q/p)^n} & \text{if } p \neq q
\end{cases}
\]

Probability hit 0 before \(n\):

\[
jQ^n_0 = 1-jP^n_0 = \begin{cases} 
\frac{n-j}{n} & \text{if } p = q \\
\frac{(q/p)^j - (q/p)^n}{1-(q/p)^n} & \text{if } p \neq q
\end{cases}
\]

Probability eventually hit 0:

\[
jP_0 = \begin{cases} 
1 & \text{if } q \geq p \\
(q/p)^j & \text{if } p > q
\end{cases}
\]

Probability eventually hit \(n\):

\[
jP^n = \begin{cases} 
1 & \text{if } p \geq q \\
(p/q)^{n-j} & \text{if } p < q
\end{cases}
\]

Average number of steps to hit 0 or \(n\):

\[
E[jS^n_0] = \begin{cases} 
j(n-j) & \text{if } p = q = 1/2 \\
\frac{n}{p-q} \times \left( \frac{1-(q/p)^j}{1-(q/p)^n} \right) - \frac{j}{p-q} & \text{if } p \neq q
\end{cases}
\]

Average number of steps to hit 0 (for \(q \geq p\)):

\[
E[jT_0] = \begin{cases} 
\infty & \text{if } p = q \\
\frac{j}{q-p} & \text{if } p < q
\end{cases}
\]

Example 2. Let \(j = 4\) be already queued, with \(M = 3\) minutes average service time, and \(L = 5\) minutes as the average time of new arrivals.

(a) What is the probability that the server clears before 6 are in the queue?
(b) What is the probability that the queue ever gets up to 7?
(c) What is the average number of steps needed for the queue to clear when \(j = 4\)?
**Answers.** Let \( p = \frac{M}{L+M} = \frac{3}{8} \) and \( q = \frac{L}{L+M} = \frac{5}{8} \). (a) \( 4Q_0 = \frac{(5/3)^4 - (5/3)^6}{1 - (5/3)^6} \approx 0.67132 \)

(b) \( 4P_7 = (3/8)^7 \approx 0.0527 \)  
(c) \( E[4S_0] = \frac{4}{5/8 - 3/8} = 16 \)

**The Mean Time to Clear**

Provided \( M \leq L \), then we know the queue will clear with probability 1. We also know the mean number of steps (arrivals/services) required. But each step will take a different amount of time. After an arrival, the time of the next step is the minimum time of the next arrival \( \exp(L) \) or the additional service time which is still \( \exp(M) \). After a service completion, the time of the next step is the minimum of the next service time \( \exp(M) \) or the additional time of the next arrival which is still \( \exp(L) \). In either case, the time of each next step is

\[
X = \min(\exp(M), \exp(L)) = \exp(R), \quad \text{where} \quad R = \frac{LM}{L+M}.
\]

Moreover, by Theorem 2, the average time between steps while the queue is occupied is \( R \), regardless if the step is due to an arrival or a service completion. Because the time length of each step is independent of how many steps are actually made, the average time of clearing \( C \) is just the product of the two averages. So when \( M < L \), which makes \( p < q \), we have

\[
E[C] = E[\text{Number of Steps}] \times E[\text{Time of Steps}]
= \frac{j}{q - p} \times R = \frac{j}{L} \frac{L}{M} \frac{LM}{L+M} = \frac{jLM}{L-M}.
\]

But when \( M = L \) and thus \( q = p \), then the average number of steps and average amount of time are both \( +\infty \). Thus we can state:

**Theorem 3.** When the queue is occupied, the average time needed to clear is

\[
E[C] = \begin{cases}  
\frac{jML}{L-M} & \text{if } M < L \\
\infty & \text{if } M = L
\end{cases}
\]

**Example 3.** Let \( j = 4 \), \( M = 3 \), and \( L = 5 \). What is the mean time of each step (i.e., arrival or job completion)? What is the mean time to clear the queue?

**Answer.** Each incremental step averages \( 1/(1/M + 1/L) = 1.875 \) minutes. As in Example 2, the average number of steps is \( j/(q-p) = 16 \). Hence, the average time needed to clear the queue is \( 16 \times 1.875 = 30 \) minutes, also given by \( jML / (L-M) \).
The Number in the Queue at time $t$

We now consider a stochastic process $Z_t$ that gives the exact number in the queue at time $t$. By our initial conditions, we know $Z_0 = j$. Here is one way to account for the mixture of services/arrivals:

Let $S_1$ be the service time of the first person and let $S_i = S_{i-1} + \text{Service time of } i\text{th person for } i \geq 2$. Suppose the first person requires 3.5 minutes, the second requires 2 minutes, and the third requires 4 minutes. Then $S_1 = 3.5$, $S_2 = 5.5$, and $S_3 = 9.5$.

Let $Y_1$ be the time of the first new arrival and let $Y_i = Y_{i-1} + \text{additional time of } i\text{th arrival}$. For instance, suppose the first new arrival is in 6 minutes, the next is 4 minutes later, and the next is 8 minutes later. Then $Y_1 = 6$, $Y_2 = 10$, and $Y_3 = 18$.

The number in the queue at time $t$ is then

$$Z_t = j - \sum_{i=1}^{\infty} i 1_{[S_i, S_{i+1})}(t) + \sum_{i=1}^{\infty} i 1_{[Y_i, Y_{i+1})}(t)$$

For example, consider the cumulative service times $S_1 = 3.5$, $S_2 = 5.5$, $S_3 = 9.5$, $S_4 = 15$ and the cumulative arrival times $Y_1 = 6$, $Y_2 = 10$, and $Y_3 = 18$.

On the time interval $[0, 3.5)$ there are 4 in the queue. On $[3.5, 5.5)$ only one has been served so we subtract 1. But on $[S_2, S_3) = [5.5, 9.5)$ we can subtract 2 and on $[9.5, 15)$ we can subtract 3. But there are also overlaps in time with the arrivals. On $[6, 10)$ we must add 1 and on $[10, 18)$ we must add 2. Thus we have the following result:

$Z_t = 4 \text{ for } t \text{ in } [0, 3.5)$ \quad Service Over at $t = 3.5 \rightarrow Z_t = 3 \text{ for } t \text{ in } [3.5, 5.5)$

Service Over at $t = 5.5$, Arrival at $t = 6$, Service Over at $t = 9.5$, Arrival at $t = 10$, etc.

$\rightarrow Z_t = 2 \text{ in } [5.5, 6) \rightarrow Z_t = 3 \text{ in } [6, 9.5) \rightarrow Z_t = 2 \text{ in } [9.5, 10)$ etc.

Below are some randomly generated paths of this stochastic process:
Unfortunately, there are no closed-form expressions for the probabilities $P(Z_t = k)$ of there being exactly $k$ in the queue at time $t$, but they can be approximated through simulations. Below are sample distribution results from 2000 trials at two specific times with our example of $j = 4, M = 3,$ and $L = 5$, where $Z_t$ remains 0 upon clearing.

<table>
<thead>
<tr>
<th># in Queue</th>
<th>Sample Prop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.211</td>
</tr>
<tr>
<td>1</td>
<td>0.1085</td>
</tr>
<tr>
<td>2</td>
<td>0.153</td>
</tr>
<tr>
<td>3</td>
<td>0.173</td>
</tr>
<tr>
<td>4</td>
<td>0.148</td>
</tr>
<tr>
<td>5</td>
<td>0.104</td>
</tr>
<tr>
<td>6</td>
<td>0.06</td>
</tr>
<tr>
<td>7</td>
<td>0.028</td>
</tr>
<tr>
<td>8</td>
<td>0.0105</td>
</tr>
<tr>
<td>9</td>
<td>0.003</td>
</tr>
<tr>
<td>10</td>
<td>0.001</td>
</tr>
</tbody>
</table>

$t = 10$ minutes \{Avg, 2.7225\}
Note: We are stopping the process when the queue clears, so no more arrivals occur after that time. Hence, we have an increasing proportion of trials with empty queues as time increases.

Although we cannot give an exact analysis of the distribution at a specific time $t$, we can give a closed-form description of the long-term distribution over large amounts of time when we continue the queue after clearing (see Exercise 3). There also is a closed-form description of the queue at time $t$ assuming that it stays empty upon clearing.

The Average Number in the Queue at time $t$

A very difficult problem is to find the average in the queue $E[Z_t]$ at time $t$. Like a random walk with $q \geq p$, $E[Z_t]$ will decrease as time increases. However, because $Z_t \geq 0$ and there is always some possibility that $Z_t > 0$, we must have $E[Z_t] > 0$. Below is a graph of simulated average queue amounts from 2000 trials at times $t = 0, 5, 10, 15, \ldots, 50$ with $j = 4, M = 3, \text{and } L = 5$, where $Z_t$ remains 0 upon clearing.

It appears that $E[Z_t]$ is close to an exponential decay function of the form $4 \times a^t$. In fact, an exponential regression on the data will yield a close fit with a high $r^2$ value.
Exercises

1. Let $j = 3$ be already queued, with $M = 4$ minutes average service time, and $L = 6$ minutes as the average time of new arrivals.

   (a) What is the next new arrival’s mean time of waiting for service?
   (b) What is the probability that the next new arrival must wait more than 12 minutes to be served?
   (c) What is the probability that the queue reaches 5 before it clears?
   (d) What is the probability that the queue ever gets up to 8?
   (e) What is the mean time needed to clear the queue?

2. Suppose we start with a cleared queue ($j = 0$). Arrivals occur on average every $L = 6$ minutes and service time averages $M = 4$ minutes. Wait for the first arrival, then wait for the queue to clear. Let $A$ be the time of the first arrival and let $T$ be the total time for the first arrival and re-clearing.

   (a) Derive $E[T]$ in general, then give its value in this case.
   (b) What is the probability that the queue gets to 3 before re-clearing?
   (c) Suppose we conduct $n$ sample trials. If we add up all the times $A_i$ of the first arrival for all $i$ trials and divide by the sum of the total times $T_i$, then we approximate the proportion $p_c$ of time for which the server was clear. But $p_c$ can be approximated in terms of $E[A]$ and $E[T]$ as follows:

   $$p_c \approx \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} T_i} = \frac{1}{n} \sum_{i=1}^{n} A_i = \frac{E[A]}{E[T]}$$

   Simplify this fraction in terms of $L$ and $M$ to give an approximation of $p_c$.

3. The limiting behavior of the M/M/1 queue over long periods of time (not at a specific single time), can be described as follows provided $M < L$:

   Let $r = \frac{M}{L}$ which is the traffic rate. The proportion of time that has $k$ elements in the queue is $p_k = r^k (1 - r)$. The average number in the queue is $E[N] = \sum_{k=0}^{\infty} k \times p_k$.

   Consider the limiting behavior for $M = 3$ and $L = 5$. Let $N$ be the number in the queue.

   (a) What proportion of time is the queue occupied?
   (b) What proportion of time does the queue have at least 3 elements?
   (c) What is the average number in the queue?

   (d) Derive a general formula for $E[N]$ in terms of $r$ then in terms of $M$ and $L$.
   (e) Use Part (d) to derive the average waiting time of a new arrival.